Chan, Shu-Park "Section I – Circuits"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

The Intel Pentium® processor, introduced at speeds of up to 300 MHz, combines the architectural advances in the Pentium Pro processor with the instruction set extensions of Intel MMX™ media enhancement technology. This combination delivers new levels of performance and the fastest Intel processor to workstations.

The Pentium II processor core, with 7.5 million transistors, is based on Intel's advanced P6 architecture and is manufactured on .35-micron process technology. First implemented in the Pentium Pro processor, the Dual Independent Bus architecture is made up of the L2 cache bus and the processor-to-main-memory system bus. The latter enables simultaneous parallel transactions instead of single, sequential transactions of previous generation processors.

The types of applications that will benefit from the speed of the Pentium II processor and the media enhancement of MMX technology include scanning, image manipulation, video conferencing, Internet browsers and plug-ins, video editing and playback, printing, faxing, compression, and encryption.

The Pentium II processor is the newest member of the P6 processor family, but certainly not the last in the line of high performance processors. (Courtesy of Intel Corporation.)

I

Circuits

1 Passive Components *M. Pecht, P. Lall, G. Ballou, C. Sankaran, N. Angelopoulos*
Resistors • Capacitors and Inductors • Transformers • Electrical Fuses

2 Voltage and Current Sources *R.C. Dorf, Z. Wan, C.R. Paul, J.R. Cogdell*
Step, Impulse, Ramp, Sinusoidal, Exponential, and DC Signals • Ideal and Practical Sources • Controlled Sources

3 Linear Circuit Analysis *M.D. Ciletti, J.D. Irwin, A.D. Kraus, N. Balabanian, T.A. Bickart, S.P. Chan, N.S. Nise*
Voltage and Current Laws • Node and Mesh Analysis • Network Theorems • Power and Energy • Three-Phase Circuits • Graph Theory • Two Port Parameters and Transformations

4 Passive Signal Processing *W.J. Kerwin*
Low-Pass Filter Functions • Low-Pass Filters • Filter Design

5 Nonlinear Circuits *J.L. Hudgins, T.F. Bogart, Jr., K. Mayaram, M.P. Kennedy, G. Kolumbán*
Diodes and Rectifiers • Limiters • Distortion • Communicating with Chaos

6 Laplace Transform *R.C. Dorf, Z. Wan, D.E. Johnson*
Definitions and Properties • Applications

7 State Variables: Concept and Formulation *W.K. Chen*
State Equations in Normal Form • The Concept of State and State Variables and Normal Tree • Systematic Procedure in Writing State Equations • State Equations for Networks Described by Scalar Differential Equations • Extension to Time-Varying and Nonlinear Networks

8 The z-Transform *R.C. Dorf, Z. Wan*
Properties of the *z*-Transform • Unilateral *z*-Transform • *z*-Transform Inversion • Sampled Data

9 T-P Equivalent Networks *Z. Wan, R.C. Dorf*
Three-Phase Connections • Wye ⇔ Delta Transformations

10 Transfer Functions of Filters *R.C. Dorf, Z. Wan*
Ideal Filters • The Ideal Linear-Phase Low-Pass Filter • Ideal Linear-Phase Bandpass Filters • Causal Filters • Butterworth Filters • Chebyshev Filters

11 Frequency Response *P. Neudorfer*
Linear Frequency Response Plotting • Bode Diagrams • A Comparison of Methods

12 Stability Analysis *F. Szidarovszky, A.T. Bahill*
Using the State of the System to Determine Stability • Lyapunov Stability Theory • Stability of Time-Invariant Linear Systems • BIBO Stability • Physical Examples

13 Computer Software for Circuit Analysis and Design *J.G. Rollins, P. Bendix*
Analog Circuit Simulation • Parameter Extraction for Analog Circuit Simulation

Shu-Park Chan
International Technological University

THIS SECTION PROVIDES A BRIEF REVIEW of the definitions and fundamental concepts used in the study of linear circuits and systems. We can describe a *circuit* or *system*, in a broad sense, as a collection of objects called *elements* (*components, parts,* or *subsystems*) which form an entity governed by certain laws or constraints. Thus, a physical system is an entity made up of physical objects as its elements or components. A subsystem of a given system can also be considered as a system itself.

A mathematical model describes the behavior of a physical system or device in terms of a set of equations, together with a schematic diagram of the device containing the symbols of its elements, their connections, and numerical values. As an example, a physical electrical system can be represented graphically by a network which includes resistors, inductors, and capacitors, etc. as its components. Such an illustration, together with a set of linear differential equations, is referred to as a model system.

Electrical circuits may be classified into various categories. Four of the more familiar classifications are (a) linear and nonlinear circuits, (b) time-invariant and time-varying circuits, (c) passive and active circuits, and (d) lumped and distributed circuits. A *linear* circuit can be described by a set of linear (differential) equations; otherwise it is a nonlinear circuit. A *time-invariant* circuit or system implies that none of the components of the circuit have parameters that vary with time; otherwise it is a *time-variant* system. If the total energy delivered to a given circuit is nonnegative at any instant of time, the circuit is said to be *passive*; otherwise it is *active*. Finally, if the dimensions of the components of the circuit are small compared to the wavelength of the highest of the signal frequencies applied to the circuit, it is called a *lumped* circuit; otherwise it is referred to as a *distributed* circuit.

There are, of course, other ways of classifying circuits. For example, one might wish to classify circuits according to the number of accessible terminals or terminal pairs (ports). Thus, terms such as *n-terminal circuit* and *n-port* are commonly used in circuit theory. Another method of classification is based on circuit configurations (topology),[1] which gives rise to such terms as *ladders, lattices, bridged-T circuits,* etc.

As indicated earlier, although the words *circuit* and *system* are synonymous and will be used interchangeably throughout the text, the terms *circuit theory* and *system theory* sometimes denote different points of view in the study of circuits or systems. Roughly speaking, *circuit theory* is mainly concerned with interconnections of components (circuit topology) within a given system, whereas *system theory* attempts to attain generality by means of abstraction through a generalized (input-output state) model.

One of the goals of this section is to present a unified treatment on the study of linear circuits and systems. That is, while the study of linear circuits with regard to their topological properties is treated as an important phase of the entire development of the theory, a generality can be attained from such a study.

The subject of circuit theory can be divided into two main parts, namely, analysis and synthesis. In a broad sense, *analysis* may be defined as "the separating of any material or abstract entity [system] into its constituent elements;" on the other hand, *synthesis* is "the combining of the constituent elements of separate materials or abstract entities into a single or unified entity [system]."[2]

It is worth noting that in an analysis problem, the solution is always *unique* no matter how difficult it may be, whereas in a synthesis problem there might exist an infinite number of solutions or, sometimes, *none at all!*

It should also be noted that in some network theory texts the words *synthesis* and *design* might be used interchangeably throughout the entire discussion of the subject. However, the term *synthesis* is generally used to describe *analytical* procedures that can usually be carried out step by step, whereas the term *design* includes practical (design) procedures (such as trial-and-error techniques which are based, to a great extent, on the experience of the designer) as well as analytical methods.

In analyzing the behavior of a given physical system, the first step is to establish a mathematical model. This model is usually in the form of a set of either differential or difference equations (or a combination of them),

[1]Circuit topology or graph theory deals with the way in which the circuit elements are interconnected. A detailed discussion on elementary applied graph theory is given in Chapter 3.6.

[2]The definitions of analysis and synthesis are quoted directly from *The Random House Dictionary of the English Language,* 2nd ed., Unabridged, New York: Random House, 1987.

the solution of which accurately describes the motion of the physical systems. There is, of course, no exception to this in the field of electrical engineering. A physical electrical system such as an amplifier circuit, for example, is first represented by a circuit drawn on paper. The circuit is composed of resistors, capacitors, inductors, and voltage and/or current sources,[1] and each of these circuit elements is given a symbol together with a mathematical expression (i.e., the voltage-current or simply *v-i* relation) relating its terminal voltage and current at every instant of time. Once the network and the *v-i* relation for each element is specified, Kirchhoff's voltage and current laws can be applied, possibly together with the physical principles to be introduced in Chapter 3.1, to establish the mathematical model in the form of differential equations.

In Section I, focus is on analysis only (leaving coverage of synthesis and design to Section III, "Electronics"). Specifically, the passive circuit elements—resistors, capacitors, inductors, transformers, and fuses—as well as voltage and current sources (active elements) are discussed. This is followed by a brief discussion on the elements of linear circuit analysis. Next, some popularly used passive filters and nonlinear circuits are introduced. Then, Laplace transform, state variables, *z*-transform, and T and π configurations are covered. Finally, transfer functions, frequency response, and stability analysis are discussed.

Nomenclature

Symbol	Quantity	Unit	Symbol	Quantity	Unit
A	area	m^2	ω	angular frequency	rad/s
B	magnetic flux density	Tesla	P	power	W
C	capacitance	F	PF	power factor	
e	induced voltage	V	q	charge	C
ε	dielectric constant	F/m	Q	selectivity	
ε	ripple factor		R	resistance	Ω
f	frequency	Hz	$R(T)$	temperature coefficient of resistance	Ω/°C
F	force	Newton			
ϕ	magnetic flux	weber	ρ	resistivity	Ωm
I	current	A	s	Laplace operator	
J	Jacobian		τ	damping factor	
k	Boltzmann constant	1.38×10^{-23} J/K	θ	phase angle	degree
k	dielectric coefficient		v	velocity	m/s
K	coupling coefficient		V	voltage	V
L	inductance	H	W	energy	J
λ	eigenvalue		X	reactance	Ω
M	mutual inductance	H	Y	admittance	S
n	turns ratio		Z	impedance	Ω
n	filter order				

[1]Here, of course, active elements such as transistors are represented by their equivalent circuits as combinations of resistors and dependent sources.

Pecht, M., Lall, P., Ballou, G., Sankaran, C., Angelopoulos, N. "Passive Components"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

1

Passive Components

Michael Pecht
University of Maryland

Pradeep Lall
Motorola

Glen Ballou
Ballou Associates

C. Sankaran
Electro-Test

Nick Angelopoulos
Gould Shawmut Company

1.1 Resistors
Resistor Characteristics • Resistor Types
1.2 Capacitors and Inductors
Capacitors • Types of Capacitors • Inductors
1.3 Transformers
Types of Transformers • Principle of Transformation • Electromagnetic Equation • Transformer Core • Transformer Losses • Transformer Connections • Transformer Impedance
1.4 Electrical Fuses
Ratings • Fuse Performance • Selective Coordination • Standards • Products • Standard—Class H • HRC • Trends

1.1 Resistors

Michael Pecht and Pradeep Lall

The resistor is an electrical device whose primary function is to introduce resistance to the flow of electric current. The magnitude of opposition to the flow of current is called the resistance of the resistor. A larger resistance value indicates a greater opposition to current flow.

The resistance is measured in ohms. An ohm is the resistance that arises when a current of one ampere is passed through a resistor subjected to one volt across its terminals.

The various uses of resistors include setting biases, controlling gain, fixing time constants, matching and loading circuits, voltage division, and heat generation. The following sections discuss resistor characteristics and various resistor types.

Resistor Characteristics

Voltage and Current Characteristics of Resistors

The resistance of a resistor is directly proportional to the **resistivity** of the material and the length of the resistor and inversely proportional to the cross-sectional area perpendicular to the direction of current flow. The resistance R of a resistor is given by

$$R = \frac{\rho l}{A} \tag{1.1}$$

where ρ is the resistivity of the resistor material ($\Omega \cdot$ cm), l is the length of the resistor along direction of current flow (cm), and A is the cross-sectional area perpendicular to current flow (cm^2) (Fig. 1.1). Resistivity is an inherent property of materials. Good resistor materials typically have resistivities between 2×10^{-6} and 200×10^{-6} $\Omega \cdot$ cm.

The resistance can also be defined in terms of sheet resistivity. If the sheet resistivity is used, a standard sheet thickness is assumed and factored into resistivity. Typically, resistors are rectangular in shape; therefore the length l divided by the width w gives the number of squares within the resistor (Fig. 1.2). The number of squares multiplied by the resistivity is the resistance.

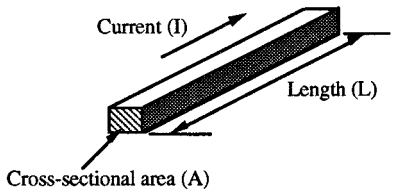


FIGURE 1.1 Resistance of a rectangular cross-section resistor with cross-sectional area A and length L.

$$R_{\text{sheet}} = \rho_{\text{sheet}} \frac{l}{w} \tag{1.2}$$

where ρ_{sheet} is the sheet resistivity (Ω/square), l is the length of resistor (cm), w is the width of the resistor (cm), and R_{sheet} is the sheet resistance (Ω).

The resistance of a resistor can be defined in terms of the **voltage drop** across the resistor and current through the resistor related by Ohm's law,

$$R = \frac{V}{I} \tag{1.3}$$

where R is the resistance (Ω), V is the voltage across the resistor (V), and I is the current through the resistor (A). Whenever a current is passed through a resistor, a voltage is dropped across the ends of the resistor. Figure 1.3 depicts the symbol of the resistor with the Ohm's law relation.

All resistors dissipate power when a voltage is applied. The power dissipated by the resistor is represented by

$$P = \frac{V^2}{R} \tag{1.4}$$

where P is the power dissipated (W), V is the voltage across the resistor (V), and R is the resistance (Ω). An ideal resistor dissipates electric energy without storing electric or magnetic energy.

Resistor Networks

Resistors may be joined to form networks. If resistors are joined in series, the effective resistance (R_T) is the sum of the individual resistances (Fig. 1.4).

$$R_T = \sum_{i=1}^{n} R_i \tag{1.5}$$

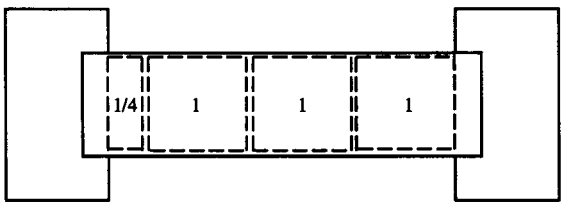


FIGURE 1.2 Number of squares in a rectangular resistor.

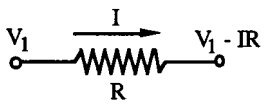


FIGURE 1.3 A resistor with resistance R having a current I flowing through it will have a voltage drop of IR across it.

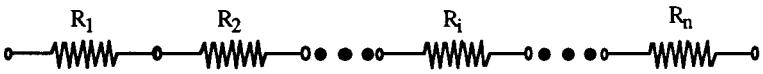


FIGURE 1.4 Resistors connected in series.

If resistors are joined in parallel, the effective resistance (R_T) is the reciprocal of the sum of the reciprocals of individual resistances (Fig. 1.5).

$$\frac{1}{R_T} = \sum_{i=1}^{n} \frac{1}{R_i} \tag{1.6}$$

Temperature Coefficient of Electrical Resistance

The resistance for most resistors changes with temperature. The temperature coefficient of electrical resistance is the change in electrical resistance of a resistor per unit change in temperature. The **temperature coefficient of resistance** is measured in Ω/°C. The temperature coefficient of resistors may be either positive or negative. A positive temperature coefficient denotes a rise in resistance with a rise in temperature; a negative temperature coefficient of resistance denotes a decrease in resistance with a rise in temperature. Pure metals typically have a positive temperature coefficient of resistance, while some metal alloys such as constantin and manganin have a zero temperature coefficient of resistance. Carbon and graphite mixed with binders usually exhibit negative temperature coefficients, although certain choices of binders and process variations may yield positive temperature coefficients. The temperature coefficient of resistance is given by

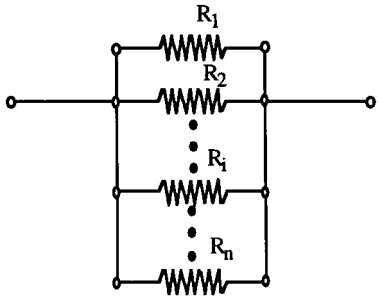


FIGURE 1.5 Resistors connected in parallel.

$$R(T_2) = R(T_1)[1 + \alpha_{T1}(T_2 - T_1)] \tag{1.7}$$

where α_{T1} is the temperature coefficient of electrical resistance at reference temperature T_1, $R(T_2)$ is the resistance at temperature T_2 (Ω), and $R(T_1)$ is the resistance at temperature T_1 (Ω). The reference temperature is usually taken to be 20°C. Because the variation in resistance between any two temperatures is usually not linear as predicted by Eq. (1.7), common practice is to apply the equation between temperature increments and then to plot the resistance change versus temperature for a number of incremental temperatures.

High-Frequency Effects

Resistors show a change in their resistance value when subjected to ac voltages. The change in resistance with voltage frequency is known as the *Boella effect*. The effect occurs because all resistors have some inductance and capacitance along with the resistive component and thus can be approximated by an equivalent circuit shown in Fig. 1.6. Even though the definition of useful frequency range is application dependent, typically, the useful range of the resistor is the highest frequency at which the impedance differs from the resistance by more than the tolerance of the resistor.

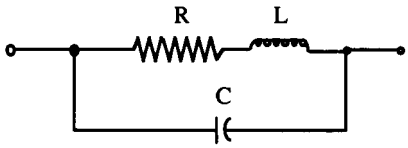


FIGURE 1.6 Equivalent circuit for a resistor.

The frequency effect on resistance varies with the resistor construction. Wire-wound resistors typically exhibit an increase in their impedance with frequency. In composition resistors the capacitances are formed by the many conducting particles which are held in contact by a dielectric binder. The ac impedance for film resistors remains constant until 100 MHz (1 MHz = 10^6 Hz) and then decreases at higher frequencies (Fig. 1.7). For film resistors, the decrease in dc resistance at higher frequencies decreases with increase in resistance. Film resistors have the most stable high-frequency performance.

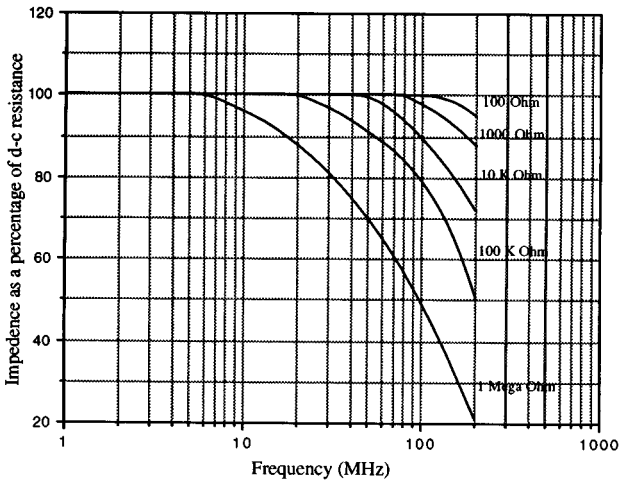


FIGURE 1.7 Typical graph of impedance as a percentage of dc resistance versus frequency for film resistors.

The smaller the diameter of the resistor the better is its frequency response. Most high-frequency resistors have a length to diameter ratio between 4:1 to 10:1. Dielectric losses are kept to a minimum by proper choice of base material.

Voltage Coefficient of Resistance

Resistance is not always independent of the applied voltage. The **voltage coefficient of resistance** is the change in resistance per unit change in voltage, expressed as a percentage of the resistance at 10% of rated voltage. The voltage coefficient is given by the relationship

$$\text{Voltage coefficient} = \frac{100(R_1 - R_2)}{R_2(V_1 - V_2)} \tag{1.8}$$

where R_1 is the resistance at the rated voltage V_1 and R_2 is the resistance at 10% of rated voltage V_2.

Noise

Resistors exhibit electrical noise in the form of small ac voltage fluctuations when dc voltage is applied. Noise in a resistor is a function of the applied voltage, physical dimensions, and materials. The total noise is a sum of Johnson noise, current flow noise, noise due to cracked bodies, and loose end caps and leads. For variable resistors the noise can also be caused by the jumping of a moving contact over turns and by an imperfect electrical path between the contact and resistance element.

The Johnson noise is temperature-dependent thermal noise (Fig. 1.8). Thermal noise is also called "white noise" because the noise level is the same at all frequencies. The magnitude of thermal noise, E_{RMS} (V), is dependent on the resistance value and the temperature of the resistance due to thermal agitation.

$$E_{RMS} = \sqrt{4kRT\Delta f} \tag{1.9}$$

where E_{RMS} is the root-mean-square value of the noise voltage (V), R is the resistance (Ω), K is the Boltzmann constant (1.38×10^{-23} J/K), T is the temperature (K), and Δf is the bandwidth (Hz) over which the noise energy is measured.

Figure 1.8 shows the variation in current noise versus voltage frequency. Current noise varies inversely with frequency and is a function of the current flowing through the resistor and the value of the resistor. The magnitude of current noise is directly proportional to the square root of current. The current noise magnitude is usually expressed by a noise index given as the ratio of the root-mean-square current noise voltage (E_{RMS})

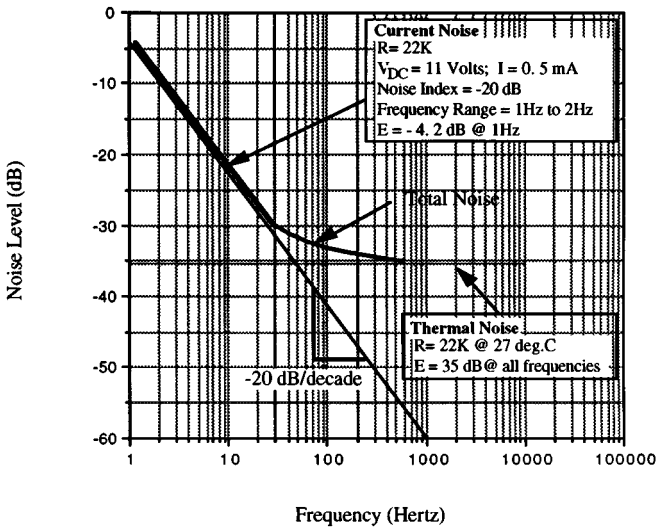


FIGURE 1.8 The total resistor noise is the sum of current noise and thermal noise. The current noise approaches the thermal noise at higher frequencies. (*Source:* Phillips Components, Discrete Products Division, *1990–91 Resistor/Capacitor Data Book,* 1991. With permission.)

over one decade bandwidth to the average voltage caused by a specified constant current passed through the resistor at a specified hot-spot temperature [Phillips, 1991].

$$\text{N.I.} = 20 \log_{10}\left(\frac{\text{Noise voltage}}{\text{dc voltage}}\right) \tag{1.10}$$

$$E_{\text{RMS}} = V_{\text{dc}} \times 10^{\text{N.I.}/20} \sqrt{\log\left(\frac{f_2}{f_1}\right)} \tag{1.11}$$

where N.I. is the noise index, V_{dc} is the dc voltage drop across the resistor, and f_1 and f_2 represent the frequency range over which the noise is being computed. Units of noise index are μV/V. At higher frequencies, the current noise becomes less dominant compared to Johnson noise.

Precision film resistors have extremely low noise. Composition resistors show some degree of noise due to internal electrical contacts between the conducting particles held together with the binder. Wire-wound resistors are essentially free of electrical noise unless resistor terminations are faulty.

Power Rating and Derating Curves

Resistors must be operated within specified temperature limits to avoid permanent damage to the materials. The temperature limit is defined in terms of the maximum power, called the *power rating,* and derating curve. The power rating of a resistor is the maximum power in watts which the resistor can dissipate. The maximum power rating is a function of resistor material, maximum voltage rating, resistor dimensions, and maximum allowable hot-spot temperature. The maximum hot-spot temperature is the temperature of the hottest part on the resistor when dissipating full-rated power at rated ambient temperature.

The maximum allowable power rating as a function of the ambient temperature is given by the derating curve. Figure 1.9 shows a typical power rating curve for a resistor. The derating curve is usually linearly drawn from the full-rated load temperature to the maximum allowable no-load temperature. A resistor may be operated at ambient temperatures above the maximum full-load ambient temperature if operating at lower than full-rated power capacity. The maximum allowable no-load temperature is also the maximum storage temperature for the resistor.

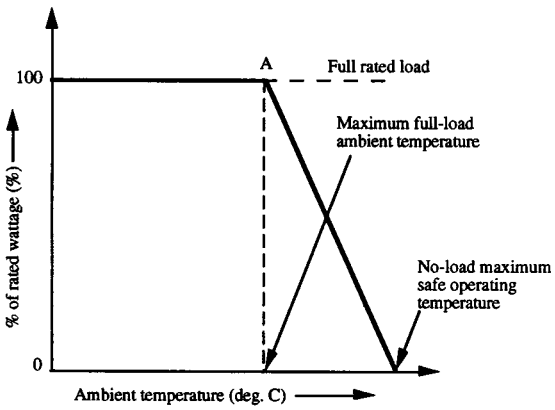


FIGURE 1.9 Typical derating curve for resistors.

Voltage Rating of Resistors

The maximum voltage that may be applied to the resistor is called the **voltage rating** and is related to the power rating by

$$V = \sqrt{PR} \tag{1.12}$$

where V is the voltage rating (V), P is the power rating (W), and R is the resistance (Ω). For a given value of voltage and power rating, a critical value of resistance can be calculated. For values of resistance below the critical value, the maximum voltage is never reached; for values of resistance above the critical value, the power dissipated is lower than the rated power (Fig. 1.10).

Color Coding of Resistors

Resistors are generally identified by color coding or direct digital marking. The color code is given in Table 1.1. The color code is commonly used in composition resistors and film resistors. The color code essentially consists of four bands of different colors. The first band is the most significant figure, the second band is the second significant figure, the third band is the multiplier or the number of zeros that have to be added after the first two significant figures, and the fourth band is the tolerance on the resistance value. If the fourth band is not present, the resistor tolerance is the standard 20% above and below the rated value. When the color code is used on fixed wire-wound resistors, the first band is applied in double width.

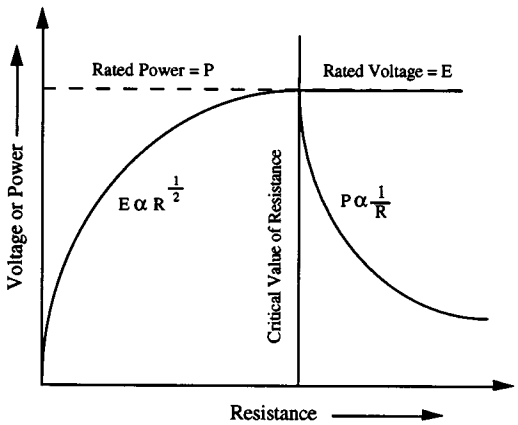


FIGURE 1.10 Relationship of applied voltage and power above and below the critical value of resistance.

TABLE 1.1 Color Code Table for Resistors

Color	First Band	Second Band	Third Band	Fourth Band Tolerance, %
Black	0	0	1	
Brown	1	1	10	
Red	2	2	100	
Orange	3	3	1,000	
Yellow	4	4	10,000	
Green	5	5	100,000	
Blue	6	6	1,000,000	
Violet	7	7	10,000,000	
Gray	8	8	100,000,000	
White	9	9	1,000,000,000	
Gold			0.1	5%
Silver			0.01	10%
No band				20%

Blanks in the table represent situations which do not exist in the color code.

Resistor Types

Resistors can be broadly categorized as fixed, variable, and special-purpose. Each of these resistor types is discussed in detail with typical ranges of their characteristics.

Fixed Resistors

The fixed resistors are those whose value cannot be varied after manufacture. Fixed resistors are classified into composition resistors, wire-wound resistors, and metal-film resistors. Table 1.2 outlines the characteristics of some typical fixed resistors.

Wire-Wound Resistors. Wire-wound resistors are made by winding wire of nickel-chromium alloy on a ceramic tube covering with a vitreous coating. The spiral winding has inductive and capacitive characteristics that make it unsuitable for operation above 50 kHz. The frequency limit can be raised by noninductive winding so that the magnetic fields produced by the two parts of the winding cancel.

Composition Resistors. Composition resistors are composed of carbon particles mixed with a binder. This mixture is molded into a cylindrical shape and hardened by baking. Leads are attached axially to each end, and the assembly is encapsulated in a protective encapsulation coating. Color bands on the outer surface indicate the resistance value and tolerance. Composition resistors are economical and exhibit low noise levels for resistances above 1 MΩ. Composition resistors are usually rated for temperatures in the neighborhood of 70°C for power ranging from 1/8 to 2 W. Composition resistors have end-to-end shunted capacitance that may be noticed at frequencies in the neighborhood of 100 kHz, especially for resistance values above 0.3 MΩ.

Metal-Film Resistors. Metal-film resistors are commonly made of nichrome, tin-oxide, or tantalum nitride, either hermetically sealed or using molded-phenolic cases. Metal-film resistors are not as stable as the

TABLE 1.2 Characteristics of Typical Fixed Resistors

Resistor Types	Resistance Range	Watt Range	Operating Temp. Range	α, ppm/°C
Wire-wound resistor				
Precision	0.1 to 1.2 MΩ	1/8 to 1/4	–55 to 145	10
Power	0.1 to 180 kΩ	1 to 210	–55 to 275	260
Metal-film resistor				
Precision	1 to 250 MΩ	1/20 to 1	–55 to 125	50–100
Power	5 to 100 kΩ	1 to 5	–55 to 155	20–100
Composition resistor				
General purpose	2.7 to 100 MΩ	1/8 to 2	–55 to 130	1500

wire-wound resistors. Depending on the application, fixed resistors are manufactured as precision resistors, semiprecision resistors, standard general-purpose resistors, or power resistors. Precision resistors have low voltage and power coefficients, excellent temperature and **time stabilities**, low noise, and very low reactance. These resistors are available in metal-film or wire constructions and are typically designed for circuits having very close resistance tolerances on values. Semiprecision resistors are smaller than precision resistors and are primarily used for current-limiting or voltage-dropping functions in circuit applications. Semiprecision resistors have long-term temperature stability. General-purpose resistors are used in circuits that do not require tight resistance tolerances or long-term stability. For general-purpose resistors, initial resistance variation may be in the neighborhood of 5% and the variation in resistance under full-rated power may approach 20%. Typically, general-purpose resistors have a high coefficient of resistance and high noise levels. Power resistors are used for power supplies, control circuits, and voltage dividers where operational stability of 5% is acceptable. Power resistors are available in wire-wound and film constructions. Film-type power resistors have the advantage of stability at high frequencies and have higher resistance values than wire-wound resistors for a given size.

Variable Resistors

Potentiometers. The potentiometer is a special form of variable resistor with three terminals. Two terminals are connected to the opposite sides of the resistive element, and the third connects to a sliding contact that can be adjusted as a voltage divider.

Potentiometers are usually circular in form with the movable contact attached to a shaft that rotates. Potentiometers are manufactured as carbon composition, metallic film, and wire-wound resistors available in single-turn or multiturn units. The movable contact does not go all the way toward the end of the resistive element, and a small resistance called the *hop-off* resistance is present to prevent accidental burning of the resistive element.

Rheostat. The rheostat is a current-setting device in which one terminal is connected to the resistive element and the second terminal is connected to a movable contact to place a selected section of the resistive element into the circuit. Typically, rheostats are wire-wound resistors used as speed controls for motors, ovens, and heater controls and in applications where adjustments on the voltage and current levels are required, such as voltage dividers and bleeder circuits.

Special-Purpose Resistors

Integrated Circuit Resistors. Integrated circuit resistors are classified into two general categories: semiconductor resistors and deposited film resistors. Semiconductor resistors use the bulk resistivity of **doped** semiconductor regions to obtain the desired resistance value. Deposited film resistors are formed by depositing resistance films on an insulating substrate which are etched and patterned to form the desired resistive network. Depending on the thickness and dimensions of the deposited films, the resistors are classified into thick-film and thin-film resistors.

Semiconductor resistors can be divided into four types: diffused, bulk, pinched, and ion-implanted. Table 1.3 shows some typical resistor properties for semiconductor resistors. Diffused semiconductor resistors use resistivity of the diffused region in the semiconductor substrate to introduce a resistance in the circuit. Both *n*-type and *p*-type diffusions are used to form the diffused resistor.

A bulk resistor uses the bulk resistivity of the semiconductor to introduce a resistance into the circuit. Mathematically the sheet resistance of a bulk resistor is given by

$$R_{\text{sheet}} = \frac{\rho_e}{d} \tag{1.13}$$

where R_{sheet} is the sheet resistance in (Ω/square), ρ_e is the sheet resistivity (Ω/square), and d is the depth of the *n*-type **epitaxial layer**.

Pinched resistors are formed by reducing the effective cross-sectional area of diffused resistors. The reduced cross section of the diffused length results in extremely high sheet resistivities from ordinary diffused resistors.

TABLE 1.3 Typical Characteristics of Integrated Circuit Resistors

Resistor Type	Sheet Resistivity (per square)	Temperature Coefficient (ppm/°C)
Semiconductor		
Diffused	0.8 to 260 Ω	1100 to 2000
Bulk	0.003 to 10 kΩ	2900 to 5000
Pinched	0.001 to 10 kΩ	3000 to 6000
Ion-implanted	0.5 to 20 kΩ	100 to 1300
Deposited resistors		
Thin-film		
Tantalum	0.01 to 1 kΩ	∓100
SnO_2	0.08 to 4 kΩ	–1500 to 0
Ni-Cr	40 to 450 Ω	∓100
Cermet (Cr-SiO)	0.03 to 2.5 kΩ	∓150
Thick-film		
Ruthenium-silver	10 Ω to 10 MΩ	∓200
Palladium-silver	0.01 to 100 kΩ	–500 to 150

Ion-implanted resistors are formed by implanting ions on the semiconductor surface by bombarding the silicon lattice with high-energy ions. The implanted ions lie in a very shallow layer along the surface (0.1 to 0.8 μm). For similar thicknesses ion-implanted resistors yield sheet resistivities 20 times greater than diffused resistors. Table 1.3 shows typical properties of diffused, bulk, pinched, and ion-implanted resistors. Typical sheet resistance values range from 80 to 250 Ω/square.

Varistors. Varistors are voltage-dependent resistors that show a high degree of nonlinearity between their resistance value and applied voltage. They are composed of a nonhomogeneous material that provides a rectifying action. Varistors are used for protection of electronic circuits, semiconductor components, collectors of motors, and relay contacts against overvoltage.

The relationship between the voltage and current of a varistor is given by

$$V = kI^{\beta} \tag{1.14}$$

where V is the voltage (V), I is the current (A), and k and β are constants that depend on the materials and manufacturing process. The electrical characteristics of a varistor are specified by its β and k values.

Varistors in Series. The resultant k value of n varistors connected in series is nk. This can be derived by considering n varistors connected in series and a voltage nV applied across the ends. The current through each varistor remains the same as for V volts over one varistor. Mathematically, the voltage and current are expressed as

$$nV = k_1 I^{\beta} \tag{1.15}$$

Equating the expressions (1.14) and (1.15), the equivalent constant k_1 for the series combination of varistors is given as

$$k_1 = nk \tag{1.16}$$

Varistors in Parallel. The equivalent k value for a parallel combination of varistors can be obtained by connecting n varistors in parallel and applying a voltage V across the terminals. The current through the varistors will still be n times the current through a single varistor with a voltage V across it. Mathematically the current and voltage are related as

$$V = k_2(nI)^{\beta} \tag{1.17}$$

From Eqs. (1.14) and (1.17) the equivalent constant k_2 for the series combination of varistors is given as

$$k_2 = \frac{k}{n^{\beta}} \tag{1.18}$$

Thermistors. Thermistors are resistors that change their resistance exponentially with changes in temperature. If the resistance decreases with increase in temperature, the resistor is called a negative temperature coefficient (NTC) resistor. If the resistance increases with temperature, the resistor is called a positive temperature coefficient (PTC) resistor.

NTC thermistors are ceramic semiconductors made by sintering mixtures of heavy metal oxides such as manganese, nickel, cobalt, copper, and iron. The resistance temperature relationship for NTC thermistors is

$$R_T = A e^{B/T} \tag{1.19}$$

where T is temperature (K), R_T is the resistance (Ω), and A, B are constants whose values are determined by conducting experiments at two temperatures and solving the equations simultaneously.

PTC thermistors are prepared from $BaTiO_3$ or solid solutions of $PbTiO_3$ or $SrTiO_3$. The resistance temperature relationship for PTC thermistors is

$$R_T = A + Ce^{BT} \tag{1.20}$$

where T is temperature (K), R_T is the resistance (Ω), and A, B are constants determined by conducting experiments at two temperatures and solving the equations simultaneously. Positive thermistors have a PTC only between certain temperature ranges. Outside this range the temperature is either zero or negative. Typically, the absolute value of the temperature coefficient of resistance for PTC resistors is much higher than for NTC resistors.

Defining Terms

Doping: The intrinsic carrier concentration of semiconductors (e.g., Si) is too low to allow controlled charge transport. For this reason some impurities called dopants are purposely added to the semiconductor. The process of adding dopants is called doping. Dopants may belong to group IIIA (e.g., boron) or group VA (e.g., phosphorus) in the periodic table. If the elements belong to the group IIIA, the resulting semiconductor is called a *p*-type semiconductor. On the other hand, if the elements belong to the group VA, the resulting semiconductor is called an *n*-type semiconductor.

Epitaxial layer: Epitaxy refers to processes used to grow a thin crystalline layer on a crystalline substrate. In the epitaxial process the wafer acts as a seed crystal. The layer grown by this process is called an epitaxial layer.

Resistivity: The resistance of a conductor with unit length and unit cross-sectional area.

Temperature coefficient of resistance: The change in electrical resistance of a resistor per unit change in temperature.

Time stability: The degree to which the initial value of resistance is maintained to a stated degree of certainty under stated conditions of use over a stated period of time. Time stability is usually expressed as a percent or parts per million change in resistance per 1000 hours of continuous use.

Voltage coefficient of resistance: The change in resistance per unit change in voltage, expressed as a percentage of the resistance at 10% of rated voltage.

Voltage drop: The difference in potential between the two ends of the resistor measured in the direction of flow of current. The voltage drop is $V = IR$, where V is the voltage across the resistor, I is the current through the resistor, and R is the resistance.

Voltage rating: The maximum voltage that may be applied to the resistor.

Related Topics

22.1 Physical Properties • 25.1 Integrated Circuit Technology • 51.1 Introduction

References

Phillips Components, Discrete Products Division, *1990–91 Resistor/Capacitor Data Book,* 1991.
C.C. Wellard, *Resistance and Resistors,* New York: McGraw-Hill, 1960.

Further Information

IEEE Transactions on Electron Devices and *IEEE Electron Device Letters:* Published monthly by the Institute of Electrical and Electronics Engineers.

IEEE Components, Hybrids and Manufacturing Technology: Published quarterly by the Institute of Electrical and Electronics Engineers.

G.W.A. Dummer, *Materials for Conductive and Resistive Functions*, New York: Hayden Book Co., 1970.
H.F. Littlejohn and C.E. Burckel, *Handbook of Power Resistors*, Mount Vernon, N.Y.: Ward Leonard Electric Company, 1951.
I.R. Sinclair, *Passive Components: A User's Guide*, Oxford: Heinmann Newnes, 1990.

1.2 Capacitors and Inductors

Glen Ballou

Capacitors

If a potential difference is found between two points, an electric **field** exists that is the result of the separation of unlike charges. The strength of the field will depend on the amount the charges have been separated.

Capacitance is the concept of energy storage in an electric field and is restricted to the area, shape, and spacing of the **capacitor** plates and the property of the material separating them.

When electrical current flows into a capacitor, a force is established between two parallel plates separated by a **dielectric.** This energy is stored and remains even after the input is removed. By connecting a **conductor** (a resistor, hard wire, or even air) across the capacitor, the charged capacitor can regain electron balance, that is, discharge its stored energy.

The value of a parallel-plate capacitor can be found with the equation

$$C = \frac{x\epsilon[(N-1)A]}{d} \times 10^{-13} \tag{1.21}$$

where C = capacitance, F; ϵ = dielectric constant of insulation; d = spacing between plates; N = number of plates; A = area of plates; and $x = 0.0885$ when A and d are in centimeters, and $x = 0.225$ when A and d are in inches.

The work necessary to transport a unit charge from one plate to the other is

$$e = kg \tag{1.22}$$

where e = volts expressing energy per unit charge, g = coulombs of charge already transported, and k = proportionality factor between work necessary to carry a unit charge between the two plates and charge already transported. It is equal to $1/C$, where C is the capacitance, F.

The value of a capacitor can now be calculated from the equation

$$C = \frac{q}{e} \tag{1.23}$$

where q = charge (C) and e is found with Eq. (1.22).

The energy stored in a capacitor is

$$W = \frac{CV^2}{2} \tag{1.24}$$

where W = energy, J; C = capacitance, F; and V = applied voltage, V.

The **dielectric constant** of a material determines the electrostatic energy which may be stored in that material per unit volume for a given voltage. The value of the dielectric constant expresses the ratio of a capacitor in a vacuum to one using a given dielectric. The dielectric of air is 1, the reference unit employed for expressing the dielectric constant. As the dielectric constant is increased or decreased, the capacitance will increase or decrease, respectively. Table 1.4 lists the dielectric constants of various materials.

The dielectric constant of most materials is affected by both temperature and frequency, except for quartz, Styrofoam, and Teflon, whose dielectric constants remain essentially constant.

The equation for calculating the *force of attraction* between two plates is

$$F = \frac{AV^2}{k(1504S)^2} \tag{1.25}$$

TABLE 1.4 Comparison of Capacitor Dielectric Constants

Dielectric	K (Dielectric Constant)
Air or vacuum	1.0
Paper	2.0–6.0
Plastic	2.1–6.0
Mineral oil	2.2–2.3
Silicone oil	2.7–2.8
Quartz	3.8–4.4
Glass	4.8–8.0
Porcelain	5.1–5.9
Mica	5.4–8.7
Aluminum oxide	8.4
Tantalum pentoxide	26
Ceramic	12–400,000

Source: G. Ballou, *Handbook for Sound Engineers, The New Audio Cyclopedia,* Carmel, Ind.: Macmillan Computer Publishing Company, 1991. With permission.

where F = attraction force, dyn; A = area of one plate, cm^2; V = potential energy difference, V; k = dielectric coefficient; and S = separation between plates, cm.

The Q for a capacitor when the resistance and capacitance is in series is

$$Q = \frac{1}{2\pi fRC} \tag{1.26}$$

where Q = ratio expressing the factor of merit; f = frequency, Hz; R = resistance, Ω; and C = capacitance, F.

When capacitors are connected in *series,* the total capacitance is

$$C_T = \frac{1}{1/C_1 + 1/C_2 + \cdots + 1/C_n} \tag{1.27}$$

and is always less than the value of the smallest capacitor.

When capacitors are connected in *parallel*, the total capacitance is

$$C_T = C_1 + C_2 + \cdots + C_n \tag{1.28}$$

and is always larger than the largest capacitor.

When a voltage is applied across a group of capacitors connected in series, the voltage drop across the combination is equal to the applied voltage. The drop across each individual capacitor is inversely proportional to its capacitance.

$$V_C = \frac{V_A C_X}{C_T} \tag{1.29}$$

where V_C = voltage across the individual capacitor in the series (C_1, C_2, ...,C_n), V; V_A = applied voltage, V; C_T = total capacitance of the series combination, F; and C_X = capacitance of individual capacitor under consideration, F.

In an ac circuit, the **capacitive reactance,** or the **impedance**, of the capacitor is

$$X_C = \frac{1}{2\pi fC} \tag{1.30}$$

where X_C = capacitive reactance, Ω; f = frequency, Hz; and C = capacitance, F. The current will lead the voltage by 90° in a circuit with a pure capacitor.

When a dc voltage is connected across a capacitor, a time t is required to charge the capacitor to the applied voltage. This is called a **time constant** and is calculated with the equation

$$t = RC \tag{1.31}$$

where t = time, s; R = resistance, Ω; and C = capacitance, F.

In a circuit consisting of pure resistance and capacitance, the *time constant t* is defined as the time required to charge the capacitor to 63.2% of the applied voltage.

During the next time constant, the capacitor charges to 63.2% of the remaining difference of full value, or to 86.5% of the full value. The charge on a capacitor can never actually reach 100% but is considered to be 100% after five time constants. When the voltage is removed, the capacitor discharges to 63.2% of the full value.

Capacitance is expressed in microfarads (μF, or 10^{-6} F) or picofarads (pF, or 10^{-12} F) with a stated accuracy or tolerance. Tolerance may also be stated as GMV (guaranteed minimum value), sometimes referred to as MRV (minimum rated value).

All capacitors have a *maximum working voltage* that must not be exceeded and is a combination of the dc value plus the peak ac value which may be applied during operation.

Quality Factor (*Q*)

Quality factor is the ratio of the capacitor's **reactance** to its resistance at a specified frequency and is found by the equation

$$Q = \frac{1}{2\pi fCR} = \frac{1}{PF} \tag{1.32}$$

where Q = quality factor; f = frequency, Hz; C = value of capacitance, F; R = internal resistance, Ω; and PF = power factor

Power Factor (PF)

Power factor is the preferred measurement in describing capacitive losses in ac circuits. It is the fraction of input volt-amperes (or power) dissipated in the capacitor dielectric and is virtually independent of the capacitance, applied voltage, and frequency.

Equivalent Series Resistance (ESR)

Equivalent series resistance is expressed in ohms or milliohms (Ω, mΩ) and is derived from lead resistance, termination losses, and dissipation in the dielectric material.

Equivalent Series Inductance (ESL)

The *equivalent series inductance* can be useful or detrimental. It reduces high-frequency performance; however, it can be used in conjunction with the internal capacitance to form a resonant circuit.

Dissipation Factor (DF)

The **dissipation factor** in percentage is the ratio of the effective series resistance of a capacitor to its reactance at a specified frequency. It is the reciprocal of *quality factor* (Q) and an indication of power loss within the capacitor. It should be as low as possible.

Insulation Resistance

Insulation resistance is the resistance of the dielectric material and determines the time a capacitor, once charged, will hold its charge. A discharged capacitor has a low insulation resistance; however once charged to its rated value, it increases to megohms. The leakage in electrolytic capacitors should not exceed

$$I_L = 0.04C + 0.30 \tag{1.33}$$

where I_L = leakage current, μA, and C = capacitance, μF.

Dielectric Absorption (DA)

The *dielectric absorption* is a reluctance of the dielectric to give up stored electrons when the capacitor is discharged. This is often called "memory" because if a capacitor is discharged through a resistance and the resistance is removed, the electrons that remained in the dielectric will reconvene on the electrode, causing a voltage to appear across the capacitor. DA is tested by charging the capacitor for 5 min, discharging it for 5 s, then having an open circuit for 1 min after which the recovery voltage is read. The percentage of DA is defined as the ratio of recovery to charging voltage times 100.

Types of Capacitors

Capacitors are used to filter, couple, tune, block dc, pass ac, bypass, shift phase, compensate, feed through, isolate, store energy, suppress noise, and start motors. They must also be small, lightweight, reliable, and withstand adverse conditions.

Capacitors are grouped according to their dielectric material and mechanical configuration.

Ceramic Capacitors

Ceramic capacitors are used most often for bypass and coupling applications (Fig. 1.11). Ceramic capacitors can be produced with a variety of K values (dielectric constant). A high K value translates to small size and less stability. High-K capacitors with a dielectric constant >3000 are physically small and have values between 0.001 to several microfarads.

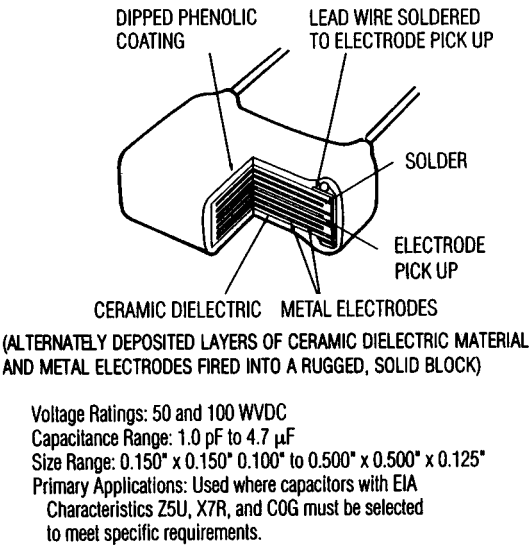


FIGURE 1.11 Monolythic® multilayer ceramic capacitors. (Courtesy of Sprague Electric Company.)

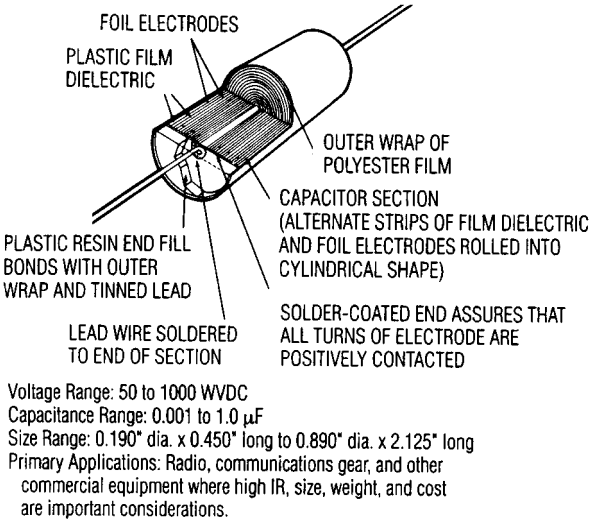


FIGURE 1.12 Film-wrapped film capacitors. (Courtesy of Sprague Electric Company.)

Good temperature stability requires capacitors to have a K value between 10 and 200. If high Q is also required, the capacitor will be physically larger. Ceramic capacitors with a zero temperature change are called **negative-positive-zero (NPO)** and come in a capacitance range of 1.0 pF to 0.033 μF.

An N750 temperature-compensated capacitor is used when accurate capacitance is required over a large temperature range. The 750 indicates a 750-ppm decrease in capacitance with a 1°C increase in temperature (750 ppm/°C). This equates to a 1.5% decrease in capacitance for a 20°C temperature increase. N750 capacitors come in values between 4.0 and 680 pF.

Film Capacitors

Film capacitors consist of alternate layers of metal foil and one or more layers of a flexible plastic insulating material (dielectric) in ribbon form rolled and encapsulated (see Fig. 1.12).

Mica Capacitors

Mica capacitors have small capacitance values and are usually used in high-frequency circuits. They are constructed as alternate layers of metal foil and mica insulation, which are stacked and encapsulated, or are silvered mica, where a silver electrode is screened on the mica insulators.

Paper-Foil-Filled Capacitors

Paper-foil-filled capacitors are often used as motor capacitors and are rated at 60 Hz. They are made of alternate layers of aluminum and paper saturated with oil that are rolled together. The assembly is mounted in an oil-filled, hermetically sealed metal case.

Electrolytic Capacitors

Electrolytic capacitors provide high capacitance in a tolerable size; however, they do have drawbacks. Low temperatures reduce performance, while high temperatures dry them out. The **electrolytes** themselves can leak and corrode the equipment. Repeated surges above the rated working voltage, excessive ripple currents, and high operating temperature reduce performance and shorten capacitor life.

Electrolytic capacitors are manufactured by an electrochemical formation of an oxide film on a metal surface. The metal on which the oxide film is formed serves as the **anode** or positive terminal of the capacitor; the oxide film is the dielectric, and the **cathode** or negative terminal is either a conducting liquid or a gel.

The equivalent circuit of an electrolytic capacitor is shown in Fig. 1.13, where A and B are the capacitor terminals, C is the effective capacitance, and L is the self-inductance of the capacitor caused by terminals, electrodes, and geometry.

The shunt resistance (insulation resistance) R_s accounts for the dc leakage current. Heat is generated in the ESR from ripple current and in the shunt resistance by voltage. The ESR is due to the spacer-electrolyte-oxide system and varies only slightly except at low temperature, where it increases greatly.

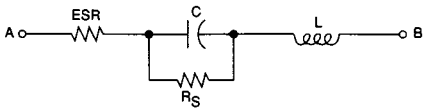


FIGURE 1.13 Simplified equivalent circuit of an electrolytic capacitor.

The *impedance* of a capacitor (Fig. 1.14) is frequency-dependent. The initial downward slope is caused by the capacitive reactance X_C. The trough (lowest impedance) is almost totally resistive, and the upward slope is due to the capacitor's self-inductance X_L. An ESR plot would show an ESR decrease to about 5–10 kHz, remaining relatively constant thereafter.

Leakage current is the direct current that passes through a capacitor when a correctly polarized dc voltage is applied to its terminals. It is proportional to temperature, becoming increasingly important at elevated **ambient temperatures**. Leakage current decreases slowly after voltage is applied, reaching steady-state conditions in about 10 min.

If a capacitor is connected with reverse polarity, the oxide film is forward-biased, offering very little resistance to current flow. This causes overheating and self-destruction of the capacitor.

The total heat generated within a capacitor is the sum of the heat created by the $I_{\text{leakage}} \times V_{\text{applied}}$ and the I^2R losses in the ESR.

The ac **ripple current** rating is very important in filter applications because excessive current produces temperature rise, shortening capacitor life. The maximum permissible rms ripple current is limited by the internal temperature and the rate of heat dissipation from the capacitor. Lower ESR and longer enclosures increase the ripple current rating.

Capacitor life expectancy is doubled for each 10°C decrease in operating temperature, so a capacitor operating at room temperature will have a life expectancy 64 times that of the same capacitor operating at 85°C (185°F).

The *surge voltage* specification of a capacitor determines its ability to withstand high transient voltages that generally occur during the starting up period of equipment. Standard tests generally specify a short on and long off period for an interval of 24 h or more, and the allowable surge voltage levels are generally 10% above the rated voltage of the capacitor.

Figure 1.15 shows how temperature, frequency, time, and applied voltage affect electrolytic capacitors.

Aluminum Electrolytic Capacitors. *Aluminum electrolytic capacitors* use aluminum as the base material (Fig. 1.16). The surface is often etched to increase the surface area as much as 100 times that of unetched foil, resulting in higher capacitance in the same volume.

Aluminum electrolytic capacitors can withstand up to 1.5 V of reverse voltage without detriment. Higher reverse voltages, when applied over extended periods, lead to loss of capacitance. Excess reverse voltages applied for short periods cause some change in capacitance but not to capacitor failure.

Large-value capacitors are often used to filter dc power supplies. After a capacitor is charged, the rectifier stops conducting and the capacitor discharges into the load, as shown in Fig. 1.17, until the next cycle. Then the capacitor recharges again to the peak voltage. The Δe is equal to the total peak-to-peak ripple voltage and is a complex wave containing many harmonics of the fundamental ripple frequency, causing the noticeable heating of the capacitor.

Tantalum Capacitors. *Tantalum electrolytics* are the preferred type where high reliability and long service life are paramount considerations.

Tantalum capacitors have as much as three times better capacitance per volume efficiency than aluminum electrolytic capacitors, because tantalum pentoxide has a dielectric constant three times greater than that of aluminum oxide (see Table 1.4).

The capacitance of any capacitor is determined by the surface area of the two conducting plates, the

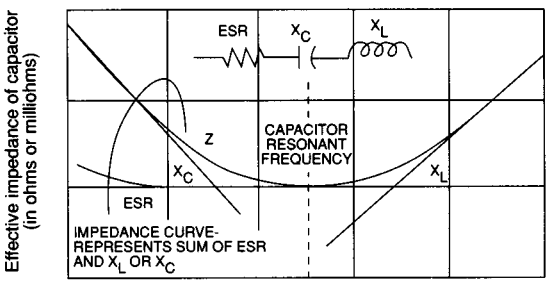


Figure 1.14 Impedance characteristics of a capacitor.

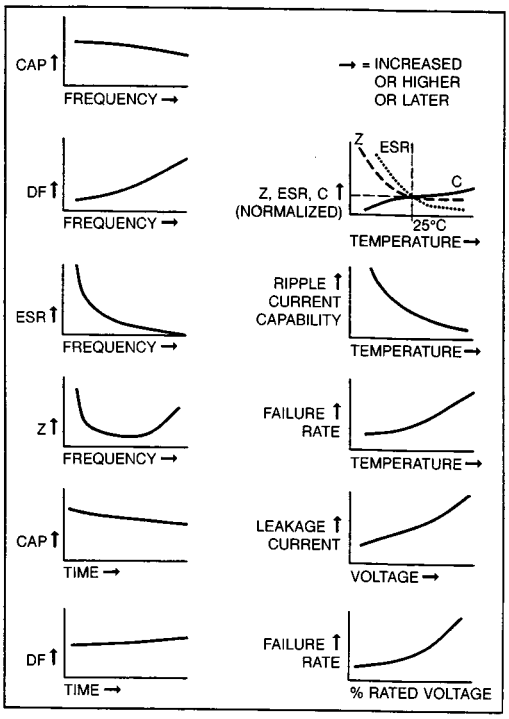
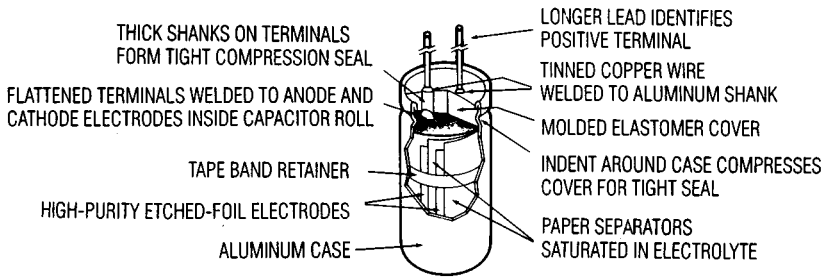


FIGURE 1.15 Variations in aluminum electrolytic characteristics caused by temperature, frequency, time, and applied voltage. (Courtesy of Sprague Electric Company.)



Voltage Range: 6.3 to 63 WVDC
Capacitance Range: 0.47 to 3300 μF
Size Range: 0.197" dia. x 0.433" long to 0.630" dia. x 1.614" long
Primary Applications: Coupling, decoupling, bypass, and filtering.
Vertical installation on high-density printed wiring boards in transistorized radios, portable TV sets, auto radios, tape recorders, etc.

(Courtesy

FIGURE 1.16 Verti-lytic® miniature single-ended aluminum electrolytic capacitor. (Courtesy of Sprague Electric Company.)

distance between the plates, and the dielectric constant of the insulating material between the plates [see Eq. (1.21)].

In tantalum electrolytics, the distance between the plates is the thickness of the tantalum pentoxide film, and since the dielectric constant of the tantalum pentoxide is high, the capacitance of a tantalum capacitor is high.

Tantalum capacitors contain either liquid or solid electrolytes. The liquid electrolyte in wet-slug and foil capacitors, generally sulfuric acid, forms the cathode (negative) plate. In solid-electrolyte capacitors, a dry material, manganese dioxide, forms the cathode plate.

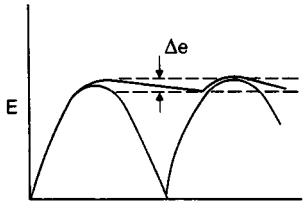
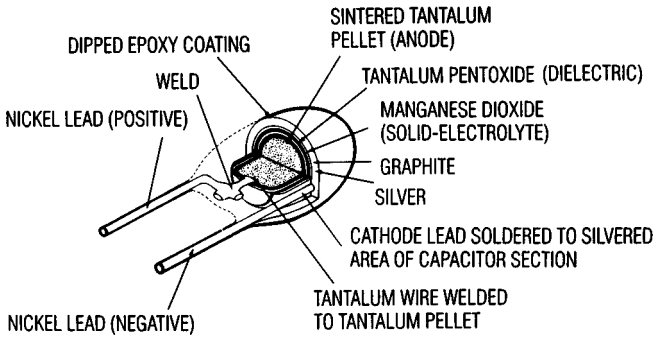


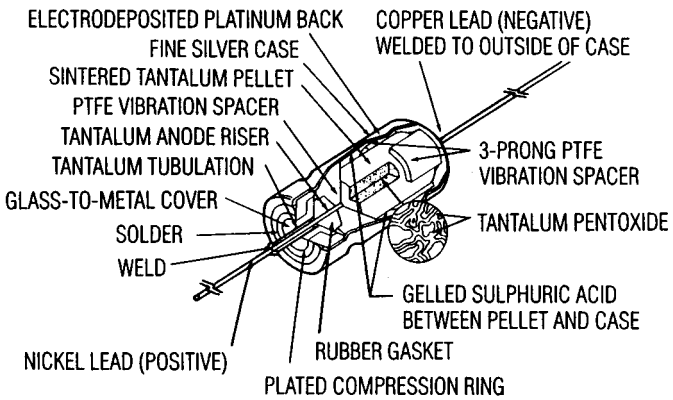
FIGURE 1.17 Full-wave capacitor charge and discharge.

Foil Tantalum Capacitors. *Foil tantalum capacitors* can be designed to voltage values up to 300 V dc. Of the three types of tantalum electrolytic capacitors, the foil design has the lowest capacitance per unit volume and is best suited for the higher voltages primarily found in older designs of equipment. It is expensive and used only where neither a solid-electrolyte (Fig. 1.18) nor a wet-slug (Fig. 1.19) tantalum capacitor can be employed.



Voltage Range: 3 to 50 WVDC
Capacitance Range: 0.10 to 680 μF
Size Range: 0.175" dia. x 0.280" high to 0.400" dia. x 0.750" high
Primary Applications: For printed wiring boards applications where low cost, small size, high stability, low d-c leakage , and low dissipation factor are important.

FIGURE 1.18 Tantalex® solid electrolyte tantalum capacitor. (Courtesy of Sprague Electric Company.)



Voltage Range: 6 to 125 WVDC
Capacitance Range: 1.7 to 1200 μF
Size Range: 0.188" dia. x 0.453" long to 0.375" dia. x 1.062" long
Primary Applications: Industrial and military equipment where reliability and premium performance with respect to low d-c leakage current, high inrush current capability, and high volumetric efficiency.

FIGURE 1.19 Hermetically sealed sintered-anode tantalum capacitor. (Courtesy of Sprague Electric Company.)

Foil tantalum capacitors are generally designed for operation over the temperature range of –55 to +125°C (–67 to +257°F) and are found primarily in industrial and military electronics equipment.

Solid-electrolyte sintered-anode tantalum capacitors differ from the wet versions in their electrolyte, which is manganese dioxide.

Another variation of the solid-electrolyte tantalum capacitor encases the element in plastic resins, such as epoxy materials offering excellent reliability and high stability for consumer and commercial electronics with the added feature of low cost.

Still other designs of "solid tantalum" capacitors use plastic film or sleeving as the encasing material, and others use metal shells that are backfilled with an epoxy resin. Finally, there are small tubular and rectangular molded plastic encasements.

Wet-electrolyte sintered-anode tantalum capacitors, often called "wet-slug" tantalum capacitors, use a pellet of sintered tantalum powder to which a lead has been attached, as shown in Fig. 1.19. This anode has an enormous surface area for its size.

Wet-slug tantalum capacitors are manufactured in a voltage range to 125 V dc.

Use Considerations. Foil tantalum capacitors are used only where high-voltage constructions are required or where there is substantial reverse voltage applied to a capacitor during circuit operation.

Wet sintered-anode capacitors, or "wet-slug" tantalum capacitors, are used where low dc leakage is required. The conventional "silver can" design will not tolerate reverse voltage. In military or aerospace applications where utmost reliability is desired, tantalum cases are used instead of silver cases. The tantalum-cased wet-slug units withstand up to 3 V reverse voltage and operate under higher ripple currents and at temperatures up to 200°C (392°F).

Solid-electrolyte designs are the least expensive for a given rating and are used where their very small size is important. They will typically withstand a reverse voltage up to 15% of the rated dc working voltage. They also have good low-temperature performance characteristics and freedom from corrosive electrolytes.

Inductors

Inductance is used for the storage of magnetic energy. Magnetic energy is stored as long as current keeps flowing through the inductor. In a perfect **inductor**, the current of a sine wave lags the voltage by 90°.

Impedance

Inductive reactance X_L, the impedance of an inductor to an ac signal, is found by the equation

$$X_L = 2\pi fL \tag{1.34}$$

where X_L = inductive reactance, Ω; f = frequency, Hz; and L = inductance, H.

The type of wire used for its construction does not affect the inductance of a **coil**. Q of the coil will be governed by the resistance of the wire. Therefore coils wound with silver or gold wire have the highest Q for a given design.

To increase inductance, inductors are connected in series. The total inductance will always be greater than the largest inductor.

$$L_T = L_1 + L_2 + \cdots + L_n \tag{1.35}$$

To reduce inductance, inductors are connected in parallel.

$$L_T = \frac{1}{1/L_1 + 1/L_2 + \cdots + 1/L_n} \tag{1.36}$$

The total inductance will always be less than the value of the lowest inductor.

Mutual Inductance

Mutual inductance is the property that exists between two conductors carrying current when their magnetic lines of force link together.

The mutual inductance of two coils with fields interacting can be determined by the equation

$$M = \frac{L_A - L_B}{4} \tag{1.37}$$

where M = mutual inductance of L_A and L_B, H; L_A = total inductance, H, of coils L_1 and L_2 with fields aiding; and L_B = total inductance, H, of coils L_1 and L_2 with fields opposing.

The *coupled inductance* can be determined by the following equations. In parallel with fields aiding,

$$L_T = \frac{1}{\dfrac{1}{L_1 + M} + \dfrac{1}{L_2 + M}} \tag{1.38}$$

In parallel with fields opposing,

$$L_T = \frac{1}{\dfrac{1}{L_1 - M} - \dfrac{1}{L_2 - M}} \tag{1.39}$$

In series with fields aiding,

$$L_T = L_1 + L_2 + 2M \tag{1.40}$$

In series with fields opposing,

$$L_T = L_1 + L_2 - 2M \tag{1.41}$$

where L_T = total inductance, H; L_1 and L_2 = inductances of the individual coils, H; and M = mutual inductance, H.

When two coils are inductively coupled to give transformer action, the coupling coefficient is determined by

$$K = \frac{M}{\sqrt{L_1 L_2}} \tag{1.42}$$

where K = coupling coefficient; M = mutual inductance, H; and L_1 and L_2 = inductances of the two coils, H.

An inductor in a circuit has a reactance equal to $j2\pi fL\ \Omega$. Mutual inductance in a circuit has a reactance equal to $j2\pi fL\ \Omega$. The operator j denotes that the reactance dissipates no energy; however, it does oppose current flow.

The energy stored in an inductor can be determined by the equation

$$W = \frac{LI^2}{2} \tag{1.43}$$

where W = energy, J (W · s); L = inductance, H; and I = current, A.

Coil Inductance

Inductance is related to the turns in a coil as follows:

1. The inductance is proportional to the square of the turns.
2. The inductance increases as the length of the **winding** is increased.
3. A shorted turn decreases the inductance, affects the frequency response, and increases the insertion loss.
4. The inductance increases as the permeability of the core material increases.
5. The inductance increases with an increase in the cross-sectional area of the core material.
6. Inductance is increased by inserting an iron core into the coil.
7. Introducing an air gap into a choke reduces the inductance.

A conductor moving at any angle to the lines of force cuts a number of lines of force proportional to the sine of the angles. Thus,

$$V = \beta L v \sin \theta \times 10^{-8} \tag{1.44}$$

where β = flux density; L = length of the conductor, cm; and v = velocity, cm/s, of conductor moving at an angle θ.

The maximum voltage induced in a conductor moving in a magnetic field is proportional to the number of magnetic lines of force cut by that conductor. When a conductor moves parallel to the lines of force, it cuts no lines of force; therefore, no current is generated in the conductor. A conductor that moves at right angles to the lines of force cuts the maximum number of lines per inch per second, therefore creating a maximum voltage. The right-hand rule determines direction of the induced electromotive force (emf). The emf is in the direction in which the axis of a right-hand screw, when turned with the velocity vector, moves through the smallest angle toward the flux density vector.

The **magnetomotive force** (mmf) in **ampere-turns** produced by a coil is found by multiplying the number of turns of wire in the coil by the current flowing through it.

$$\begin{aligned} \text{Ampere-turns} &= T\left(\frac{V}{R}\right) \\ &= TI \end{aligned} \tag{1.45}$$

where T = number of turns; V = voltage, V; and R = resistance, Ω.

The inductance of a single layer, a spiral, and multilayer coils can be calculated by using either Wheeler's or Nagaoka's equations. The accuracy of the calculation will vary between 1 and 5%. The inductance of a single-layer coil can be calculated using Wheeler's equation:

$$L = \frac{B^2 N^2}{9B + 10A} \quad \mu\text{H} \tag{1.46}$$

For the multilayer coil,

$$L = \frac{0.8 B^2 N^2}{6B + 9A + 10C} \quad \mu\text{H} \tag{1.47}$$

For the spiral coil,

$$L = \frac{B^2 N^2}{8B + 11C} \quad \mu\text{H} \tag{1.48}$$

where B = radius of the winding, N = number of turns in the coil, A = length of the winding, and C = thickness of the winding.

Q

Q is the ratio of the inductive reactance to the internal resistance of the coil and is affected by frequency, inductance, dc resistance, inductive reactance, the type of winding, the core losses, the distributed capacity, and the permeability of the core material.

The Q for a coil where R and L are in series is

$$Q = \frac{2\pi fL}{R} \tag{1.49}$$

where f = frequency, Hz; L = inductance, H; and R = resistance, Ω.

The Q of the coil can be measured using the circuit of Fig. 1.20 for frequencies up to 1 MHz. The voltage across the inductance (L) at resonance equals $Q(V)$ (where V is the voltage developed by the oscillator); therefore, it is only necessary to measure the output voltage from the oscillator and the voltage across the inductance.

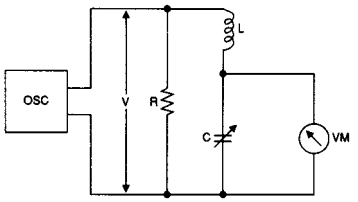


FIGURE 1.20 Circuit for measuring the Q of a coil.

The oscillator voltage is driven across a low value of resistance, R, about 1/100 of the anticipated rf resistance of the LC combination, to assure that the measurement will not be in error by more than 1%. For most measurements, R will be about 0.10 Ω and should have a voltage of 0.1 V. Most oscillators cannot be operated into this low impedance, so a step-down matching transformer must be employed. Make C as large as convenient to minimize the ratio of the impedance looking from the voltmeter to the impedance of the test circuit. The LC circuit is then tuned to resonate and the resultant voltage measured. The value of Q may then be equated

$$Q = \frac{\text{resonant voltage across } C}{\text{voltage across } R} \tag{1.50}$$

The Q of any coil may be approximated by the equation

$$\begin{aligned} Q &= \frac{2\pi fL}{R} \\ &= \frac{X_L}{R} \end{aligned} \tag{1.51}$$

where f = the frequency, Hz; L = the inductance, H; R = the dc resistance, Ω (as measured by an ohmmeter); and X_L = the inductive reactance of the coil.

Time Constant

When a dc voltage is applied to an RL circuit, a certain amount of time is required to change the circuit [see text with Eq. (1.31)]. The time constant can be determined with the equation

$$T = \frac{L}{R} \tag{1.52}$$

where R = resistance, Ω; L = inductance, H; and T = time, s.

The *right-hand rule* is used to determine the direction of a magnetic field around a conductor carrying a direct current. Grasp the conductor in the right hand with the thumb extending along the conductor pointing in the direction of the current. With the fingers partly closed, the finger tips will point in the direction of the magnetic field.

Maxwell's rule states, "If the direction of travel of a right-handed corkscrew represents the direction of the current in a straight conductor, the direction of rotation of the corkscrew will represent the direction of the magnetic lines of force."

Impedance

The total impedance created by resistors, capacitors, and inductors in circuits can be determined with the following equations.

For resistance and capacitance in series,

$$Z = \sqrt{R^2 + X_C^2} \tag{1.53}$$

$$\theta = \arctan \frac{X_C}{R} \tag{1.54}$$

For resistance and inductance in series,

$$Z = \sqrt{R^2 + X_L^2} \tag{1.55}$$

$$\theta = \arctan \frac{X_L}{R} \tag{1.56}$$

For inductance and capacitance in series,

$$Z = \begin{cases} X_L - X_C & \text{when } X_L > X_C \quad (1.57) \\ X_C - X_L & \text{when } X_C > X_L \quad (1.58) \end{cases}$$

For resistance, inductance, and capacitance in series,

$$Z = \sqrt{R^2 + (X_L - X_C)^2} \tag{1.59}$$

$$\theta = \arctan \frac{X_L - X_C}{R} \tag{1.60}$$

For capacitance and resistance in parallel,

$$Z = \frac{RX_C}{\sqrt{R^2 + X_C^2}} \tag{1.61}$$

For resistance and inductance in parallel,

$$Z = \frac{RX_L}{\sqrt{R^2 + X_L^2}} \tag{1.62}$$

For capacitance and inductance in parallel,

$$Z = \begin{cases} \dfrac{X_L X_C}{X_L - X_C} & \text{when } X_L > X_C \quad (1.63) \\ \dfrac{X_C X_L}{X_C - X_L} & \text{when } X_C > X_L \quad (1.64) \end{cases}$$

For inductance, capacitance, and resistance in parallel,

$$Z = \frac{RX_L X_C}{\sqrt{X_L^2 X_C^2 + R^2 (X_L - X_C)^2}} \quad (1.65)$$

$$\theta = \arctan \frac{R(X_L - X_C)}{X_L X_C} \quad (1.66)$$

For inductance and series resistance in parallel with resistance,

$$Z = R_2 \sqrt{\frac{R_1^2 + X_L^2}{(R_1 + R_2)^2 + X_L^2}} \quad (1.67)$$

$$\theta = \arctan \frac{X_L R_2}{R_1^2 + X_L^2 + R_1 R_2} \quad (1.68)$$

For inductance and series resistance in parallel with capacitance,

$$Z = X_C \sqrt{\frac{R^2 + X_L^2}{R^2 + (X_L - X_C)^2}} \quad (1.69)$$

$$\theta = \arctan \frac{X_L (X_C - X_L) - R^2}{RX_C} \quad (1.70)$$

For capacitance and series resistance in parallel with inductance and series resistance,

$$Z = \sqrt{\frac{(R_1^2 + X_L^2)(R_2^2 + X_C^2)}{(R_1 + R_2)^2 + (X_L - X_C)^2}} \quad (1.71)$$

$$\theta = \arctan \frac{X_L (R_2^2 + X_C^2) - X_C (R_1^2 + X_L^2)}{R_1 (R_2^2 + X_C^2) + R_2 (R_1^2 + X_L^2)} \quad (1.72)$$

where Z = impedance, Ω; R = resistance, Ω; L = inductance, H; X_L = inductive reactance, Ω; X_C = capacitive reactance, Ω; and θ = **phase** angle, degrees, by which current leads voltage in a capacitive circuit or lags voltage in an inductive circuit (0° indicates an in-phase condition).

Resonant Frequency

When an inductor and capacitor are connected in series or parallel, they form a resonant circuit. The **resonant frequency** can be determined from the equation

$$\begin{aligned} f &= \frac{1}{2\pi\sqrt{LC}} \\ &= \frac{1}{2\pi C X_C} \\ &= \frac{X_L}{2\pi L} \end{aligned} \tag{1.73}$$

where f = frequency, Hz; L = inductance, H; C = capacitance, F; and X_L, X_C = impedance, Ω.

The resonant frequency can also be determined through the use of a reactance chart developed by the Bell Telephone Laboratories (Fig. 1.21). This chart can be used for solving problems of inductance, capacitance, frequency, and impedance. If two of the values are known, the third and fourth values may be found with its use.

Defining Terms

Air capacitor: A fixed or variable capacitor in which air is the dielectric material between the capacitor's plates.

Ambient temperature: The temperature of the air or liquid surrounding any electrical part or device. Usually refers to the effect of such temperature in aiding or retarding removal of heat by radiation and convection from the part or device in question.

Ampere-turns: The magnetomotive force produced by a coil, derived by multiplying the number of turns of wire in a coil by the current (A) flowing through it.

Anode: The positive electrode of a capacitor.

Capacitive reactance: The opposition offered to the flow of an alternating or pulsating current by capacitance measured in ohms.

Capacitor: An electrical device capable of storing electrical energy and releasing it at some predetermined rate at some predetermined time. It consists essentially of two conducting surfaces (electrodes) separated by an insulating material or dielectric. A capacitor stores electrical energy, blocks the flow of direct current, and permits the flow of alternating current to a degree dependent essentially upon capacitance and frequency. The amount of energy stored, $E = 0.5\ CV^2$.

Cathode: The capacitor's negative electrode.

Coil: Anumber of turns of wire in the form of a spiral. The spiral may be wrapped around an iron core or an insulating form, or it may be self-supporting. A coil offers considerable opposition to ac current but very little to dc current.

Conductor: Abare or insulated wire or combination of wires not insulated from one another, suitable for carrying an electric current.

Dielectric: The insulating (nonconducting) medium between the two electrodes (plates) of a capacitor.

Dielectric constant: The ratio of the capacitance of a capacitor with a given dielectric to that of the same capacitor having a vacuum dielectric.

Disk capacitor: A small single-layer ceramic capacitor with a dielectric insulator consisting of conductively silvered opposing surfaces.

Dissipation factor (DF): The ratio of the effective series resistance of a capacitor to its reactance at a specified frequency measured in percent.

Electrolyte: Current-conducting solution between two electrodes or plates of a capacitor, at least one of which is covered by a dielectric.

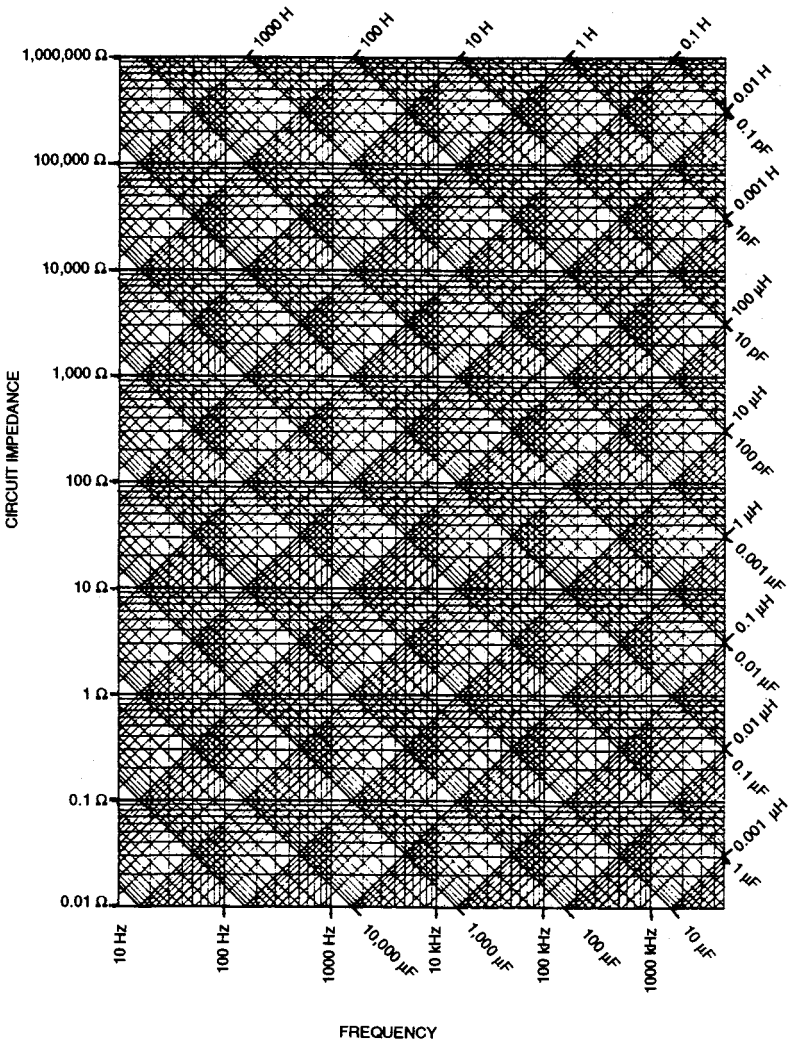


FIGURE 1.21 Reactance chart. (Courtesy AT&T Bell Laboratories.)

Electrolytic capacitor: A capacitor solution between two electrodes or plates of a capacitor, at least one of which is covered by a dielectric.

Equivalent series resistance (ESR): All internal series resistance of a capacitor concentrated or "lumped" at one point and treated as one resistance of a capacitor regardless of source, i.e., lead resistance, termination losses, or dissipation in the dielectric material.

Farad: The basic unit of measure in capacitors. Acapacitor charged to 1 volt with a charge of 1 coulomb (1 ampere flowing for 1 second) has a capacitance of 1 farad.

Field: Ageneral term referring to the region under the influence of a physical agency such as electricity, magnetism, or a combination produced by an electrical charged object.

Impedance (Z): Total opposition offered to the flow of an alternating or pulsating current measured in ohms. (Impedance is the vector sum of the resistance and the capacitive and inductive reactance, i.e., the ratio of voltage to current.)

Inductance: The property which opposes any change in the existing current. Inductance is present only when the current is changing.

Inductive reactance (X_L): The opposition to the flow of alternating or pulsating current by the inductance of a circuit.

Inductor: Aconductor used to introduce inductance into a circuit.

Leakage current: Stray direct current of relatively small value which flows through a capacitor when voltage is impressed across it.

Magnetomotive force: The force by which the magnetic field is produced, either by a current flowing through a coil of wire or by the proximity of a magnetized body. The amount of magnetism produced in the first method is proportional to the current through the coil and the number of turns in it.

Mutual inductance: The property that exists between two current-carrying conductors when the magnetic lines of force from one link with those from another.

Negative-positive-zero (NPO): An ultrastable temperature coefficient (±30 ppm/°C from –55 to 125°C) temperature-compensating capacitor.

Phase: The angular relationship between current and voltage in an ac circuit. The fraction of the period which has elapsed in a periodic function or wave measured from some fixed origin. If the time for one period is represented as 360° along a time axis, the phase position is called phase angle.

Polarized capacitor: An electrolytic capacitor in which the dielectric film is formed on only one metal electrode. The impedance to the flow of current is then greater in one direction than in the other. Reversed polarity can damage the part if excessive current flow occurs.

Power factor (PF): The ratio of effective series resistance to impedance of a capacitor, expressed as a percentage.

Quality factor (*Q*): The ratio of the reactance to its equivalent series resistance.

Reactance (*X*): Opposition to the flow of alternating current. Capacitive reactance (X_c) is the opposition offered by capacitors at a specified frequency and is measured in ohms.

Resonant frequency: The frequency at which a given system or object will respond with maximum amplitude when driven by an external sinusoidal force of constant amplitude.

Reverse leakage current: Anondestructive current flowing through a capacitor subjected to a voltage of polarity opposite to that normally specified.

Ripple current: The total amount of alternating and direct current that may be applied to an electrolytic capacitor under stated conditions.

Temperature coefficient (TC): A capacitor's change in capacitance per degree change in temperature. May be positive, negative, or zero and is usually expressed in parts per million per degree Celsius (ppm/°C) if the characteristics are linear. For nonlinear types, TC is expressed as a percentage of room temperature (25°C)capacitance.

Time constant: In a capacitor-resistor circuit, the number of seconds required for the capacitor to reach 63.2% of its full charge after a voltage is applied. The time constant of a capacitor with a capacitance (C) in farads in series with a resistance (R) in ohms is equal to $R \times C$ seconds.

Winding: Aconductive path, usually wire, inductively coupled to a magnetic core or cell.

Related Topic

55.5 Dielectric Materials

References

Exploring the capacitor, *Hewlett-Packard Bench Briefs,* September/October 1979. Sections reprinted with permission from *Bench Briefs,* a Hewlett-Packard service publication.

Capacitors, *1979 Electronic Buyer's Handbook,* vol. 1, November 1978. Copyright 1978 by CMP Publications, Inc. Reprinted with permission.

W. G. Jung and R. March, "Picking capacitors," *Audio,* March 1980.

"Electrolytic capacitors: Past, present and future," and "What is an electrolytic capacitor," *Electron. Des.,* May 28, 1981.

R.F. Graf, "Introduction To Aluminum Capacitors," Sprague Electric Company. Parts reprinted with permission.

"Introduction To Aluminum Capacitors," Sprague Electric Company. Parts reprinted with permission.

Handbook of Electronics Tables and Formulas, 6th ed., Indianapolis: Sams, 1986.

1.3 Transformers

C. Sankaran

The electrical transformer was invented by an American electrical engineer, William Stanley, in 1885 and was used in the first ac lighting installation at Great Barrington, Massachusetts. The first transformer was used to step up the power from 500 to 3000 V and transmitted for a distance of 1219 m (4000 ft). At the receiving end the voltage was stepped down to 500 V to power street and office lighting. By comparison, present transformers are designed to transmit hundreds of megawatts of power at voltages of 700 kV and beyond for distances of several hundred miles.

Transformation of power from one voltage level to another is a vital operation in any transmission, distribution, and utilization network. Normally, power is generated at a voltage that takes into consideration the cost of generators in relation to their operating voltage. Generated power is transmitted by overhead lines many miles and undergoes several voltage transformations before it is made available to the actual user. Figure 1.22 shows a typical power flow line diagram.

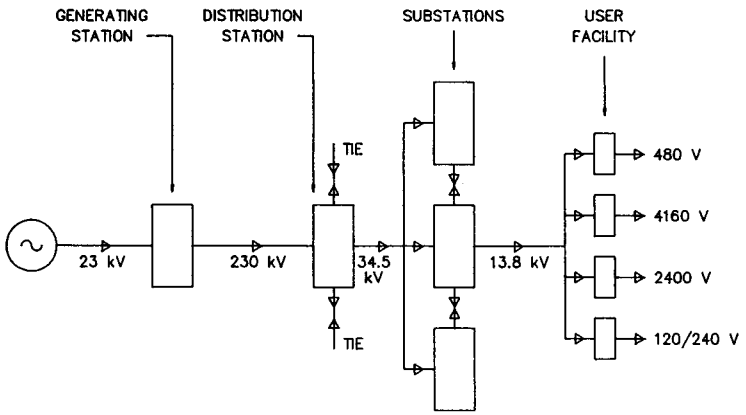


FIGURE 1.22 Power flow line diagram.

Types of Transformers

Transformers are broadly grouped into two main categories: dry-type and liquid-filled transformers. Dry-type transformers are cooled by natural or forced circulation of air or inert gas through or around the transformer enclosure. Dry-type transformers are further subdivided into ventilated, sealed, or encapsulated types depending upon the construction of the transformer. Dry transformers are extensively used in industrial power distribution for rating up to 5000 kVA and 34.5 kV.

Liquid-filled transformers are cooled by natural or forced circulation of a liquid coolant through the windings of the transformer. This liquid also serves as a **dielectric** to provide superior voltage-withstand characteristics. The most commonly used liquid in a transformer is a mineral oil known as transformer oil that has a continuous operating temperature rating of 105°C, a flash point of 150°C, and a fire point of 180°C. A good grade transformer oil has a **breakdown strength** of 86.6 kV/cm (220 kV/in.) that is far higher than the breakdown strength of air, which is 9.84 kV/cm (25 kV/in.) at atmospheric pressure.

Silicone fluid is used as an alternative to mineral oil. The breakdown strength of silicone liquid is over 118 kV/cm (300 kV/in.) and it has a flash point of 300°C and a fire point of 360°C. Silicone-fluid-filled transformers are classified as less flammable. The high dielectric strengths and superior thermal conductivities of liquid coolants make them ideally suited for large high-voltage power transformers that are used in modern power generation and distribution.

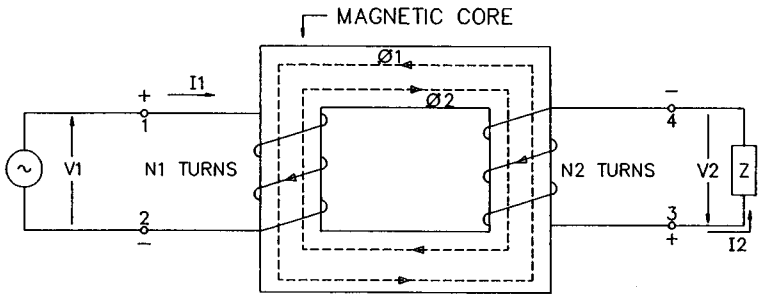


FIGURE 1.23 Electrical power transfer.

Principle of Transformation

The actual process of transfer of electrical power from a voltage of V_1 to a voltage of V_2 is explained with the aid of the simplified transformer representation shown in Fig. 1.23. Application of voltage across the primary winding of the transformer results in a **magnetic field** of ϕ_1 Wb in the magnetic core, which in turn induces a voltage of V_2 at the secondary terminals. V_1 and V_2 are related by the expression $V_1/V_2 = N_1/N_2$, where N_1 and N_2 are the number of turns in the primary and secondary windings, respectively. If a load current of I_2 A is drawn from the secondary terminals, the load current establishes a magnetic field of ϕ_2 Wb in the core and in the direction shown. Since the effect of load current is to reduce the amount of primary magnetic field, the reduction in ϕ_1 results in an increase in the primary current I_1 so that the net magnetic field is almost restored to the initial value and the slight reduction in the field is due to leakage **magnetic flux**. The currents in the two windings are related by the expression $I_1/I_2 = N_2/N_1$. Since $V_1/V_2 = N_1/N_2 = I_2/I_1$, we have the expression $V_1 \cdot I_1 = V_2 \cdot I_2$. Therefore, the voltamperes in the two windings are equal in theory. In reality, there is a slight loss of power during transformation that is due to the energy necessary to set up the magnetic field and to overcome the losses in the transformer core and windings. Transformers are static power conversion devices and are therefore highly efficient. Transformer efficiencies are about 95% for small units (15 kVA and less), and the efficiency can be higher than 99% for units rated above 5 MVA.

Electromagnetic Equation

Figure 1.24 shows a magnetic core with the area of cross section $A = W \cdot D$ m^2. The transformer primary winding that consists of N turns is excited by a sinusoidal voltage $v = V \sin(\omega t)$, where ω is the angular frequency given by the expression $\omega = 2\pi f$ and f is the frequency of the applied voltage waveform. ϕ is magnetic field in the core due to the excitation current i:

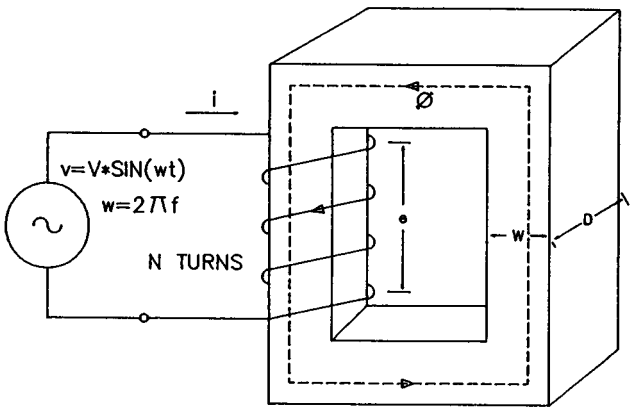


FIGURE 1.24 Electromagnetic relation.

$$\phi = \Phi \sin\left(\omega t - \frac{\pi}{2}\right) = -\Phi \cos(\omega t)$$

Induced voltage in the winding

$$e = -N\frac{d\phi}{dt} = N\frac{d[\Phi \cos(\omega t)]}{dt} = -N\omega\Phi \sin(\omega t)$$

Maximum value of the induced voltage

$$E = N\omega\Phi$$

The root-mean-square value

$$E_{\text{rms}} = \frac{E}{\sqrt{2}} = \frac{2\pi f N\Phi}{\sqrt{2}} = 4.44 f \text{ NBA}$$

where flux Φ (webers) is replaced by the product of the flux density B (teslas) and the area of cross section of the core.

This fundamental design equation determines the size of the transformer for any given voltage and frequency. Power transformers are normally operated at flux density levels of 1.5 T.

Transformer Core

The transformer core is the medium that enables the transfer of power from the primary to the secondary to occur in a transformer. In order that the transformation of power may occur with the least amount of loss, the magnetic core is made up of laminations which have the highest permeability, permeability being a measure of the ease with which the magnetic field is set up in the core.

The magnetic field reverses direction every one half cycle of the applied voltage and energy is expended in the core to accomplish the cyclic reversals of the field. This loss component is known as the hysteresis loss P_h:

$$P_h = 150.7 V_e f B^{1.6} \quad \text{W}$$

where V_e is the volume of the core in cubic meters, f is the frequency, and B is the maximum flux density in teslas.

As the magnetic field reverses direction and cuts across the core structure, it induces a voltage in the laminations known as eddy voltages. This phenomenon causes eddy currents to circulate in the laminations. The loss due to eddy currents is called the eddy current loss P_e:

$$P_e = 1.65 V_e B^2 f^2 t^2 / r$$

where V_e is the volume of the core in cubic meters, f is the frequency, B is the maximum flux density in teslas, t is thickness of the laminations in meters, and r is the resistivity of the core material in ohm-meters.

Hysteresis losses are reduced by operating the core at low flux densities and using core material of high permeability. Eddy current losses are minimized by low flux levels, reduction in thickness of the laminations, and high resistivity core material.

Cold-rolled, grain-oriented silicon steel laminations are exclusively used in large power transformers to reduce core losses. A typical silicon steel used in transformers contains 95% iron, 3% silicon, 1% manganese, 0.2% phosphor, 0.06% carbon, 0.025% sulphur, and traces of other impurities.

Transformer Losses

The heat developed in a transformer is a function of the losses that occur during transformation. Therefore, the transformer losses must be minimized and the heat due to the losses must be efficiently conducted away from the core, the windings, and the cooling medium. The losses in a transformer are grouped into two categories: (1) no-load losses and (2) load losses. The no-load losses are the losses in the core due to excitation and are mostly composed of hysteresis and eddy current losses. The load losses are grouped into three categories: (1) winding I^2R losses, (2) winding eddy current losses, and (3) other stray losses. The winding I^2R losses are the result of the flow of load current through the resistance of the primary and secondary windings. The winding eddy current losses are caused by the magnetic field set up by the winding current, due to formation of eddy voltages in the conductors. The winding eddy losses are proportional to the square of the rms value of the current and to the square of the frequency of the current. When transformers are required to supply loads that are rich in **harmonic frequency** components, the eddy loss factor must be given extra consideration. The other stray loss component is the result of induced currents in the buswork, core clamps, and tank walls by the magnetic field set up by the load current.

Transformer Connections

A single-phase transformer has one input (primary) winding and one output (secondary) winding. A conventional three-phase transformer has three input and three output windings. The three windings can be connected in one of several different configurations to obtain three-phase connections that are distinct. Each form of connection has its own merits and demerits.

Y Connection (Fig. 1.25)

In the Y connection, one end of each of the three windings is connected together to form a Y, or a neutral point. This point is normally grounded, which limits the maximum potential to ground in the transformer to the line to neutral voltage of the power system. The grounded neutral also limits transient overvoltages in the transformer when subjected to lightning or switching surges. Availability of the neutral point allows the transformer to supply line to neutral single-phase loads in addition to normal three-phase loads. Each phase of the Y-connected winding must be designed to carry the full line current, whereas the phase voltages are only 57.7% of the line voltages.

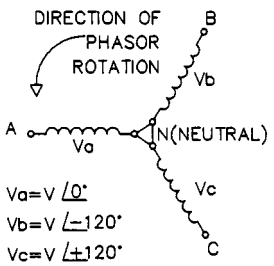


FIGURE 1.25 Y connection.

Delta Connection (Fig. 1.26)

In the delta connection, the finish point of each winding is connected to the start point of the adjacent winding to form a closed triangle, or delta. A delta winding in the transformer tends to balance out unbalanced loads that are present on the system. Each phase of the delta winding only carries 57.7% of the line current, whereas the phase voltages are equal to the line voltages.

Large power transformers are designed so that the high-voltage side is connected in Y and the low-voltage side is connected in delta. Distribution transformers that are required to supply single-phase loads are designed in the opposite configuration so that the neutral point is available at the low-voltage end.

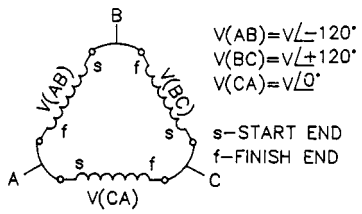


FIGURE 1.26 Delta connection.

Open-Delta Connection (Fig. 1.27)

An open-delta connection is used to deliver three-phase power if one phase of a three-phase bank of transformers fails in service. When the failed unit is removed from service, the remaining units can still supply three-phase power but at a reduced rating. An open-delta connection is also used as an economical means to deliver three-phase power using only two single-phase transformers. If P

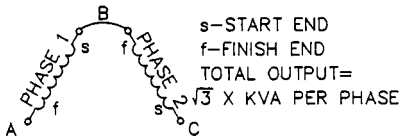


FIGURE 1.27 Open-delta connection.

is the total three-phase kVA, then each transformer of the open-delta bank must have a rating of $P/\sqrt{3}$ kVA. The disadvantage of the open-delta connection is the unequal **regulation** of the three phases of the transformer.

T Connection (Fig. 1.28)

The T connection is used for three-phase power transformation when two separate single-phase transformers with special configurations are available. If a voltage transformation from V_1 to V_2 volts is required, one of the units (main transformer) must have a voltage ratio of V_1/V_2 with the midpoint of each winding brought out. The other unit must have a ratio of $0.866V_1/0.866V_2$ with the neutral point brought out, if needed.

The Scott connection is a special type of T connection used to transform three-phase power to two-phase power for operation of electric furnaces and two-phase motors. It is shown in Fig. 1.29.

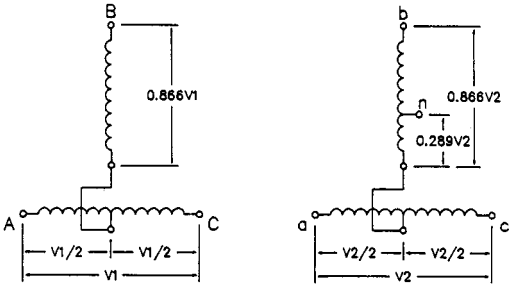


FIGURE 1.28 T connection.

Zigzag Connection (Fig. 1.30)

This connection is also called the interconnected star connection where the winding of each phase is divided into two halves and interconnected to form a zigzag configuration. The zigzag connection is mostly used to derive a neutral point for grounding purposes in three-phase, three-wire systems. The neutral point can be used to (1) supply single-phase loads, (2) provide a safety ground, and (3) sense and limit ground fault currents.

Transformer Impedance

Impedance is an inherent property in a transformer that results in a voltage drop as power is transferred from the primary to the secondary side of the power system. The impedance of a transformer consists of two parts: resistance (R) and reactance (X). The resistance component is due to the resistance of the material of the winding and the percentage value of the voltage drop due to resistance becomes less as the rating of the transformer increases. The reactive component, which is also known as leakage reactance, is the result of incomplete linkage of the magnetic field set up by the secondary winding with the turns of the primary winding, and vice versa. The net impedance of the transformer is given by $Z = \sqrt{R^2 + X^2}$. The impedance value marked on the transformer is the percentage voltage drop due to this impedance under full-load operating conditions:

$$\%\ \text{impedance}\ z = IZ\left(\frac{100}{V}\right)$$

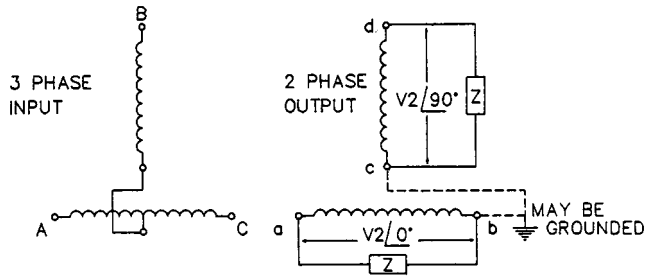


FIGURE 1.29 Three-phase–two-phase transformation.

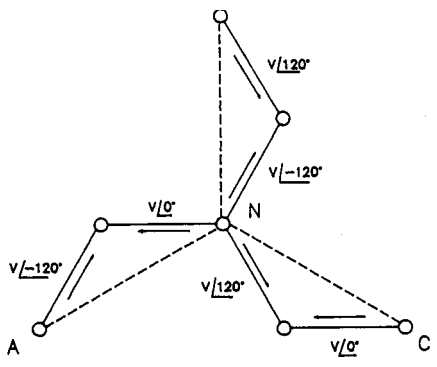


FIGURE 1.30 Zigzag connection.

where I is the full-load current of the transformer, Z is the impedance in ohms of the transformer, and V is the voltage rating of the transformer winding. It should be noted that the values of I and Z must be referred to the same side of the transformer as the voltage V.

Transformers are also major contributors of impedance to limit the fault currents in electrical power systems.

Defining Terms

Breakdown strength: Voltage gradient at which the molecules of medium break down to allow passage of damaging levels of electric current.

Dielectric: Solid, liquid, or gaseous substance that acts as an insulation to the flow of electric current.

Harmonic frequency: Integral multiples of fundamental frequency. For example, for a 60-Hz supply the harmonic frequencies are 120, 180, 240, 300, . . .

Magnetic field: Magnetic force field where lines of magnetism exist.

Magnetic flux: Term for lines of magnetism.

Regulation: The change in voltage from no-load to full-load expressed as a percentage of full-load voltage.

Related Topics

9.3 Wye ⇔ Delta Transformations • 36.1 Magnetism • 61.6 Protection • 64.1 Transformer Construction

References and Further Information

Bean, Chackan, Moore and Wentz, *Transformers for the Electric Power Industry*, New York: McGraw-Hill, 1966.

General Electric, *Transformer Connections,* 1960.

A. Gray, *Electrical Machine Design*, New York: McGraw-Hill.

IEEE, *C57 Standards on Transformers*, New York: IEEE Press, 1992.

IEEE Transactions on Industry Applications.

R. R. Lawrence, *Principles of Alternating Current Machinery,* New York: McGraw-Hill, 1920.

Power Engineering Review.

C. Sankaran, *Introduction to Transformers,* New York: IEEE Press, 1992.

S. A. Stigant and A.C. Franklin, T*he J & P Transformer Book*, London: Newnes-Butterworths, 1973.

1.4 Electrical Fuses

Nick Angelopoulos

The fuse is a simple and reliable safety device. It is second to none in its ease of application and its ability to protect people and equipment.

The fuse is a current-sensitive device. It has a conductor with a reduced cross section (element) normally surrounded by an arc-quenching and heat-conducting material (filler). The entire unit is enclosed in a body fitted with end contacts. A basic fuse element design is illustrated in Fig. 1.32.

Ratings

Most fuses have three electrical ratings: ampere rating, voltage rating, and **interrupting rating**. The ampere rating indicates the current the fuse can carry without melting or exceeding specific temperature rise limits. The voltage rating, ac or dc, usually indicates the maximum system voltage that can be applied to the fuse. The interrupting rating (I.R.) defines the maximum short-circuit current that a fuse can safely interrupt. If a fault current higher than the interrupting rating causes the fuse to operate, the high internal pressure may cause the fuse to rupture. It is imperative, therefore, to install a fuse, or any other type of protective device, that has an interrupting rating not less than the available short-circuit current. A violent explosion may occur if the interrupting rating of any protective device is inadequate.

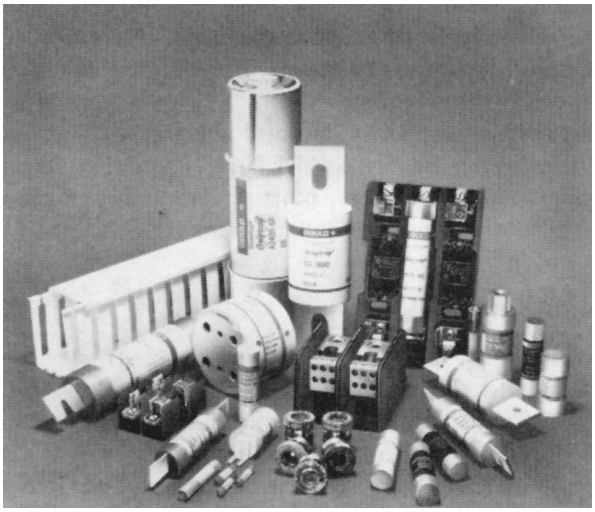


FIGURE 1.31 A variety of plug, cartridge, and blade type fuses.

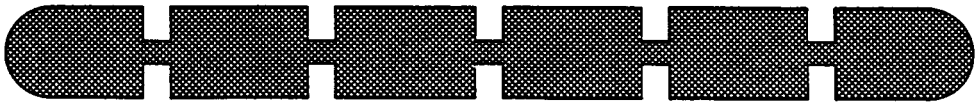


FIGURE 1.32 Basic fuse element.

A fuse must perform two functions. The first, the "passive" function, is one that tends to be taken for granted. In fact, if the fuse performs the passive function well, we tend to forget that the fuse exists at all. The passive function simply entails that the fuse can carry up to its normal load current without aging or overheating. Once the current level exceeds predetermined limits, the "active" function comes into play and the fuse operates. It is when the fuse is performing its active function that we become aware of its existence.

In most cases, the fuse will perform its active function in response to two types of circuit conditions. The first is an overload condition, for instance, when a hair dryer, teakettle, toaster, and radio are plugged into the same circuit. This overload condition will eventually cause the element to melt. The second condition is the overcurrent condition, commonly called the short circuit or the fault condition. This can produce a drastic, almost instantaneous, rise in current, causing the element to melt usually in less than a quarter of a cycle. Factors that can lead to a fault condition include rodents in the electrical system, loose connections, dirt and moisture, breakdown of insulation, foreign contaminants, and personal mistakes. Preventive maintenance and care can reduce these causes. Unfortunately, none of us are perfect and faults can occur in virtually every electrical system—we must protect against them.

Fuse Performance

Fuse performance characteristics under overload conditions are published in the form of *average melting time–current characteristic curves,* or simply *time-current curves.* Fuses are tested with a variety of currents, and the melting times are recorded. The result is a graph of time versus current coordinates that are plotted on log-log scale, as illustrated in Fig. 1.33.

Under short-circuit conditions the fuse operates and fully opens the circuit in less than 0.01 s. At 50 or 60 Hz, this represents operation within the first half cycle. The current waveform let-through by the fuse is the shaded, almost triangular, portion shown in Fig. 1.34(a). This depicts a fraction of the current that would have been let through into the circuit had a fuse not been installed.

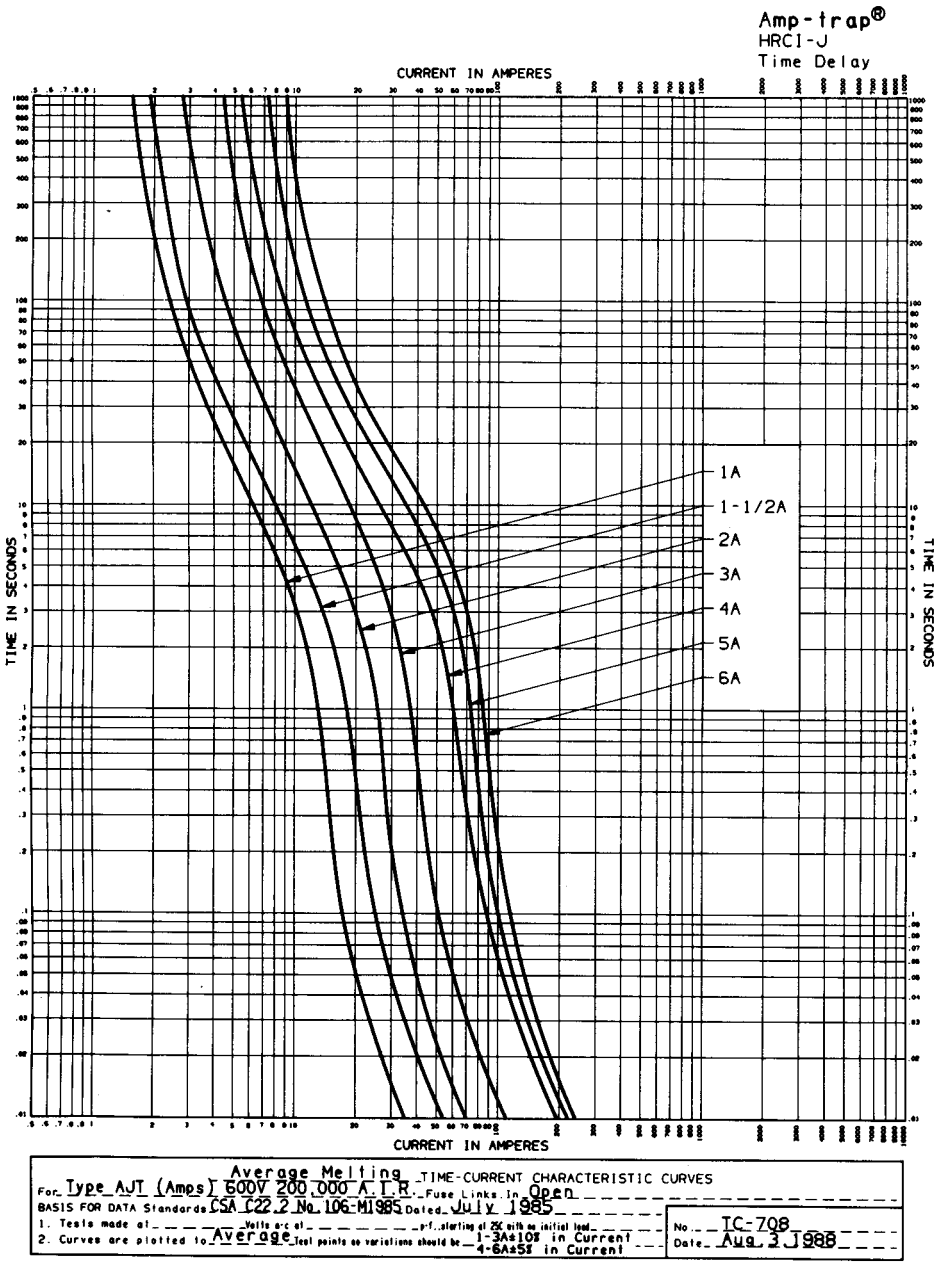


FIGURE 1.33 Time-current characteristic curves.

Fuse short-circuit performance characteristics are published in the form of peak let-through (I_p) graphs and I^2t graphs. I_p (peak current) is simply the peak of the shaded triangular waveform, which increases as the fault current increases, as shown in Fig. 1.34(b). The electromagnetic forces, which can cause mechanical damage to equipment, are proportional to I_p^2.

I^2t represents heat energy measured in units of A^2 s (ampere squared seconds) and is documented on I^2t graphs. These I^2t graphs, as illustrated in Fig. 1.34(c), provide three values of I^2t: minimum melting I^2t, arcing I^2t, and total clearing I^2t. I^2t and I_p short-circuit performance characteristics can be used to coordinate fuses and other equipment. In particular, I^2t values are often used to selectively coordinate fuses in a distribution system.

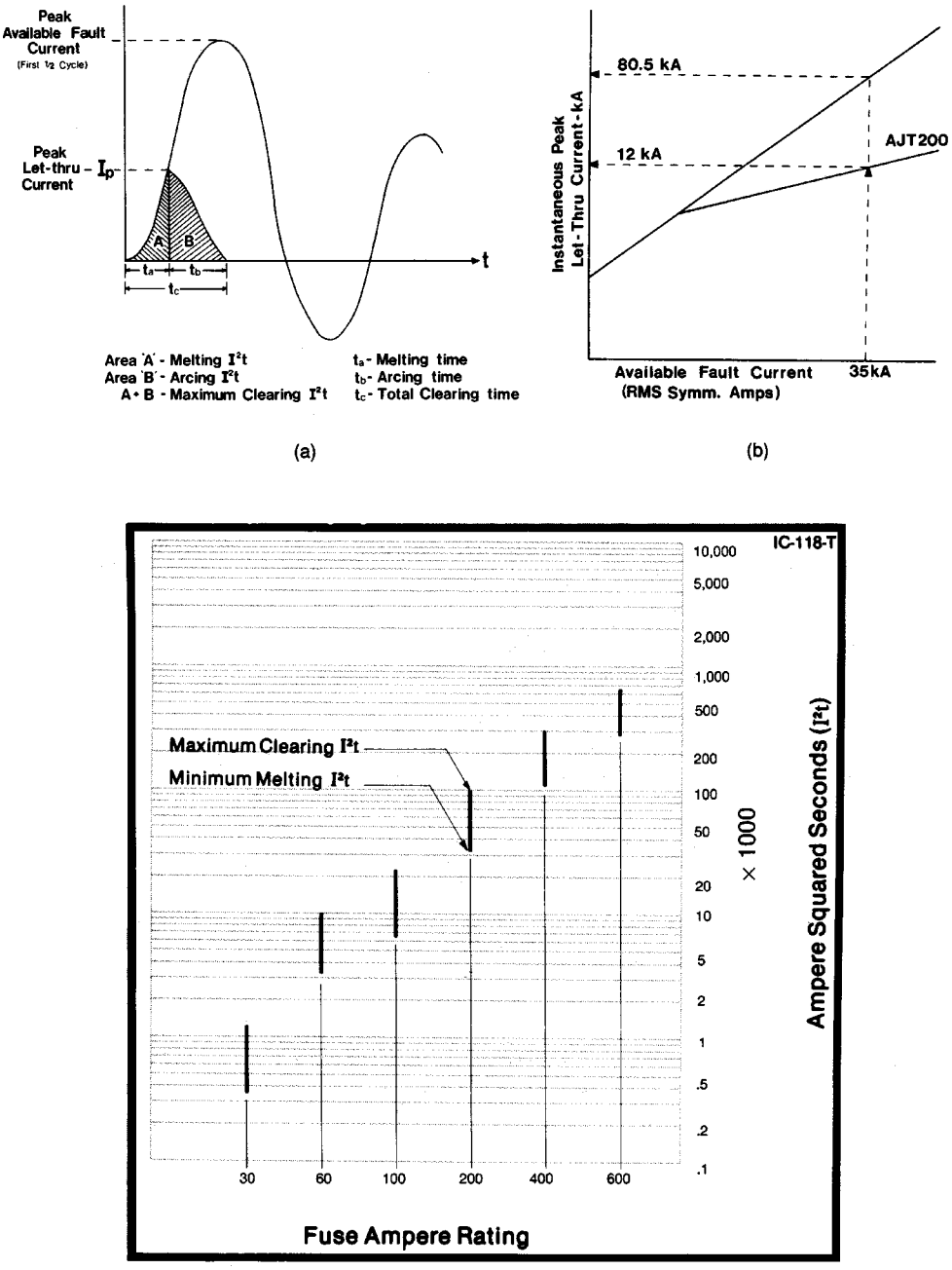


FIGURE 1.34 (a) Fuse short-circuit operation. (b) Variation of fuse peak let-through current I_p. (c) I^2t graph.

Selective Coordination

In any power distribution system, selective coordination exists when the fuse immediately upstream from a fault operates, leaving all other fuses further upstream unaffected. This increases system reliability by isolating the faulted branch while maintaining power to all other branches. Selective coordination is easily assessed by

comparing the I^2t characteristics for feeder and branch circuit fuses. The branch fuse should have a total clearing I^2t value that is less than the melting I^2t value of the feeder or upstream fuse. This ensures that the branch fuse will melt, arc, and clear the fault before the feeder fuse begins to melt.

Standards

Overload and short-circuit characteristics are well documented by fuse manufacturers. These characteristics are standardized by product standards written in most cases by safety organizations such as CSA (Canadian Standards Association) and UL (Underwriters Laboratories). CSA standards and UL specify product designations, dimensions, performance characteristics, and temperature rise limits. These standards are used in conjunction with national code regulations such as CEC (Canadian Electrical Code) and NEC (National Electrical Code) that specify how the product is applied.

IEC (International Electrotechnical Commission—Geneva, Switzerland) was founded to harmonize electrical standards to increase international trade in electrical products. Any country can become a member and participate in the standards-writing activities of IEC. Unlike CSA and UL, IEC is not a certifying body that certifies or approves products. IEC publishes consensus standards for national standards authorities such as CSA (Canada), UL (USA), BSI (UK) and DIN (Germany) to adopt as their own national standards.

Products

North American low-voltage distribution fuses can be classified under two types: Standard or Class H, as referred to in the United States, and **HRC (high rupturing capacity)** or current-limiting fuses, as referred to in Canada. It is the interrupting rating that essentially differentiates one type from the other.

Most Standard or Class H fuses have an interrupting rating of 10,000 A. They are not classified as HRC or current-limiting fuses, which usually have an interrupting rating of 200,000 A. Selection is often based on the calculated available short-circuit current.

In general, short-circuit currents in excess of 10,000 A do not exist in residential applications. In commercial and industrial installations, short-circuit currents in excess of 10,000 A are very common. Use of HRC fuses usually means that a fault current assessment is not required.

Standard—Class H

In North America, Standard or Class H fuses are available in 250- and 600-V ratings with ampere ratings up to 600 A. There are primarily three types: one-time, time-delay, and renewable. Rating for rating, they are all constructed to the same dimensions and are physically interchangeable in standard-type fusible switches and fuse blocks.

One-time fuses are not reusable once blown. They are used for general-purpose resistive loads such as lighting, feeders, and cables.

Time-delay fuses have a specified delay in their overload characteristics and are designed for motor circuits. When started, motors typically draw six times their full load current for approximately 3 to 4 seconds. This surge then decreases to a level within the motor full-load current rating. Time-delay fuse overload characteristics are designed to allow for motor starting conditions.

Renewable fuses are constructed with replaceable links or elements. This feature minimizes the cost of replacing fuses. However, the concept of replacing fuse elements in the field is not acceptable to most users today because of the potential risk of improper replacement.

HRC

HRC or current-limiting fuses have an interrupting rating of 200 kA and are recognized by a letter designation system common to North American fuses. In the United States they are known as Class J, Class L, Class R, etc., and in Canada they are known as HRCI-J, HRC-L, HRCI-R, and so forth. HRC fuses are available in ratings up to 600 V and 6000 A. The main differences among the various types are their dimensions and their short-circuit performance (I_p and I^2t) characteristics.

One type of HRC fuse found in Canada, but not in the United States, is the HRCII-C or Class C fuse. This fuse was developed originally in England and is constructed with bolt-on-type blade contacts. It is available in a voltage rating of 600 V with ampere ratings from 2 to 600 A. Some higher ampere ratings are also available but are not as common. HRCII-C fuses are primarily regarded as providing short-circuit protection only. Therefore, they should be used in conjunction with an overload device.

HRCI-R or Class R fuses were developed in the United States. Originally constructed to Standard or Class H fuse dimensions, they were classified as Class K and are available in the United States with two levels of short-circuit performance characteristics: Class K1 and Class K5. However, they are not recognized in Canadian Standards. Under fault conditions, Class K1 fuses limit the I_p and I^2t to lower levels than do Class K5 fuses. Since both Class K1 and K5 are constructed to Standard or Class H fuse dimensions, problems with interchangeability occur. As a result, a second generation of these K fuses was therefore introduced with a rejection feature incorporated in the end caps and blade contacts. This rejection feature, when used in conjunction with rejection-style fuse clips, prevents replacement of these fuses with Standard or Class H 10-kA I.R. fuses. These rejection style fuses are known as Class RK1 and Class RK5. They are available with time-delay or non-time-delay characteristics and with voltage ratings of 250 or 600 V and ampere ratings up to 600 A. In Canada, CSA has only one classification for these fuses, HRCI-R, which have the same maximum I_p and I^2t current-limiting levels as specified by UL for Class RK5 fuses.

HRCI-J or Class J fuses are a more recent development. In Canada, they have become the most popular HRC fuse specified for new installations. Both time-delay and non-time-delay characteristics are available in ratings of 600 V with ampere ratings up to 600 A. They are constructed with dimensions much smaller than HRCI-R or Class R fuses and have end caps or blade contacts which fit into 600-V Standard or Class H-type fuse clips.

However, the fuse clips must be mounted closer together to accommodate the shorter fuse length. Its shorter length, therefore, becomes an inherent rejection feature that does not allow insertion of Standard or HRCI-R fuses. The blade contacts are also drilled to allow bolt-on mounting if required. CSA and UL specify these fuses to have maximum short-circuit current-limiting I_p and I^2t limits lower than those specified for HRCI-R and HRCII-C fuses. HRCI-J fuses may be used for a wide variety of applications. The time-delay type is commonly used in motor circuits sized at approximately 125 to 150% of motor full-load current.

HRC-L or Class L fuses are unique in dimension but may be considered as an extension of the HRCI-J fuses for ampere ratings above 600 A. They are rated at 600 V with ampere ratings from 601 to 6000 A. They are physically larger and are constructed with bolt-on-type blade contacts. These fuses are generally used in low-voltage distribution systems where supply transformers are capable of delivering more than 600 A.

In addition to Standard and HRC fuses, there are many other types designed for specific applications. For example, there are medium- or high-voltage fuses to protect power distribution transformers and medium-voltage motors. There are fuses used to protect sensitive semiconductor devices such as diodes, SCRs, and triacs. These fuses are designed to be extremely fast under short-circuit conditions. There is also a wide variety of dedicated fuses designed for protection of specific equipment requirements such as electric welders, capacitors, and circuit breakers, to name a few.

Trends

Ultimately, it is the electrical equipment being protected that dictates the type of fuse needed for proper protection. This equipment is forever changing and tends to get smaller as new technology becomes available. Present trends indicate that fuses also must become smaller and faster under fault conditions, particularly as available short-circuit fault currents are tending to increase.

With free trade and the globalization of industry, a greater need for harmonizing product standards exists. The North American fuse industry is taking big steps toward harmonizing CSA and UL fuse standards, and at the same time is participating in the IEC standards process. Standardization will help the electrical industry to identify and select the best fuse for the job—anywhere in the world.

Defining Terms

HRC (high rupturing capacity): A term used to denote fuses having a high interrupting rating. Most low-voltage HRC-type fuses have an interrupting rating of 200 kA rms symmetrical.

***I*²*t* (ampere squared seconds):** A convenient way of indicating the heating effect or thermal energy which is produced during a fault condition before the circuit protective device has opened the circuit. As a protective device, the HRC or current-limiting fuse lets through far less damaging I^2t than other protective devices.

Interrupting rating (I.R.): The maximum value of short-circuit current that a fuse can safely interrupt.

Related Topic

1.1 Resistors

References

R.K. Clidero and K.H. Sharpe, *Application of Electrical Construction*, Ontario, Canada: General Publishing Co. Ltd., 1982.

Gould Inc., *Shawmut Advisor*, Circuit Protection Division, Newburyport, Mass.

C. A. Gross, *Power Systems Analysis*, 2nd ed., New York: Wiley, 1986.

E. Jacks, *High Rupturing Capacity Fuses*, New York: Wiley, 1975.

A. Wright and P.G. Newbery, *Electric Fuses*, London: Peter Peregrinus Ltd., 1984.

Further Information

For greater detail the "Shawmut Advisor" (Gould, Inc., 374 Merrimac Street, Newburyport MA 01950) or the "Fuse Technology Course Notes" (Gould Shawmut Company, 88 Horner Avenue, Toronto, Canada M8Z-5Y3) may be referred to for fuse performance and application.

Dorf, R.C., Wan, Z., Paul, C.R., Cogdell, J.R. “Voltage and Current Sources”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

2

Voltage and Current Sources

Richard C. Dorf
University of California, Davis

Zhen Wan
University of California, Davis

Clayton R. Paul
University of Kentucky, Lexington

J. R. Cogdell
University of Texas at Austin

2.1 Step, Impulse, Ramp, Sinusoidal, Exponential, and DC Signals
Step Function • The Impulse • Ramp Function • Sinusoidal Function • DCSignal

2.2 Ideal and Practical Sources
Ideal Sources • Practical Sources

2.3 Controlled Sources
What Are Controlled Sources? • What Is the Significance of Controlled Sources? • How Does the Presence of Controlled Sources Affect Circuit Analysis?

2.1 Step, Impulse, Ramp, Sinusoidal, Exponential, and DC Signals

Richard C. Dorf and Zhen Wan

The important signals for circuits include the step, impulse, ramp, sinusoid, and dc signals. These signals are widely used and are described here in the time domain. All of these signals have a Laplace transform.

Step Function

The **unit-step** function $u(t)$ is defined mathematically by

$$u(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

Here *unit step* means that the amplitude of $u(t)$ is equal to 1 for $t \geq 0$. Note that we are following the convention that $u(0) = 1$. From a strict mathematical standpoint, $u(t)$ is not defined at $t = 0$. Nevertheless, we usually take $u(0) = 1$. If A is an arbitrary nonzero number, $Au(t)$ is the step function with amplitude A for $t \geq 0$. The unit step function is plotted in Fig. 2.1.

The Impulse

The **unit impulse** $\delta(t)$, also called the *delta function* or the *Dirac distribution*, is defined by

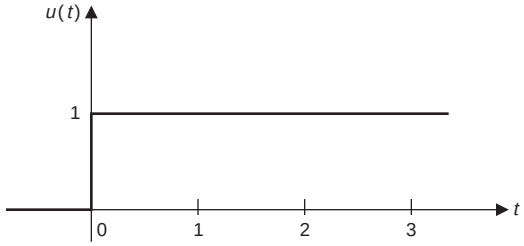


FIGURE 2.1 Unit-step function.

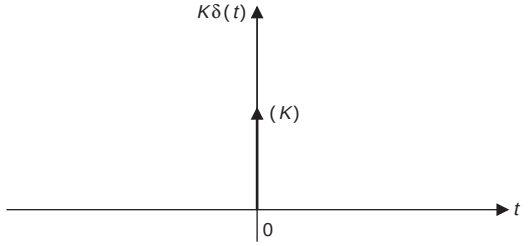


FIGURE 2.2 Graphical representation of the impulse $K\delta(t)$

$$\delta(t) = 0, \qquad t \neq 0$$
$$\int_{-\varepsilon}^{\varepsilon} \delta(\lambda)\, d\lambda = 1, \qquad \text{for any real number } \varepsilon > 0$$

The first condition states that $\delta(t)$ is zero for all nonzero values of t, while the second condition states that the area under the impulse is 1, so $\delta(t)$ has unit area. It is important to point out that the value $\delta(0)$ of $\delta(t)$ at $t = 0$ is not defined; in particular, $\delta(0)$ is not equal to infinity. For any real number K, $K\delta(t)$ is the impulse with area K. It is defined by

$$K\delta(t) = 0, \qquad t \neq 0$$
$$\int_{-\varepsilon}^{\varepsilon} K\delta(\lambda)\, d\lambda = K, \qquad \text{for any real number } \varepsilon > 0$$

The graphical representation of $K\delta(t)$ is shown in Fig. 2.2. The notation K in the figure refers to the area of the impulse $K\delta(t)$.

The unit-step function $u(t)$ is equal to the integral of the unit impulse $\delta(t)$; more precisely, we have

$$u(t) = \int_{-\infty}^{t} \delta(\lambda)\, d\lambda, \qquad \text{all } t \text{ except } t = 0$$

Conversely, the first derivative of $u(t)$, with respect to t, is equal to $\delta(t)$, except at $t = 0$, where the derivative of $u(t)$ is not defined.

Ramp Function

The *unit-ramp function* $r(t)$ is defined mathematically by

$$r(t) = \begin{cases} t, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

Note that for $t \geq 0$, the slope of $r(t)$ is 1. Thus, $r(t)$ has *unit slope,* which is the reason $r(t)$ is called the unit-ramp function. If K is an arbitrary nonzero scalar (real number), the ramp function $Kr(t)$ has slope K for $t \geq 0$. The unit-ramp function is plotted in Fig. 2.3.

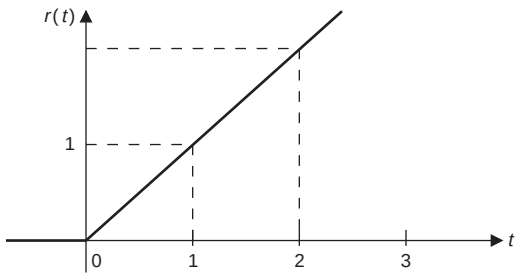


FIGURE 2.3 Unit-ramp function

The unit-ramp function $r(t)$ is equal to the integral of the unit-step function $u(t)$; that is,

$$r(t) = \int_{-\infty}^{t} u(\lambda)\, d\lambda$$

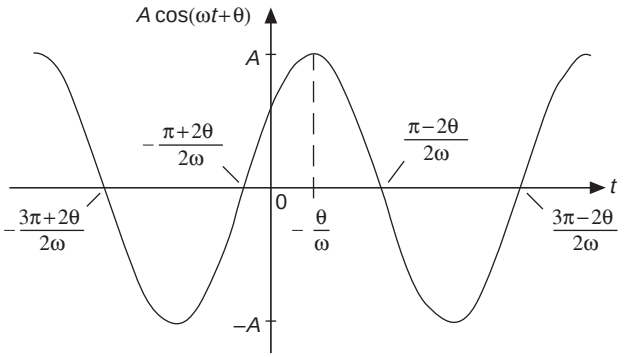


FIGURE 2.4 The sinusoid $A\cos(\omega t + \theta)$ with $-\pi/2 < \theta < 0$.

Conversely, the first derivative of $r(t)$ with respect to t is equal to $u(t)$, except at $t = 0$, where the derivative of $r(t)$ is not defined.

Sinusoidal Function

The sinusoid is a continuous-time signal: $A\cos(\omega t + \theta)$.

Here A is the amplitude, ω is the frequency in radians per second (rad/s), and θ is the phase in radians. The frequency f in cycles per second, or hertz (Hz), is $f = \omega/2\pi$. The sinusoid is a periodic signal with period $2\pi/\omega$. The sinusoid is plotted in Fig. 2.4.

Decaying Exponential

In general, an exponentially decaying quantity (Fig. 2.5) can be expressed as

$$a = A\,e^{-t/\tau}$$

where a = instantaneous value
A = amplitude or maximum value
e = base of natural logarithms = 2.718 …
τ = time constant in seconds
t = time in seconds

The current of a discharging capacitor can be approximated by a decaying exponential function of time.

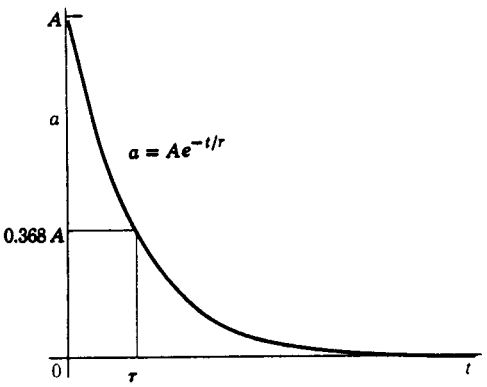


FIGURE 2.5 The decaying exponential.

Time Constant

Since the exponential factor only *approaches* zero as t increases without limit, such functions theoretically last forever. In the same sense, all radioactive disintegrations last forever. In the case of an exponentially decaying current, it is convenient to use the value of time that makes the exponent –1. When $t = \tau$ = the *time constant,* the value of the exponential factor is

$$e^{-t/\tau} = e^{-1} = \frac{1}{e} = \frac{1}{2.718} = 0.368$$

In other words, after a time equal to the time constant, the exponential factor is reduced to approximatly 37% of its initial value.

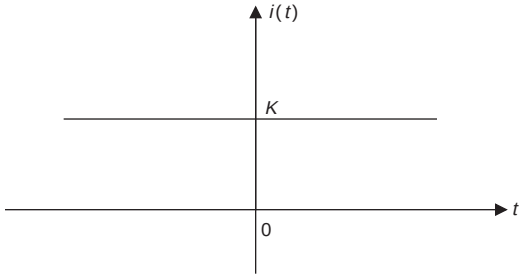


FIGURE 2.6 The dc signal with amplitude K.

DC Signal

The direct current signal (dc signal) can be defined mathematically by

$$i(t) = K \qquad -\infty < t < +\infty$$

Here, K is any nonzero number. The dc signal remains a constant value of K for any $-\infty < t < \infty$. The dc signal is plotted in Fig. 2.6.

Defining Terms

Ramp: A continually growing signal such that its value is zero for $t \leq 0$ and proportional to time t for $t > 0$.
Sinusoid: A periodic signal $x(t) = A\cos(\omega t + \theta)$ where $\omega = 2\pi f$ with frequency in hertz.
Unit impulse: A very short pulse such that its value is zero for $t \neq 0$ and the integral of the pulse is 1.
Unit step: Function of time that is zero for $t < t_0$ and unity for $t > t_0$. At $t = t_0$ the magnitude changes from zero to one. The unit step is dimensionless.

Related Topic

11.1 Introduction

References

R.C. Dorf, *Introduction to Electric Circuits,* 3rd ed., New York: Wiley, 1996.
R.E. Ziemer, *Signals and Systems,* 2nd ed., New York: Macmillan, 1989.

Further Information

IEEE Transactions on Circuits and Systems
IEEE Transactions on Education

2.2 Ideal and Practical Sources

Clayton R. Paul

A *mathematical model* of an electric circuit contains *ideal models* of physical circuit elements. Some of these ideal circuit elements (e.g., the resistor, capacitor, inductor, and transformer) were discussed previously. Here we will define and examine both *ideal* and *practical voltage and current sources.* The terminal characteristics of these models will be compared to those of actual sources.

All-Plastic Battery

Researchers at the U.S. Air Force's Rome Laboratory and Johns Hopkins University have developed an all-plastic battery using polymers instead of conventional electrode materials. All-plastic power cells could be a safer, more flexible substitute for use in electronic devices and other commercial applications. In addition, all-polymer cells reduce toxic waste disposal, negate environmental concerns, and can meet EPA and FAA requirements.

Applications include powering GPS receivers, communication transceivers, remote sensors, backup power systems, cellular phones, pagers, computing products and other portable equipment. Potential larger applications include remote monitoring stations, highway communication signs and electric vehicles.

The Johns Hopkins scientists are among the first to create a potentially practical battery in which both of the electrodes and the electrolyte are made of polymers. Fluoro-substituted thiophenes polymers have been developed with potential differences of up to 2.9 volts, and with potential specific energy densities of 30 to 75 watt hours/kg.

All plastic batteries can be recharged hundreds of times and operate under extreme hot and cold temperature conditions without serious performance degradation. The finished cell can be as thin as a business card and malleable, allowing battery manufacturers to cut a cell to a specific space or make the battery the actual case of the device to be powered. (Reprinted with permission from *NASA Tech Briefs,* 20(10), 26, 1996.)

Ideal Sources

The *ideal independent voltage source* shown in Fig. 2.7 constrains the terminal voltage across the element to a prescribed function of time, $v_s(t)$, as $v(t) = v_s(t)$. The polarity of the source is denoted by ± signs within the circle which denotes this as an ideal *independent* source. Controlled or *dependent* ideal voltage sources will be discussed in Section 2.3. The current through the element will be determined by the circuit that is attached to the terminals of this source.

The *ideal independent current source* in Fig. 2.8 constrains the terminal current through the element to a prescribed function of time, $i_s(t)$, as $i(t) = i_s(t)$. The polarity of the source is denoted by an arrow within the

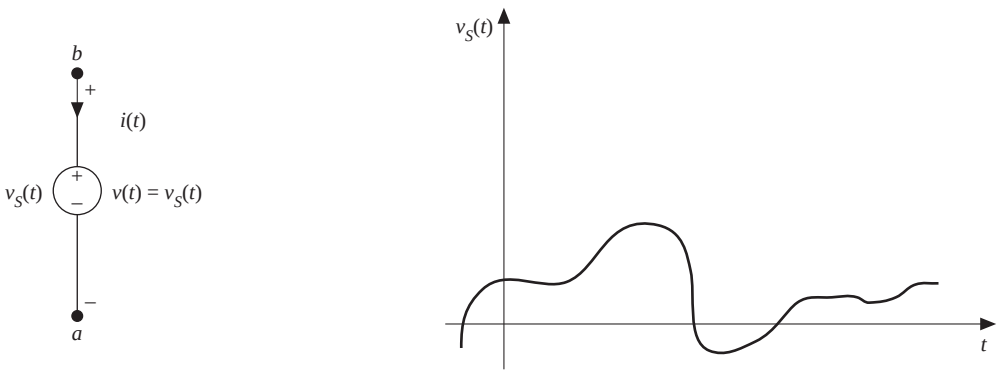


FIGURE 2.7 Ideal independent voltage source.

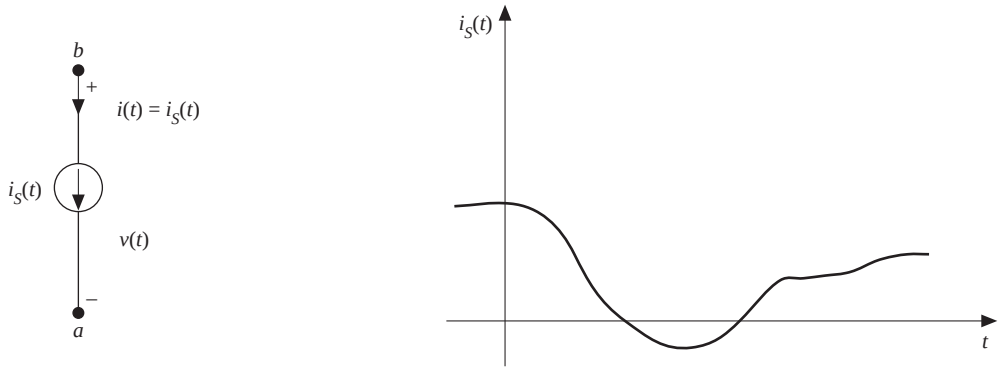


FIGURE 2.8 Ideal independent current source.

circle which also denotes this as an ideal *independent* source. The voltage across the element will be determined by the circuit that is attached to the terminals of this source.

Numerous functional forms are useful in describing the source variation with time. These were discussed in Section 2.1—the step, impulse, ramp, sinusoidal, and dc functions. For example, an ideal independent dc voltage source is described by $v_S(t) = V_S$, where V_S is a constant. An ideal independent sinusoidal current source is described by $i_S(t) = I_S\sin(\omega t + \phi)$ or $i_S(t) = I_S\cos(\omega t + \phi)$, where I_S is a constant, $\omega = 2\pi f$ with f the *frequency* in hertz and ϕ is a phase angle. Ideal sources may be used to model actual sources such as temperature transducers, phonograph cartridges, and electric power generators. Thus usually the time form of the output cannot generally be described with a simple, basic function such as dc, sinusoidal, ramp, step, or impulse waveforms. We often, however, represent the more complicated waveforms as a linear combination of more basic functions.

Practical Sources

The preceding ideal independent sources constrain the terminal voltage or current to a *known* function of time *independent of the circuit that may be placed across its terminals.* Practical sources, such as batteries, have their terminal voltage (current) dependent upon the terminal current (voltage) caused by the circuit attached to the source terminals. A simple example of this is an automobile storage battery. The battery's terminal voltage is approximately 12 V when no load is connected across its terminals. When the battery is applied across the terminals of the starter by activating the ignition switch, a large current is drawn from its terminals. During starting, its terminal voltage drops as illustrated in Fig. 2.9(a). How shall we construct a *circuit model* using the ideal elements discussed thus far to model this nonideal behavior? A model is shown in Fig. 2.9(b) and consists of the *series* connection of an ideal resistor, R_S, and an ideal independent voltage source, $V_S = 12$ V. To determine the terminal voltage–current relation, we sum Kirchhoff's voltage law around the loop to give

$$v = V_S - R_S i \tag{2.1}$$

This equation is plotted in Fig. 2.9(b) and approximates that of the actual battery. The equation gives a straight line with slope $-R_S$ that intersects the v axis ($i = 0$) at $v = V_S$. The resistance R_S is said to be the *internal resistance* of this nonideal source model. It is a fictitious resistance but the model nevertheless gives an equivalent *terminal behavior.*

Although we have derived an approximate model of an actual source, another equivalent form may be obtained. This alternative form is shown in Fig. 2.9(c) and consists of the *parallel* combination of an ideal independent current source, $I_S = V_S/R_S$, and the same resistance, R_S, used in the previous model. Although it may seem strange to model an automobile battery using a current source, the model is completely equivalent to the series voltage source–resistor model of Fig. 2.9(b) *at the output terminals a–b.* This is shown by writing Kirchhoff's current law at the upper node to give

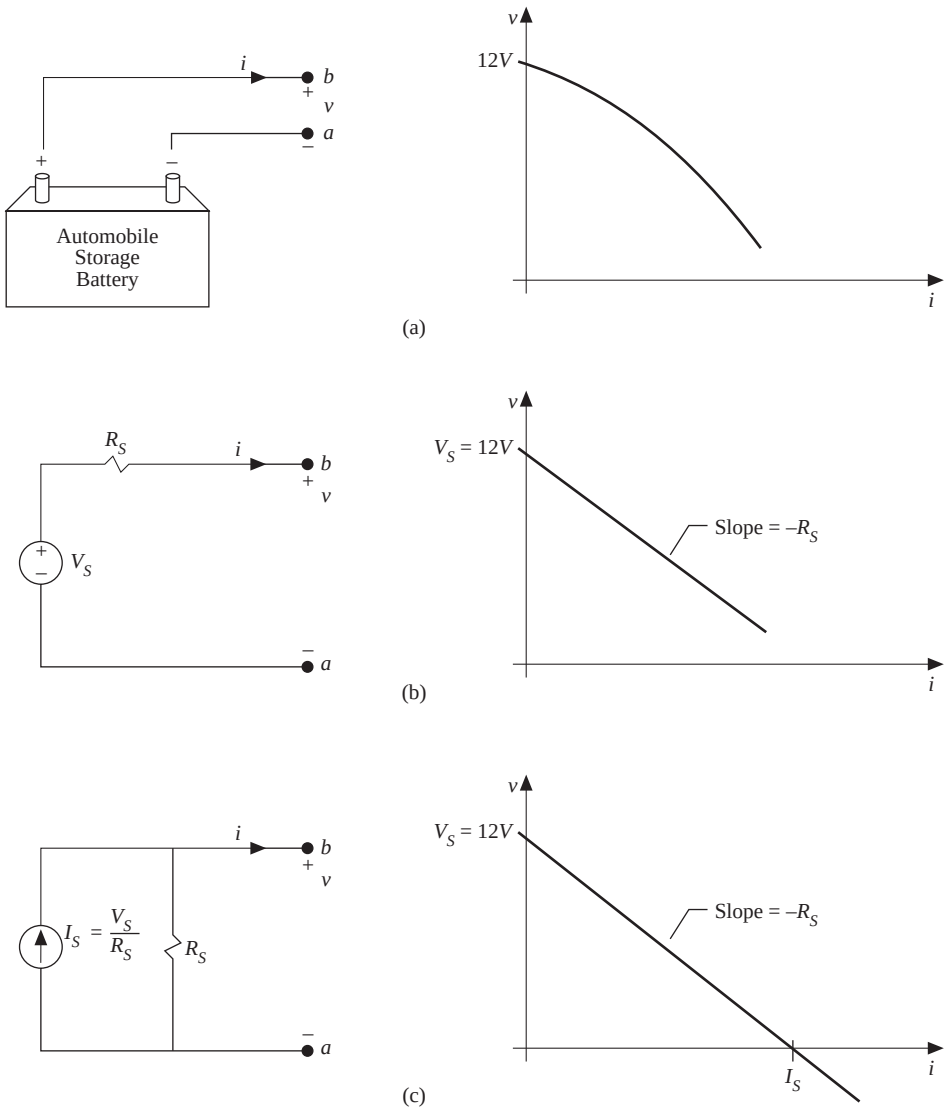


FIGURE 2.9 Practical sources. (a) Terminal *v-i* characteristic; (b) approximation by a voltage source; (c) approximation by a current source.

$$i = I_S - \frac{1}{R_S} v \tag{2.2}$$

Rewriting this equation gives

$$v = R_S I_S - R_S i \tag{2.3}$$

Comparing Eq. (2.3) to Eq. (2.1) shows that

$$V_S = R_S I_S \tag{2.4}$$

Therefore, we can convert from one form (voltage source in series with a resistor) to another form (current source in parallel with a resistor) very simply.

An ideal voltage source is represented by the model of Fig. 2.9(b) with $R_S = 0$. An actual battery therefore provides a close approximation of an ideal voltage source since the source resistance R_S is usually quite small. An ideal current source is represented by the model of Fig. 2.9(c) with $R_S = \infty$. This is very closely represented by the bipolar junction transistor (BJT).

Related Topic

1.1 Resistors

Defining Term

Ideal source: An ideal model of an actual source that assumes that the parameters of the source, such as its magnitude, are independent of other circuit variables.

Reference

C.R. Paul, *Analysis of Linear Circuits,* New York: McGraw-Hill, 1989.

2.3 Controlled Sources

J. R. Cogdell

When the analysis of electronic (nonreciprocal) circuits became important in circuit theory, controlled sources were added to the family of circuit elements. Table 2.1 shows the four types of controlled sources. In this section, we will address the questions: What are controlled sources? Why are controlled sources important? How do controlled sources affect methods of circuit analysis?

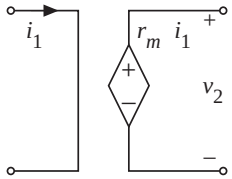
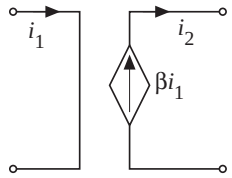
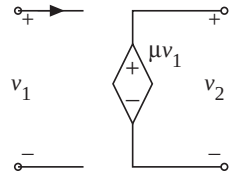
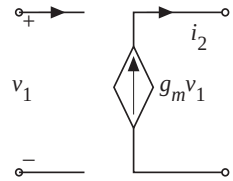
What Are Controlled Sources?

By *source* we mean a voltage or current source in the usual sense. By *controlled* we mean that the strength of such a source is controlled by some circuit variable(s) elsewhere in the circuit. Figure 2.10 illustrates a simple circuit containing an (independent) current source, i_s, two resistors, and a controlled voltage source, whose magnitude is controlled by the current i_1. Thus, i_1 determines two voltages in the circuit, the voltage across R_1 via Ohm's law and the controlled voltage source via some unspecified effect.

A controlled source may be controlled by more than one circuit variable, but we will discuss those having a single controlling variable since multiple controlling variables require no new ideas. Similarly, we will deal only with resistive elements, since inductors and capacitors introduce no new concepts. The controlled voltage or current source may depend on the controlling variable in a linear or nonlinear manner. When the relationship is nonlinear, however, the equations are frequently linearized to examine the effects of small variations about some dc values. When we linearize, we will use the customary notation of small letters to represent general and time-variable voltages and currents and large letters to represent constants such as the dc value or the peak value of a sinusoid. On subscripts, large letters represent the total voltage or current and small letters represent the **small-signal** component. Thus, the equation $i_B = I_B + I_b \cos \omega t$ means that the total base current is the sum of a constant and a small-signal component, which is sinusoidal with an amplitude of I_b.

To introduce the context and use of controlled sources we will consider a circuit model for the bipolar junction transistor (BJT). In Fig. 2.11 we show the standard symbol for an *npn* BJT with base (B), emitter (E), and collector (C) identified, and voltage and current variables defined. We have shown the common emitter configuration, with the emitter terminal shared to make input and output terminals. The base current, i_B, ideally depends upon the base-emitter voltage, v_{BE}, by the relationship

TABLE 2.1 Names, Circuit Symbols, and Definitions for the Four Possible Types of Controlled Sources

Name	Circuit Symbol	Definition and Units
Current-controlled voltage source (CCVS)		$v_2 = r_m i_1$ r_m = transresistance units, ohms
Current-controlled current source (CCCS)		$i_2 = \beta i_1$ β, current gain, dimensionless
Voltage-controlled voltage source (VCVS)		$v_2 = \mu v_1$ μ, voltage gain, dimensionless
Voltage-controlled current source (VCCS)		$i_2 = g_m v_1$ g_m, transconductance units, Siemans (mhos)

$$i_B = I_0\left\{\exp\left[\frac{v_{BE}}{V_T}\right] - 1\right\} \tag{2.5}$$

where I_0 and V_T are constants. We note that the base current depends on the base-emitter voltage only, but in a nonlinear manner. We can represent this current by a voltage-controlled current source, but the more common representation would be that of a nonlinear conductance, $G_{BE}(v_{BE})$, where

$$G_{BE}(v_{BE}) = \frac{i_B}{v_{BE}}$$

Let us model the effects of small changes in the base current. If the changes are small, the nonlinear nature of the conductance can be ignored and the circuit model becomes a linear conductance (or resistor). Mathematically this conductance arises from a first-order expansion of the nonlinear function. Thus, if $v_{BE} = V_{BE} + v_{be}$, where v_{BE} is the total base-emitter voltage, V_{BE} is a (large) constant voltage and v_{be} is a (small) variation in the base-emitter voltage, then the first two terms in a Taylor series expansion are

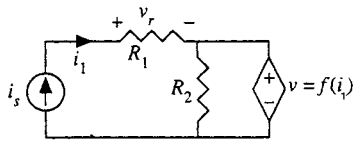


FIGURE 2.10 A simple circuit containing a controlled source.

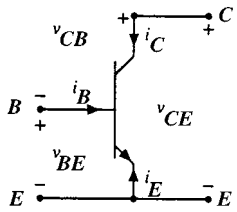


FIGURE 2.11 An *npn* BJT in the common emitter configuration.

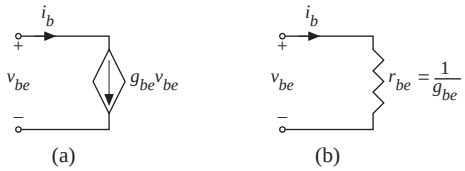


FIGURE 2.12 Equivalent circuits for the base circuit: (a) uses a controlled source and (b) uses a resistor.

$$i_B = I_0\left\{\exp\left[\frac{V_{BE} + v_{be}}{V_T}\right] - 1\right\} \cong I_0\left\{\exp\left[\frac{V_{BE}}{V_T}\right] - 1\right\} + \frac{I_0}{V_T}\exp\left[\frac{V_{BE}}{V_T}\right]v_{be} \tag{2.6}$$

We note that the base current is approximated by the sum of a constant term and a term that is first order in the small variation in base-emitter voltage, v_{be}. The multiplier of this small voltage is the linearized conductance, g_{be}. If we were interested only in small changes in currents and voltages, only this conductance would be required in the model. Thus, the input (base-emitter) circuit can be represented for the small-signal base variables, i_b and v_{be}, by either equivalent circuit in Fig. 2.12.

The voltage-controlled current source, $g_{be}v_{be}$, can be replaced by a simple resistor because the small-signal voltage and current associate with the same branch. The process of **linearization** is important to the modeling of the collector-emitter characteristic, to which we now turn.

The collector current, i_C, can be represented by one of the Eber and Moll equations

$$i_C = \beta I_0\left\{\exp\left[\frac{v_{BE}}{V_T}\right] - 1\right\} - I_0'\left\{\exp\left[\frac{v_{BC}}{V_T}\right] - 1\right\} \tag{2.7}$$

where β and I_0' are constants. If we restrict our model to the amplifying region of the transistor, the second term is negligible and we may express the collector current as

$$i_C = \beta I_0\left\{\exp\left[\frac{v_{BE}}{V_T}\right] - 1\right\} = \beta i_B \tag{2.8}$$

Thus, for the ideal transistor, the collector-emitter circuit may be modeled by a current-controlled current source, which may be combined with the results expressed in Eq. (2.5) to give the model shown in Fig. 2.13.

Using the technique of small-signal analysis, we may derive either of the small-signal equivalent circuits shown in Fig. 2.14.

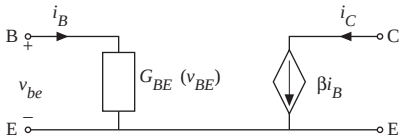


FIGURE 2.13 Equivalent circuit for BJT.

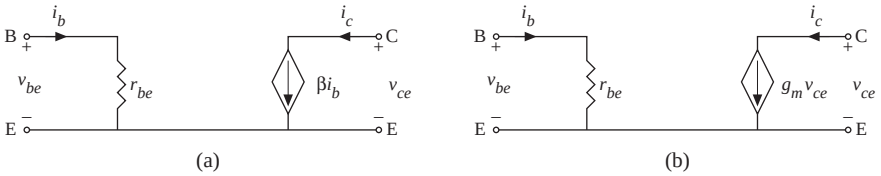


FIGURE 2.14 Two BJT small-signal equivalent circuits ($g_m = \beta/r_{be}$): (a) uses a CCCS and (b) uses a VCCS.

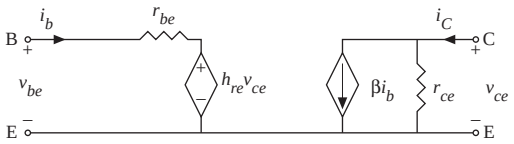


FIGURE 2.15 Full hybrid parameter model for small-signal BJT.

The small-signal characteristics of the *npn* transistor in its amplifying region is better represented by the equivalent circuit shown in Fig. 2.15. Note we have introduced a voltage-controlled voltage source to model the influence of the (output) collector-emitter voltage on the (input) base-emitter voltage, and we have placed a resistor, r_{ce}, in parallel with the collector current source to model the influence of the collector-emitter voltage on the collector current.

The four parameters in Fig. 2.15 (r_{be}, h_{re}, β, and r_{ce}) are the hybrid parameters describing the transistor properties, although our notation differs from that commonly used. The parameters in the small-signal equivalent circuit depend on the operating point of the device, which is set by the time-average voltages and currents (V_{BE}, I_C, etc.) applied to the device. All of the parameters are readily measured for a given transistor and operating point, and manufacturers commonly specify ranges for the various parameters for a type of transistor.

What Is the Significance of Controlled Sources?

Commonplace wisdom in engineering education and practice is that information and techniques that are presented visually are more useful than abstract, mathematical forms. Equivalent circuits are universally used in describing electrical engineering systems and devices because circuits portray interactions in a universal, pictorial language. This is true generally, and it is doubly necessary when circuit variables interact through the mysterious coupling modeled by controlled sources. This is the primary significance of controlled sources: that they represent unusual couplings of circuit variables in the universal, visual language of circuits.

A second significance is illustrated by our equivalent circuit of the *npn* bipolar transistor, namely, the characterization of a class of similar devices. For example, the parameter β in Eq. (2.8) gives important information about a single transistor, and similarly for the range of β for a type of transistor. In this connection, controlled sources lead to a vocabulary for discussing some property of a class of systems or devices, in this case the current gain of an *npn* BJT.

How Does the Presence of Controlled Sources Affect Circuit Analysis?

The presence of nonreciprocal elements, which are modeled by controlled sources, affects the analysis of the circuit. Simple circuits may be analyzed through the direct application of Kirchhoff's laws to branch circuit variables. Controlled sources enter this process similar to the constitutive relations defining *R*, *L*, and *C*, i.e., in defining relationships between branch circuit variables. Thus, controlled sources add no complexity to this basic technique.

The presence of controlled sources negates the advantages of the method that uses series and parallel combinations of resistors for voltage and current dividers. The problem is that the couplings between circuit variables that are expressed by controlled sources make all the familiar formulas unreliable.

When superposition is used, the controlled sources are left on in all cases as independent sources are turned on and off, thus reflecting the kinship of controlled sources to the circuit elements. In principle, little complexity is added; in practice, the repeated solutions required by superposition entail much additional work when controlled sources are involved.

The classical methods of nodal and loop (mesh) analysis incorporate controlled sources without great difficulty. For purposes of determining the number of independent variables required, that is, in establishing the topology of the circuit, the controlled sources are treated as ordinary voltage or current sources. The equations are then written according to the usual procedures. Before the equations are solved, however, the controlling variables must be expressed in terms of the unknowns of the problem. For example, let us say we

are performing a nodal analysis on a circuit containing a current-controlled current source. For purposes of counting independent nodes, the controlled current source is treated as an open circuit. After equations are written for the unknown node voltages, the current source will introduce into at least one equation its controlling current, which is not one of the nodal variables. The additional step required by the controlled source is that of expressing the controlling current in terms of the nodal variables.

The parameters introduced into the circuit equations by the controlled sources end up on the left side of the equations with the resistors rather than on the right side with the independent sources. Furthermore, the symmetries that normally exist among the coefficients are disturbed by the presence of controlled sources.

The methods of Thévenin and Norton equivalent circuits continue to be very powerful with controlled sources in the circuits, but some complications arise. The controlled sources must be left on for calculation of the Thévenin (open-circuit) voltage or Norton (short-circuit) current and also for the calculation of the output impedance of the circuit. This usually eliminates the method of combining elements in series or parallel to determine the output impedance of the circuit, and one must either determine the output impedance from the ratio of the Thévenin voltage to the Norton current or else excite the circuit with an external source and calculate the response.

Defining Terms

Controlled source (dependent source): A voltage or current source whose intensity is controlled by a circuit voltage or current elsewhere in the circuit.

Linearization: Approximating nonlinear relationships by linear relationships derived from the first-order terms in a power series expansion of the nonlinear relationships. Normally the linearized equations are useful for a limited range of the voltage and current variables.

Small-signal: Small-signal variables are those first-order variables used in a linearized circuit. A small-signal equivalent circuit is a linearized circuit picturing the relationships between the small-signal voltages and currents in a linearized circuit.

Related Topics

2.2 Ideal and Practical Sources • 22.3 Electrical Equivalent Circuit Models and Device Simulators for Semiconductor Devices

References

E. J. Angelo, Jr., *Electronic Circuits*, 2nd ed., New York: McGraw-Hill, 1964.
N. Balabanian and T. Bickart, *Linear Network Theory*, Chesterland, Ohio: Matrix Publishers, 1981.
L. O. Chua, *Introduction to Nonlinear Network Theory*, New York: McGraw-Hill, 1969.
B. Friedland, O. Wing, and R. Ash, *Principles of Linear Networks,* New York: McGraw-Hill, 1961.
L. P. Huelsman, *Basic Circuit Theory*, 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Ciletti, M.D., Irwin, J.D., Kraus, A.D., Balabanian, N., Bickart, T.A., Chan, S.P., Nise, N.S. "Linear Circuit Analysis"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

3

Linear Circuit Analysis

Michael D. Ciletti
University of Colorado

J. David Irwin
Auburn University

Allan D. Kraus
Allan D. Kraus Associates

Norman Balabanian
University of Florida

Theodore A. Bickart
Michigan State University

Shu-Park Chan
International Technological University

Norman S. Nise
California State Polytechnic University

3.1 Voltage and Current Laws
Kirchhoff's Current Law • Kirchhoff's Current Law in the Complex Domain • Kirchhoff's Voltage Law • Kirchhoff's Voltage Law in the Complex Domain • Importance of KVL and KCL

3.2 Node and Mesh Analysis
Node Analysis • Mesh Analysis • Summary

3.3 Network Theorems
Linearity and Superposition • The Network Theorems of Thévenin and Norton • Tellegen's Theorem • Maximum Power Transfer • The Reciprocity Theorem • The Substitution and Compensation Theorem

3.4 Power and Energy
Tellegen's Theorem • AC Steady-State Power • Maximum Power Transfer • Measuring AC Power and Energy

3.5 Three-Phase Circuits

3.6 Graph Theory
The *k*-Tree Approach • The Flowgraph Approach • The *k*-Tree Approach Versus the Flowgraph Approach • Some Topological Applications in Network Analysis and Design

3.7 Two-Port Parameters and Transformations
Introduction • Defining Two-Port Networks • Mathematical Modeling of Two-Port Networsk via *z* Parameters • Evaluating Two-Port Network Characteristics in Terms of *z* Parameters • An Example Finding *z* Parameters and Network Characteristics • Additional Two-Port Parameters and Conversions • Two Port Parameter Selection

3.1 Voltage and Current Laws

Michael D. Ciletti

Analysis of linear circuits rests on two fundamental physical laws that describe how the voltages and currents in a circuit must behave. This behavior results from whatever voltage sources, current sources, and energy storage elements are connected to the circuit. A voltage source imposes a constraint on the evolution of the voltage between a pair of nodes; a current source imposes a constraint on the evolution of the current in a branch of the circuit. The energy storage elements (capacitors and inductors) impose initial conditions on currents and voltages in the circuit; they also establish a dynamic relationship between the voltage and the current at their terminals.

Regardless of how a linear circuit is stimulated, every node voltage and every branch current, at every instant of time, must be consistent with Kirchhoff's voltage and current laws. These two laws govern even the most complex linear circuits. (They also apply to a broad category of nonlinear circuits that are modeled by point models of voltage and current.)

A circuit can be considered to have a topological (or graph) view, consisting of a labeled set of nodes and a labeled set of edges. Each edge is associated with a pair of nodes. A node is drawn as a *dot* and represents a

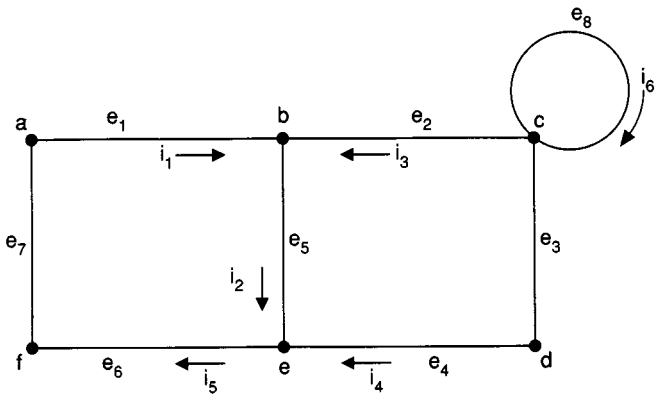


FIGURE 3.1 Graph representation of a linear circuit.

connection between two or more physical components; an edge is drawn as a *line* and represents a path, or branch, for current flow through a component (see Fig. 3.1).

The edges, or branches, of the graph are assigned current labels, $i_1, i_2, \ldots, i_m$. Each current has a designated direction, usually denoted by an *arrow* symbol. If the arrow is drawn toward a node, the associated current is said to be *entering* the node; if the arrow is drawn away from the node, the current is said to be *leaving* the node. The current i_1 is entering node b in Fig. 3.1; the current i_5 is leaving node e.

Given a branch, the pair of nodes to which the branch is attached defines the convention for measuring voltages in the circuit. Given the ordered pair of nodes (a, b), a voltage measurement is formed as follows:

$$v_{ab} = v_a - v_b$$

where v_a and v_b are the absolute electrical potentials (voltages) at the respective nodes, taken relative to some reference node. Typically, one node of the circuit is labeled as *ground,* or reference node; the remaining nodes are assigned voltage labels. The measured quantity, v_{ab}, is called the *voltage drop* from node a to node b. We note that

$$v_{ab} = -v_{ba}$$

and that

$$v_{ba} = v_b - v_a$$

is called the *voltage rise* from a to b. Each node voltage implicitly defines the voltage drop between the respective node and the ground node.

The pair of nodes to which an edge is attached may be written as (a,b) or (b,a). Given an ordered pair of nodes (a, b), a *path from a to b* is a directed sequence of edges in which the first edge in the sequence contains node label a, the last edge in the sequence contains node label b, and the node indices of any two adjacent members of the sequence have at least one node label in common. In Fig. 3.1, the edge sequence $\{e_1, e_2, e_4\}$ is not a path, because e_2 and e_4 do not share a common node label. The sequence $\{e_1, e_2\}$ is a path from node a to node c.

A path is said to be *closed* if the first node index of its first edge is identical to the second node index of its last edge. The following edge sequence forms a closed path in the graph given in Fig. 3.1: $\{e_1, e_2, e_3, e_4, e_6, e_7\}$. Note that the edge sequences $\{e_8\}$ and $\{e_1, e_1\}$ are closed paths.

Kirchhoff's Current Law

Kirchhoff's current law (KCL) imposes constraints on the currents in the branches that are attached to each node of a circuit. In simplest terms, KCL states that the sum of the currents that are entering a given node

must equal the sum of the currents that are leaving the node. Thus, the set of currents in branches attached to a given node can be partitioned into two groups whose orientation is away from (into) the node. The two groups must contain the same net current. Applying KCL at node b in Fig. 3.1 gives

$$i_1(t) + i_3(t) = i_2(t)$$

A connection of water pipes that has no leaks is a physical analogy of this situation. The net rate at which water is flowing into a joint of two or more pipes must equal the net rate at which water is flowing away from the joint. The joint itself has the property that it only connects the pipes and thereby imposes a structure on the flow of water, but it cannot store water. This is true regardless of when the flow is measured. Likewise, the nodes of a circuit are modeled as though they cannot store charge. (Physical circuits are sometimes modeled for the purpose of simulation as though they store charge, but these nodes implicitly have a capacitor that provides the physical mechanism for *storing* the charge. Thus, KCL is ultimately satisfied.)

KCL can be stated alternatively as: "the algebraic sum of the branch currents entering (or leaving) any node of a circuit at any instant of time must be zero." In this form, the label of any current whose orientation is away from the node is preceded by a minus sign. The currents *entering* node b in Fig. 3.1 must satisfy

$$i_1(t) - i_2(t) + i_3(t) = 0$$

In general, the currents entering or leaving each node m of a circuit must satisfy

$$\sum i_{km}(t) = 0$$

where $i_{km}(t)$ is understood to be the current in branch k attached to node m. The currents used in this expression are understood to be the currents that would be measured in the branches attached to the node, and their values include a magnitude and an algebraic sign. If the measurement convention is oriented for the case where currents are entering the node, then the actual current in a branch has a positive or negative sign, depending on whether the current is truly flowing toward the node in question.

Once KCL has been written for the nodes of a circuit, the equations can be rewritten by substituting into the equations the voltage-current relationships of the individual components. If a circuit is resistive, the resulting equations will be algebraic. If capacitors or inductors are included in the circuit, the substitution will produce a differential equation. For example, writing KCL at the node for v_3 in Fig. 3.2 produces

$$i_2 + i_1 - i_3 = 0$$

and

$$C_1 \frac{dv_1}{dt} + \frac{v_4 - v_3}{R_2} - C_2 \frac{dv_2}{dt} = 0$$

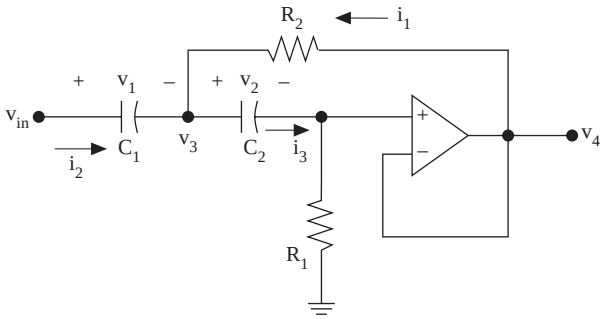


FIGURE 3.2 Example of a circuit containing energy storage elements.

KCL for the node between C_2 and R_1 can be written to eliminate variables and lead to a solution describing the capacitor voltages. The capacitor voltages, together with the applied voltage source, determine the remaining voltages and currents in the circuit. Nodal analysis (see Section 3.2) treats the systematic modeling and analysis of a circuit under the influence of its sources and energy storage elements.

Kirchhoff's Current Law in the Complex Domain

Kirchhoff's current law is ordinarily stated in terms of the real (time-domain) currents flowing in a circuit, because it actually describes physical quantities, at least in a macroscopic, statistical sense. It also applied, however, to a variety of purely mathematical models that are commonly used to analyze circuits in the so-called complex domain.

For example, if a linear circuit is in the sinusoidal steady state, all of the currents and voltages in the circuit are sinusoidal. Thus, each voltage has the form

$$v(t) = A \sin(\omega t + \phi)$$

and each current has the form

$$i(t) = B \sin(\omega t + \theta)$$

where the positive coefficients A and B are called the magnitudes of the signals, and ϕ and θ are the phase angles of the signals. These mathematical models describe the physical behavior of electrical quantities, and instrumentation, such as an oscilloscope, can display the actual waveforms represented by the mathematical model. Although methods exist for manipulating the models of circuits to obtain the magnitude and phase coefficients that uniquely determine the waveform of each voltage and current, the manipulations are cumbersome and not easily extended to address other issues in circuit analysis.

Steinmetz [Smith and Dorf, 1992] found a way to exploit complex algebra to create an elegant framework for representing signals and analyzing circuits when they are in the steady state. In this approach, a model is developed in which each physical sign is replaced by a "complex" mathematical signal. This complex signal in polar, or exponential, form is represented as

$$v_c(t) = Ae^{(j\omega t + \phi)}$$

The algebra of complex exponential signals allows us to write this as

$$v_c(t) = Ae^{j\phi}e^{j\omega t}$$

and Euler's identity gives the equivalent rectangular form:

$$v_c(t) = A[\cos(\omega t + \phi) + j \sin(\omega t + \phi)]$$

So we see that a physical signal is either the real (cosine) or the imaginary (sine) component of an abstract, complex mathematical signal. The additional mathematics required for treatment of complex numbers allows us to associate a phasor, or complex amplitude, with a sinusoidal signal. The time-invariant phasor associated with $v(t)$ is the quantity

$$\mathbf{V}_c = Ae^{j\phi}$$

Notice that the phasor $\mathbf{v}_c$ is an algebraic constant and that in incorporates the parameters A and ϕ of the corresponding time-domain sinusoidal signal.

Phasors can be thought of as being vectors in a two-dimensional plane. If the vector is allowed to rotate about the origin in the counterclockwise direction with frequency ω, the projection of its tip onto the horizontal

(real) axis defines the time-domain signal corresponding to the real part of $v_c(t)$, i.e., $A\cos[\omega t + \phi]$, and its projection onto the vertical (imaginary) axis defines the time-domain signal corresponding to the imaginary part of $v_c(t)$, i.e., $A\sin[\omega t + \phi]$.

The composite signal $v_c(t)$ is a mathematical entity; it cannot be seen with an oscilloscope. Its value lies in the fact that when a circuit is in the steady state, its voltages and currents are uniquely determined by their corresponding phasors, and these in turn satisfy Kirchhoff's voltage and current laws! Thus, we are able to write

$$\sum \mathbf{I}_{km} = 0$$

where $\boldsymbol{I_{km}}$ is the phasor of $i_{km}(t)$, the sinusoidal current in branch k attached to node m. An equation of this form can be written at each node of the circuit. For example, at node b in Fig. 3.1 KCL would have the form

$$\mathbf{I}_1 - \mathbf{I}_2 + \mathbf{I}_3 = 0$$

Consequently, a set of linear, algebraic equations describe the phasors of the currents and voltages in a circuit in the sinusoidal steady state, i.e., the notion of time is suppressed (see Section 3.2). The solution of the set of equations yields the phasor of each voltage and current in the circuit, from which the actual time-domain expressions can be extracted.

It can also be shown that KCL can be extended to apply to the Fourier transforms and the Laplace transforms of the currents in a circuit. Thus, a single relationship between the currents at the nodes of a circuit applies to all of the known mathematical representations of the currents [Ciletti, 1988].

Kirchhoff's Voltage Law

Kirchhoff's voltage law (KVL) describes a relationship among the voltages measured across the branches in any closed, connected path in a circuit. Each branch in a circuit is connected to two nodes. For the purpose of applying KVL, a path has an orientation in the sense that in "walking" along the path one would enter one of the nodes and exit the other. This establishes a direction for determining the voltage across a branch in the path: the voltage is the difference between the potential of the node entered and the potential of the node at which the path exits. Alternatively, the voltage drop along a branch is the difference of the node voltage at the entered node and the node voltage at the exit node. For example, if a path includes a branch between node "a" and node "b", the voltage drop measured along the path in the direction from node "a" to node "b" is denoted by v_{ab} and is given by $v_{ab} = v_a - v_b$. Given v_{ab}, branch voltage along the path in the direction from node "b" to node "a" is $v_{ba} = v_b - v_a = -v_{ab}$.

Kirchhoff's voltage law, like Kirchhoff's current law, is true at any time. KVL can also be stated in terms of voltage rises instead of voltage drops.

KVL can be expressed mathematically as "the algebraic sum of the voltages drops around any closed path of a circuit at any instant of time is zero." This statement can also be cast as an equation:

$$\sum v_{km}(t) = 0$$

where $v_{km}(t)$ is the instantaneous voltage drop measured across branch k of path m. By convention, the voltage drop is taken in the direction of the edge sequence that forms the path.

The edge sequence $\{e_1, e_2, e_3, e_4, e_6, e_7\}$ forms a closed path in Fig. 3.1. The sum of the voltage drops taken around the path must satisfy KVL:

$$v_{ab}(t) + v_{bc}(t) + v_{cd}(t) + v_{de}(t) + v_{ef}(t) + v_{fa}(t) = 0$$

Since $v_{af}(t) = -v_{fa}(t)$, we can also write

$$v_{af}(t) = v_{ab}(t) + v_{bc}(t) + v_{cd}(t) + v_{de}(t) + v_{ef}(t)$$

Had we chosen the path corresponding to the edge sequence $\{e_1, e_5, e_6, e_7\}$ for the path, we would have obtained

$$v_{af}(t) = v_{ab}(t) + v_{be}(t) + v_{ef}(t)$$

This demonstrates how KCL can be used to determine the voltage between a pair of nodes. It also reveals the fact that the voltage between a pair of nodes is independent of the path between the nodes on which the voltages are measured.

Kirchhoff's Voltage Law in the Complex Domain

Kirchhoff's voltage law also applies to the phasors of the voltages in a circuit in steady state and to the Fourier transforms and Laplace transforms of the voltages in a circuit.

Importance of KVL and KCL

Kirchhoff's current law is used extensively in nodal analysis because it is amenable to computer-based implementation and supports a systematic approach to circuit analysis. Nodal analysis leads to a set of algebraic equations in which the variables are the voltages at the nodes of the circuit. This formulation is popular in CAD programs because the variables correspond directly to physical quantities that can be measured easily.

Kirchhoff's voltage law can be used to completely analyze a circuit, but it is seldom used in large-scale circuit simulation programs. The basic reason is that the currents that correspond to a loop of a circuit do not necessarily correspond to the currents in the individual branches of the circuit. Nonetheless, KVL is frequently used to troubleshoot a circuit by measuring voltage drops across selected components.

Defining Terms

Branch: A symbol representing a path for current through a component in an electrical circuit.
Branch current: The current in a branch of a circuit.
Branch voltage: The voltage across a branch of a circuit.
Independent source: A voltage (current) source whose voltage (current) does not depend on any other voltage or current in the circuit.
Node: A symbol representing a physical connection between two electrical components in a circuit.
Node voltage: The voltage between a node and a reference node (usually ground).

Related Topic

3.6 Graph Theory

References

M.D. Ciletti, *Introduction to Circuit Analysis and Design,* New York: Holt, Rinehart and Winston, 1988.
R.H. Smith and R.C. Dorf, *Circuits, Devices and Systems,* New York: Wiley, 1992.

Further Information

Kirchhoff's laws form the foundation of modern computer software for analyzing electrical circuits. The interested reader might consider the use of determining the minimum number of algebraic equations that fully characterizes the circuit. It is determined by KCL, KVL, or some mixture of the two?

3.2 Node and Mesh Analysis

J. David Irwin

In this section Kirchhoff's current law (KCL) and Kirchhoff's voltage law (KVL) will be used to determine currents and voltages throughout a network. For simplicity, we will first illustrate the basic principles of both node analysis and mesh analysis using only dc circuits. Once the fundamental concepts have been explained and illustrated, we will demonstrate the generality of both analysis techniques through an ac circuit example.

Node Analysis

In a node analysis, the node voltages are the variables in a circuit, and KCL is the vehicle used to determine them. One node in the network is selected as a reference node, and then all other node voltages are defined with respect to that particular node. This reference node is typically referred to as *ground* using the symbol (⏚), indicating that it is at ground-zero potential.

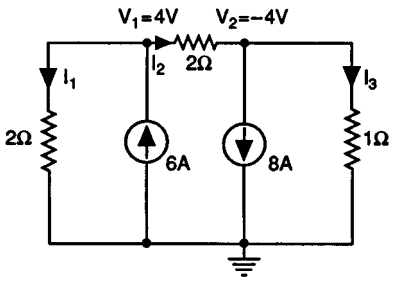


FIGURE 3.3 A three-node network.

Consider the network shown in Fig. 3.3. The network has three nodes, and the nodes at the bottom of the circuit has been selected as the reference node. Therefore the two remaining nodes, labeled V_1 and V_2, are measured with respect to this reference node.

Suppose that the node voltages V_1 and V_2 have somehow been determined, i.e., $V_1 = 4$ V and $v_2 = -4$ V. Once these node voltages are known, Ohm's law can be used to find all branch currents. For example,

$$I_1 = \frac{V_1 - 0}{2} = 2\,\text{A}$$

$$I_2 = \frac{V_1 - V_2}{2} = \frac{4 - (-4)}{2} = 4\,\text{A}$$

$$I_3 = \frac{V_2 - 0}{1} = \frac{-4}{1} = -4\,\text{A}$$

Note that KCL is satisfied at every node, i.e.,

$$I_1 - 6 + I_2 = 0$$

$$-I_2 + 8 + I_3 = 0$$

$$-I_1 + 6 - 8 - I_3 = 0$$

Therefore, as a general rule, if the node voltages are known, all branch currents in the network can be immediately determined.

In order to determine the node voltages in a network, we apply KCL to every node in the network except the reference node. Therefore, given an N-node circuit, we employ $N-1$ linearly independent simultaneous equations to determine the $N-1$ unknown node voltages. Graph theory, which is covered in Section 3.6, can be used to prove that exactly $N-1$ linearly independent KCL equations are required to find the $N-1$ unknown node voltages in a network.

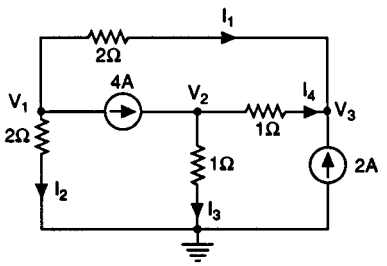


FIGURE 3.4 A four-node network.

Let us now demonstrate the use of KCL in determining the node voltages in a network. For the network shown in Fig. 3.4, the bottom

node is selected as the reference and the three remaining nodes, labeled V_1, V_2, and V_3, are measured with respect to that node. All unknown branch currents are also labeled. The KCL equations for the three nonreference nodes are

$$I_1 + 4 + I_2 = 0$$

$$-4 + I_3 + I_4 = 0$$

$$-I_1 - I_4 - 2 = 0$$

Using Ohm's law these equations can be expressed as

$$\frac{V_1 - V_3}{2} + 4 + \frac{V_1}{2} = 0$$

$$-4 + \frac{V_2}{1} + \frac{V_2 - V_3}{1} = 0$$

$$-\frac{(V_1 - V_3)}{2} - \frac{(V_2 - V_3)}{1} - 2 = 0$$

Solving these equations, using any convenient method, yields $V_1 = -8/3$ V, $V_2 = 10/3$ V, and $V_3 = 8/3$ V. Applying Ohm's law we find that the branch currents are $I_1 = -16/6$ A, $I_2 = -8/6$ A, $I_3 = 20/6$ A, and $I_4 = 4/6$ A. A quick check indicates that KCL is satisfied at every node.

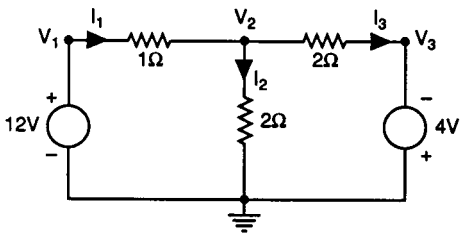


FIGURE 3.5 A four-node network containing voltage sources.

The circuits examined thus far have contained only current sources and resistors. In order to expand our capabilities, we next examine a circuit containing voltage sources. The circuit shown in Fig. 3.5 has three nonreference nodes labeled V_1, V_2, and V_3. However, we do not have three unknown node voltages. Since known voltage sources exist between the reference node and nodes V_1 and V_3, these two node voltages are known, i.e., $V_1 = 12$ V and $V_3 = -4$ V. Therefore, we have only one unknown node voltage, V_2. The equations for this network are then

$$V_1 = 12$$

$$V_3 = -4$$

and

$$-I_1 + I_2 + I_3 = 0$$

The KCL equation for node V_2 written using Ohm's law is

$$-\frac{(12 - V_2)}{1} + \frac{V_2}{2} + \frac{V_2 - (-4)}{2} = 0$$

Solving this equation yields $V_2 = 5$ V, $I_1 = 7$ A, $I_2 = 5/2$ A, and $I_3 = 9/2$ A. Therefore, KCL is satisfied at every node.

Thus, the presence of a voltage source in the network actually simplifies a node analysis. In an attempt to generalize this idea, consider the network in Fig. 3.6. Note that in this case $V_1 = 12$ V and the difference between node voltages V_3 and V_2 is constrained to be 6 V. Hence, two of the three equations needed to solve for the node voltages in the network are

$$V_1 = 12$$

$$V_3 - V_2 = 6$$

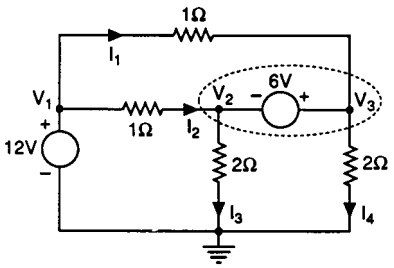


FIGURE 3.6 A four-node network used to illustrate a supernode.

To obtain the third required equation, we form what is called a supernode, indicated by the dotted enclosure in the network. Just as KCL must be satisfied at any node in the network, it must be satisfied at the supernode as well. Therefore, summing all the currents leaving the supernode yields the equation

$$\frac{V_2 - V_1}{1} + \frac{V_2}{2} + \frac{V_3 - V_1}{1} + \frac{V_3}{2} = 0$$

The three equations yield the node voltages $V_1 = 12$ V, $V_2 = 5$ V, and $V_3 = 11$ V, and therefore $I_1 = 1$ A, $I_2 = 7$ A, $I_3 = 5/2$ A, and $I_4 = 11/2$ A.

Mesh Analysis

In a mesh analysis the mesh currents in the network are the variables and KVL is the mechanism used to determine them. Once all the mesh currents have been determined, Ohm's law will yield the voltages anywhere in e circuit. If the network contains N independent meshes, then graph theory can be used to prove that N independent linear simultaneous equations will be required to determine the N mesh currents.

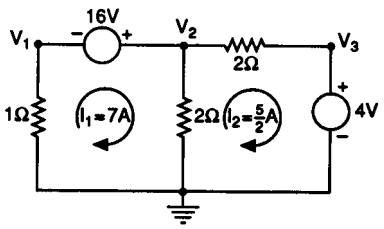


FIGURE 3.7 A network containing two independent meshes.

The network shown in Fig. 3.7 has two independent meshes. They are labeled I_1 and I_2, as shown. If the mesh currents are known to be $I_1 = 7$ A and $I_2 = 5/2$ A, then all voltages in the network can be calculated. For example, the voltage V_1, i.e., the voltage across the 1-Ω resistor, is $V_1 = -I_1R = -(7)(1) = -7$ V. Likewise $V = (I_1 - I_2)R = (7 - 5/2)(2) = 9$ V. Furthermore, we can check our analysis by showing that KVL is satisfied around every mesh. Starting at the lower left-hand corner and applying KVL to the left-hand mesh we obtain

$$-(7)(1) + 16 - (7 - 5/2)(2) = 0$$

where we have assumed that increases in energy level are positive and decreases in energy level are negative.

Consider now the network in Fig. 3.8. Once again, if we assume that an increase in energy level is positive and a decrease in energy level is negative, the three KVL equations for the three meshes defined are

$$-I_1(1) - 6 - (I_1 - I_2)(1) = 0$$

$$+12 - (I_2 - I_1)(1) - (I_2 - I_3)(2) = 0$$

$$-(I_3 - I_2)(2) + 6 - I_3(2) = 0$$

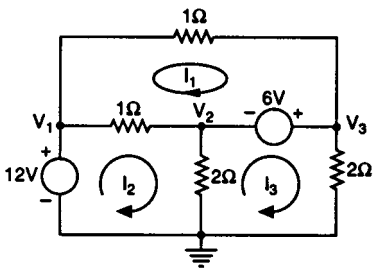


FIGURE 3.8 A three-mesh network.

These equations can be written as

$$2I_1 - I_2 = -6$$

$$-I_{12} + 3I_2 - 2I_3 = 12$$

$$-2I_2 + 4I_3 = 6$$

Solving these equations using any convenient method yields $I1 = 1$ A, $I_2 = 8$ A, and $I_3 = 11/2$ A. Any voltage in the network can now be easily calculated, e.g., $V_2 = (I_2 - I_3)(2) = 5$ V and $V_3 = I_3(2) = 11$ V.

Just as in the node analysis discussion, we now expand our capabilities by considering circuits which contain current sources. In this case, we will show that for mesh analysis, the presence of current sources makes the solution easier.

The network in Fig. 3.9 has four meshes which are labeled I_1, I_2, I_3, and I_4. However, since two of these currents, i.e., I_3 and I_4, pass directly through a current source, two of the four linearly independent equations required to solve the network are

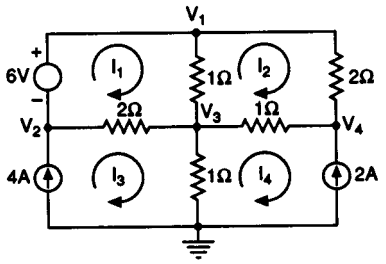


FIGURE 3.9 A four-mesh network containing current sources.

$$I_3 = 4$$

$$I_4 = -2$$

The two remaining KVL equations for the meshes defined by I_1 and I_2 are

$$+6 - (I_1 - I_2)(1) - (I_1 - I_3)(2) = 0$$

$$-(I_2 - I_1)(1) - I_2(2) - (I_2 - I_4)(1) = 0$$

Solving these equations for I_1 and I_2 yields $I_1 = 54/11$ A and $I_2 = 8/11$ A. A quick check will show that KCL is satisfied at every node. Furthermore, we can calculate any node voltage in the network. For example, $V_3 = (I_3 - I_4)(1) = 6$ V and $V_1 = V_3 + (I_1 - I_2)(1) = 112/11$ V.

Summary

Both node analysis and mesh analysis have been presented and discussed. Although the methods have been presented within the framework of dc circuits with only independent sources, the techniques are applicable to ac analysis and circuits containing dependent sources.

To illustrate the applicability of the two techniques to ac circuit analysis, consider the network in Fig. 3.10. All voltages and currents are phasors and the impedance of each passive element is known.

In the node analysis case, the voltage V_4 is known and the voltage between V_2 and V_3 is constrained. Therefore, two of the four required equations are

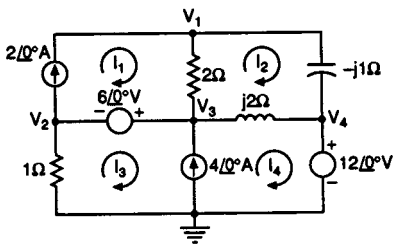


FIGURE 3.10 A network containing five nodes and four meshes.

$$V_4 = 12\ \angle 0°$$

$$V_2 + 6\ \angle 0° = V_3$$

KCL for the node labeled V_1 and the supernode containing the nodes labeled V_2 and V_3 is

$$\frac{V_1 - V_3}{2} + \frac{V_1 - V_4}{-j1} = 2\angle 0°$$

$$\frac{V_2}{1} + 2\angle 0° + \frac{V_3 - V_1}{2} + \frac{V_3 - V_4}{j2} = 4\angle 0° = 0$$

Solving these equations yields the remaining unknown node voltages.

$$\mathbf{V}_1 = 11.9 - j0.88 = 11.93\angle -4.22° \text{ V}$$

$$\mathbf{V}_2 = 3.66 - j1.07 = 3.91\angle -16.34° \text{ V}$$

$$\mathbf{V}_3 = 9.66 - j1.07 = 9.72\angle -6.34° \text{ V}$$

In the mesh analysis case, the currents I_1 and I_3 are constrained to be

$$\mathbf{I}_1 = 2\angle 0°$$

$$\mathbf{I}_4 - \mathbf{I}_3 = -4\angle 0°$$

The two remaining KVL equations are obtained from the mesh defined by mesh current I_2 and the loop which encompasses the meshes defined by mesh currents I_3 and I_4.

$$-2(I_2 - I_1) - (-j1)I_2 - j2(I_2 - I4) = 0$$

$$-(1I_3 + 6\angle 0° - j2(I_4 - I_2) - 12\angle 0° = 0$$

Solving these equations yields the remaining unknown mesh currents

$$\mathbf{I}_2 = 0.88\angle -6.34° \text{ A}$$

$$\mathbf{I}_3 = 3.91\angle 163.66° \text{ A}$$

$$\mathbf{I}_4 = 1.13\angle 72.35° \text{ A}$$

As a quick check we can use these currents to compute the node voltages. For example, if we calculate

$$\mathrm{V}_2 = -1(\mathrm{I}_3)$$

and

$$\mathrm{V}_1 = -j1(\mathrm{I}_2) + 12\angle 0°$$

we obtain the answers computed earlier.

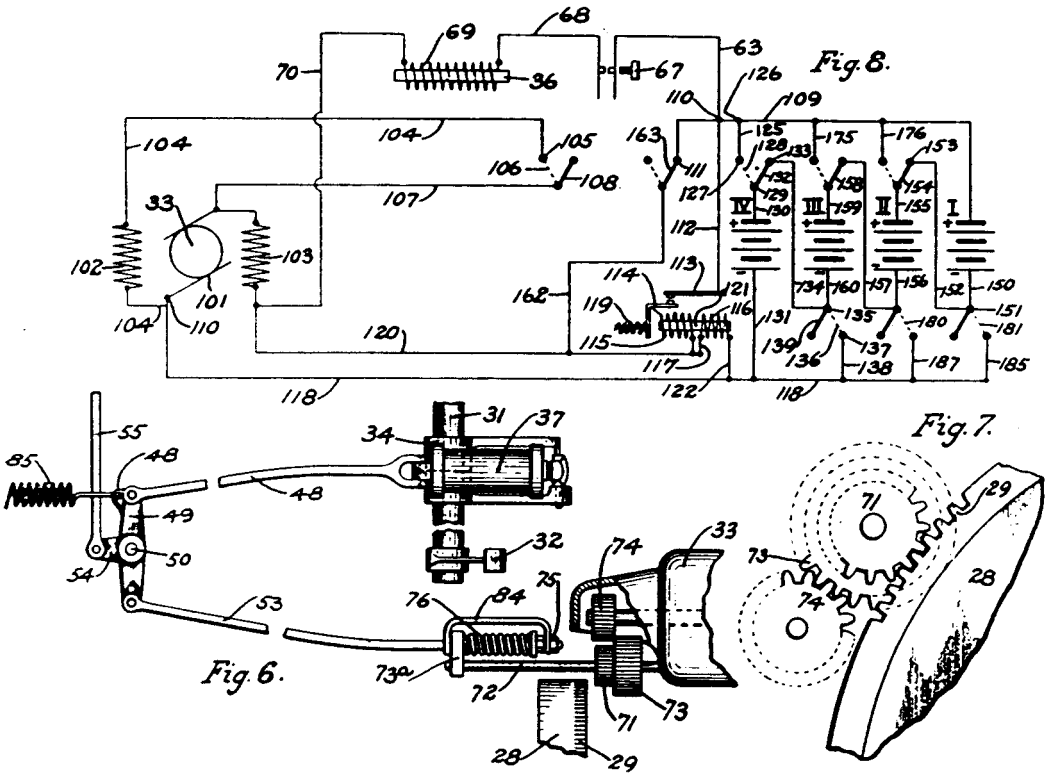
As a final point, because both node and mesh analysis will yield all currents and voltages in a network, which technique should be used? The answer to this question depends upon the network to be analyzed. If the network contains more voltage sources than current sources, node analysis might be the easier technique. If, however, the network contains more current sources than voltage sources, mesh analysis may be the easiest approach.

Engine-Starting Device

Charles F. Kettering
Patented August 17, 1915
#1,150,523

Early automobiles were all started with a crank, or arm-strong starters, as they were known. This backbreaking process was difficult for everyone, especially women. And it was dangerous. Backfires often resulted in broken wrists. Worse yet, if accidentally left in gear, the car could advance upon the person cranking. Numerous deaths and injuries were reported.

In 1910, Henry Leland, Cadillac Motors president, commissioned Charles Kettering and his Dayton Engineering Laboratories Company to develop an electric self-starter to replace the crank. Kettering had to overcome two large problems: (1) making a motor small enough to fit in a car yet powerful enough to crank the engine, and (2) finding a battery more powerful than any yet in existence. Electric Storage Battery of Philadelphia supplied an experimental 65-lb battery and Delco unveiled a working prototype electric "self-starter" system installed in a 1912 Cadillac on February 17, 1911. Leland immediately ordered 12,000 units for Cadillac. Within a few years, almost all new cars were equipped with electric starters. (Copyright © 1995, DewRay Products, Inc. Used with permission.)



Defining Terms

ac: An abbreviation for alternating current.
dc: An abbreviation for direct current.
Kirchhoff's current law (KCL): This law states that the algebraic sum of the currents either entering or leaving a node must be zero. Alternatively, the law states that the sum of the currents entering a node must be equal to the sum of the currents leaving that node.
Kirchhoff's voltage law (KVL): This law states that the algebraic sum of the voltages around any loop is zero. A loop is any closed path through the circuit in which no node is encountered more than once.
Mesh analysis: A circuit analysis technique in which KVL is used to determine the mesh currents in a network. A mesh is a loop that does not contain any loops within it.
Node analysis: A circuit analysis technique in which KCL is used to determine the node voltages in a network.
Ohm's law: A fundamental law which states that the voltage across a resistance is directly proportional to the current flowing through it.
Reference node: One node in a network that is selected to be a common point, and all other node voltages are measured with respect to that point.
Supernode: A cluster of node, interconnected with voltage sources, such that the voltage between any two nodes in the group is known.

Related Topics

3.1 Voltage and Current Laws • 3.6 Graph Theory

Reference

J.D. Irwin, *Basic Engineering Circuit Analysis,* 5th ed., Upper Saddle River, N.J.: Prentice-Hall, 1996.

3.3 Network Theorems

Allan D. Kraus

Linearity and Superposition

Linearity

Consider a system (which may consist of a single network element) represented by a block, as shown in Fig. 3.11, and observe that the system has an input designated by e (for excitation) and an output designated by r (for response). The system is considered to be *linear* if it satisfies the *homogeneity* and *superposition* conditions.

The homogeneity condition: If an arbitrary input to the system, e, causes a response, r, then if ce is the input, the output is cr where c is some arbitrary constant.

The superposition condition: If the input to the system, e_1, causes a response, r_1, and if an input to the system, e_2, causes a response, r_2, then a response, $r_1 + r_2$, will occur when the input is $e_1 + e_2$.

If neither the homogeneity condition nor the superposition condition is satisfied, the system is said to be *nonlinear.*

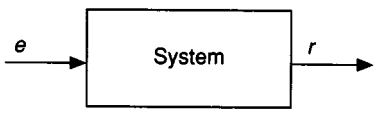


FIGURE 3.11 A simple system.

The Superposition Theorem

While both the homogeneity and superposition conditions are necessary for linearity, the superposition condition, in itself, provides the basis for the superposition theorem:

> If cause and effect are linearly related, the total effect due to several causes acting simultaneously is equal to the sum of the individual effects due to each of the causes acting one at a time.

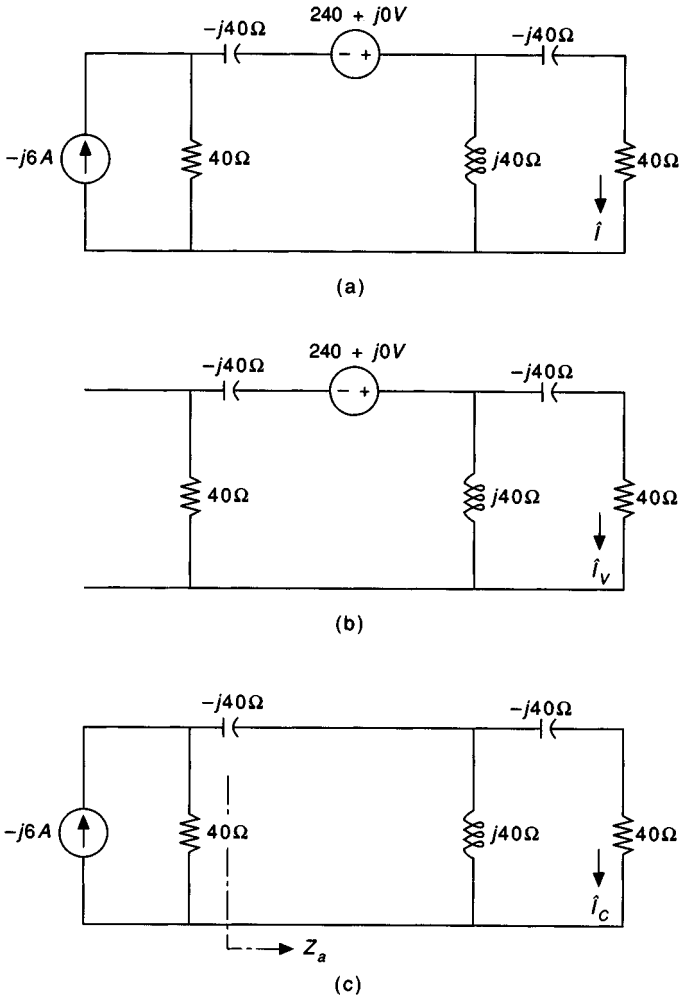


FIGURE 3.12 (a) A network to be solved by using superposition; (b) the network with the current source nulled; and (c) the network with the voltage source nulled.

Example 3.1. Consider the network driven by a current source at the left and a voltage source at the top, as shown in Fig. 3.12(a). The current phasor indicated by $\hat{I}$ is to be determined. According to the superposition theorem, the current $\hat{I}$ will be the sum of the two current components $\hat{I}_V$ due to the voltage source acting alone as shown in Fig. 3.12(b) and $\hat{I}_C$ due to the current source acting alone shown in Fig. 3.12(c).

$$\hat{I} = \hat{I}_V + \hat{I}_C$$

Figures 3.12(b) and (c) follow from the methods of removing the effects of independent voltage and current sources. Voltage sources are nulled in a network by replacing them with short circuits and current sources are nulled in a network by replacing them with open circuits.

The networks displayed in Figs. 3.12(b) and (c) are simple ladder networks in the phasor domain, and the strategy is to first determine the equivalent impedances presented to the voltage and current sources. In Fig. 3.12(b), the group of three impedances to the right of the voltage source are in series-parallel and possess an impedance of

$$Z_P = \frac{(40 - j40)(j40)}{40 + j40 - j40} = 40 + j40\ \Omega$$

and the total impedance presented to the voltage source is

$$Z = Z_P + 40 - j40 = 40 + j40 + 40 - j40 = 80\ \Omega$$

Then $\hat{I}_1$, the current leaving the voltage source, is

$$\hat{I}_1 = \frac{240 + j0}{80} = 3 + j0\,\text{A}$$

and by a current division

$$\hat{I}_V = \left[\frac{j40}{40 - j40 + j40}\right](3 + j0) = j(3 + j0) = 0 + j3\,\text{A}$$

In Fig. 3.12(b), the current source delivers current to the 40-Ω resistor and to an impedance consisting of the capacitor and Z_p. Call this impedance Z_a so that

$$Z_a = -j40 + Z_P = -j40 + 40 + j40 = 40\ \Omega$$

Then, two current divisions give $\hat{I}_C$

$$\hat{I}_C = \left[\frac{40}{40 + 40}\right]\left[\frac{j40}{40 - j40 + j40}\right](0 - j6) = \frac{j}{2}(0 - j6) = 3 + j0\,\text{A}$$

The current $\hat{I}$ in the circuit of Fig. 3.12(a) is

$$\hat{I} = \hat{I}_V + \hat{I}_C = 0 + j3 + (3 + j0) = 3 + j3\ \text{A}$$

The Network Theorems of Thévenin and Norton

If interest is to be focused on the voltages and across the currents through a small portion of a network such as network *B* in Fig. 3.13(a), it is convenient to replace network *A*, which is complicated and of little interest, by a simple equivalent. The simple equivalent may contain a single, equivalent, voltage source in series with an equivalent impedance in series as displayed in Fig. 3.13(b). In this case, the equivalent is called a *Thévenin equivalent.* Alternatively, the simple equivalent may consist of an equivalent current source in parallel with an equivalent impedance. This equivalent, shown in Fig. 3.13(c), is called a *Norton equivalent.* Observe that as long as Z_T (subscript *T* for Thévenin) is equal to Z_N (subscript *N* for Norton), the two equivalents may be obtained from one another by a simple source transformation.

Conditions of Application

The Thévenin and Norton network equivalents are only valid at the terminals of network *A* in Fig. 3.13(a) and they do not extend to its interior. In addition, there are certain restrictions on networks *A* and *B*. Network *A* may contain only linear elements but may contain both independent and dependent sources. Network *B*, on the other hand, is not restricted to linear elements; it may contain nonlinear or time-varying elements and may

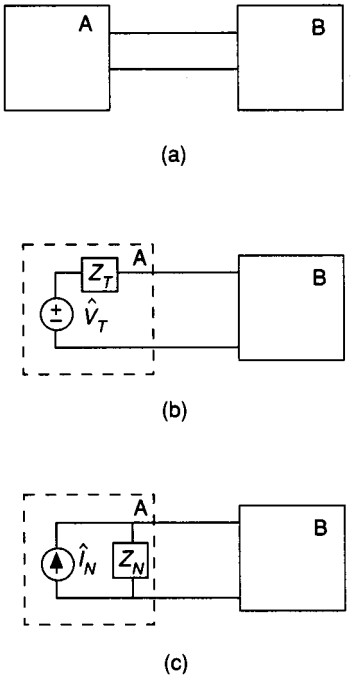


FIGURE 3.13 (a) Two one-port networks; (b) the Thévenin equivalent for network *a*; and (c) the Norton equivalent for network *a*.

also contain both independent and dependent sources. Together, there can be no controlled source coupling or magnetic coupling between networks *A* and *B*.

The Thévenin Theorem

The statement of the Thévenin theorem is based on Fig. 3.13(b):

> Insofar as a load which has no magnetic or controlled source coupling to a one-port is concerned, a network containing linear elements and both independent and controlled sources may be replaced by an ideal voltage source of strength, $\hat{V}_T$, and an equivalent impedance Z_T, in series with the source. The value of $\hat{V}_T$ is the open-circuit voltage, $\hat{V}_{OC}$, appearing across the terminals of the network and Z_T is the driving point impedance at the terminals of the network, obtained with all independent sources set equal to zero.

The Norton Theorem

The Norton theorem involves a current source equivalent. The statement of the Norton theorem is based on Fig. 3.13(c):

> Insofar as a load which has no magnetic or controlled source coupling to a one-port is concerned, the network containing linear elements and both independent and controlled sources may be replaced by an ideal current source of strength, $\hat{I}_N$, and an equivalent impedance, Z_N, in parallel with the source. The value of $\hat{I}_N$ is the short-circuit current, $\hat{I}_{SC}$, which results when the terminals of the network are shorted and Z_N is the driving point impedance at the terminals when all independent sources are set equal to zero.

The Equivalent Impedance, $Z_T = Z_N$

Three methods are available for the determination of Z_T. All of them are applicable at the analyst's discretion. When controlled sources are present, however, the first method cannot be used.

The first method involves the direct calculation of $Z_{eq} = Z_T = Z_N$ by looking into the terminals of the network after all independent sources have been nulled. Independent sources are nulled in a network by replacing all independent voltage sources with a short circuit and all independent current sources with an open circuit.

The second method, which may be used when controlled sources are present in the network, requires the computation of both the Thévenin equivalent voltage (the open-circuit voltage at the terminals of the network) and the Norton equivalent current (the current through the short-circuited terminals of the network). The equivalent impedance is the ratio of these two quantities

$$Z_T = Z_N = Z_{eq} = \frac{\hat{V}_T}{\hat{I}_N} = \frac{\hat{V}_{OC}}{\hat{I}_{SC}}$$

The third method may also be used when controlled sources are present within the network. A test voltage may be placed across the terminals with a resulting current calculated or measured. Alternatively, a test current may be injected into the terminals with a resulting voltage determined. In either case, the equivalent resistance can be obtained from the value of the ratio of the test voltage $\hat{V}_o$ to the resulting current $\hat{I}_o$

$$Z_T = \frac{\hat{V}_o}{\hat{I}_o}$$

Example 3.2. The current through the capacitor with impedance $-j35\ \Omega$ in Fig. 3.14(a) may be found using Thévenin's theorem. The first step is to remove the $-j35$-Ω capacitor and consider it as the load. When this is done, the network in Fig. 3.14(b) results.

The Thévenin equivalent voltage is the voltage across the 40–Ω resistor. The current through the 40-Ω resistor was found in Example 3.1 to be $I = 3 + j3\ \Omega$ Thus,

$$\hat{V}_T = 40(3 + j3) = 120 + j120\ \text{V}$$

The Thévenin equivalent impedance may be found by looking into the terminals of the network in Fig. 3.14(c). Observe that both sources in Fig. 3.14(a) have been nulled and that, for ease of computation, impedances Z_a and Z_b have been placed on Fig. 3.14(c). Here,

$$Z_a = \frac{(40 - j40)(j40)}{40 + j40 - j40} = 40 + j40\ \Omega$$

$$Z_b = \frac{(40)(40)}{40 + 40} = 20\ \Omega$$

and

$$Z_T = Z_b + j15 = 20 + j15\ \Omega$$

Both the Thévenin equivalent voltage and impedance are shown in Fig. 3.14(d), and when the load is attached, as in Fig. 3.14(d), the current can be computed as

$$\hat{I} = \frac{\hat{V}_T}{20 + j15 - j35} = \frac{120 + j120}{20 - j20} = 0 + j6\ \text{A}$$

The Norton equivalent circuit is obtained via a simple voltage-to-current source transformation and is shown in Fig. 3.15. Here it is observed that a single current division gives

$$\hat{I} = \left[\frac{20 + j15}{20 + j15 - j35}\right](6.72 + j0.96) = 0 + j6\ \text{A}$$

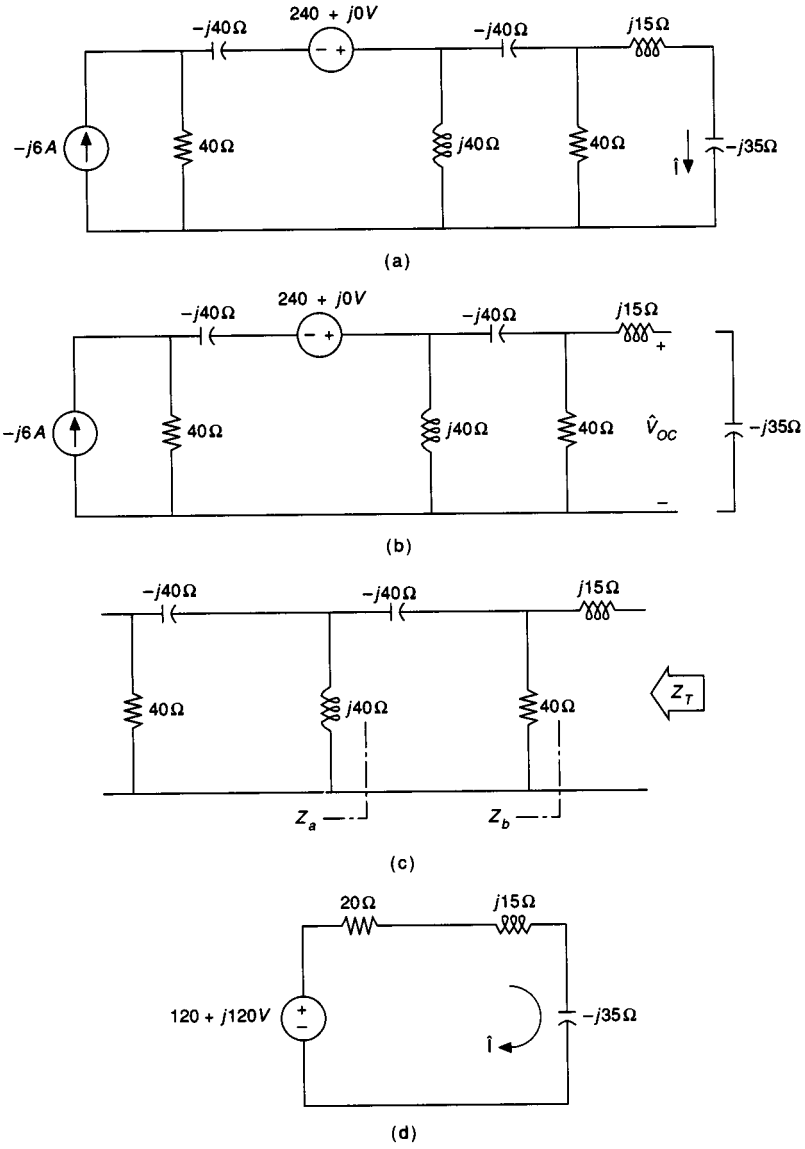


FIGURE 3.14 (a) A network in the phasor domain; (b) the network with the load removed; (c) the network for the computation of the Thévenin equivalent impedance; and (d) the Thévenin equivalent.

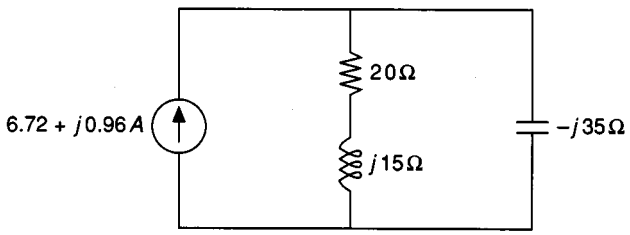


FIGURE 3.15 The Norton equivalent of Fig. 3.14(d).

Tellegen's Theorem

Tellegen's theorem states:

> In an arbitrarily lumped network subject to KVL and KCL constraints, with reference directions of the branch currents and branch voltages associated with the KVL and KCL constraints, the product of all branch currents and branch voltages must equal zero.

Tellegen's theorem may be summarized by the equation

$$\sum_{k=1}^{b} v_k j_k = 0$$

where the lower case letters v and j represent instantaneous values of the branch voltages and branch currents, respectively, and where b is the total number of branches. A matrix representation employing the branch current and branch voltage vectors also exists. Because **V** and **J** are column vectors

$$\mathbf{V} \cdot \mathbf{J} = \mathbf{V}^T\mathbf{J} = \mathbf{J}^T\mathbf{V}$$

The prerequisite concerning the KVL and KCL constraints in the statement of Tellegen's theorem is of crucial importance.

Example 3.3. Figure 3.16 displays an oriented graph of a particular network in which there are six branches labeled with numbers within parentheses and four nodes labeled by numbers within circles. Several known branch currents and branch voltages are indicated. Because the type of elements or their values is not germane to the construction of the graph, the other branch currents and branch voltages may be evaluated from repeated applications of KCL and KVL. KCL may be used first at the various nodes.

$$\text{node 3:} \quad j_2 = j_6 - j_4 = 4 - 2 = 2 \text{ A}$$

$$\text{node 1:} \quad j_3 = -j_1 - j_2 = -8 - 2 = -10 \text{ A}$$

$$\text{node 2:} \quad j_5 = j_3 - j_4 = -10 - 2 = -12 \text{ A}$$

Then KVL gives

$$v_3 = v_2 - v_4 = 8 - 6 = 2 \text{ V}$$

$$v_6 = v_5 - v_4 = -10 - 6 = -16 \text{ V}$$

$$v_1 = v_2 + v_6 = 8 - 16 = -8 \text{ V}$$

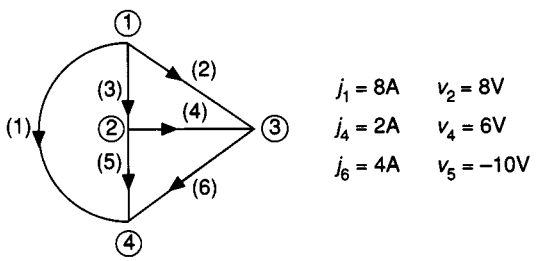


FIGURE 3.16 An oriented graph of a particular network with some known branch currents and branch voltages.

The transpose of the branch voltage and current vectors are

$$\mathbf{V}_T = [-8 \quad 8 \quad 2 \quad 6 \quad -10 \quad -16] \text{ V}$$

and

$$\mathbf{J}_T = [8 \quad 2 \quad -10 \quad 2 \quad -12 \quad 4] \text{ V}$$

The scalar product of **V** and **J** gives

$$-8(8) + 8(2) + 2(-10) + 6(2) + (-10)(-12) + (-16)(4) = -148 + 148 = 0$$

and Tellegen's theorem is confirmed.

Maximum Power Transfer

The maximum power transfer theorem pertains to the connections of a load to the Thévenin equivalent of a source network in such a manner as to transfer maximum power to the load. For a given network operating at a prescribed voltage with a Thévenin equivalent impedance

$$Z_T = |Z_T| \angle \theta_T$$

the real power drawn by any load of impedance

$$Z_o = |Z_o| \angle \theta_o$$

is a function of just two variables, $|Z_o|$ and θ_o. If the power is to be a maximum, there are three alternatives to the selection of $|Z_o|$ and θ_o:

(1) Both $|Z_o|$ and θ_o are at the designer's discretion and both are allowed to vary in any manner in order to achieve the desired result. In this case, the load should be selected to be the complex conjugate of the Thévenin equivalent impedance

$$Z_o = Z_T^*$$

(2) The angle, θ_o, is fixed but the magnitude, $|Z_o|$, is allowed to vary. For example, the analyst may select and fix $\theta_o = 0°$. This requires that the load be resistive (Z is entirely real). In this case, the value of the load resistance should be selected to be equal to the magnitude of the Thévenin equivalent impedance

$$R_o = |Z_T|$$

(3) The magnitude of the load impedance, $|Z_o|$, can be fixed, but the impedance angle, θ_o, is allowed to vary. In this case, the value of the load impedance angle should be

$$\theta_o = \arcsin\left[-\frac{2|Z_o||Z_T|\sin\theta_T}{|Z_o|^2 + |Z_T|^2}\right]$$

Example 3.4. Figure 3.17(a) is identical to Fig. 3.14(b) with the exception of a load, Z_o, substituted for the capacitive load. The Thévenin equivalent is shown in Fig. 3.17(b). The value of Z_o to transfer maximum power

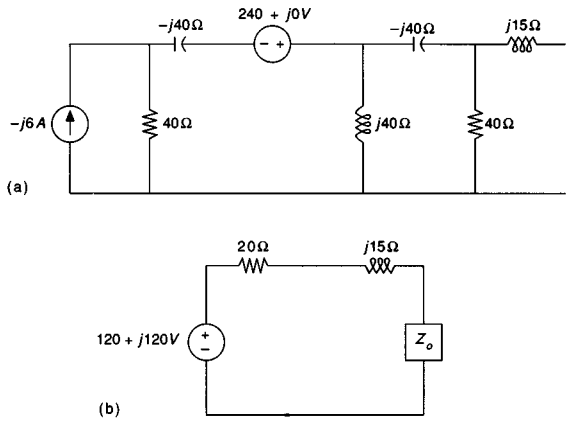


FIGURE 3.17 (a) A network for which the load, Z_o, is to be selected for maximum power transfer, and (b) the Thévenin equivalent of the network.

is to be found if its elements are unrestricted, if it is to be a single resistor, or if the magnitude of Z_o must be 20 Ω but its angle is adjustable.

For maximum power transfer to Z_o when the elements of Z_o are completely at the discretion of the network designer, Z_o must be the complex conjugate of Z_T

$$Z_o = Z_T^* = 20 - j15\ \Omega$$

If Z_o is to be a single resistor, R_o, then the magnitude of $Z_o = R_o$ must be equal to the magnitude of Z_T. Here

$$Z_T = 20 + j15 = 25\ \underline{/36.87°}$$

so that

$$R_o = |Z_o| = 25\ \Omega$$

If the magnitude of Z_o must be 20 Ω but the angle is adjustable, the required angle is calculated from

$$\begin{aligned}\theta_o &= \arcsin\left[-\frac{2|Z_o||Z_T|}{|Z_o|^2 + |Z_T|^2}\sin\theta_T\right] \\ &= \arcsin\left[-\frac{2(20)(25)}{(20)^2 + (25)^2}\sin\ \underline{/36.87°}\right] \\ &= \arcsin(-0.585) = -35.83°\end{aligned}$$

This makes Z_o

$$Z_o = 20\ \underline{/-35.83°} = 16.22 - j11.71\ \Omega$$

The Reciprocity Theorem

The reciprocity theorem is a useful general theorem that applies to all linear, passive, and bilateral networks. However, it applies only to cases where current and voltage are involved.

The ratio of a single excitation applied at one point to an observed response at another is invariant with respect to an interchange of the points of excitation and observation.

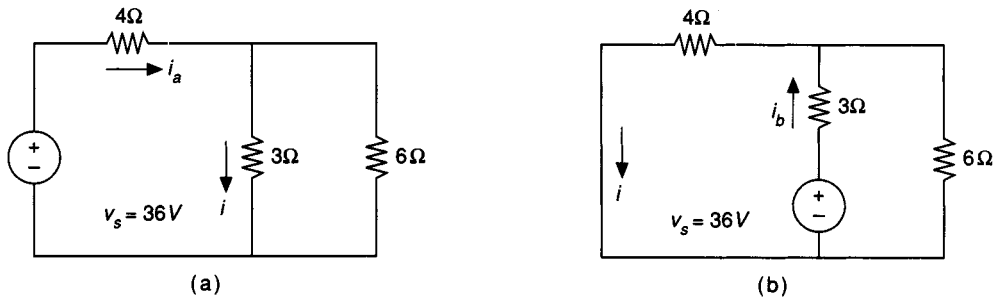


FIGURE 3.18 Two networks which can be used to illustrate the reciprocity principle.

The reciprocity principle also applies if the excitation is a current and the observed response is a voltage. It will not apply, in general, for voltage–voltage and current–current situations, and, of course, it is not applicable to network models of nonlinear devices.

Example 3.5. It is easily shown that the positions of v_s and i in Fig. 3.18(a) may be interchanged as in Fig. 3.18(b) without changing the value of the current i.

In Fig. 3.18(a), the resistance presented to the voltage source is

$$R = 4 + \frac{3(6)}{3+6} = 4 + 2 = 6\ \Omega$$

Then

$$i_a = \frac{v_s}{R} = \frac{36}{6} = 6\ \text{A}$$

and by current division

$$i_a = \frac{6}{6+3} i_a = \left(\frac{2}{3}\right) 6 = 4\ \text{A}$$

In Fig. 3.18(b), the resistance presented to the voltage source is

$$R = 3 + \frac{6(4)}{6+4} = 3 + \frac{12}{5} = \frac{27}{5}\ \Omega$$

Then

$$i_b = \frac{v_s}{R} = \frac{36}{27/5} = \frac{180}{27} = \frac{20}{3}\ \text{A}$$

and again, by current division

$$i = \frac{6}{4+6} i_b = \left(\frac{3}{5}\right)\frac{20}{3} = 4\ \text{A}$$

The network is reciprocal.

The Substitution and Compensation Theorems

The Substitution Theorem

Any branch in a network with branch voltage, v_k, and branch current, i_k, can be replaced by another branch provided it also has branch voltage, v_k, and branch current, i_k.

The Compensation Theorem

In a linear network, if the impedance of a branch carrying a current $\hat{I}$ is changed from Z to $Z + \Delta Z$, then the corresponding change of any voltage or current elsewhere in the network will be due to a compensating voltage source, $\Delta Z\hat{I}$, placed in series with $Z + \Delta Z$ with polarity such that the source, $\Delta Z\hat{I}$, is opposing the current $\hat{I}$.

Defining Terms

Linear network: A network in which the parameters of resistance, inductance, and capacitance are constant with respect to voltage or current or the rate of change of voltage or current and in which the voltage or current of sources is either independent of or proportional to other voltages or currents, or their derivatives.

Maximum power transfer theorem: In any electrical network which carries direct or alternating current, the maximum possible power transferred from one section to another occurs when the impedance of the section acting as the load is the complex conjugate of the impedance of the section that acts as the source. Here, both impedances are measured across the pair of terminals in which the power is transferred with the other part of the network disconnected.

Norton theorem: The voltage across an element that is connected to two terminals of a linear, bilateral network is equal to the short-circuit current between these terminals in the absence of the element, divided by the admittance of the network looking back from the terminals into the network, with all generators replaced by their internal admittances.

Principle of superposition: In a linear electrical network, the voltage or current in any element resulting from several sources acting together is the sum of the voltages or currents from each source acting alone.

Reciprocity theorem: In a network consisting of linear, passive impedances, the ratio of the voltage introduced into any branch to the current in any other branch is equal in magnitude and phase to the ratio that results if the positions of the voltage and current are interchanged.

Thévenin theorem: The current flowing in any impedance connected to two terminals of a linear, bilateral network containing generators is equal to the current flowing in the same impedance when it is connected to a voltage generator whose voltage is the voltage at the open-circuited terminals in question and whose series impedance is the impedance of the network looking back from the terminals into the network, with all generators replaced by their internal impedances.

Related Topics

2.2 Ideal and Practical Sources • 3.4 Power and Energy

References

J. D. Irwin, *Basic Engineering Circuit Analysis,* 4th ed., New York: Macmillan, 1993.
A. D. Kraus, *Circuit Analysis,* St. Paul: West Publishing, 1991.
J. W. Nilsson, *Electric Circuits,* 5th ed., Reading, Mass.: Addison-Wesley, 1995.

Further Information

Three texts listed in the References have achieved widespread usage and contain more details on the material contained in this section.

3.4 Power and Energy

Norman Balabanian and Theodore A. Bickart

The concept of the voltage v between two points was introduced in Section 3.1 as the energy w expended per unit charge in moving the charge between the two points. Coupled with the definition of current i as the time rate of charge motion and that of power p as the time rate of change of energy, this leads to the following fundamental relationship between the power delivered to a two-terminal electrical component and the voltage and current of that component, with standard references (meaning that the voltage reference plus is at the tail of the current reference arrow) as shown in Fig. 3.19:

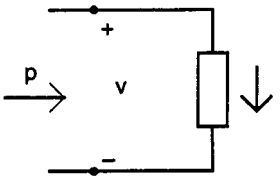


FIGURE 3.19 Power delivered to a circuit.

$$p = vi \tag{3.1}$$

Assuming that the voltage and current are in volts and amperes, respectively, the power is in *watts.* This relationship applies to any two-terminal component or network, whether linear or nonlinear.

The power delivered to the basic linear resistive, inductive, and capacitive elements is obtained by inserting the v-i relationships into this expression. Then, using the relationship between power and energy (power as the time derivative of energy and energy, therefore, as the integral of power), the energy stored in the capacitor and inductor is also obtained:

$$\begin{aligned} p_R &= v_R i_R = Ri^2 \\ p_C &= v_C i_C = Cv_C \frac{dv_C}{dt} \qquad & w_C(t) &= \int_0^t Cv_C \frac{dv_C}{dt}\,dt = \frac{1}{2}Cv_C^2(t) \\ p_L &= v_L i_L = Li_L \frac{di_L}{dt} & w_L(t) &= \int_0^t Li_L \frac{di_L}{dt}\,dt = \frac{1}{2}Li_L^2(t) \end{aligned} \tag{3.2}$$

where the origin of time ($t = 0$) is chosen as the time when the capacitor voltage (respectively, the inductor current) is zero.

Tellegen's Theorem

A result that has far-reaching consequences in electrical engineering is Tellegen's theorem. It will be stated in terms of the networks shown in Fig. 3.20. These two are said to be topologically equivalent; that is, they are represented by the same graph but the components that constitute the branches of the graph are not necessarily the same in the two networks. they can even be nonlinear, as illustrated by the diode in one of the networks. Assuming all branches have standard references, including the source branches, Tellegen's theorem states that

$$\begin{aligned} \sum_{\text{all } j} v_{bj} i_{aj} &= 0 \\ \mathbf{v}_b' \mathbf{i}_a &= 0 \end{aligned} \tag{3.3}$$

In the second line, the variables are vectors and the prime stands for the transpose. The a and b subscripts refer to the two networks.

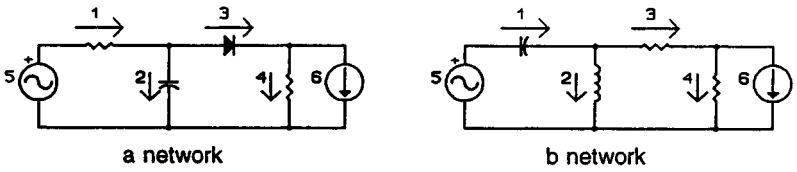


FIGURE 3.20 Topologically equivalent networks.

This is an amazing result. It can be easily proved with the use of Kirchhoff's two laws.[1] The products of v and i are reminiscent of power as in Eq. (3.1). However, the product of the voltage of a branch in one network and the current of its topologically corresponding branch (which may not even be the same type of component) in another network does not constitute power in either branch. furthermore, the variables in one network might be functions of time, while those of the other network might be steady-state phasors or Laplace transforms.

Nevertheless, some conclusions about power can be derived from Tellegen's theorem. Since a network is topologically equivalent to itself, the b network can be the same as the a network. In that case each vi product in Eq. (3.3) represents the power delivered to the corresponding branch, including the sources. The equation then says that if we add the power delivered to all the branches of a network, the result will be zero.

This result can be recast if the sources are separated from the other branches and one of the references of each source (current reference for each v-source and voltage reference for each i-source) is reversed. Then the vi product for each source, with new references, will enter Eq. (3.3) with a negative sign and will represent the power supplied by this source. When these terms are transposed to the right side of the equation, their signs are changed. The new equation will state in mathematical form that

> In any electrical network, the sum of the power supplied by the sources is equal to the sum of the power delivered to all the nonsource branches.

This is not very surprising since it is equivalent to the law of conservation of energy, a fundamental principle of science.

AC Steady-State Power

Let us now consider the ac steady-state case, where all voltages and currents are sinusoidal. Thus, in the two-terminal circuit of Fig. 3.19:

$$\begin{aligned} v(t) &= \sqrt{2}\,|V|\cos(\omega t + \alpha) \leftrightarrow V = |V|\,e^{j\alpha} \\ i(t) &= \sqrt{2}\,|I|\cos(\omega t + \beta) \leftrightarrow I = |I|\,e^{j\beta} \end{aligned} \tag{3.4}$$

The capital V and I are phasors representing the voltage and current, and their magnitudes are the corresponding rms values. The power delivered to the network at any instant of time is given by:

$$\begin{aligned} p(t) &= v(t)i(t) = 2\,|V|\,|I|\cos(\omega t + \alpha)\cos(\omega t + \beta) \\ &= \Big[|V|\,|I|\cos(\alpha - \beta)\Big] + \Big[|V|\,|I|\cos(2\omega t + \alpha + \beta)\Big] \end{aligned} \tag{3.5}$$

The last form is obtained by using trigonometric identities for the sum and difference of two angles. Whereas both the voltage and the current are sinusoidal, the instantaneous power contains a constant term (independent

[1]See, for example, N. Balabanian and T. A. Bickart, *Linear Network Theory,* Matrix Publishers, Chesterland, Ohio, 1981, chap. 9.

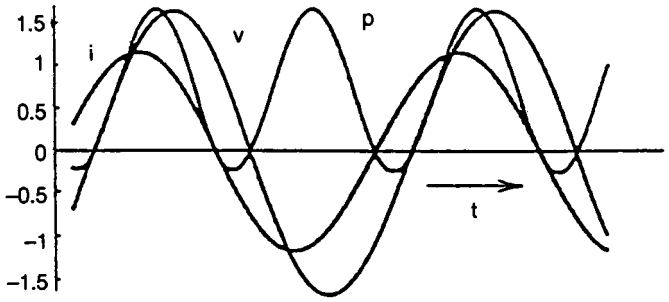


FIGURE 3.21 Instantaneous voltage, current, and power.

of time) in addition to a sinusoidal term. Furthermore, the frequency of the sinusoidal term is twice that of the voltage or current. Plots of *v*, *i*, and *p* are shown in Fig. 3.21 for specific values of α and β. The power is sometimes positive, sometimes negative. This means that power is sometimes delivered to the terminals and sometimes extracted from them.

The energy which is transmitted into the network over some interval of time is found by integrating the power over this interval. If the area under the positive part of the power curve were the same as the area under the negative part, the net energy transmitted over one cycle would be zero. For the values of α and β used in the figure, however, the positive area is greater, so there is a net transmission of energy toward the network. The energy flows back from the network to the source over part of the cycle, but on the average, more energy flows toward the network than away from it.

In Terms of RMS Values and Phase Difference

Consider the question from another point of view. The preceding equation shows the power to consist of a constant term and a sinusoid. The average value of a sinusoid is zero, so this term will contribute nothing to the net energy transmitted. Only the constant term will contribute. This constant term is the average value of the power, as can be seen either from the preceding figure or by integrating the preceding equation over one cycle. Denoting the average power by *P* and letting $\theta = \alpha - \beta$, which is the angle of the network impedance, the average power becomes:

$$\begin{aligned} P &= |V||I| \cos\theta \\ &= |V||I| \operatorname{Re}\left(e^{j\theta}\right) = \operatorname{Re}\left[|V||I| e^{j(\alpha-\beta)}\right] \\ &= \operatorname{Re}\left[\left(|V| e^{j\alpha}\right)\left(|I| e^{-j\beta}\right)\right] \\ &= \operatorname{Re}\left(VI^*\right) \end{aligned} \tag{3.6}$$

The third line is obtained by breaking up the exponential in the previous line by the law of exponents. The first factor between square brackets in this line is identified as the phasor voltage and the second factor as the conjugate of the phasor current. The last line then follows. It expresses the average power in terms of the voltage and current phasors and is sometimes more convenient to use.

Complex and Reactive Power

The average ac power is found to be the real part of a complex quantity VI^*, labeled *S*, that in rectangular form is

$$\begin{aligned} S = VI^* &= |V||I| e^{j\theta} = |V||I| \cos\theta + j|V||I| \sin\theta \\ &= P + jQ \end{aligned} \tag{3.7}$$

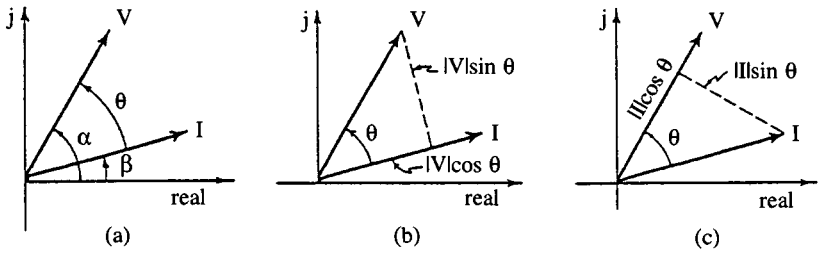


FIGURE 3.22 In-phase and quadrature components of V and I.

where

$$P = |V||I|\cos\theta \quad (a)$$
$$Q = |V||I|\sin\theta \quad (b) \qquad (3.8)$$
$$|S| = |V||I| \quad (c)$$

We already know P to be the average power. Since it is the real part of some complex quantity, it would be reasonable to call it the **real power**. The complex quantity S of which P is the real part is, therefore, called the *complex power*. Its magnitude is the product of the rms values of voltage and current: $|S| = |V|\,|I|$. It is called the *apparent power* and its unit is the volt-ampere (VA). To be consistent, then we should call Q the *imaginary power*. This is not usually done, however; instead, Q is called the **reactive power** and its unit is a VAR (volt-ampere reactive).

Phasor and Power Diagrams

An interpretation useful for clarifying and understanding the preceding relationships and for the calculation of power is a graphical approach. Figure 3.22(a) is a phasor diagram of V and I in a particular case. The phasor voltage can be resolved into two components, one parallel to the phasor current (or in phase with I) and another perpendicular to the current (or in quadrature with it). This is illustrated in Fig. 3.22(b). Hence, the average power P is the magnitude of phasor I multiplied by the in-phase component of V; the reactive power Q is the magnitude of I multiplied by the quadrature component of V.

Alternatively, one can imagine resolving phasor I into two components, one in phase with V and one in quadrature with it, as illustrated in Fig. 3.22(c). Then P is the product of the magnitude of V with the in-phase component of I, and Q is the product of the magnitude of V with the quadrature component of I. Real power is produced only by the in-phase components of V and I. The quadrature components contribute only to the reactive power.

The in-phase or quadrature components of V and I do not depend on the specific values of the angles of each, but on their phase difference. One can imagine the two phasors in the preceding diagram to be rigidly held together and rotated around the origin by any angle. As long as the angle θ is held fixed, all of the discussion of this section will still apply. It is common to take the current phasor as the reference for angle; that is, to choose $\beta = 0$ so that phasor I lies along the real axis. Then $\theta = \alpha$.

Power Factor

For any given circuit it is useful to know what part of the total complex power is real (average) power and what part is reactive power. This is usually expressed in terms of the **power factor** F_p, defined as the ratio of real power to apparent power:

$$\text{Power factor} \doteq F_p = \frac{P}{|S|} = \frac{P}{|V||I|} \qquad (3.9)$$

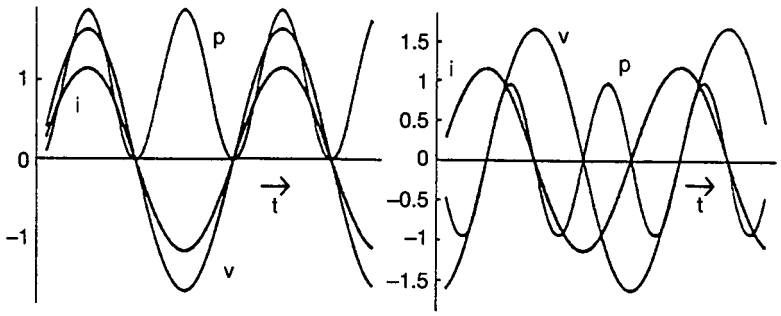


FIGURE 3.23 Power waveform for unity and zero power factors.

Not counting the right side, this is a general relationship, although we developed it here for sinusoidal excitations. With $P = |V||I|\cos\theta$, we find that the power factor is simply $\cos\theta$. Because of this, θ itself is called the power factor angle.

Since the cosine is an even function [$\cos(-\theta) = \cos\theta$], specifying the power factor does not reveal the sign of θ. Remember that θ is the angle of the impedance. If θ is positive, this means that the current lags the voltage; we say that the power factor is a *lagging* power factor. On the other hand, if θ is negative, the current leads the voltage and we say this represent a *leading* power factor.

The power factor will reach its maximum value, unity, when the voltage and current are in phase. This will happen in a purely resistive circuit, of course. It will also happen in more general circuits for specific element values and a specific frequency.

We can now obtain a physical interpretation for the reactive power. When the power factor is unity, the voltage and current are in phase and $\sin\theta = 0$. Hence, the reactive power is zero. In this case, the instantaneous power is never negative. This case is illustrated by the current, voltage, and power waveforms in Fig. 3.23; the power curve never dips below the axis, and there is no exchange of energy between the source and the circuit. At the other extreme, when the power factor is zero, the voltage and current are 90° out of phase and $\sin\theta = 1$. Now the reactive power is a maximum and the average power is zero. In this case, the instantaneous power is positive over half a cycle (of the voltage) and negative over the other half. All the energy delivered by the source over half a cycle is returned to the source by the circuit over the other half.

It is clear, then, that the reactive power is a measure of the exchange of energy between the source and the circuit without being used by the circuit. Although none of this exchanged energy is dissipated by or stored in the circuit, and it is returned unused to the source, nevertheless it is temporarily made available to the circuit by the source.[1]

Average Stored Energy

The average ac energy stored in an inductor or a capacitor can be established by using the expressions for the instantaneous stored energy for arbitrary time functions in Eq. (3.2), specifying the time function to be sinusoidal, and taking the average value of the result.

$$W_L = \frac{1}{2} L |I|^2 \qquad W_C = \frac{1}{2} C |V|^2 \tag{3.10}$$

[1]Power companies charge their industrial customers not only for the average power they use but for the reactive power they return. There is a reason for this. Suppose a given power system is to deliver a fixed amount of average power at a constant voltage amplitude. Since $P = |V||I|\cos\theta$, the current will be inversely proportional to the power factor. If the reactive power is high, the power factor will be low and a high current will be required to deliver the given power. To carry a large current, the conductors carrying it to the customer must be correspondingly larger and better insulated, which means a larger capital investment in physical plant and facilities. It may be cost effective for customers to try to reduce the reactive power they require, even if they have to buy additional equipment to do so.

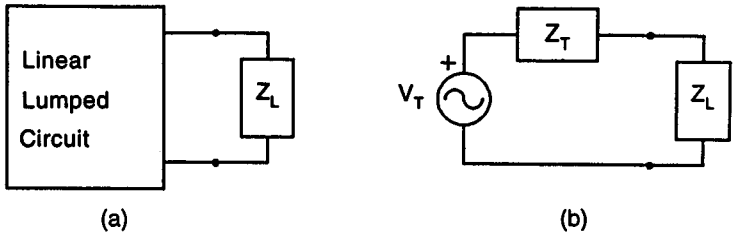


FIGURE 3.24 A linear circuit delivering power to a load in the steady state.

Application of Tellegen's Theorem to Complex Power

An example of two topologically equivalent networks was shown in Fig. 3.20. Let us now specify that two such networks are linear, all sources are same-frequency sinusoids, they are operating in the steady state, and all variables are phasors. Furthermore, suppose the two networks are the same, except that the sources of network *b* have phasors that are the complex conjugates of those of network *a*. Then, if **V** and **I** denote the vectors of branch voltages and currents of network *a*, Tellegen's theorem in Eq. (3.3) becomes:

$$\sum_{\text{all } j} \dot{V}_j^{\star} I_j = V^{\star} I = 0 \tag{3.11}$$

where $\mathbf{V}^{\star}$ is the conjugate transpose of vector **V**.

This result states that the sum of the complex power delivered to all branches of a linear circuit operating in the ac steady state is zero. Alternatively stated, the total complex power delivered to a network by its sources equals the sum of the complex power delivered to its nonsource branches. Again, this result is not surprising. Since, if a complex quantity is zero, both the real and imaginary parts must be zero, the same result can be stated for the average power and for the reactive power.

Maximum Power Transfer

The diagram in Fig. 3.24 illustrates a two-terminal linear circuit at whose terminals an impedance Z_L is connected. The circuit is assumed to be operating in the ac steady state. The problem to be addressed is this: given the two-terminal circuit, how can the impedance connected to it be adjusted so that the maximum possible average power is transferred from the circuit to the impedance?

The first step is to replace the circuit by its Thévenin equivalent, as shown in Fig. 3.24(b). The current phasor in this circuit is $I = V_T/(Z_T + Z_L)$. The average power transferred by the circuit to the impedance is:

$$\begin{aligned} P &= |I|^2 \,\text{Re}(Z_L) = \frac{|V_T|^2 \,\text{Re}(Z_L)}{|Z_T + Z_L|^2} \\ &= \frac{|V_T|^2 \, R_L}{(R_T + R_L)^2 + (X_T + X_L)^2} \end{aligned} \tag{3.12}$$

In this expression, only the load (that is, R_L and X_L) can be varied. The preceding equation, then, expresses a dependent variable (P) in terms of two independent ones (R_L and X_L).

What is required is to maximize *P*. For a function of more than one variable, this is done by setting the partial derivatives with respect to each of the independent variables equal to zero; that is, $\partial P/\partial R_L = 0$ and $\partial P/\partial X_L = 0$. Carrying out these differentiations leads to the result that maximum power will be transferred when the load impedance is the conjugate of the Thévenin impedance of the circuit: $Z_L = Z_T^{\star}$. If the Thévenin impedance is purely resistive, then the load resistance must equal the Thévenin resistance.

In some cases, both the load impedance and the Thévenin impedance of the source may be fixed. In such a case, the matching for maximum power transfer can be achieved by using a transformer, as illustrated in Fig. 3.25, where the impedances are both resistive. The transformer is assumed to be ideal, with turns ratio n. Maximum power is transferred if $n^2 = R_T/R_L$.

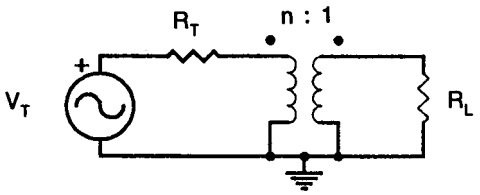


FIGURE 3.25 Matching with an ideal transformer.

Measuring AC Power and Energy

With ac steady-state average power given in the first line of Eq. (3.6), measuring the average power requires measuring the rms values of voltage and current, as well as the power factor. This is accomplished by the arrangement shown in Fig. 3.26, which includes a breakout of an electrodynamometer-type wattmeter. The current in the high-resistance pivoted coil is proportional to the voltage across the load. The current to the load and the pivoted coil together through the energizing coil of the electromagnet establishes a proportional magnetic field across the cylinder of rotation of the pivoted coil. the torque on the pivoted coil is proportional to the product of the magnetic field strength and the current in the pivoted coil. If the current in the pivoted coil is negligible compared to that in the load, then the torque becomes essentially proportional to the product of the voltage across the load (equal to that across the pivoted coil) and the current in the load (essentially equal to that through the energizing coil of the electromagnet). The dynamics of the pivoted coil together with the restraining spring, at ac power frequencies, ensures that the angular displacement of the pivoted coil becomes proportional to the average of the torque or, equivalently, the average power.

One of the most ubiquitous of electrical instruments is the induction-type watthour meter, which measures the energy delivered to a load. Every customer of an electrical utility has one, for example. In this instance the pivoted coil is replaced by a rotating conducting (usually aluminum) disk as shown in Fig. 3.27. An induced eddy current in the disk replaces the pivoted coil current interaction with the load-current-established magnetic field. After compensating for the less-than-ideal nature of the electrical elements making up the meter as just described, the result is that the disk rotates at a rate proportional to the average power to the load and the rotational count is proportional to the energy delivered to the load.

At frequencies above the ac power frequencies and, in some instances, at the ac power frequencies, electronic instruments are available to measure power and energy. They are not a cost-effective substitute for these meters in the monitoring of power and energy delivered to most of the millions upon millions of homes and businesses.

Defining Terms

AC steady-state power: Consider an ac source connected at a pair of terminals to an otherwise isolated network. Let $\sqrt{2}\cdot|V|$ and $\sqrt{2}\cdot|I|$ denote the peak values, respectively, of the ac steady-state voltage and current at the terminals. Furthermore, let θ denote the phase angle by which the voltage leads the current. Then the average power delivered by the source to the network would be expressed as $P = |V|\cdot|I|\cos(\theta)$.

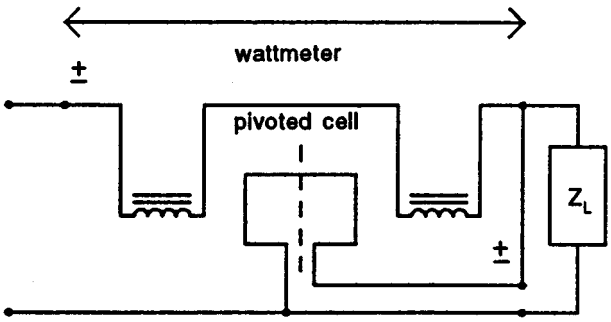


FIGURE 3.26 A wattmeter connected to a load.

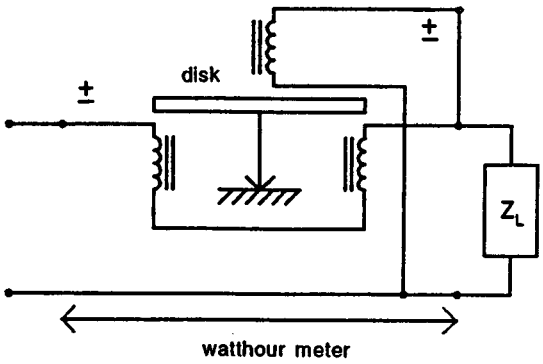


FIGURE 3.27 A watthour meter connected to a load.

Power and energy: Consider an electrical source connected at a pair of terminals to an otherwise isolated network. Power, denoted by p, is the time rate of change in the energy delivered to the network by the source. This can be expressed as $p = vi$, where v, the voltage across the terminals, is the energy expended per unit charge in moving the charge between the pair of terminals and i, the current through the terminals, is the time rate of charge motion.

Power factor: Consider an ac source connected at a pair of terminals to an otherwise isolated network. The power factor, the ratio of the real power to the apparent power $|V| \cdot |I|$, is easily established to be $\cos(\theta)$, where θ is the power factor angle.

Reactive power: Consider an ac source connected at a pair of terminals to an otherwise isolated network. The reactive power is a measure of the energy exchanged between the source and the network without being dissipated in the network. The reactive power delivered would be expressed as $Q = |V| \cdot |I| \sin(\theta)$.

Real power: Consider an ac source connected at a pair of terminals to an otherwise isolated network. The real power, equal to the average power, is the power dissipated by the source in the network.

Tellegen's theorem: Two networks, here including all sources, are topologically equivalent if they are similar structurally, component by component. Tellegen's theorem states that the sum over all products of the product of the current of a component of ne network, network a, and of the voltage of the corresponding component of the other network, network b, is zero. This would be expressed as $\Sigma_{\text{all } j}\, v_{bj}\, i_{aj} = 0$. From this general relationship it follows that in any electrical network, the sum of the power supplied by the sources is equal to the sum of the power delivered to all the nonsource components.

Related Topic

3.3 Network Theorems

References

N. Balabanian, *Electric Circuits,* New York: McGraw-Hill, 1994.

A. E. Fitzgerald, D. E. Higginbotham, and A. Grabel, *Basic Electrical Engineering,* 5th ed., New York: McGraw-Hill, 1981.

W. H. Hayt, Jr. and J. E. Kemmerly, *Engineering Circuit Analysis,* 4th ed., New York: McGraw-Hill, 1986.

J. D. Irwin, *Basic Engineering Circuit Analysis,* New York: Macmillan, 1995.

D. E. Johnson, J. L. Hilburn, and J. R. Johnson, *Basic Electric Circuit Analysis,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1990.

T. N. Trick, *Introduction to Circuit Analysis,* New York: John Wiley, 1977.

3.5 Three-Phase Circuits

Norman Balabanian

Figure 3.28(a) represents the basic circuit for considering the flow of power from a single sinusoidal source to a load. The power can be thought to cross an imaginary boundary surface (represented by the dotted line in the figure) separating the source from the load. Suppose that:

$$\begin{aligned} v(t) &= \sqrt{2}\,|V|\cos(\omega t + \alpha) \\ i(t) &= \sqrt{2}\,|I|\cos(\omega t + \beta) \end{aligned} \tag{3.13}$$

Then the power to the load at any instant of time is

$$p(t) = |V||I|\,[\cos(\alpha - \beta) + \cos\,(2\omega t + \alpha + \beta)] \tag{3.14}$$

The instantaneous power has a constant term and a sinusoidal term at twice the frequency. The quantity in brackets fluctuates between a minimum value of $\cos(\alpha - \beta) - 1$ and a maximum value of $\cos(\alpha - \beta) + 1$. This fluctuation of power delivered to the load has certain disadvantages in some situations where the transmission of power is the purpose of a system. An electric motor, for example, operates by receiving electric power and transmitting mechanical (rotational) power at its shaft. If the electric power is delivered to the motor in spurts, the motor is likely to vibrate. In order to run satisfactorily, a physically larger motor will be needed, with a larger shaft and flywheel, to provide inertia than would be the case if the delivered power were constant.

This problem is overcome in practice by the use of what is called a *three-phase* system. This section will provide a brief discussion of three-phase systems.

Consider the circuit in Fig. 3.28(b). This arrangement is similar to a combination of three of the simple circuits in Fig. 3.28(a) connected in such a way that each one shares the return connection from *O* to *N*. The three sources can be viewed collectively as a single source and the three loads—which are assumed to be

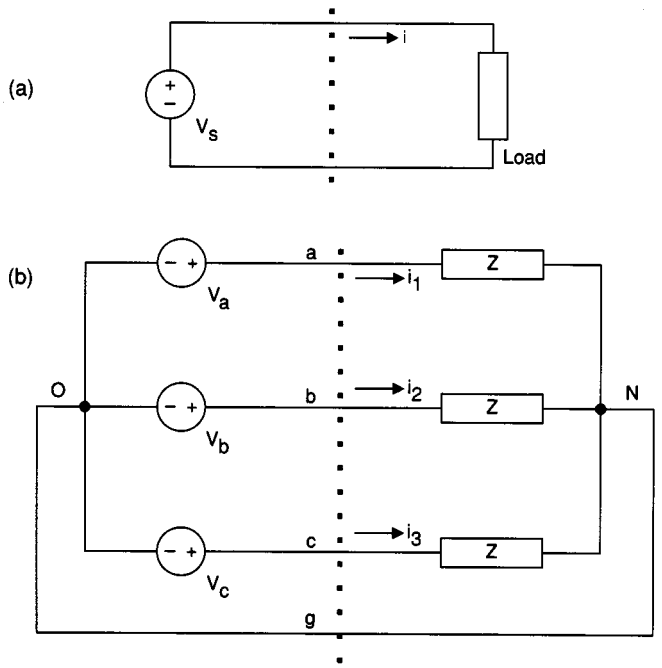


FIGURE 3.28 Flow of power from source to load.

identical—can be viewed collectively as a single load. Then, as before, the dotted line represents a surface separating the source from the load. Each of the individual sources and loads is referred to as one *phase* of the three-phase system.

The three sources are assumed to have the same frequency; they are said to be *synchronized*. It is also assumed that the three voltages have the same rms values and the phase difference between each pair of voltages is ±120° (2π/3 rad). Thus, they can be written:

$$\begin{aligned} v_a &= \sqrt{2}\,|V|\cos(\omega t + \alpha_1) \quad \leftrightarrow \quad V_a = |V|\,e^{j0^\circ} \\ v_b &= \sqrt{2}\,|V|\cos(\omega t + \alpha_2) \quad \leftrightarrow \quad V_b = |V|\,e^{-j120^\circ} \\ v_c &= \sqrt{2}\,|V|\cos(\omega t + \alpha_3) \quad \leftrightarrow \quad V_c = |V|\,e^{j120^\circ} \end{aligned} \tag{3.15}$$

The **phasors** representing the sinusoids have also been shown. For convenience, the angle of v_a has been chosen as the reference for angles; v_b *lags* v_a by 120° and v_c *leads* v_a by 120°.

Because the loads are identical, the rms values of the three currents shown in the figure will also be the same and the phase difference between each pair of them will be ±120°. Thus, the currents can be written:

$$\begin{aligned} i_1 &= \sqrt{2}\,|I|\cos(\omega t + \beta_1) \leftrightarrow I_1 = |I|\,e^{j\beta_1} \\ i_2 &= \sqrt{2}\,|I|\cos(\omega t + \beta_2) \leftrightarrow I_2 = |I|\,e^{j(\beta_1 - 120^\circ)} \\ i_3 &= \sqrt{2}\,|I|\cos(\omega t + \beta_3) \leftrightarrow I_3 = |I|\,e^{j(\beta_1 + 120^\circ)} \end{aligned} \tag{3.16}$$

Perhaps a better form for visualizing the voltages and currents is a graphical one. Phasor diagrams for the voltages separately and the currents separately are shown in Fig. 3.29. The value of angle β_1 will depend on the load. An interesting result is clear from these diagrams. First, V_2 and V_3 are each other's conjugates. So if we add them, the imaginary parts cancel and the sum will be real, as illustrated by the construction in the voltage diagram. Furthermore, the construction shows that this real part is negative and equal in size to V_1. Hence, the sum of the three voltages is zero. The same is true of the sum of the three currents, as can be established graphically by a similar construction.

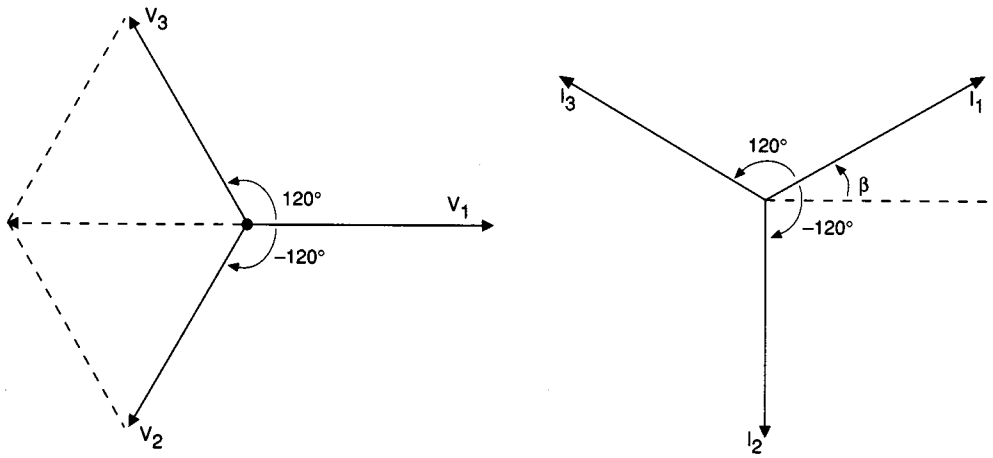


FIGURE 3.29 Voltage and current phasor diagrams.

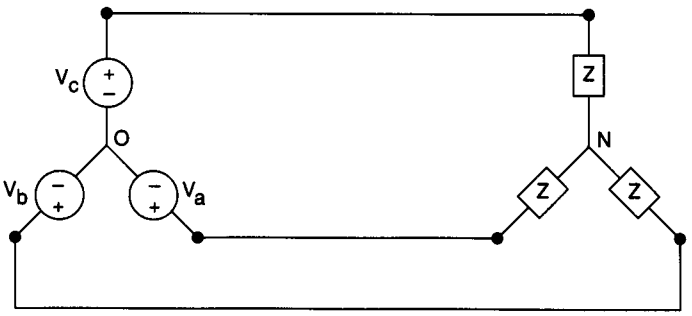


FIGURE 3.30 Wye-connected three-phase system.

By Kirchhoff's current law applied at node *N* in Fig. 3.28(b), we find that the current in the return line is the sum of the three currents in Eq. (3.16). However, since this sum was found to be zero, the return line carries no current. Hence it can be removed entirely without affecting the operation of the system. The resulting circuit is redrawn in Fig. 3.30. Because of its geometrical form, this connection of both the sources and the loads is said to be a wye (Y) connection.

The instantaneous power delivered by each of the sources has the form given in Eq. (3.14), consisting of a constant term representing the average power and a double-frequency sinusoidal term. The latter, being sinusoidal, can be represented by a phasor also. The only caution is that a different frequency is involved here, so this power phasor should not be mixed with the voltage and current phasors in the same diagram or calculations. Let $|S| = |V||I|$ be the apparent power delivered by each of the three sources and let the three power phasors be S_a, S_b, and S_c, respectively. Then:

$$\begin{aligned} S_a &= |S|\, e^{j(\alpha_1+\beta_1)} = |S|\, e^{j\beta_1} \\ S_b &= |S|\, e^{j(\alpha_2+\beta_2)} = |S|\, e^{j(-120°+\beta_1-120°)} = |S|\, e^{j(\beta_1+120°)} \\ S_c &= |S|\, e^{j(\alpha_3+\beta_3)} = |S|\, e^{j(+120°+\beta_1+120°)} = |S|\, e^{j(\beta_1-120°)} \end{aligned} \tag{3.17}$$

It is evident that the phase relationships among these three phasors are the same as the ones among the voltages and the currents. That is, the second leads the first by 120° and the third lags the first by 120°. Hence, just like the voltages and the currents, the sum of these three phasors will also be zero. This is a very significant result. Although the instantaneous power delivered by each source has a constant component and a sinusoidal component, when the three powers are added, the sinusoidal components add to zero, leaving only the constants. Thus, the total power delivered to the three loads is constant.

To determine the value of this constant power, use Eq. (3.14) as a model. The contribution of the *k*th source to the total (constant) power is $|S| \cos(\alpha_k - \beta_k)$. One can easily verify that $\alpha_k - \beta_k = \alpha_1 - \beta_1 = -\beta_1$. The first equality follows from the relationships among the α's from Eq. (3.15) and among the β's from Eq. (3.16). The choice of $\alpha_1 = 0$ leads to the last equality. Hence, the constant terms contributed to the power by each source are the same. If *P* is the total average power, then:

$$P = P_a + P_b + P_c = 3P_a = 3\,|V||I|\cos(\alpha_1 - \beta_1) \tag{3.18}$$

Although the angle α_1 has been set equal to zero, for the sake of generality we have shown it explicitly in this equation.

What has just been described is a *balanced* three-phase three-wire power system. The three sources in practice are not three independent sources but consist of three different parts of the same generator. The same is true

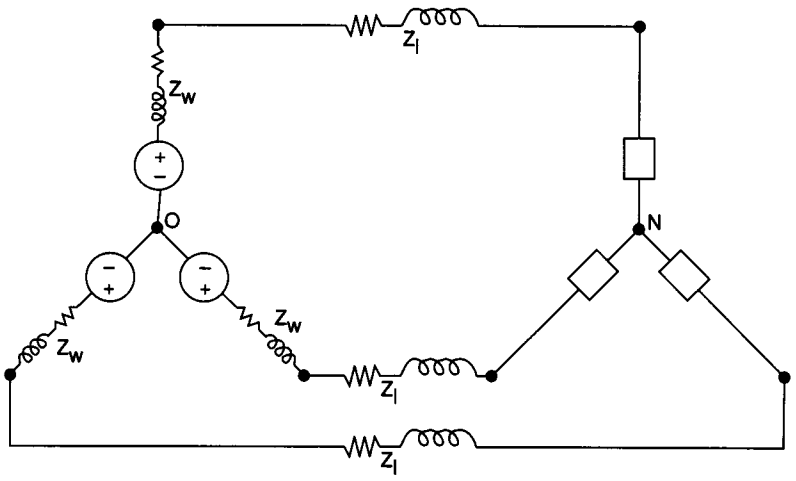


FIGURE 3.31 Three-phase circuit with nonzero winding and line impedances.

of the loads.[1] What has been described is ideal in a number of ways. First, the circuit can be *unbalanced*—for example, by the loads being somewhat unequal. Second, since the real devices whose ideal model is a voltage source are coils of wire, each source should be accompanied by a branch consisting of the coil inductance and resistance. Third, since the power station (or the distribution transformer at some intermediate point) may be at some distance from the load, the parameters of the physical line carrying the power (the line inductance and resistance) must also be inserted in series between the source and the load.

For an unbalanced system, the analysis of this section does not apply. An entirely new analytical technique is required to do full justice to such a system.[2] However, an understanding of balanced circuits is a prerequisite for tackling the unbalanced case.

The last two of the conditions that make the circuit less than ideal (line and source impedances) introduce algebraic complications, but nothing fundamental is changed in the preceding theory. If these two conditions are taken into account, the appropriate circuit takes the form shown in Fig. 3.31. Here the internal impedance of a source and the line impedance connecting that source to its load are both connected in series with the corresponding load. Thus, instead of the impedance in each phase being Z, it is $Z + Z_w + Z_l$, where w and l are subscripts standing for "winding" and "line," respectively. Hence, the rms value of each current is

$$|I| = \frac{|V|}{|Z + Z_w + Z_l|} \tag{3.19}$$

instead of $|V|/|Z|$. All other results we had arrived at remain unchanged. namely, that the sum of the phase currents is zero and that the sum of the phase powers is a constant. The detailed calculations simply become a little more complicated.

One other point, illustrated for the loads in Fig. 3.32, should be mentioned. Given wye-connected sources or loads, the wye and the **delta** can be made equivalent by proper selection of the arms of the delta. Thus,

[1]An ac power generator consists of (a) a rotor, which produces a magnetic field and which is rotated by a prime mover (say a turbine), and (b) a stator on which are wound one or more coils of wire. In three-phase systems, the number of coils is three. The rotating magnetic field induces a voltage in each of the coils. The 120° leading and lagging phase relationships among these voltages are obtained by distributing the conductors of the coils around the circumference of the stator so that they are separated geometrically by 120°. Thus, the three sources described in the text are in reality a single physical device, a single generator. Similarly, the three loads might be the three windings on a three-phase motor, again a single physical device.

[2]The technique for analyzing unbalanced circuits utilizes what are called *symmetrical components.*

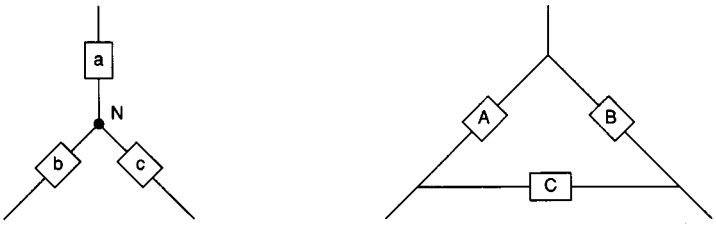


FIGURE 3.32 Wye connection and delta connection.

either the sources in Fig. 3.30 or the loads, or both, can be replaced by a delta equivalent; thus we can conceive of four different three-phase circuits; wye-wye, delta-wye, wye-delta, and delta-delta. Not only can we conceive of them, they are extensively used in practice.

It is not worthwhile to carry out detailed calculations for these four cases. Once the basic properties described here are understood, one should be able to make the calculations. Observe, however, that in the delta structure, there is no neutral connection, so the phase voltages cannot be measured. The only voltages that can be measured are the *line-to-line* or simply the *line* voltages. These are the differences of the phase voltages taken in pairs, as is evident from Fig. 3.31.

Defining Terms

Delta connection: The sources or loads in a three-phase system connected end-to-end, forming a closed path, like the Greek letter Δ.

Phasor: A complex number representing a sinusoid; its magnitude and angle are the rms value and phase of the sinusoid, respectively.

Wye connection: The three sources or loads in a three-phase system connected to have one common point, like the letter Y.

Related Topic

9.2 Three-Phase Connections

References

V. del Toro, *Electric Power Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1992.
B.R. Gungor, *Power Systems,* San Diego: Harcourt Brace Jovanovich, 1988.
P.Z. Peebles and T.A. Giuma, *Principles of Electrical Engineering,* New York: McGraw-Hill, 1991.
J.J. Grainger and W.D. Stevenson, Jr., *Power Systems Analysis,* New York: McGraw-Hill, 1994.
G.T. Heydt, *Electric Power Quality,* Stars in a Circle Publications, 1996.
B.S. Guru and H.R. Hirizoglu, *Electric Machinery and Transformers,* Saunders, 1996.

3.6 Graph Theory[1]

Shu-Park Chan

Topology is a branch of mathematics; it may be described as "the study of those properties of geometric forms that remain invariant under certain transformations, as bending, stretching, etc."[2] Network topology (or

[1]Based on S.-P. Chan, "Graph theory and some of its applications in electrical network theory," in *Mathematical Aspects of Electrical Network Analysis,* vol. 3, *SIAM/AMS Proceedings,* American Mathematical Society, Providence, R.I., 1971. With permission.

[2]This brief description of topology is quoted directly from the *Random House Dictionary of the English Language,* Random House, New York, 1967.

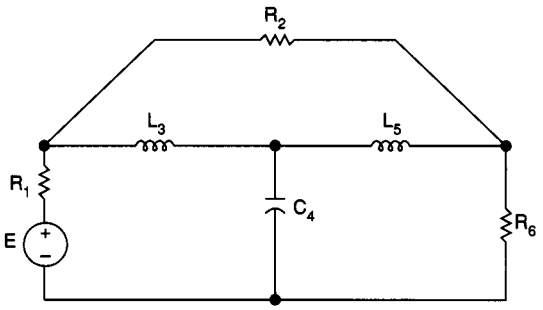


FIGURE 3.33 A passive network N with a voltage driver E.

network graph theory) is a study of (electrical) networks in connection with their nonmetric geometrical (namely topological) properties by investigating the interconnections between the branches and the nodes of the networks. Such a study will lead to important results in network theory such as algorithms for formulating network equations and the proofs of various basic network theorems [Chan, 1969; Seshu and Reed, 1961].

The following are some basic definitions in network graph theory, which will be needed in the development of topological formulas in the analysis of linear networks and systems.

A **linear graph** (or simply a *graph*) is a set of line segments called *edges* and points called *vertices*, which are the endpoints of the edges, interconnected in such a way that the edges are connected to (or *incident* with) the vertices. The *degree* of a vertex of a graph is the number of edges incident with that vertex.

A subset G_i of the edges of a given graph G is called a **subgraph** of G. If G_i does not contain all of the edges of G, it is a **proper subgraph** of G. A **path** is a subgraph having all vertices of degree 2 except for the two endpoints, which are of degree 1 and are called the *terminals* of the path. The set of all edges in a path constitutes a **path-set**. If the two terminals of a path coincide, the path is a closed path and is called a **circuit** (or **loop**). The set of all edges contained in a circuit is called a **circuit-set** (or **loop-set**).

A graph or subgraph is said to be **connected** if there is at least one path between *every* pair of its vertices. A **tree** of a connected graph G is a connected subgraph which contains all the vertices of G but no circuits. The edges contained in a tree are called the **branches of the tree**. A 2-tree of a connected graph G is a (proper) subgraph of G consisting of two unconnected circuitless subgraphs, each subgraph itself being connected, which together contain all the vertices of G. Similarly, a k-tree is a subgraph of k unconnected circuitless subgraphs, each subgraph being connected, which together include all the vertices of G. The **k-tree admittance product of a k-tree** is the product of the admittances of all the branches of the k-tree.

Example 3.5. The graph G shown in Fig. 3.34 is the graph of the network N of Fig. 3.33. The edges of G are e_1, e_2, e_4, e_5, and e_6; the vertices of G are V_1, V_2, and V_3, and V_4. A path of G is the subgraph G_1 consisting of edges e_2, e_3, and e_6 with vertices V_2 and V_4 as terminals. Thus, the set $\{e_2, e_3, e_6\}$ is a path-set. With edge e_4 added to G_1, we form another subgraph G_2, which is a circuit since as far as G_2 is concerned all its vertices are of degree 2. Hence the set $\{e_2, e_3, e_4, e_6\}$ is a circuit-set. Obviously, G is a connected graph since there exists a path between every pair of vertices of G. A tree of G may be the subgraph consisting of edges e_1, e_4, and e_6. Two other trees of G are $\{e_2, e_5, e_6\}$ and $\{e_3, e_4, e_5\}$. A 2-tree of G is $\{e_2, e_4\}$; another one is $\{e_3, e_6\}$; and still another one is $\{e_3, e_5\}$. Note that both $\{e_2, e_4\}$ and $\{e_3, e_6\}$ are subgraphs which obviously satisfy the definition of a 2-tree in the sense that each contains two disjoint circuitless connected subgraphs, both of which include all the four vertices of G. Thus, $\{e_3, e_5\}$ does not seem to be a 2-tree. However, if we agree to consider $\{e_3, e_5\}$ as a subgraph which contains edges e_3 and e_5 plus the isolated vertex V_4, we see that $\{e_3, e_5\}$ will satisfy the definition of a 2-tree since it now has two circuitless connected subgraphs with e_3 and e_5 forming one of them and the vertex V_4 alone forming the other. Moreover, both subgraphs together indeed

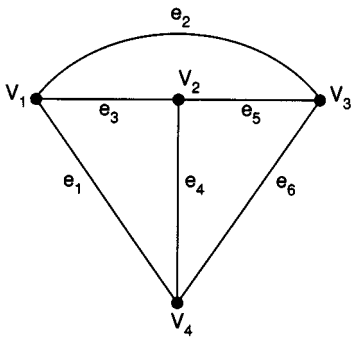


FIGURE 3.34 The graph G of the network N of Fig. 3.33.

include all the four vertices of G. It is worth noting that a 2-tree is obtained from a tree by removing *any one* of the branches from the tree; in general, a k-tree is obtained from a $(k-1)$ tree by removing from it any one of its branches. Finally, the tree admittance product of the tree $\{e_2, e_5, e_6\}$ is $\frac{1}{2}\frac{1}{5}\frac{1}{6}$; the 2-tree admittance product of the 2-tree $\{e_3, e_5\}$ is $\frac{1}{3}\frac{1}{5}$ (with the admittance of a vertex defined to be 1).

The *k*-Tree Approach

The development of the analysis of passive electrical networks using topological concepts may be dated back to 1847 when Kirchhoff formulated his set of topological formulas in terms of resistances and the branch-current system of equations. In 1892, Maxwell developed another set of topological formulas based on the k-tree concept, which are the duals of Kirchhoff's. These two sets of formulas were supported mainly by heuristic reasoning and no formal proofs were then available.

In the following we shall discuss only Maxwell's topological formulas for linear networks without mutual inductances.

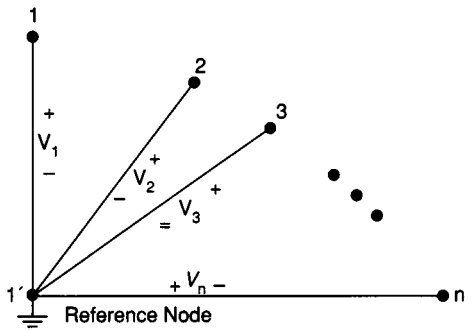


FIGURE 3.35 A network N with n independent nodes.

Consider a network N with n independent nodes as shown in Fig. 3.35. The node 1′ is taken as reference (datum) node. the voltages $V_1, V_2, \ldots, V_n$ (which are functions of s) are the transforms of the node-pair voltages (or simply node voltages) $v_1, v_2, \ldots, v_n$ (which are function s of t) between the n nodes and the reference node 1′ with the plus polarity marks at the n nodes. It can be shown [Aitken, 1956] that the matrix equation for the n independent nodes of N is given by

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{bmatrix} \tag{3.20}$$

or, in abbreviated matrix notation,

$$Y_n V_n = I_n \tag{3.21}$$

where Y_n is the node admittance matrix, V_n the $n \times 1$ matrix of the node voltage transforms, and I_n the $n \times 1$ matrix of the transforms of the known current sources.

For a relaxed passive one-port (with zero initial conditions) shown in Fig. 3.36, the driving-point impedance function $Z_d(s)$ and its reciprocal, namely driving-point admittance function $Y_d(s)$, are given by

$$Z_d(s) = V_1/I_1 = \Delta_{11}/\Delta$$

and

$$Y_d(s) = 1/Z_d(s) = \Delta/\Delta_{11}$$

respectively, where Δ is the determinant of the node admittance matrix Y_n, and Δ_{11} is the (1,1)-cofactor of Δ.

Similarly, for a passive reciprocal *RLC* two-port (Fig. 3.37), the open-circuit impedances and the short-circuit admittances are seen to be

$$z_{11} = \Delta_{11}/\Delta \tag{3.22a}$$

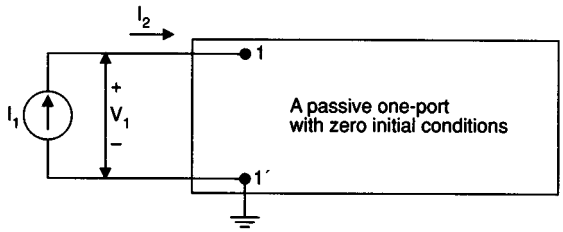


FIGURE 3.36 The network *N* driven by a single current source.

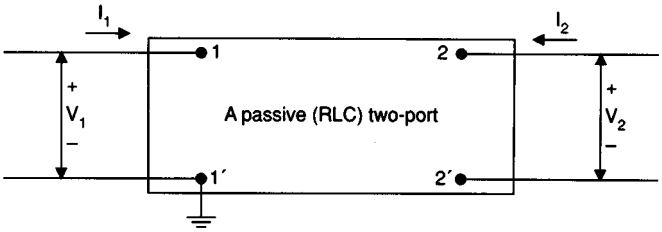


FIGURE 3.37 A passive two-port.

$$z_{12} = z_{21} = (\Delta_{12} - \Delta_{12'})/\Delta \qquad (3.22b)$$

$$z_{22} = (\Delta_{22} + \Delta_{2'2'} - 2\Delta_{22'})/\Delta \qquad (3.22c)$$

and

$$y_{11} = (\Delta_{22} + \Delta_{2'2'} - 2\Delta_{22'})/(\Delta_{1122} + \Delta_{112'2'} - 2\Delta_{1122'}) \qquad (3.23a)$$

$$y_{12} = y_{21} = \Delta_{12'} - \Delta_{12}/(\Delta_{1122} + \Delta_{112'2'} - 2\Delta_{1122'}) \qquad (3.23b)$$

$$y_{22} = \Delta_{11}/(\Delta_{1122} + \Delta_{112'2'} - 2\Delta_{1122'}) \qquad (3.23c)$$

respectively, where Δ_{ij} is the (i,j)-cofactor of Δ, and Δ_{ijkm} is the cofactor of Δ by deleting rows i and k and columns j and m from Δ [Aitken, 1956].

Expressions in terms of network determinants and cofactors for other network transfer functions are given by (Fig. 3.38):

$$z_{12} = \frac{V_2}{I_1} = \frac{\Delta_{12} - \Delta_{12'}}{\Delta} \qquad \text{(transfer impedance function)} \qquad (3.24a)$$

$$G_{12} = \frac{V_2}{V_1} = \frac{\Delta_{12} - \Delta_{12'}}{\Delta_{11}} \qquad \text{(voltage-ratio transfer function)} \qquad (3.24b)$$

$$Y_{12} = Y_L G_{12} = Y_L\left(\frac{\Delta_{12} - \Delta_{12'}}{\Delta_{11}}\right) \qquad \text{(transfer admittance function)} \qquad (3.24c)$$

$$\alpha_{12} = Y_1 Z_{12} = Y_L\left(\frac{\Delta_{12} - \Delta_{12'}}{\Delta}\right) \qquad \text{(current-ratio transfer function)} \qquad (3.24d)$$

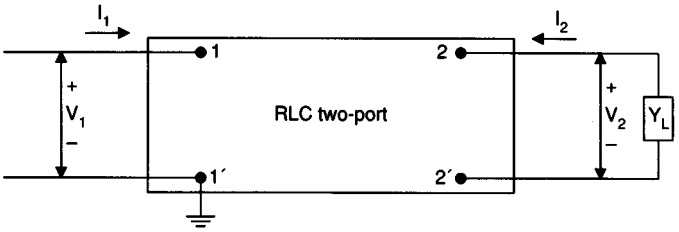


FIGURE 3.38 A loaded passive two-port.

The topological formulas for the various network functions of a passive one-port or two-port are derived from the following theorems which are stated without proof [Chan, 1969].

Theorem 3.1. Let N be a passive network without mutual inductances. The determinant Δ of the node admittance matrix Y_n is equal to the sum of all tree-admittances of N, where a tree-admittance product $t^{(i)}(y)$ is defined to be the product of the admittance of all the branches of the tree $T^{(i)}$. That is,

$$\Delta = \det Y_n = \sum_i T^{(i)}(y) \tag{3.25}$$

Theorem 3.2. Let Δ be the determinant of the node admittance matrix Y_n of a passive network N with $n + 1$ nodes and without mutual inductances. Also let the reference node be denoted by $1'$. Then the (j,j)-cofactor Δ_{jj} of Δ is equal to the sum of all the 2-tree-admittance products $T_{2j,1'}(y)$ of N, each of which contains node j in one part and node $1'$ as the reference node) and without mutual inductances is given by

$$\Delta_{jj} = \sum_k T^{(k)}_{2_{j,1'}}(y) \tag{3.26}$$

where the summation is taken over all the 2-tree-admittance products of the form $T_{2j,1'}(y)$.

Theorem 3.3. The (i,j)-cofactor Δ_{ij} of Δ of a relaxed passive network N with n independent nodes (with node $1'$ as the reference node) and without mutual inductances is given by

$$\Delta_{ij} = \sum_k T^{(k)}_{2_{ij,1'}}(y) \tag{3.27}$$

where the summation is taken over all the 2-tree-admittance products of the form $T_{2ij,1'}(y)$ with each containing nodes i and j in one connected port and the reference node $1'$ in the other.

For example, the topological formulas for the driving-point function of a passive one-port can be readily obtained from Eqs. (3.25) and (3.26) in Theorems 3.1 and 3.2 as stated in the next theorem.

Theorem 3.4. With the same notation as in Theorems 3.1 and 3.2, the driving-point admittance $Y_d(s)$ and the driving-point impedance $Z_d(s)$ of a passive one-port containing no mutual inductances at terminals 1 and $1'$ are given by

$$Y_d(s) = \frac{\Delta}{\Delta_{11}} = \frac{\sum_i T^{(i)}(y)}{\sum_k T^{(k)}_{2_{1,1}}(y)} \quad \text{and} \quad Z_d(s) = \frac{\Delta_{11}}{\Delta} = \frac{\sum_k T^{(k)}_{2_{1,1}}(y)}{\sum_i T^{(i)}(y)} \tag{3.28}$$

respectively.

For convenience we define the following shorthand notation:

$$(a)V(Y) \equiv \sum_i T^{(i)}(y) = \text{sum of all tree-admittance products, and}$$

$$(b)W_{j,r}(y) \equiv \sum_k T_{2\,j,r}(y) = \text{sum of all 2-tree-admittance products with node } j \text{ and the reference node } r \text{ contained in different parts.} \tag{3.29}$$

Thus Eq. (3.28) may be written as

$$Y_d(s) = V(Y)/W_{1,1'}(Y) \qquad \text{and} \qquad Z_d(s) = W_{1,1'}(Y)/V(Y) \tag{3.30}$$

In a two-port network N, there are four nodes to be specified, namely, 1 and 1′ at the input port (1,1′) and nodes 2 and 2′ at the output port (2,2′), as illustrated in Fig. 3.38. However, for a 2-tree of the type $T_{2\,ij,1'}$, only three nodes have been used, thus leaving the fourth one unidentified.

With very little effort, it can be shown that, in general, the following relationship holds:

$$W_{ij,1'}(Y) = W_{ijk,1'}(Y) + W_{ij,k1'}(Y)$$

or simply

$$W_{ij,1'} = W_{ijk,1'} + W_{ij,k1'} \tag{3.31}$$

where i, j, k, and 1′ are the four terminals of N with 1′ denoting the datum (reference) node. The symbol $W_{ijk,1'}$ denotes the sum of all the 2-tree-admittance products, each containing nodes i, j, and k in one connected part and the reference node, 1′, in the other.

We now state the next theorem.

Theorem 3.5. With the same hypothesis and notation as stated earlier in this section,

$$\Delta_{12} - \Delta_{12'} = W_{12,12'}(Y) - W_{12',1'2}(Y) \tag{3.32}$$

It is interesting to note that Eq. (3.32) is stated by Percival [1953] in the following descriptive fashion:

$$\Delta_{12} - \Delta_{12'} = W_{12,12'} - W_{12',1'2} = \left(\begin{array}{c} 1 \circ\!\!-\!\!-\!\!\circ\ 2 \\ \\ 1' \circ\!\!-\!\!-\!\!\circ\ 2' \end{array} \right) - \left(\begin{array}{c} 1 \circ \quad \circ\ 2 \\ \quad \times \\ 1' \circ \quad \circ\ 2' \end{array} \right)$$

which illustrates the two types of 2-trees involved in the formula. Hence, we state the topological formulas for z_{11}, z_{12}, and z_{22} in the following theorem.

Theorem 3.6. With the same hypothesis and notation as stated earlier in this section

$$z_{11} = W_{1,1'}(Y)/V(Y) \tag{3.33a}$$

$$z_{12} = z_{21} = \{W_{12,1'2'}(Y) - W_{12',1'2}(Y)\}/V(Y) \tag{3.33b}$$

$$z_{22} = W_{2,2'}(Y)/V(Y) \tag{3.33c}$$

We shall now develop the topological expressions for the short-circuit admittance functions. Let us denote by $U_{a,b,c}(Y)$ the sum of all 3-tree admittance products of the form $T_{3_{a,b,c}}(Y)$ with identical subscripts in both

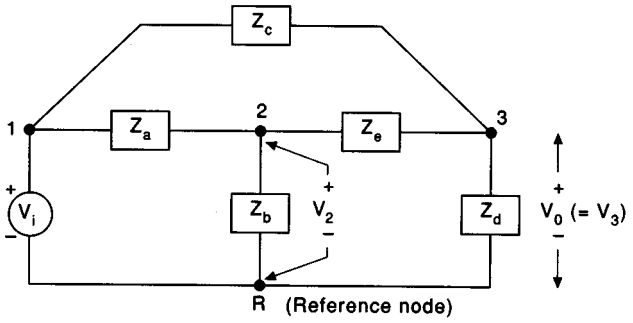


FIGURE 3.39 The network N of Example 3.7.

symbols to represent the same specified distribution of vertices. Then, following arguments similar to those of Theorem 3.5, we readily see that

$$\Delta_{1122} = \sum_i T^{(i)}_{3_{1,2,1'}}(y) \equiv U_{1,2,1'}(Y) \tag{3.34a}$$

$$\Delta_{112'2'} = \sum_j T^{(j)}_{3_{1,2',1'}}(y) \equiv U_{1,2',1'}(Y) \tag{3.34b}$$

$$\Delta_{1122'} = \sum_k T^{(k)}_{3_{1,22',1'}}(y) \equiv U_{1,22',1'}(U) \tag{3.34c}$$

where 1,1′,2,2′ are the four terminals of the two-port with 1′ denoting the reference node (Fig. 3.39). However, we note that in Eqs. (3.34a) and (3.34b) only three of the four terminals have been specified. We can therefore further expand $U_{1,2,1'}$ and $U_{1,2',1'}$ to obtain the following:

$$\Delta_{1122} + \Delta_{112'2'} - 2\Delta_{1122'} = U_{12',2,1'} + U_{1,2,1'2'} + U_{12,2',1'} + U_{1,2',1'2} \tag{3.35}$$

For convenience, we shall use the shorthand notation ΣU to denote the sum of the right of Eq. (3.35). Thus, we define

$$\sum U = U_{12',2,1'} + U_{1,2,1'2'} + U_{12,2',1'} + U_{1,2',1'2} \tag{3.36}$$

Hence, we obtain the topological formulas for the short-circuit admittances as stated in the following theorem.

Theorem 3.7. The short-circuit admittance functions y_{11}, y_{12}, and y_{22} of a passive two-port network with no mutual inductances are given by

$$y_{11} = W_{2,2'} \Big/ \sum U \tag{3.37a}$$

$$y_{12} = y_{21} = \left(W_{12,1'2} - W_{12,1'2'}\right) \Big/ \sum U \tag{3.37b}$$

$$y_{22} = W_{1,1'} \Big/ \sum U \tag{3.37c}$$

where ΣU is defined in Eq. (3.36) above.

Finally, following similar developments, other network functions are stated in Theorem 3.8.

Theorem 3.8. With the same notation as before,

$$Z_{12}(s) = \frac{W_{12,1'2'} \sqcap W_{12',1'2}}{V} \tag{3.38a}$$

$$G_{12}(s) = \frac{W_{12,1'2'} \sqcap W_{12',1'2}}{W_{1,1'}} \tag{3.38b}$$

$$Y_{12}(s) = Y_L \frac{W_{12,1'2'} \sqcap W_{12',1'2}}{W_{1,1'}} \tag{3.38c}$$

$$\alpha_{12}(s) = Y_L \frac{W_{12,1'2'} \sqcap W_{12',1'2}}{V} \tag{3.38d}$$

The Flowgraph Approach

Mathematically speaking, a linear electrical network or, more generally, a linear system can be described by a set of simultaneous linear equations. Solutions to these equations can be obtained either by the method of successive substitutions (elimination theory), by the method of determinants (Cramer's rule), or by any of the topological techniques such as Maxwell's *k*-tree approach discussed in the preceding subsection and the flowgraph techniques represented by the works of Mason [1953, 1956], and Coates [1959].

Although the methods using algebraic manipulations can be amended and executed by a computer, they do not reveal the physical situations existing in the system. The flowgraph techniques, on the other hand, show intuitively the causal relationships between the variables of the system of interest and hence enable the network analyst to have an excellent physical insight into the problem.

In the following, two of the more well-known flowgraph techniques are discussed, namely, the **signal-flowgraph** technique devised by Mason and the method based on the flowgraph of Coates and recently modified by Chan and Bapna [1967].

A *signal-flowgraph* G_m of a system S of n independent linear (algebraic) equations in n unknowns

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad i = 1, 2, \ldots, n \tag{3.39}$$

is a graph with junction points called *nodes* which are connected by directed line segments called *branches* with signals traveling along the branches only in the direction described by the arrows of the branches. A signal x_k traveling along a branch between x_k and x_j is multiplied by the gain of the branches g_{kj}, so that a signal of $g_{kj}x_k$ is delivered at node x_j. An *input node* (*source*) is a node which contains only outgoing branches; an *output node* (*sink*) is a node which has only incoming branches. A *path* is a continuous unidirectional succession of branches, all of which are traveling in the same direction; a *forward path* is a path from the input node to the output node along which all nodes are encountered exactly once; and a *feedback path* (*loop*) is a closed path which originates from and terminates at the same node, and along which all other nodes are encountered exactly once (the trivial case is a *self-loop* which contains exactly one node and one branch). A *path gain* is the product of all the branch gains of the path; similarly, a loop gain is the product of all the branch gains of the branches in a loop.

The procedure for obtaining the Mason graph from a system of linear algebraic equations may be described in the following steps:

a. Arrange all the equations of the system in such a way that the jth dependent (output) variable x_j in the jth equation is expressed explicitly in terms of the other variables. Thus, if the system under study is given by Eq. (3.39), namely

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{3.40}$$

where $b_1, b_2, \ldots, b_n$ are inputs (sources) and $x_1, x_2, \ldots, x_n$ are outputs, the equations may be rewritten as

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}b_1 - \frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \cdots - \frac{a_{1n}}{a_{11}}x_n \\ x_2 &= \frac{1}{a_{22}}b_2 - \frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_3 - \cdots - \frac{a_{2n}}{a_{22}}x_n \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ x_n &= \frac{1}{a_{nn}}b_n - \frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots - \frac{a_{n-1,n-1}}{a_{nn}}x_{n-1} \end{aligned} \tag{3.41}$$

b. The number of input nodes in the flowgraph is equal to the number of nonzero b's. That is, each of the source nodes corresponds to a nonzero b_j.
c. To each of the output nodes is associated one of the dependent variables $x_1, x_2, \ldots, x_n$.
d. The value of the variable represented by a node is equal to the sum of all the incoming signals.
e. The value of the variable represented by any node is transmitted onto all branches leaving the node.

It is a simple matter to write the equations from the flowgraph since every node, except the source nodes of the graph, represents an equation, and the equation associated with node k, for example, is obtained by equating to x_k the sum of all incoming branch gains multiplied by the values of the variables from which these branches originate.

Mason's general gain formula is now stated in the following theorem.

Theorem 3.9. Let G be the overall graph gain and G_k be the gain of the kth forward path from the source to the sink. Then

$$G = \frac{1}{\Delta}\sum_k G_k\Delta_k \tag{3.42}$$

where

$$\Delta = 1 - \sum_m p_{m1} + \sum_m p_{m2} - \sum_m p_{m3} + \cdots + (-1)^j \sum_m p_{mj}$$

p_{m1} = loop gain (the product of all the branch gains around a loop)
p_{m2} = product of the loop gains of the mth set of two nontouching loops
p_{m3} = product of the loop gains of the mth set of three nontouching loops, and in general

p_{mj} = product of the loop gains of the mth set of j nontouching loops
Δ_k = the value of Δ for that subgraph of the graph obtained by removing the kth forward path along with those branches touching the path

Mason's signal-flowgraphs constitute a very useful graphical technique for the analysis of linear systems. This technique not only retains the intuitive character of the block diagrams but at the same time allows one to obtain the gain between an input node and an output node of a signal-flowgraph by inspection. However, the derivation of the gain formula [Eq. (3.42)] is by no means simple, and, more importantly, if more than one input is present in the system, the gain cannot be obtained directly; that is, the principle of superposition must be applied to determine the gain due to the presence of more than one input. Thus, by slight modification of the conventions involved in Mason's signal-flowgraph, Coates [1959] was able to introduce the so-called "flowgraphs" which are suitable for direct calculation of gain.

Recently, Chan and Bapna [1967] further modified Coates's flowgraphs and developed a simpler gain formula based on the modified graphs. The definitions and the gain formula based on the modified Coates graphs are presented in the following discussion.

The **flowgraph G_l** (called the *modified Coates graph*) of a system S of n independent linear equations in n unknowns

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad i = 1, 2, \ldots, n$$

is an oriented graph such that the variable x_j in S is represented by a *node* (also denoted by x_j) in G_l, and the coefficient a_{ij} of the variable x_j in S by a *branch* with a branch gain a_{ij} connected between nodes x_i and x_j in G_l and directed from x_j to x_i. Furthermore, a *source node* is included in G_l such that for each constant b_k in S there is a node with gain b_k in G_l from node 1 to node s_k. *Graph* G_{l0} is the subgraph of G_l obtained by deleting the source node l and all the branches connected to it. Graph G_{ij} is the subgraph of G_l obtained by first removing all the outgoing branches from node x_j and then short-circuiting node l to node x_j. A *loop set* l is a subgraph of G_{l0} that contains all the nodes of G_{l0} with each node having exactly one incoming and one outgoing branch. The product p of the gains of all the branches in l is called a *loop-set product.* A 2-*loop-set* I_2 is a subgraph of G_{lj} containing all the nodes of G_{lj} with each node having exactly one incoming and one outgoing branch. The product p_2 of the gains of all the branches in l_2 is called a 2-*loop-set product.*

The modified Coates gain formula is now stated in the following theorem.

Theorem 3.10. In a system of n independent linear equations in n unknowns

$$a_{ij}x_j = b_i \qquad i = 1, 2, \ldots, n$$

the value of the variable x_j is given by

$$x_j = \sum_{(\text{all } p_2)} (-1)^{N_{l_2}} p_2 \Bigg/ \sum_{(\text{all } p_2)} (-1)^{N_l} p \tag{3.43}$$

where N_{l_2} is the number of loops in a 2-loop-set l_2 and N_l is the number of loops in a loop set l.

Since both the Mason graph G_m and the modified Coates graph G_l are topological representations of a system of equations it is logical that certain interrelationships exist between the two graphs so that one can be transformed into the other. Such interrelationships have been noted [Chan, 1969], and the transformations are briefly stated as follows:

A. *Transformation of G_m into G_l.* Graph G_m can be transformed into an equivalent Coates graph G_l (representing an equivalent system of equations) by the following steps:

a. Subtract 1 from the gain of each existing self-loop.
b. Add a self-loop with a gain of –1 to each branch devoid of self-loop.
c. Multiply by $-b_k$ the gain of the branch at the kth source node b_k ($k = 1, 2, \ldots, r$, r being the number of source nodes) and then combine all the (r) nodes into one source node (now denoted by 1).

B. *Transformation of G_l into G_m.* Graph G_l can be transformed into G_m by the following steps:
a. Add 1 to the gain of each existing self-loop.
b. Add a self-loop with a gain of 1 to each node devoid of self-loop except the source node l.
c. Break the source node l into r source nodes (r being the number of branches connected to the source node l before breaking), and identify the r new sources nodes by $b_1, b_2, \ldots, b_r$, with the gain of the corresponding r branches multiplied by $-1/b_1, -1/b_2, \ldots, -1/b_r$, respectively, so that the new gains of these branches are all equal to l, keeping the edge orientations unchanged.

The gain formulas of Mason and Coates are the classical ones in the theory of flowgraphs. From the systems viewpoint, the Mason technique provides an excellent physical insight as one can visualize the signal flow through the subgraphs (forward paths and feedback loops) of G_m. The graph reduction technique based on the Mason graph enables one to obtain the gain expression using a step-by-step approach and at the same time observe the cause-and-effect relationships in each step. However, since the Mason formula computes the ratio of a specified output over *one* particular input, the principle of superposition must be used in order to obtain the overall gain of the system if more than one input is present. The Coates formula, on the other hand, computes the output directly regardless of the number of inputs present in the system, but because of such a direct computation of a given output, the graph reduction rules of Mason cannot be applied to a Coates graph since the Coates graph is *not* based on the same cause-effect formulation of equations as Mason's.

The *k*-Tree Approach Versus the Flowgraph Approach

When a linear network is given, loop or node equations can be written from the network, and the analysis of the network can be accomplished by means of either Coates's or Mason's technique.

However, it has been shown [Chan, 1969] that if the Maxwell k-tree approach is employed in solving a linear network, the redundancy inherent either in the direct expansion of determinants or in the flowgraph techniques described above can be either completely eliminated for passive networks or greatly reduced for active networks. This point and others will be illustrated in the following example.

Example 3.7. Consider the network N as shown in Fig. 3.39. Let us determine the voltage gain, $G_{12} = V_0/V_1$, using (1) Mason's method, (2) Coates's method, and (3) the k-tree method.

The two node equations for the network are given by

$$\begin{aligned} \text{for node 2:} \quad & (Y_a + Y_b + Y_e)V_2 + (-Y_3)V_0 = Y_a V_i \\ \text{for node 3:} \quad & (-Y_e)V_2 + (Y_c + Y_d + Y_e)V_0 = Y_c V_i \end{aligned} \tag{3.44}$$

where

$$Y_a = 1/Z_a,\ Y_b = 1/Z_b,\ Y_c = 1/Z_c,\ Y_d = 1/Z_d \text{ and } Y_e = 1/Z_e$$

(1) *Mason's approach.* Rewrite the system of two equations (3.44) as follows:

$$\begin{aligned} V_2 &= \left(\frac{Y_a}{Y_a + Y_b + Y_e}\right)V_i + \left(\frac{Y_e}{Y_a + Y_b + Y_e}\right)V_0 \\ V_0 &= \left(\frac{Y_c}{Y_c + Y_d + Y_e}\right)V_i + \left(\frac{Y_e}{Y_c + Y_d + Y_e}\right)V_2 \end{aligned} \tag{3.45}$$

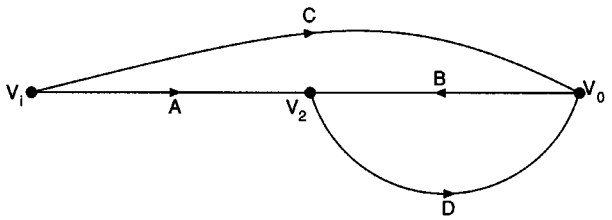


FIGURE 3.40 The Mason graph of *N*.

or

$$V_2 = AV_i + BV_0 \quad V_0 = CV_i + DV_2 \tag{3.46}$$

where

$$A = \frac{Y_a}{Y_a + Y_b + Y_e} \qquad B = \frac{Y_e}{Y_a + Y_b + Y_e}$$
$$C = \frac{Y_c}{Y_c + Y_d + Y_e} \qquad D = \frac{Y_e}{Y_c + Y_d + Y_e}$$

The Mason graph of system (3.46) is shown in Fig. 3.40, and according to the Mason graph formula (3.42), we have

$$\Delta = 1 - BD$$
$$G_C = C \quad \Delta_C = 1$$
$$G_{AD} = AD \quad \Delta_{AD} = 1$$

and hence

$$G_{12} = \frac{V_0}{V_1} = \frac{1}{\Delta}\sum_k G_k\Delta_k = \frac{1}{1 - BD}(C + AD)$$
$$= \frac{Y_c/(Y_c + Y_d + Y_e) + Y_a/(Y_a + Y_b + Y_e)(Y_c + Y_d + Y_e)}{1 - Y_e^2/(Y_a + Y_b + Y_e)(Y_c + Y_d + Y_e)}$$

Upon cancellation and rearrangement of terms

$$G_{12} = \frac{Y_aY_c + Y_aY_e + Y_bY_c + Y_cY_e}{Y_aY_c + Y_aY_d + Y_aY_e + Y_bY_c + Y_bY_d + Y_bY_e + Y_cY_e + Y_dY_e} \tag{3.47}$$

(2) *Coates's approach.* From (3.44) we obtain the Coates graphs G_l, G_{l0}, and G_{l3} as shown in Fig. 3.41(a), (b), and (c), respectively. The set of all loop-sets of G_{l0} is shown in Fig. 3.42, and the set of all 2-loop-sets of G_{l3} is shown in Fig. 3.43. Thus, by Eq. (3.43),

$$V_0 = \frac{\sum_{(\text{all } p_2)}(-1)^{N_{l_2}}p_2}{\sum_{(\text{all } p)}(-1)^{N_l}p} = \frac{(-1)^1(-Y_e)(Y_aV_i) + (-1)^2(Y_a + Y_b + Y_e)(Y_cV_i)}{(-1)^1(-Y_e)(-Y_e) + (-1)^2(Y_a + Y_b + Y_e)(Y_c + Y_d + Y_e)}$$

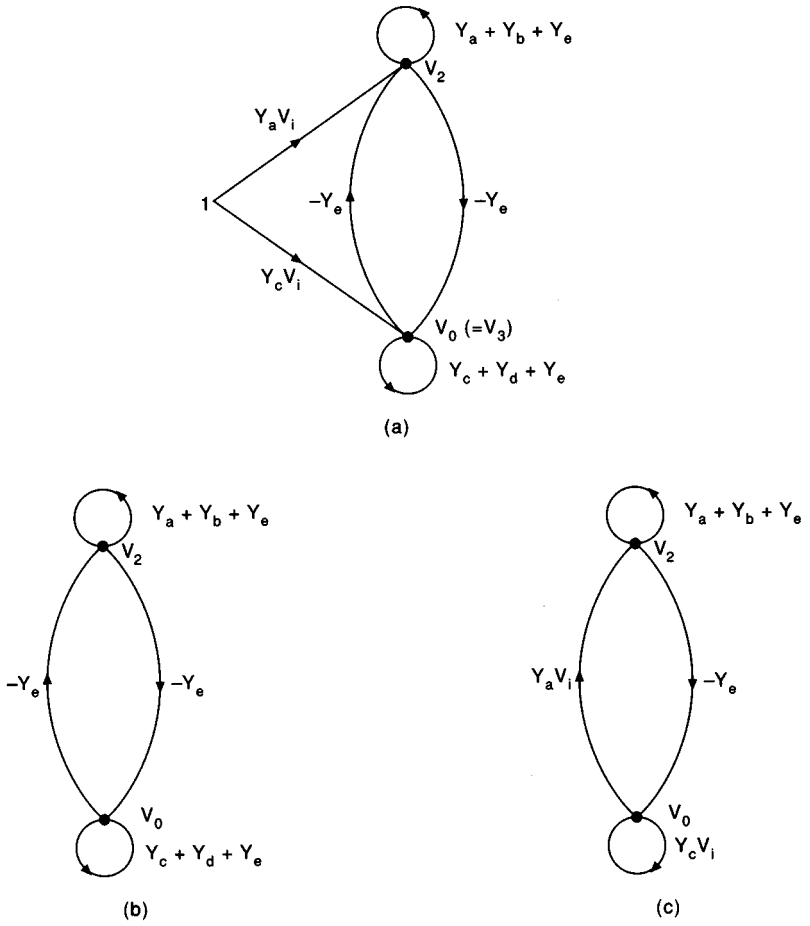


FIGURE 3.41 The Coates graphs: (a) G_l, (b) G_{l0}, and (c) G_{l3}.

Or, after simplification, we find

$$V_0 = \frac{(Y_aY_c + Y_aY_e + Y_bY_c + Y_cY_e)V_i}{Y_aY_c + Y_aY_d + Y_aY_e + Y_bY_c + Y_bY_d + Y_bY_e + Y_cY_e + Y_dY_e} \tag{3.48}$$

which gives the same ratio V_0/V_i as Eq. (3.47).

(3) *The k-tree approach.* Recall that the gain formula for V_0/V_i using the k-tree approach is given [Chan, 1969] by

$$\frac{V_0}{V_i} = \frac{\Delta_{13}}{\Delta_{11}} = \frac{W_{13,R}}{W_{1,R}} = \frac{\sum\left(\begin{array}{l}\text{all 2-tree admittance products with nodes 1 and 3 in one part}\\ \text{and the reference node } R \text{ in the other part of each of such 2-tree}\end{array}\right)}{\sum\left(\begin{array}{l}\text{all 2-tree admittance products with node 1 in one part and}\\ \text{the reference node } R \text{ in the other part of each of such 2-tree}\end{array}\right)} \tag{3.49}$$

where Δ_{13} and Δ_{11} are cofactors of the determinant Δ of the node admittance matrix of the network. Furthermore, it is noted that the 2-trees corresponding to Δ_{ii} may be obtained by finding all the trees of the modified graph G_i, which is obtained from the graph G of the network by short-circuiting node i (i being any node other than

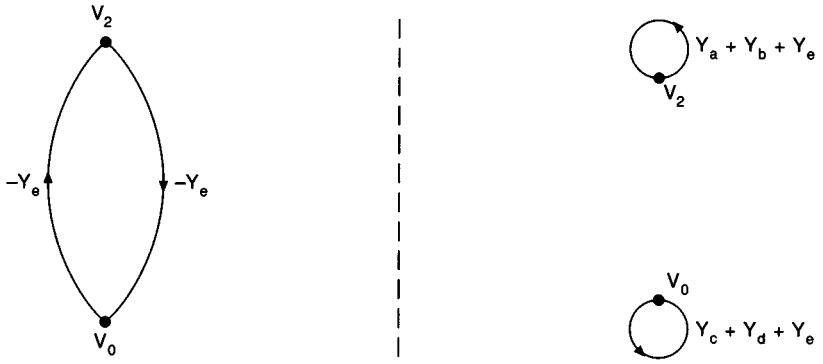


FIGURE 3.42 The set of all loop-sets of G_{l0}.

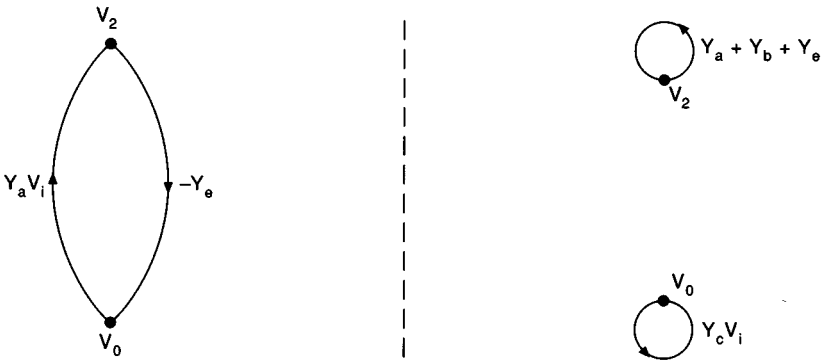


FIGURE 3.43 The set of all 2-loop-sets of G_{l3}.

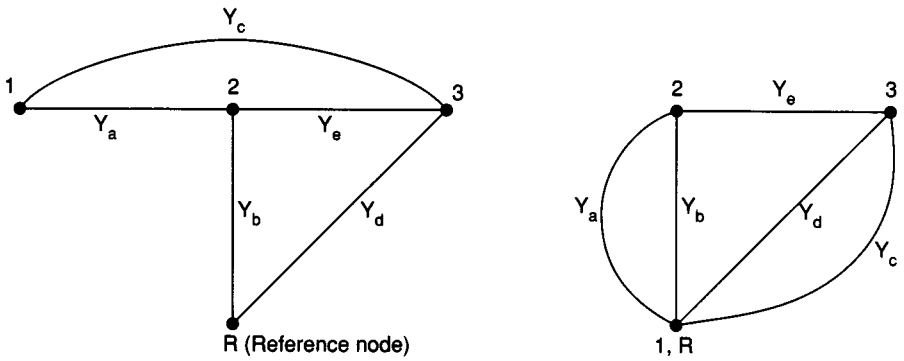


FIGURE 3.44 (a) Graph G, and (b) the modified graph G_l of G.

R) to the reference node R, and that the 2-trees corresponding to Δ_{ij} can be found by taking all those 2-trees each of which is a tree of both G_i and G_j [Chan, 1969]. Thus, for Δ_{11}, we first find G and G_1 (Fig. 3.44), and then find the set S_1 of all trees of G_1 (Fig. 3.45); then for Δ_{13}, we find G_3 (Fig. 3.46) and the set S_3 of all trees of G_3 (Fig. 3.47) and then from S_1 and S_3 we find all the terms common to both sets (which correspond to the

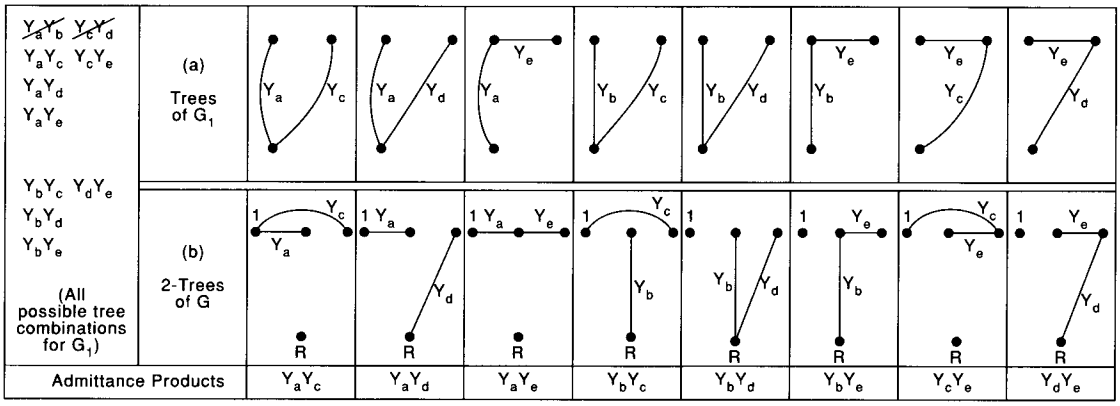


FIGURE 3.45 (a) The set of all trees of the modified graph G_1 which corresponds to (b) the set of all 2-trees of G (with nodes l and R in separate parts in each of such 2-tree).

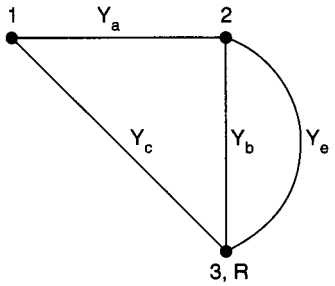


FIGURE 3.46 The modified graph G_3 of G.

set of all trees common to G_1 and G_3) as shown in Fig. 3.48. Finally we form the ratio of 2-tree admittance products according to Eq. (3.49). Thus from Figs. 3.45 and 3.48, we find

$$\frac{V_0}{V_i} = \frac{Y_aY_c + Y_aY_e + Y_bY_c + Y_cY_e}{Y_aY_c + Y_aY_d + Y_aY_e + Y_bY_c + Y_bY_d + Y_bY_e + Y_cY_e + Y_dY_e}$$

which is identical to the results obtained by the flowgraph techniques.

From the above discussions and Example 3.7 we see that the Mason approach is the best from the systems viewpoint, especially when a single source is involved. It gives an excellent physical insight to the system and reveals the cause-effect relationships at various stages when graph reduction technique is employed. While the Coates approach enables one to compute the output directly regardless of the number of inputs involved in the system, thus overcoming one of the difficulties associated with Mason's approach, it does not allow one to reduce the graph step-by-step toward the final solution as Mason's does. However, it is interesting to note that in the modified Coates technique the introduction of the loop-sets (analogous to trees) and the 2-loop-sets (analogous to 2-trees) brings together the two different concepts—the flowgraph approach and the k-tree approach.

From the networks point of view, the Maxwell k-tree approach not only enables one to express the solution in terms of the topology (namely the trees and 2-trees in Example 3.7) of the network but also avoids the cancellation problem inherent in all the flowgraph techniques since, as evident from Example 3.7, the trees and

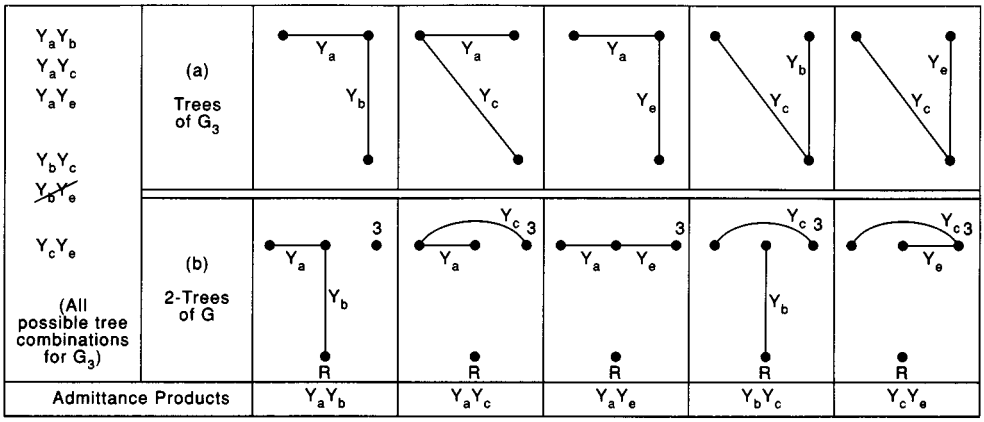


FIGURE 3.47 (a) The set of all trees of the modified graph G_3, which corresponds to (b) the set of all 2-trees of G (with nodes 3 and R in separate parts in each of such 2-tree).

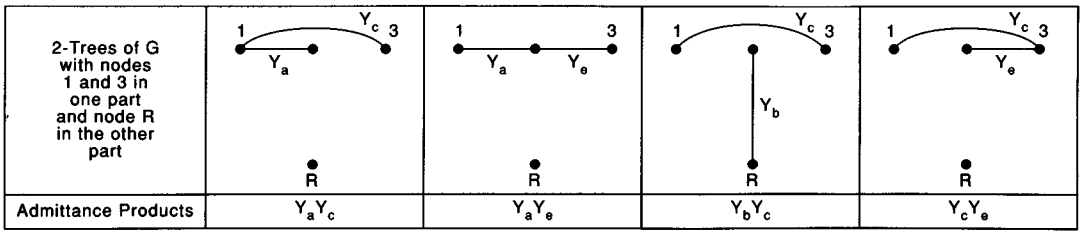


FIGURE 3.48 The set of all 2-trees of G (with nodes l and 3 in one part of the reference node R in the other part of each of such 2-tree).

the 2-trees in the gain expression by the k-tree approach correspond (one-to-one) to the *uncanceled terms* in the final expressions of the gain by the flowgraph techniques. Finally, it should be obvious that the k-tree approach depends upon the knowledge of the graph of a given network. Thus, if in a network problem only the system of (loop or node) equations is given and the network is not known, or more generally, if a system is characterized by a block diagram or a system of equations, the k-tree approach cannot be applied and one must resort to the flowgraph techniques between the two approaches.

Some Topological Applications in Network Analysis and Design

In practice a circuit designer often has to make approximations and analyze the same network structure many times with different sets of component values before the final network realization is obtained. Conventional analysis techniques which require the evaluation of high-order determinants are undesirable even on a digital computer because of the large amount of redundancy inherent in the determinant expansion process. The extra calculation in the evaluation (expansion of determinants) and simplification (cancellation of terms) is time consuming and costly and thereby contributes much to the undesirability of such methods.

The k-tree topological formulas presented in this section, on the other hand, eliminate completely the cancellation of terms. Also, they are particularly suited for digital computation when the size of the network is not exceedingly large. All of the terms involved in the formulas can be computed by means of a digital compute using a single "tree-finding" program [Chan, 1969]. Thus, the application of topological formulas in analyzing a network with the aid of a digital computer can mean a saving of a considerable amount of time and cost to the circuit designer, especially true when it is necessary to repeat the same analysis procedure a large number of times.

In a preliminary system design, the designer usually seeks one or more concepts which will meet the specifications, and in engineering practice each concept is generally subjected to some form of analysis. For linear systems, the signal flowgraph of Mason is widely used in this activity. The flowgraph analysis is popular because it depicts the relationships existing between system variables, and the graphical structure may be manipulated using Mason's formulas to obtain system transfer functions in symbolic or symbolic/numerical form.

Although the preliminary design problems are usually of limited size (several variables), hand derivation of transfer functions is nonetheless difficult and often prone to error arising from the omission of terms. The recent introduction of remote, time-shared computers into modern design areas offers a means to handle such problems swiftly and effectively.

An efficient algorithm suitable for digital compution of transfer functions from the signal flowgraph description of a linear system has been developed (Dunn and Chan, 1969] which provides a powerful analytical tool in the conceptual phases of linear system design.

In the past several decades, graph theory has been widely used in electrical engineering, computer science, social science, and in the solution of economic problems [Swamy and Thulasiraman, 1981; Chen, 1990]. finally, the application of graph theory in conjunction with symbolic network analysis and computer-aided simulation of electronic circuits has been well recognized in recent years [Lin, 1991].

Defining Terms

Branches of a tree: The edges contained in a tree.

Circuit (or loop): A closed path where all vertices are of degree 2, thus having *no* endpoints in the path.

Circuit-set (or loop-set): The set of all edges contained in a circuit (loop).

Connectedness: A graph or subgraph is said to be connected if there is at least one path between *every* pair of its vertices.

Flowgraph G_l (or modified Coates graph G_l): The flowgraph G_l (called *the modified Coates graph*) of a system S of n independent linear equations in n unknowns

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad i = 1, 2, \ldots, n$$

is an oriented graph such that the variable x_j in S is represented by a *node* (also denoted by x_j) in G_l, and the coefficient a_{ij} of the variable x_j in S by a *branch* with a branch gain a_{ij} connected between nodes x_i and x_j in G_l and directed from x_j to x_i. Furthermore, a *source node* l is included in G_l such that for each constant b_k in S there is a node with gain b_k in G_l from node l to node s_k. Graph G_{ij} is the subgraph of G_l obtained by first removing all the outgoing branches from node x_j and then short-circuiting node l to node x_j. A *loop set l* is a subgraph of G_{l0} that contains all the nodes of G_{l0} with each node having exactly one incoming and one outgoing branch. The product p of the gains of all the branches in l is called a *loop-set product.* A 2-*loop-set* l_2 is a subgraph of G_{lj} containing all the nodes of G_{lj} with each node having exactly one incoming and one outgoing branch. The product p_2 of the gains of all the branches in l_2 is called a 2-*loop-set product.*

***k*-tree admittance product of a *k*-tree:** The product of the admittances of all the branches of the *k*-tree.

***k*-tree of a connected graph *G*:** A proper subgraph of G consisting of k unconnected circuitless subgraphs, each subgraph itself being connected, which together contain all the vertices of G.

Linear graph: A set of line segments called edges and points called vertices, which are the endpoints of the edges, interconnected in such a way that the edges are connected to (or incident with) the vertices. The degree of a vertex of a graph is the number of edges incident with that vertex.

Path: A subgraph having all vertices of degree 2 except for the two endpoints which are of degree 1 and are called the terminals of the path, where the degree of a vertex is the number of edges connected to the vertex in the subgraph.

Path-set: The set of all edges in a path.

Proper subgraph: A subgraph which does not contain all of the edges of the given graph.

Signal-flowgraph G_m (or Mason's graph G_m): A signal-flowgraph G_m of a system S of n independent linear (algebraic) equations in n unknowns

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad i = 1, 2, \ldots, n$$

is a graph with junction points called *nodes* which are connected by directed line segments called *branches* with signals traveling along the branches only in the direction described by the arrows of the branches. A signal x_k traveling along a branch between x_k and x_j is multiplied by the gain of the branches g_{kj}, so that a signal $g_{kj}\,x_k$ is delivered at node x_j. An *input node* (*source*) is a node which contains only outgoing branches; an *output node* (*sink*) is a node which has only incoming branches. A *path* is a continuous unidirectional succession of branches, all of which are traveling in the same direction; a *forward path* is a path from the input node to the output node along which all nodes are encountered exactly once; and a *feedback path* (*loop*) is a closed path which originates from and terminates at the same node, and along with all other nodes are encountered exactly once (the trivial case is a *self-loop* which contains exactly one node and one branch). A *path* gain is the product of all the branch gains of the branches in a loop.

Subgraph: A subset of the edges of a given graph.

Tree: A connected subgraph of a given connected graph G which contains all the vertices of G but no circuits.

Related Topic

3.2 Node and Mesh Analysis

References

A.C. Aitken, *Determinants and Matrices,* 9th ed., New York: Interscience, 1956.

S.P. Chan, *Introductory Topological Analysis of Electrical Networks,* New York: Holt, Rinehart and Winston, 1969.

S.P. Chan and B.H. Bapna, "A modification of the Coates gain formula for the analysis of linear systems," *Inst. J. Control,* vol. 5, pp. 483–495, 1967.

S.P. Chan and S.G. Chan, "Modifications of topological formulas," *IEEE Trans. Circuit Theory,* vol. CT-15, pp. 84–86, 1968.

W.K. Chen, *Theory of Nets: Flows in Networks,* New York: Wiley Interscience, 1990.

C.L. Coates, "Flow-graph solutions of linear algebraic equations," *IRE Trans. Circuit Theory,* vol. CT-6, pp. 170–187, 1959.

W.R. Dunn, Jr., and S.P. Chan, "Flowgraph analysis of linear systems using remote timeshared computation," *J. Franklin Inst.,* vol. 288, pp. 337–349, 1969.

G. Kirchhoff, "Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme, geführt wird," *Ann. Physik Chemie,* vol. 72, pp. 497–508, 1847; English transl., *IRE Trans. Circuit Theory,* vol. CT-5, pp. 4–7, 1958.

P.M. Lin, *Symbolic Network Analysis,* New York: Elsevier, 1991.

S.J. Mason, "Feedback theory—Some properties of signal flow graphs," *Proc. IRE,* vol. 41, pp. 1144–1156, 1953.

S.J. Mason, "Feedback theory—Further properties of signal flow graphs," *Proc. IRE,* vol. 44, pp. 920–926, 1956.

J.C. Maxwell, *Electricity and Magnetism,* Oxford: Clarendon Press, 1892.

W.S. Percival, "Solution of passive electrical networks by means of mathematical trees," *Proc. IEE,* vol. 100, pp. 143–150, 1953.

S. Seshu and M.B. Reed, *Linear Graphs and Electrical Networks,* Reading, Mass.: Addison-Wesley, 1961.

M.N.S. Swamy ad K. Thulasiraman, *Graphs, Networks, and Algorithms,* New York: Wiley, 1981.

Further Information

All defining terms used in this section can be found in S.P. Chan, *Introductory Topological Analysis of Electrical Networks,* Holt, Rinehart and Winston, New York, 1969. Also an excellent reference for the applications of graph

theory in electrical engineering (i.e., network analysis and design) is S. Seshu and M.B. Reed, *Linear Graphs and Electrical Networks,* Addison-Wesley, Reading, Mass., 1961.

For applications of graph theory in computer science, see M.N.S. Swamy and K. Thulasiraman, *Graphs, Networks, and Algorithms,* Wiley, New York, 1981.

For flowgraph applications, see W.K. Chen, *Theory of Nets: Flows in Networks,* Wiley Interscience, New York, 1990.

For applications of graph theory in symbolic network analysis, see P.M. Lin, *Symbolic Network Analysis,* Elsevier, New York, 1991.

3.7 Two-Port Parameters and Transformations

Norman S. Nise

Introduction

Many times we want to model the behavior of an electric network at only two terminals as shown in Fig. 3.49. Here, only V_1 and I_1, not voltages and currents internal to the circuit, need to be described. To produce the model for a linear circuit, we use **Thévenin's** or **Norton's theorem** to simplify the network as viewed from the selected terminals. We define the pair of terminals shown in Fig. 3.49 as a **port**, where the current, I_1, entering one terminal equals the current leaving the other terminal.

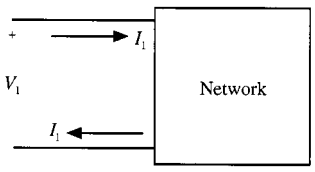


FIGURE 3.49 An electrical network port.

If we further restrict the network by stating that (1) all external connections to the circuit, such as sources and impedances, are made at the port and (2) the network can have internal **dependent sources**, but not **independent sources**, we can mathematically model the network at the port as

$$V_1 = ZI_1 \tag{3.50}$$

or

$$I_1 = YV_1 \tag{3.51}$$

where Z is the Thévenin impedance and Y is the Norton admittance at the terminals. Z and Y can be constant resistive terms, Laplace transforms $Z(s)$ or $Y(s)$, or sinusoidal steady-state functions $Z(j\omega)$ or $Y(j\omega)$.

Defining Two-Port Networks

Electrical networks can also be used to transfer signals from one port to another. Under this requirement, connections to the network are made in two places, the input and the output. For example, a transistor has an input between the base and emitter and an output between the collector and emitter. We can model such circuits as **two-port networks** as shown in Fig. 3.50. Here we see the input port, represented by V_1 and I_1, and the output port, represented by V_2 and I_2. Currents are assumed positive if they flow as shown in Fig. 3.50. The same restrictions about external connections and internal sources mentioned above for the single port also apply.

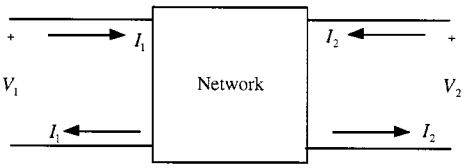


FIGURE 3.50 A two-port network.

Now that we have defined two-port networks, let us discuss how to create a mathematical model of the network by establishing relationships among all of the input and output voltages and currents. Many possibilities exist for modeling. In the next section we arbitrarily begin by introducing the z-parameter model to establish the technique. In subsequent sections we present alternative models and draw relationships among them.

Mathematical Modeling of Two-Port Networks via *z* Parameters

In order to produce a mathematical model of circuits represented by Fig. 3.50, we must find relationships among V_1, I_1, V_2, and I_2. Let us visualize placing a current source at the input and a current source at the output. Thus, we have selected two of the variables, I_1 and I_2. We call these variables the independent variables. The remaining variables, V_1 and V_2, are dependent upon the selected applied currents. We call V_1 and V_2 the dependent variables. Using **superposition** we can write each dependent variable as a function of the independent variables as follows:

$$V_1 = z_{11}I_1 + z_{12}I_2 \tag{3.52a}$$

$$V_2 = z_{21}I_1 + z_{22}I_2 \tag{3.52b}$$

We call the coefficients, z_{ij}, in Eqs. (3.52) parameters of the two-port network or, simply, **two-port parameters**.

From Eqs. (3.52), the two-port parameters are evaluated as

$$\begin{aligned} z_{11} &= \left.\frac{V_1}{I_1}\right|_{I_2=0}; \quad z_{12} = \left.\frac{V_1}{I_2}\right|_{I_1=0} \\ z_{21} &= \left.\frac{V_2}{I_1}\right|_{I_2=0}; \quad z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1=0} \end{aligned} \tag{3.53}$$

Notice that each parameter can be measured by setting a port current, I_1 or I_2, equal to zero. Since the parameters are found by setting these currents equal to zero, this set of parameters is called **open-circuit parameters**. Also, since the definitions of the parameters as shown in Eqs. (3.53) are the ratio of voltages to currents, we alternatively refer to them as **impedance parameters**, or ***z* parameters**. The parameters themselves can be impedances represented as Laplace transforms, $Z(s)$, sinusoidal steady-state impedance functions, $Z(j\omega)$, or simply pure resistance values, R.

Evaluating Two-Port Network Characteristics in Terms of *z* Parameters

The two-port parameter model can be used to find the following characteristics of a two-port network when used in some cases with a source and load as shown in Fig. 3.51:

$$\text{Input impedance} = Z_{\text{in}} = V_1/I_1 \tag{3.54a}$$

$$\text{Output impedance} = Z_{\text{out}} = V_2/I_2 \mid V_S = 0 \tag{3.54b}$$

$$\text{Network voltage gain} = V_g = V_2/V_1 \tag{3.54c}$$

$$\text{Total voltage gain} = V_{gt} = V_2/V_S \tag{3.54d}$$

$$\text{Network current gain} = I_g = I_2/I_1 \tag{3.54e}$$

To find Z_{in} of Fig. 3.51, determine V_1/I_1. From Fig. 3.51, $V_2 = -I_2 Z_L$. Substituting this value in Eq. 3.52(b) and simplifying, Eqs. (3.52) become

$$V_1 = z_{11}I_1 + z_{12}I_2 \tag{3.55a}$$

$$0 = z_{21}I_1 + (z_{22} + Z_L)I_2 \tag{3.55b}$$

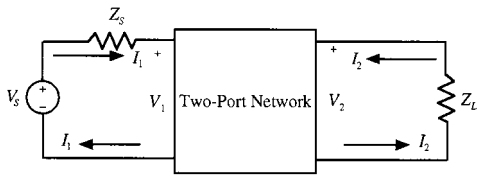


FIGURE 3.51 Terminated two-port network for finding two-port network characteristics.

Solving simultaneously for I_1 and then forming $V_1/I_1 = Z_{in}$, we obtain

$$Z_{in} = \frac{V_1}{I_1} = z_{11} - \frac{z_{12}z_{21}}{(z_{22} + Z_L)} \tag{3.56}$$

To find Z_{out}, set $V_S = 0$ in Fig. 3.51. This step terminates the input with Z_S. Next, determine V_2/I_2. From Fig. 3.51 with V_S shorted, $V_1 = -I_1 Z_S$. By substituting this value into Eq. 3.52(a) and simplifying, Eqs. (3.52) become

$$0 = (z_{11} + Z_S)I_1 + z_{12}I_2 \tag{3.57a}$$

$$V_2 = z_{21}I_1 + z_{22}I_2 \tag{3.57b}$$

By solving simultaneously for I_2 and then forming $V_2/I_2 = Z_{out}$,

$$Z_{out} = \left.\frac{V_2}{I_2}\right|_{V_S=0} = z_{22} - \frac{z_{12}z_{21}}{(z_{11} + Z_S)} \tag{3.58}$$

To find V_g, we see from Fig. 3.51 that $I_2 = -V_2/Z_L$. Substituting this value in Eqs. (3.52) and simplifying, we obtain

$$V_1 = z_{11}I_1 - \frac{z_{12}}{Z_L}V_2 \tag{3.59a}$$

$$0 = z_{21}I_1 - \left(\frac{z_{22} + Z_L}{Z_L}\right)V_2 \tag{3.59b}$$

By solving simultaneously for V_2 and then forming $V_2/V_1 = V_g$,

$$V_g = \frac{V_2}{V_1} = \frac{z_{21}Z_L}{z_{11}(z_{22} + Z_L) - z_{12}z_{21}} \tag{3.60}$$

Similarly, other characteristics, such as current gain and the total voltage gain from the source voltage to the load voltage can be found. Table 3.1 summarizes many of the network characteristics that can be found using z parameters as well as the process to arrive at the result.

TABLE 3.1 Network Characteristics Developed from *z*-Parameter Defining Eqs. (3.52)

Network Characteristic Definition	From Fig. 3.51	Substitute in Defining Eqs. (3.52) and Obtain	Solve for Network Characteristic
Input impedance $z_{in} = \dfrac{V_1}{I_1}$	$V_2 = -I_2 Z_L$	$V_1 = z_{11} I_1 + z_{12} I_2$ $0 = z_{21} I_1 + (z_{22} + Z_L) I_2$	$Z_{in} = z_{11} - \dfrac{z_{12} z_{21}}{z_{22} + Z_L}$
Output impedance $Z_{out} = \left.\dfrac{V_2}{I_2}\right\|_{V_S = 0}$	$V_1 = V_S - I_1 Z_S$ $V_S = 0$	$0 = (z_{11} + Z_S) I_1 + z_{12} I_2$ $V_2 = z_{21} I_1 + z_{22} I_2$	$Z_{out} = z_{22} - \dfrac{z_{12} z_{21}}{z_{11} + Z_S}$
Network voltage gain $V_g = \dfrac{V_2}{V_1}$	$I_2 = -\dfrac{V_2}{Z_L}$	$V_1 = z_{11} I_1 - \dfrac{z_{12} V_2}{Z_L}$ $0 = z_{21} I_1 - \dfrac{(z_{22} + Z_L) V_2}{Z_L}$	$V_g = \dfrac{z_{21} Z_L}{z_{11}(z_{22} + Z_L) - z_{12} z_{21}}$
Total voltage gain $V_{gt} = \dfrac{V_2}{V_S}$	$V_1 = V_S - I_1 Z_S$ $I_2 = -\dfrac{V_2}{Z_L}$	$V_S = (z_{11} + Z_S) I_1 - \dfrac{z_{12} V_2}{Z_L}$ $0 = z_{21} I_1 - \dfrac{(z_{22} + Z_L) V_2}{Z_L}$	$V_{gt} = \dfrac{z_{21} Z_L}{(z_{11} + Z_S)(z_{22} + Z_L) - z_{12} z_{21}}$
Network current gain $I_g = \dfrac{I_2}{I_1}$	$V_2 = -I_2 Z_L$	$V_1 = z_{11} I_1 + z_{12} I_2$ $0 = z_{21} I_1 + (z_{22} + Z_L) I_2$	$I_g = -\dfrac{z_{21}}{z_{22} + Z_L}$

To summarize the process of finding network characteristics:

1. Define the network characteristic.
2. Use approproate relationships from Fig. 3.51.
3. Substitute the relationships from Step 2 into Eqs. (3.52).
4. Solve the modified equations for the network characteristic.

An Example Finding *z* Parameters and Network Characteristics

To solve for two-port network characteristics we can first represent the network with its two-port parameters and then use these parameters to find the characteristics summarized in Table 3.1. To find the parameters, we terminate the network adhering to the definition of the parameter we are evaluating. Then, we can use mesh or nodal analysis, current or voltage division, or equivalent impedance to solve for the parameters. The following example demonstates the technique.

Consider the network of Fig. 3.52(a). The first step is to evaluate the *z* parameters. From their definition, z_{11} and z_{21} are found by open-circuiting the output and applying a voltage at the input as shown in Fig. 3.52(b). Thus, with $I_2 = 0$

$$6 I_1 - 4 I_a = V_1 \tag{3.61a}$$

$$-4 I_1 + 18 I_a = 0 \tag{3.61b}$$

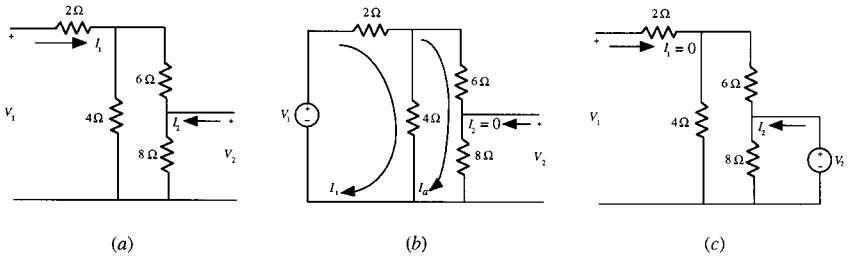


FIGURE 3.52 (a) Two-port network example; (b) two-port network modified to find z_{11} and z_{21}; (c) two-port network modified to find z_{22} and z_{12}.

Solving for I_1 yields

$$I_1 = \frac{\begin{vmatrix} V_1 & -4 \\ 0 & 18 \end{vmatrix}}{\begin{vmatrix} 6 & -4 \\ -4 & 18 \end{vmatrix}} = \frac{18V_1}{92} \tag{3.62}$$

from which

$$z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0} = \frac{46}{9} \tag{3.63}$$

We now find z_{21}. From Eq. (3.61b)

$$\frac{I_a}{I_1} = \frac{2}{9} \tag{3.64}$$

But, from Fig. 3.52(b), $I_a = V_2/8$. Thus,

$$z_{21} = \left.\frac{V_2}{I_1}\right|_{I_2=0} = \frac{16}{9} \tag{3.65}$$

Based on their definitions, z_{22} and z_{12} are found by placing a source at the output and open-circuiting the input as shown in Fig. 3.52(c). The equivalent resistance, R_{2eq}, as seen at the output with $I_1 = 0$ is

$$R_{2eq} = \frac{8 \times 10}{8 + 10} = \frac{40}{9} \tag{3.66}$$

Therefore,

$$z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1=0} = \frac{40}{9} \tag{3.67}$$

From Fig. 3.52(c), using voltage division

$$V_1 = (4/10)\,V_2 \tag{3.68}$$

But

$$V_2 = \mathrm{I}_2 \mathrm{R}_{2\,\mathrm{eq}} = \mathrm{I}_2(40/9) \tag{3.69}$$

Substituting Eq. (3.69) into Eq. (3.68) and simplifying yields

$$z_{12} = \left.\frac{V_1}{I_2}\right|_{I_1=0} = \frac{16}{9} \tag{3.70}$$

Using the z-parameter values found in Eqs. (3.63), (3.65), (3.67), and (3.70) and substituting into the network characteristic relationships shown in the last column of Table 3.1, assuming $Z_S = 20\ \Omega$ and $Z_L = 10\ \Omega$, we obtain $Z_{\mathrm{in}} = 4.89\ \Omega$, $Z_{\mathrm{out}} = 4.32\ \Omega$, $V_g = 0.252$, $V_{gt} = 0.0494$, and $I_g = -0.123$.

Additional Two-Port Parameters and Conversions

We defined the *z* parameters by establishing I_1 and I_2 as the independent variables and V_1 and V_2 as the dependent variables. Other choices of independent and dependent variables lead to definitions of alternative two-port parameters. The total number of combinations one can make with the four variables, taking two at a time as independent variables, is six. Table 3.2 defines the six possibilities as well as the names and symbols given to the parameters.

The table also presents the expressions used to calculate directly the parameters of each set based upon their definition as we did with *z* parameters. For example, consider the *y*, or **admittance, parameters**. These parameters are seen to be **short-circuit parameters**, since their evaluation requires V_1 or V_2 to be zero. Thus, to find y_{22} we short-circuit the input and find the admittance looking back from the output. For Fig. 3.52(a), $y_{22} = 23/88$. Any parameter in Table 3.2 is found either by open-circuiting or short-circuiting a terminal and then performing circuit analysis to find the defining ratio.

Another method of finding the parameters is to convert from one set to another. Using the "Definition" row in Table 3.2, we can convert the defining equations of one set to the defining equations of another set. For example, we have already found the *z* parameters. We can find the ***h* parameters** as follows:

Solve for I_2 using the second *z*-parameter equation, Eq. (3.52b), and obtain the second *h*-parameter equation as

$$I_2 = -\frac{z_{21}}{z_{22}}\,I_1 + \frac{1}{z_{22}}\,V_2 \tag{3.71}$$

which is of the form, $I_2 = h_{21}I_1 + h_{22}V_2$, the second *h*-parameter equation. Now, substitute Eq. (3.71) into the first *z*-parameter equation, Eq. (3.52a), rearrange, and obtain

$$V_1 = \frac{z_{11}z_{22} - z_{12}z_{21}}{z_{22}}\,I_1 + \frac{z_{12}}{z_{22}}\,V_2 \tag{3.72}$$

which is of the form, $V_1 = h_{11}I_1 + h_{12}V_2$, the first *h*-parameter equation. Thus, for example, $h_{21} = -z_{21}/z_{22}$ from Eq. (3.71). Other transformations are found through similar manipulations and are summarized in Table 3.2.

Finally, there are other parameter sets that are defined differently from the standard sets covered here. Specifically, they are scattering parameters used for microwave networks and image parameters used for filter design. A detailed discussion of these parameters is beyond the scope of this section. The interested reader should consult the bibliography in the "Further Information" section below, or Section 39.1 of this handbook.

TABLE 3.2 Two-Port Parameter Definitions and Conversions

	Impedance Parameters (Open-Circuit Parameters) z	Admittance Parameters (Short-Circuit Parameters) y	Hybrid Parameters h
Definition	$V_1 = z_{11}I_1 + z_{12}I_2$ $V_2 = z_{21}I_1 + z_{22}I_2$	$I_1 = y_{11}V_1 + y_{12}V_2$ $I_2 = y_{21}V_1 + y_{22}V_2$	$V_1 = h_{11}I_1 + h_{12}V_2$ $I_2 = h_{21}I_1 + h_{22}V_2$
Parameters	$z_{11} = \left.\frac{V_1}{I_1}\right\|_{I_2=0}$; $z_{12} = \left.\frac{V_1}{I_2}\right\|_{I_1=0}$ $z_{21} = \left.\frac{V_2}{I_1}\right\|_{I_2=0}$; $z_{22} = \left.\frac{V_2}{I_2}\right\|_{I_1=0}$	$y_{11} = \left.\frac{I_1}{V_1}\right\|_{V_2=0}$; $y_{12} = \left.\frac{I_1}{V_1}\right\|_{V_1=0}$ $y_{21} = \left.\frac{I_2}{V_1}\right\|_{V_2=0}$; $y_{22} = \left.\frac{I_2}{V_1}\right\|_{V_1=0}$	$h_{11} = \left.\frac{V_1}{I_1}\right\|_{V_2=0}$; $h_{12} = \left.\frac{V_1}{V_2}\right\|_{I_1=0}$ $h_{21} = \left.\frac{I_2}{I_1}\right\|_{V_2=0}$; $h_{22} = \left.\frac{I_2}{V_2}\right\|_{I_1=0}$
Conversion to z parameters		$z_{11} = \frac{y_{22}}{\Delta_y}$; $z_{12} = \frac{-y_{12}}{\Delta_y}$ $z_{21} = \frac{-y_{21}}{\Delta_y}$; $z_{22} = \frac{y_{11}}{\Delta_y}$	$z_{11} = \frac{\Delta_h}{h_{22}}$; $z_{12} = \frac{h_{12}}{h_{22}}$ $z_{21} = \frac{-h_{21}}{h_{22}}$; $z_{22} = \frac{1}{h_{22}}$
Conversion to y parameters	$y_{11} = \frac{z_{22}}{\Delta_z}$; $y_{12} = \frac{-z_{12}}{\Delta_z}$ $y_{21} = \frac{-z_{21}}{\Delta_z}$; $y_{22} = \frac{z_{11}}{\Delta_z}$		$y_{11} = \frac{1}{h_{11}}$; $y_{12} = \frac{-h_{12}}{h_{11}}$ $y_{21} = \frac{h_{21}}{h_{11}}$; $y_{22} = \frac{\Delta_h}{h_{11}}$
Conversion to h parameters	$h_{11} = \frac{\Delta_z}{z_{22}}$; $h_{12} = \frac{z_{12}}{z_{22}}$ $h_{21} = \frac{-z_{21}}{z_{22}}$; $h_{22} = \frac{1}{z_{22}}$	$h_{11} = \frac{1}{y_{11}}$; $h_{12} = \frac{-y_{12}}{y_{11}}$ $h_{21} = \frac{y_{21}}{y_{11}}$; $h_{22} = \frac{\Delta_y}{y_{11}}$	
Conversion to g parameters	$g_{11} = \frac{1}{z_{11}}$; $g_{12} = \frac{-z_{12}}{z_{11}}$ $g_{21} = \frac{z_{21}}{z_{11}}$; $g_{22} = \frac{\Delta_z}{z_{11}}$	$g_{11} = \frac{\Delta_y}{y_{22}}$; $g_{12} = \frac{y_{12}}{y_{22}}$ $g_{21} = \frac{-y_{21}}{y_{22}}$; $g_{22} = \frac{1}{y_{22}}$	$g_{11} = \frac{h_{22}}{\Delta_h}$; $g_{12} = \frac{-h_{12}}{\Delta_h}$ $g_{21} = \frac{-h_{21}}{\Delta_h}$; $g_{22} = \frac{h_{11}}{\Delta_h}$
Conversion to T parameters	$A = \frac{z_{11}}{z_{21}}$; $B = \frac{\Delta_z}{z_{21}}$ $C = \frac{1}{z_{21}}$; $D = \frac{z_{22}}{z_{21}}$	$A = \frac{-y_{22}}{y_{21}}$; $B = \frac{-1}{y_{21}}$ $C = \frac{-\Delta_y}{y_{21}}$; $D = \frac{-y_{11}}{y_{21}}$	$A = \frac{-\Delta_h}{h_{21}}$; $B = \frac{-h_{11}}{h_{21}}$ $C = \frac{-h_{22}}{h_{21}}$; $D = \frac{-1}{h_{21}}$
Conversion to T' parameters	$A' = \frac{z_{22}}{z_{12}}$; $B' = \frac{\Delta_z}{z_{12}}$ $C' = \frac{1}{z_{12}}$; $D' = \frac{z_{11}}{z_{12}}$	$A' = \frac{-y_{11}}{y_{12}}$; $B' = \frac{-1}{y_{12}}$ $C' = \frac{-\Delta_y}{y_{12}}$; $D' = \frac{-y_{22}}{y_{12}}$	$A' = \frac{1}{h_{12}}$; $B' = \frac{h_{11}}{h_{12}}$ $C' = \frac{h_{22}}{h_{12}}$; $D' = \frac{\Delta_h}{h_{12}}$
Δ	$\Delta_z = z_{11}z_{22} - z_{12}z_{21}$	$\Delta_y = y_{11}y_{22} - y_{12}y_{21}$	$\Delta_h = h_{11}h_{22} - h_{12}h_{21}$

TABLE 3.2 (continued) Two-Port Parameter Definitions and Conversions

	Inv. hybrid parameters g	Transmission parameters T	Inv. transmission par. T'
Definition	$I_1 = g_{11}V_1 + g_{12}I_2$ $V_2 = g_{21}V_1 + g_{22}I_2$	$V_1 = AV_2 - BI_2$ $I_1 = CV_2 - DI_2$	$V_2 = A'V_1 - B'I_1$ $I_2 = C'V_1 - D'I_1$
Parameters	$g_{11} = \left.\frac{I_1}{V_1}\right\|_{I_2=0}$; $g_{12} = \left.\frac{I_1}{I_2}\right\|_{V_1=0}$ $g_{21} = \left.\frac{V_2}{V_1}\right\|_{I_2=0}$; $g_{22} = \left.\frac{V_2}{I_2}\right\|_{V_1=0}$	$A = \left.\frac{V_1}{V_2}\right\|_{I_2=0}$; $B = \left.\frac{-V_1}{I_2}\right\|_{V_2=0}$ $C = \left.\frac{I_1}{V_2}\right\|_{I_2=0}$; $D = \left.\frac{-I_1}{I_2}\right\|_{V_2=0}$	$A' = \left.\frac{V_2}{V_1}\right\|_{I_1=0}$; $B' = \left.\frac{-V_2}{I_1}\right\|_{V_1=0}$ $C' = \left.\frac{I_2}{V_1}\right\|_{I_1=0}$; $D' = \left.\frac{-I_2}{I_1}\right\|_{V_1=0}$
Conversion to z parameters	$z_{11} = \frac{1}{g_{11}}$; $z_{12} = \frac{-g_{12}}{g_{11}}$ $z_{21} = \frac{g_{21}}{g_{11}}$; $z_{22} = \frac{\Delta_g}{g_{11}}$	$z_{11} = \frac{A}{C}$; $z_{12} = \frac{\Delta_T}{C}$ $z_{21} = \frac{1}{C}$; $z_{22} = \frac{D}{C}$	$z_{11} = \frac{D'}{C'}$; $z_{12} = \frac{1}{C'}$ $z_{21} = \frac{\Delta_{T'}}{C'}$; $z_{22} = \frac{A'}{C'}$
Conversion to y parameters	$y_{11} = \frac{\Delta_g}{g_{22}}$; $y_{12} = \frac{g_{12}}{g_{22}}$ $y_{21} = \frac{-g_{21}}{g_{22}}$; $y_{22} = \frac{1}{g_{22}}$	$y_{11} = \frac{D}{B}$; $y_{12} = \frac{-\Delta_T}{B}$ $y_{21} = \frac{-1}{B}$; $y_{22} = \frac{A}{B}$	$y_{11} = \frac{A'}{B'}$; $y_{12} = \frac{-1}{B'}$ $y_{21} = \frac{-\Delta_{T'}}{B'}$; $y_{22} = \frac{D'}{B'}$
Conversion to h parameters	$h_{11} = \frac{g_{22}}{\Delta_g}$; $h_{12} = \frac{-g_{12}}{\Delta_g}$ $h_{21} = \frac{-g_{21}}{\Delta_g}$; $h_{22} = \frac{g_{11}}{\Delta_g}$	$h_{11} = \frac{B}{D}$; $h_{12} = \frac{\Delta_T}{D}$ $h_{21} = \frac{-1}{D}$; $h_{22} = \frac{C}{D}$	$h_{11} = \frac{B'}{A'}$; $h_{12} = \frac{1}{A'}$ $h_{21} = \frac{-\Delta_{T'}}{A'}$; $h_{22} = \frac{C'}{A'}$
Conversion to g parameters		$g_{11} = \frac{C}{A}$; $g_{12} = \frac{-\Delta_T}{A}$ $g_{21} = \frac{1}{A}$; $g_{22} = \frac{B}{A}$	$g_{11} = \frac{C'}{D'}$; $g_{12} = \frac{-1}{D'}$ $g_{21} = \frac{\Delta_{T'}}{D'}$; $g_{22} = \frac{B'}{D'}$
Conversion to T parameters	$A = \frac{1}{g_{21}}$; $B = \frac{g_{22}}{g_{21}}$ $C = \frac{g_{11}}{g_{21}}$; $D = \frac{\Delta_g}{g_{21}}$		$A = \frac{D'}{\Delta_{T'}}$; $B = \frac{B'}{\Delta_{T'}}$ $C = \frac{C'}{\Delta_{T'}}$; $D = \frac{A'}{\Delta_{T'}}$
Conversion to T' parameters	$A' = \frac{-\Delta_g}{g_{12}}$; $B' = \frac{-g_{22}}{g_{12}}$ $C' = \frac{-g_{11}}{g_{12}}$; $D' = \frac{-1}{g_{12}}$	$A' = \frac{D}{\Delta_T}$; $B' = \frac{B}{\Delta_T}$ $C' = \frac{C}{\Delta_T}$; $D' = \frac{A}{\Delta_T}$	
Δ	$\Delta_g = g_{11}g_{22} - g_{12}g_{21}$	$\Delta_T = AD - BC$	$\Delta_{T'} = A'D' - B'C'$

Adapted from Van Valkenburg, M.E. 1974. *Network Analysis*, 3rd ed. Table 11-2, p. 337. Prentice-Hall, Englewood Cliffs, NJ. With permission.

Two-Port Parameter Selection

The choice of parameters to use for a particular analysis or design problem is based on analytical convenience or the physics of the device or network at hand. For example, an ideal transformer cannot be represented with z parameters. I_1 and I_2 are not linearly independent variables, since they are related through the turns ratio. A similar argument applies to the ***y*-parameter** representation of a transformer. Here V_1 and V_2 are not independent, since they too are related via the turns ratio. A possible choice for the transformer is the **transmission parameters**. For an ideal transformer, B and C would be zero. For a BJT transistor, there is effectively linear independence between the input current and the output voltage. Thus, the hybrid parameters are the parameters of choice for the transistor.

The choice of parameters can be based also upon the ease of analysis. For example, Table 3.3 shows that "T" networks lend themselves to easy evaluation of the z parameters, while y parameters can be easily evaluated for "Π" networks. Table 3.3 summarizes other suggested uses and selections of network parameters for a few specific cases. When electric circuits are interconnected, a judicious choice of parameters can simplify the calculations to find the overall parameter description for the interconnected networks. For example, Table 3.3 shows that the z parameters for series-connected networks are simply the sum of the z parameters of the individual circuits (see Ruston et al., [1966] for derivations of the parameters for some of the interconnected networks). The bold entries imply 2×2 matrices containing the four parameters. For example,

$$\mathbf{h} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \tag{3.73}$$

Summary

In this section, we developed two-port parameter models for two-port electrical networks. The models define interrelationships among the input and output voltages and currents. A total of six models exists, depending upon which two variables are selected as independent variables. Any model can be used to find such network characteristics as input and output impedance, and voltage and current gains. Once one model is found, other models can be obtained from transformation equations. The choice of parameter set is based upon physical reality and analytical convenience.

Defining Terms

Admittance parameters: That set of two-port parameters, such as y parameters, where all the parameters are defined to be the ratio of current to voltage. See Table 3.2 for the specific definition.

Dependent source: A voltage or current source whose value is related to another voltage or current in the network.

***g* Parameters:** See hybrid parameters.

***h* Parameters:** See hybrid parameters.

Hybrid (inverse hybrid) parameters: That set of two-port parameters, such as $h(g)$ parameters, where input current (voltage) and output voltage (current) are the independent variables. The parenthetical expressions refer to the inverse hybrid parameters. See Table 3.2 for specific definitions.

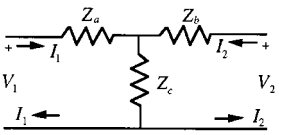
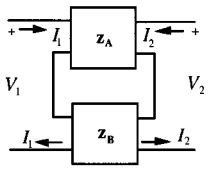
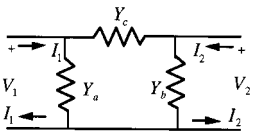
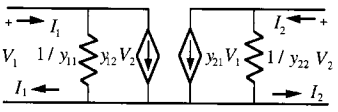
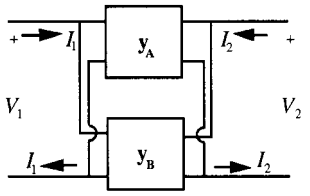
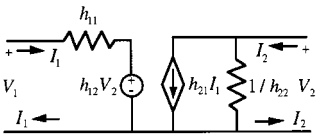
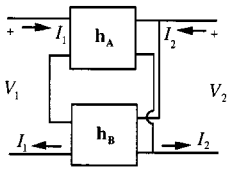
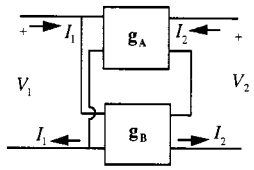
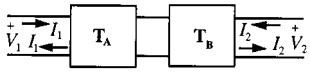
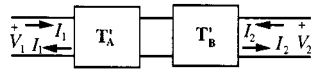
Impedance parameters: That set of two-port parameters, such as z parameters, where all the parameters are defined to be the ratio of voltage to current. See Table 3.2 for the specific definition.

Independent source: A voltage or current source whose value is not related to any other voltage or current in the network.

Norton's theorem: At a pair of terminals a linear electrical network can be replaced with a current source in parallel with an admittance. The current source is equal to the current that flows through the terminals when the terminals are short-circuited. The admittance is equal to the admittance at the terminals with all independent sources set equal to zero.

Open-circuit parameters: Two-port parameters, such as z parameters, evaluated by open-circuiting a port.

TABLE 3.3 Two-Port Parameter Set Selection

	Common Circuit Applications	Interconnected Network Applications
Impedance parameters z	• T networks  $z_{11} = Z_a + Z_c$; $z_{12} = z_{21} = Z_c$ $z_{22} = Z_b + Z_c$	• Series connected  $\mathbf{z = z_A + z_B}$
Admittance parameters y	• Π networks  $y_{11} = Y_a + Y_c$; $y_{12} = y_{21} = -Y_c$ $y_{22} = Y_b + Y_c$ • Field effect transistor equivalent circuit  where typically: $1/y_{11} = \infty$, $y_{12} = 0$, $y_{21} = g_m$, $1/y_{22} = r_d$	• Parallel connected  $\mathbf{y = y_A + y_B}$
Hybrid parameters h	• Transistor equivalent circuit  where typically for common emitter: $h_{11} = h_{ie}$, $h_{12} = h_{re}$, $h_{21} = h_{fe}$, $h_{22} = h_{oe}$	• Series–parallel connected  $\mathbf{h = h_A + h_B}$
Inverse hybrid parameters g		• Parallel-series connected  $\mathbf{g = g_A + g_B}$
Transmission parameters T	• Ideal transformer circuits	• Cascade connected  $\mathbf{T = T_A T_B}$
Inverse transmission parameters T'		• Cascade connected  $\mathbf{T' = T'_B T'_A}$

Port: Two terminals of a network where the current entering one terminal equals the current leaving the other terminal.

Short-circuit parameters: Two-port parameters, such as *y* parameters, evaluated by short-circuiting a port.

Superposition: In linear networks, a method of calculating the value of a dependent variable. First, the value of the dependent variable produced by each independent variable acting alone is calculated. Then, these values are summed to obtain the total value of the dependent variable.

Thévenin's theorem: At a pair of terminals a linear electrical network can be replaced with a voltage source in series with an impedance. The voltage source is equal to the voltage at the terminals when the terminals are open-circuited. The impedance is equal to the impedance at the terminals with all independent sources set equal to zero.

***T* parameters:** See transmission parameters.

***T′* parameters:** See transmission parameters.

Transmission (inverse transmission) parameters: That set of two-port parameters, such as the $T(T')$ parameters, where the dependent variables are the input (output) variables of the network and the independent variables are the output (input) variables. The parenthetical expressions refer to the inverse transmission parameters. See Table 3.2 for specific definitions.

Two-port networks: Networks that are modeled by specifying two ports, typically input and output ports.

Two-port parameters: A set of four constants, Laplace transforms, or sinusoidal steady-state functions used in the equations that describe a linear two-port network. Some examples are *z, y, h, g, T,* and T' parameters.

***y* Parameters:** See admittance parameters.

***z* Parameters:** See impedance parameters.

Related Topic

3.3 Network Theorems

References

H. Ruston and J. Bordogna, "Two-port networks," in *Electric Networks: Functions, Filters, Analysis,* New York: McGraw-Hill, 1966, chap. 4, pp. 244–266.

M. E. Van Valkenburg, "Two-port parameters," in *Network Analysis*, 3rd ed. Chap. 11, Englewood Cliffs, N.J.: Prentice-Hall, 1974, pp. 325–350.

Further Information

The following texts cover standard two-port parameters:

J. W. Nilsson, "Two-port circuits," in *Electric Circuits,* 4th ed., Reading Mass.: Addison-Wesley, 1995, chap. 21, pp. 755–786.

H. Ruston and J. Bordogna, "Two-port networks," in *Electric Networks: Functions, Filters, Analysis,* New York: McGraw-Hill, 1966, chap. 4, pp. 206–311.

The following texts have added coverage of scattering and image parameters:

H. Ruston and J. Bordogna, "Two-port networks," in *Electric networks: Functions, Filters, Analysis,* New York: McGraw-Hill, 1966, chap. 4, pp. 266–297.

S. Seshu and N. Balabanian, "Two-port networks," and "Image parameters and filter theory," in *Linear Network Analysis,* New York: Wiley, 1959, chaps. 8 and 11, pp. 291–342, 453–504.

The following texts show applications to electronic circuits:

F. H. Mitchell, Jr. and F. H. Mitchell, Sr., "Midrange AC amplifier design," in *Introduction to Electronics Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1992, chap. 7, pp. 335–384.

C. J. Savant, Jr., M. S. Roden, and G. L. Carpenter, "Bipolar transistors," "Design of bipolar junction transistor amplifiers," and "Field-effect transistor amplifiers," in *Electronic Design,* 2nd ed., Redwood City, Calif.: Benjamin/Cummings, 1991, chaps. 2, 3, and 4, pp. 69–212.

S. S. Sedra and K. C. Smith, "Frequency response" and "Feedback," in *Microelectronic Circuits,* 3rd ed., Philadelphia, Pa.: Saunders, 1991, chaps. 7 and 8, pp. 488–645.

Kerwin, W.J. “Passive Signal Processing”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

4

Passive Signal Processing

4.1 Introduction
Laplace Transform • Transfer Functions

4.2 Low-Pass Filter Functions
Thomson Functions • Chebyshev Functions

4.3 Low-Pass Filters
Introduction • Butterworth Filters • Thomson Filters • Chebyshev Filters

4.4 Filter Design
Scaling Laws and a Design Example • Transformation Rules, Passive Circuits

William J. Kerwin
University of Arizona

4.1 Introduction

This chapter will include detailed design information for passive RLC filters; including Butterworth, Thomson, and Chebyshev, both singly and doubly terminated. As the filter slope is increased in order to obtain greater rejection of frequencies beyond cut-off, the complexity and cost are increased and the response to a step input is worsened. In particular, the overshoot and the settling time are increased. The element values given are for normalized low pass configurations to 5th order. All higher order doubly-terminated Butterworth filter element values can be obtained using Takahasi's equation, and an example is included. In order to use this information in a practical filter these element values must be scaled. Scaling rules to denormalize in frequency and impedance are given with examples. Since all data is for low-pass filters the transformation rules to change from low-pass to high-pass and to band-pass filters are included with examples.

Laplace Transform

We will use the Laplace operator, $s = \sigma + j\omega$. Steady-state impedance is thus Ls and $1/Cs$, respectively, for an inductor (L) and a capacitor (C), and admittance is $1/Ls$ and Cs. In steady state $\sigma = 0$ and therefore $s = j\omega$.

Transfer Functions

We will consider only lumped, linear, constant, bilateral elements, and we will define the **transfer function** $T(s)$ as response over excitation.

$$T(s) = \frac{\text{signal output}}{\text{signal input}} = \frac{N(s)}{D(s)}$$

Adapted from *Instrumentation and Control: Fundamentals and Applications,* edited by Chester L. Nachtigal, pp. 487–497, copyright 1990, John Wiley and Sons, Inc. Reproduced by permission of John Wiley and Sons, Inc.

The roots of the numerator polynomial $N(s)$ are the zeros of the system, and the roots of the denominator $D(s)$ are the poles of the system (the points of infinite response). If we substitute $s = j\omega$ into $T(s)$ and separate the result into real and imaginary parts (numerator and denominator) we obtain

$$T(j\omega) = \frac{A_1 + jB_1}{A_2 + jB_2} \tag{4.1}$$

Then the magnitude of the function, $|T(j\omega)|$, is

$$|T(j\omega)| = \left(\frac{A_1^2 + B_1^2}{A_2^2 + B_2^2}\right)^{\frac{1}{2}} \tag{4.2}$$

and the phase $\underline{\overline{T(j\omega)}}$ is

$$\underline{\overline{T(j\omega)}} = \tan^{-1}\frac{B_1}{A_1} - \tan^{-1}\frac{B_2}{A_2} \tag{4.3}$$

Analysis

Although mesh or nodal analysis can always be used, since we will consider only ladder networks we will use a method commonly called *linearity*, or *working your way through*. The method starts at the output and assumes either 1 volt or 1 ampere as appropriate and uses Ohm's law and Kirchhoff's current law only.

Example 4.1. Analysis of the circuit of Fig. 4.1 for $V_o = 1$ Volt.

$$I_3 = \tfrac{3}{2}s;\quad V_1 = 1 + \left(\tfrac{3}{2}s\right)\left(\tfrac{4}{3}s\right) = 1 + 2s^2$$

$$I_2 = V_1\left(\tfrac{1}{2}s\right) = \tfrac{1}{2}s + s^3;\quad I_1 = I_2 + I_3$$

$$V_i = V_1 + I_1 = s^3 + 2s^2 + 2s + 1$$

$$T(s) = \frac{V_o}{V_i} = \frac{1}{s^3 + 2s^2 + 2s + 1}$$

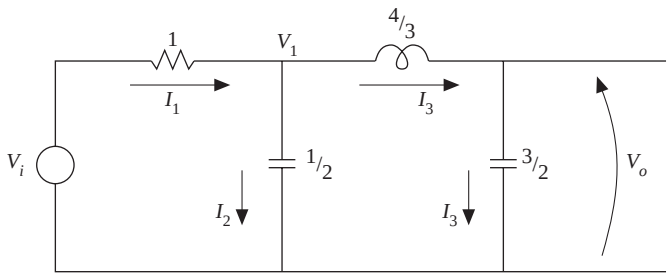


FIGURE 4.1 Singly terminated 3rd order low pass filter (Ω, H, F).

Example 4.2 Determine the magnitude and phase of T(s) in Example 4.1.

$$|T(s)| = \left| \frac{1}{s^3 + 2s^2 + 2s + 1} \right|_{s=j\omega}$$

$$|T(s)| = \sqrt{\frac{1}{\left(1 - 2\omega^2\right)^2 + \left(2\omega - \omega^3\right)^2}} = \frac{1}{\sqrt{\omega^6 + 1}}$$

$$\underline{T(s)} = \tan^{-1} 0 - \tan^{-1} \frac{2\omega - \omega^3}{1 - 2\omega^2} = -\tan^{-1} \frac{2\omega - \omega^3}{1 - 2\omega^2}$$

The values used for the circuit of Fig. 4.1 were normalized; that is, they are all near unity in ohms, henrys, and farads. These values simplify computation and, as we will see later, can easily be scaled to any desired set of actual element values. In addition, this circuit is low-pass because of the shunt capacitors and the series inductor. By low-pass we mean a circuit that passes the lower frequencies and attenuates higher frequencies. The cut-off frequency is the point at which the magnitude is 0.707 (–3 dB) of the dc level and is the dividing line between the **passband** and the **stopband**. In the above example we see that the magnitude of V_o/V_i at ω = 0 (dc) is 1.00 and that at ω = 1 rad/s we have

$$\left| T(j\omega) \right| = \left| \frac{1}{\sqrt{(\omega^6 + 1)}} \right|_{\omega=1\ rad/s} = 0.707 \tag{4.4}$$

and therefore this circuit has a cut-off frequency of 1 rad/s.

Thus, we see that the normalized element values used here give us a cut-off frequency of 1 rad/s.

4.2 Low-Pass Filter Functions[1]

The most common function in signal processing is the Butterworth. It is a function that has only poles (i.e., no finite zeros) and has the flattest magnitude possible in the passband. This function is also called **maximally flat magnitude (MFM)**. The derivation of this function is illustrated by taking a general all-pole function of third-order with a dc gain of 1 as follows:

$$T(s) = \frac{1}{as^3 + bs^2 + cs + 1} \tag{4.5}$$

The squared magnitude is

$$|T(j\omega)|^2 = \frac{1}{(1 - b\omega^2)^2 + (c\omega - a\omega^3)^2} \tag{4.6}$$

[1]Adapted from *Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.

or

$$|T(j\omega)|^2 = \frac{1}{a^2\omega^6 + (b^2 - 2ac)\omega^4 + (c^2 - 2b)\omega^2 + 1} \tag{4.7}$$

MFM requires that the coefficients of the numerator and the denominator match term by term (or be in the same ratio) except for the highest power.

Therefore

$$c^2 - 2b = 0; \qquad b^2 - 2ac = 0 \tag{4.8}$$

We will also impose a normalized cut-off (–3 dB) at $\omega = 1$ rad/s; that is,

$$|T(j\omega)|_{\omega=1} = \frac{1}{\sqrt{(a^2 + 1)}} = 0.707 \tag{4.9}$$

Thus, we find $a = 1$, then $b = 2$, $c = 2$ are solutions to the flat magnitude conditions of Eq. 4.8 and our third-order Butterworth function is

$$T(s) = \frac{1}{s^3 + 2s^2 + 2s + 1} \tag{4.10}$$

Table 4.1 gives the Butterworth denominator polynomials up to $n = 5$.

In general, for all Butterworth functions the normalized magnitude is

$$|T(j\omega)| = \frac{1}{\sqrt{(\omega^{2n} + 1)}} \tag{4.11}$$

Note that this is down 3 dB at $\omega = 1$ rad/s for all n.

This may, of course, be multiplied by any constant less than one for circuits whose dc gain is deliberately set to be less than one.

Example 4.3. A low-pass Butterworth filter is required whose cut-off frequency (–3 dB) is 3 kHz and in which the response must be down 40 dB at 12 kHz. Normalizing to a cut-off frequency of 1 rad/s, the –40-dB frequency is

$$\frac{12\text{ kHz}}{3\text{ kHz}} = 4\text{ rad/s}$$

thus

$$-40 = 20\log\frac{1}{\sqrt{4^{2n} + 1}}$$

therefore $n = 3.32$. Since n must be an integer, a fourth-order filter is required for this specification.

TABLE 4.1 Butterworth Polynomials

$s+1$
$s^2 + \sqrt{2}s + 1$
$s^3 + 2s^2 + 2s + 1$
$s^4 + 2.6131s^3 + 3.4142s^2 + 2.6131s + 1$
$s^5 + 3.2361s^4 + 5.2361s^3 + 5.2361s^2 + 3.2361s + a$

Source: Handbook of Measurement Science, edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.

There is an extremely important difference between the singly terminated (dc gain = 1) and the doubly terminated filters (dc gain = 0.5). As was shown by John Orchard, the sensitivity in the passband (ideally at maximum output) to all *L, C* components in an *L, C* filter with *equal* terminations is *zero.* This is true regardless of the circuit.

This, of course, means component tolerances and temperature coefficients are of much less importance in the equally terminated case. For this type of Butterworth low-pass filter (normalized to equal 1-Ω terminations), Takahasi has shown that the normalized element values are exactly given by

$$L,\ C = 2\ \sin\left(\frac{(2k-1)\pi}{2n}\right) \tag{4.12}$$

for any order *n,* where *k* is the *L* or *C* element from 1 to *n.*

Example 4.4. Design a normalized (ω_{-3dB} =1 rad/s) doubly terminated (i.e., source and load = 1 Ω) Butterworth low-pass filter of order 6; that is, $n = 6$.

The element values from Eq. (4.12) are

$$L_1 = 2\ \sin\frac{(2-1)\pi}{12} = 0.5176\ \text{H}$$
$$C_2 = 2\ \sin\frac{(4-1)\pi}{12} = 1.4141\ \text{F}$$
$$L_3 = 2\ \sin\frac{(6-1)\pi}{12} = 1.9319\ \text{H}$$

The values repeat for C_4, L_5, C_6 so that

$$C_4 = L_3,\ L_5 = C_2,\ C_6 = L_1$$

Thomson Functions

The Thomson function is one in which the time delay of the network is made maximally flat. This implies a linear phase characteristic since the steady-state time delay is the negative of the derivative of the phase. This function has excellent time domain characteristics and is used wherever excellent step response is required. These functions have very little overshoot to a step input and have far superior settling times compared to the Butterworth functions. The slope near cut-off is more gradual than the Butterworth. Table 4.2 gives the Thomson denominator polynomials. The numerator is a constant equal to the dc gain of the circuit multiplied by the denominator constant. The cut-off frequencies are *not* all 1 rad/s. They are given in Table 4.2.

TABLE 4.2 Thomson Polynomials

	ω_{-3dB} (rad/s)
$s + 1$	1.0000
$s^2 + 3s + 3$	1.3617
$s^3 + 6s^2 + 15s + 15$	1.7557
$s^4 + 10s^3 + 45s^2 + 105s + 105$	2.1139
$s^5 + 15s^4 + 105s^3 + 420s^2 + 945s + 945$	2.4274

Source: Handbook of Measurement Science, edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.

Chebyshev Functions

A second function defined in terms of magnitude, the Chebyshev, has an **equal ripple** character within the passband. The ripple is determined by **ϵ.**

TABLE 4.3 Chebyshev Polynomials

$s + \sinh \nu$
$s^2 + \left(\sqrt{2} \sinh \nu\right)s + \sinh^2 \nu + 1/2$
$\left(s + \sinh \nu\right)\left[s^2 + \left(\sinh \nu\right)s + \sinh^2 \nu + 3/4\right]$
$\left[s^2 + \left(0.75637 \sinh \nu\right)s + \sinh^2 \nu + 0.85355\right] \times \left[s^2 + \left(1.84776 \sinh \nu\right)s + \sinh^2 \nu + 0.14645\right]$
$\left(s + \sinh \nu\right)\left[s^2 + \left(0.61803 \sinh \nu\right)s + \sinh^2 \nu + 0.90451\right] \times \left[s^2 + \left(1.61803 \sinh \nu\right)s + \sinh^2 \nu + 0.34549\right]$

Source: Handbook of Measurement Science, edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.

$$\epsilon = \sqrt{(10^{A/10} - 1)} \tag{4.13}$$

where A = decibels of ripple; then for a given order n, we define v.

$$v = \frac{1}{n} \sinh^{-1}\left(\frac{1}{\epsilon}\right) \tag{4.14}$$

Table 4.3 gives denominator polynomials for the Chebyshev functions. In all cases, the cut-off frequency (defined as the end of the ripple) is 1 rad/s. The –3-dB frequency for the Chebyshev function is

$$\omega_{-3\text{dB}} = \cosh\left[\frac{\cosh^{-1}(1/\epsilon)}{n}\right] \tag{4.15}$$

The magnitude in the *stopband* ($\omega > 1$ rad/s) for the normalized filter is

$$\left|T(j\omega)\right|^2 = \frac{1}{1 + \epsilon^2 \cosh^2(n \cosh^{-1} \omega)} \tag{4.16}$$

for the singly terminated filter. For equal terminations the above magnitude is multiplied by one-half [1/4 in Eq. (4.16)].

Example 4.5. What order of singly terminated Chebyshev filter having 0.25-dB ripple (A) is required if the magnitude must be –60 dB at 15 kHz and the cut-off frequency (–0.25 dB) is to be 3 kHz? The normalized frequency for a magnitude of –60 dB is

$$\frac{15 \text{ kHz}}{3 \text{ kHz}} = 5 \text{ rad/s}$$

Thus, for a ripple of A = 0.25 dB, we have from Eq. (4.13)

$$\epsilon = \sqrt{(10^{A/10} - 1)} = 0.2434$$

and solving Eq. (4.16) for n with $\omega = 5$ rad/s and $\left|T(j\omega)\right|$ = –60 dB, we obtain n = 3.93. Therefore we must use n = 4 to meet these specifications.

4.3 Low-Pass Filters[1]

Introduction

Normalized element values are given here for both singly and doubly terminated filters. The source and load resistors are normalized to 1 Ω. Scaling rules will be given in Section 4.4 that will allow these values to be modified to any specified impedance value and to any cut-off frequency desired. In addition, we will cover the **transformation** of these **low-pass filters** to **high-pass** or **bandpass filters**.

Butterworth Filters

For n = 2, 3, 4, or 5, Fig. 4.2 gives the element values for the singly terminated filters and Fig. 4.3 gives the element values for the doubly terminated filters. All cut-off frequencies (–3 dB) are 1 rad/s.

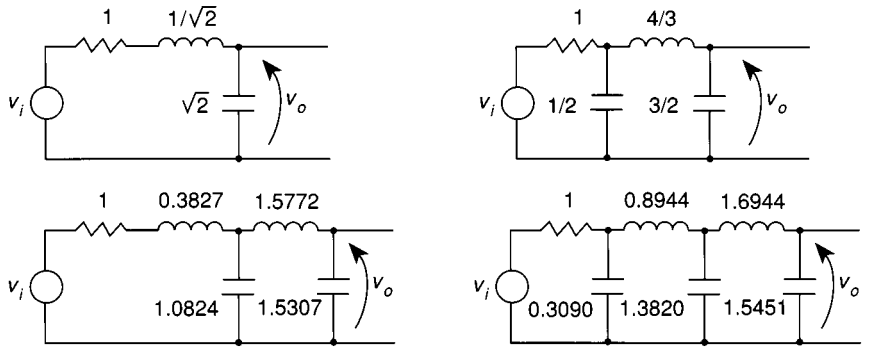


FIGURE 4.2 Singly terminated Butterworth filter element values (in Ω, H, F). *(Source: Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.)

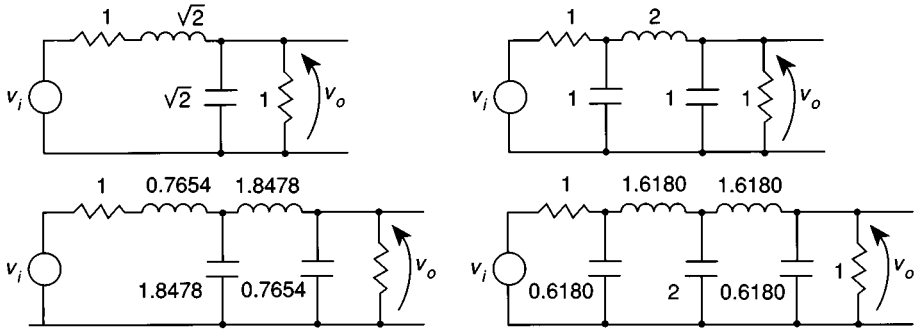


FIGURE 4.3 Doubly terminated Butterworth filter element values (in Ω, H, F). *(Source: Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.)

[1]Adapted from *Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.

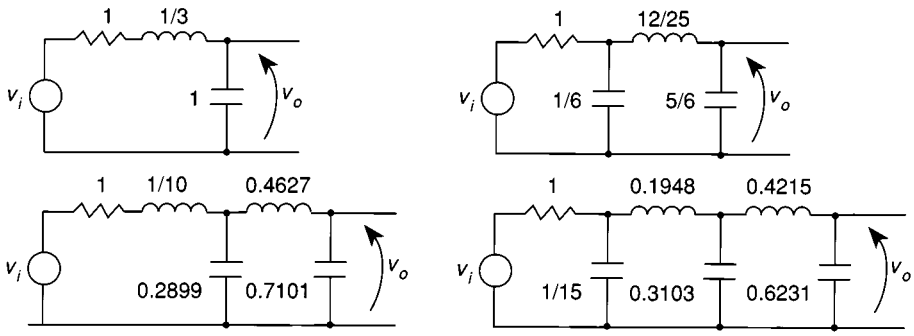


FIGURE 4.4 Singly terminated Thomson filter element values (in Ω, H, F). *(Source: Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.)

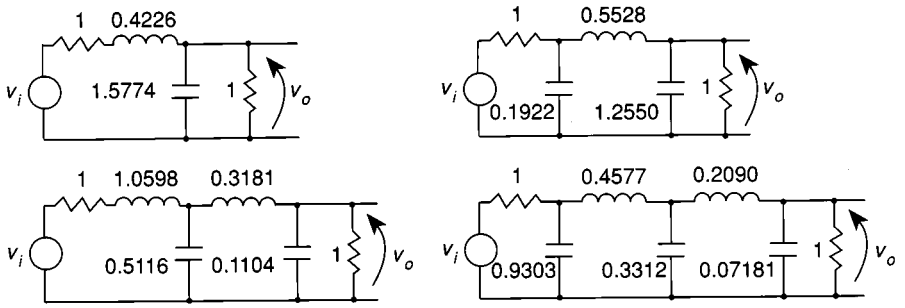


FIGURE 4.5 Doubly terminated Thomson filter element values (in Ω, H, F). (*Source: Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.)

Thomson Filters

Singly and doubly terminated Thomson filters of order $n = 2, 3, 4, 5$ are shown in Figs. 4.4 and 4.5. All time delays are 1 s. The cut-off frequencies are given in Table 4.2.

Chebyshev Filters

The amount of ripple can be specified as desired, so that only a selective sample can be given here. We will use 0.1 dB, 0.25 dB, and 0.5 dB. All cut-off frequencies (end of ripple for the Chebyshev function) are at 1 rad/s. Since the maximum power transfer condition precludes the existence of an equally terminated even-order filter, only odd orders are given for the doubly terminated case. Figure 4.6 gives the singly terminated Chebyshev filters for $n = 2, 3, 4,$ and 5 and Fig. 4.7 gives the doubly terminated Chebyshev filters for $n = 3$ and $n = 5$.

4.4 Filter Design

We now consider the steps necessary to convert normalized filters into actual filters by scaling both in frequency and in impedance. In addition, we will cover the transformation laws that convert low-pass filters to high-pass filters and low-pass to bandpass filters.

Scaling Laws and a Design Example

Since all data previously given are for normalized filters, it is necessary to use the scaling rules to design a low-pass filter for a specific signal processing application.

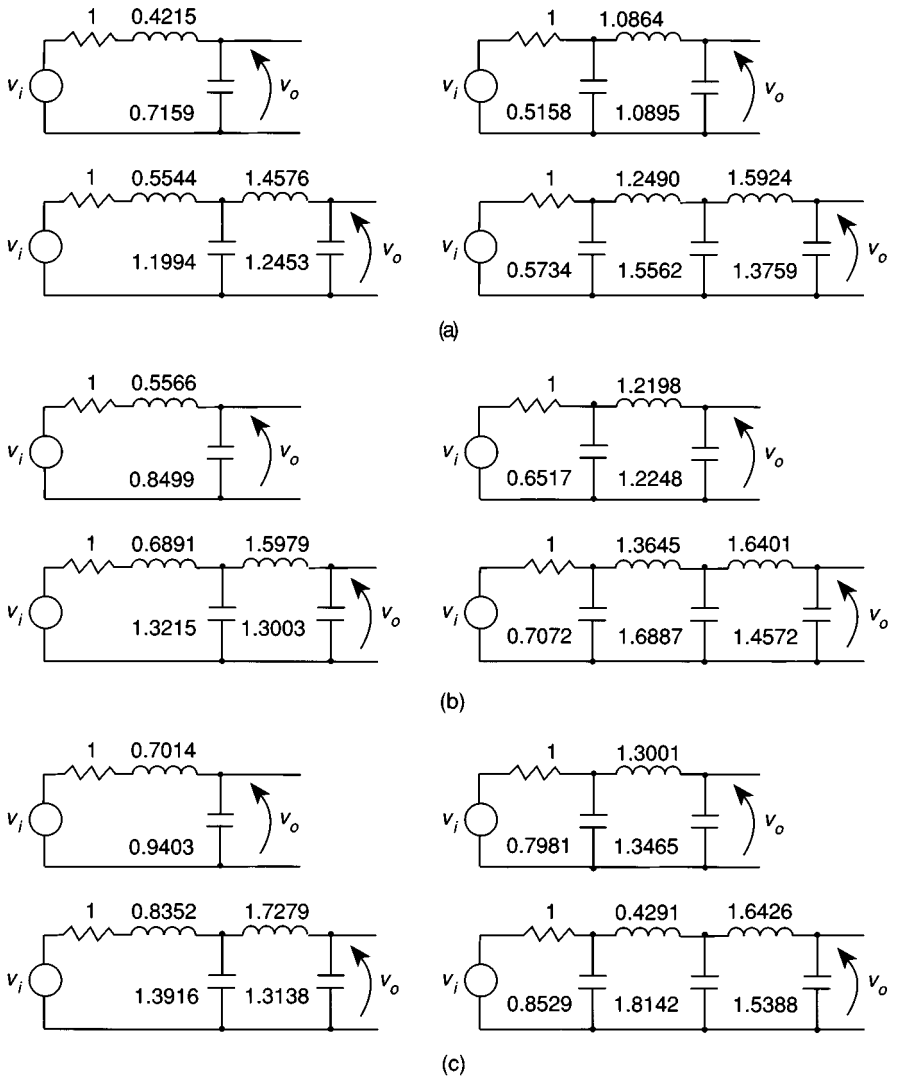


FIGURE 4.6 Singly terminated Chebyshev filter element values (in Ω, H, F): (a) 0.1-dB ripple; (b) 0.25-dB ripple; (c) 0.50-dB ripple. (*Source: Handbook of Measurement Science,* edited by Peter Sydenham, copyright 1982, John Wiley and Sons Limited. Reproduced by permission of John Wiley and Sons Limited.)

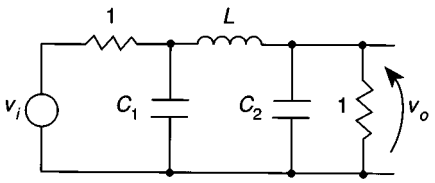
Rule 1. All impedances may be multiplied by any constant without affecting the transfer voltage ratio.

Rule 2. To modify the cut-off frequency, divide all inductors and capacitors by the ratio of the desired frequency to the normalized frequency.

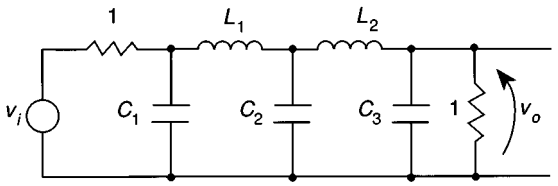
Example 4.6. Design a low-pass filter of MFMtype (Butterworth) to operate from a 600-Ω source into a 600-Ω load, with a cut-off frequency of 500 Hz. The filter must be at least 36 dB below the dc level at 2 kHz, that is, –42 dB (dc level is –6 dB).

Since 2 kHz is four times 500 Hz, it corresponds to $\omega = 4$ rad/s in the normalized filter. Thus at $\omega = 4$ rad/s we have

$$-42 \text{ dB} = 20 \log \frac{1}{2}\left[\frac{1}{\sqrt{4^{2n} + 1}}\right]$$



Ripple (dB)	C_1	L	C_2
0.10	1.0316	1.1474	1.0316
0.25	1.3034	1.1463	1.3034
0.50	1.5963	1.0967	1.5963



Ripple (dB)	C_1	L_1	C_2	L_2	C_3
0.10	1.1468	1.3712	1.9750	1.3712	1.1468
0.25	1.3824	1.3264	2.2091	1.3264	1.3824
0.50	1.7058	1.2296	2.5408	1.2296	1.7058

FIGURE 4.7 Doubly terminated Chebyshev filter element values (in Ω, H, F).

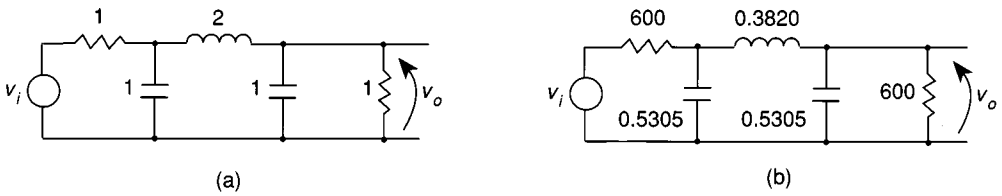


FIGURE 4.8 Third-order Butterworth low-pass filter: (a) normalized (in Ω, H, F); (b) scaled (in Ω, H, μF).

therefore, $n = 2.99$, so $n = 3$ must be chosen. The 1/2 is present because this is a doubly terminated (equal values) filter so that the dc gain is 1/2.

Thus a third-order, doubly terminated Butterworth filter is required. From Fig. 4.3 we obtain the normalized network shown in Fig. 4.8(a).

The **impedance scaling** factor is 600/1 = 600 and the **frequency scaling** factor is $2\pi 500/1 = 2\pi 500$: that is, the ratio of the desired radian cut-off frequency to the normalized cut-off frequency (1 rad/s). Note that the impedance scaling factor increases the size of the resistors and inductors, but reduces the size of the capacitors. The result is shown in Fig. 4.8(b).

Transformation Rules, Passive Circuits

All information given so far applies only to low-pass filters, yet we frequently need high-pass or bandpass filters in signal processing.

Low-Pass to High-Pass Transformation

To transform a low-pass filter to high-pass, we first scale it to a cut-off frequency of 1 rad/s if it is not already at 1 rad/s. This allows a simple frequency rotation about 1 rad/s of $s \rightarrow 1/s$. All L's become C's, all C's become L's, and all values reciprocate. The cut-off frequency does not change.

Example 4.7. Design a third-order, high-pass Butterworth filter to operate from a 600-Ω source to a 600-Ω load with a cut-off frequency of 500 Hz.

Starting with the normalized third-order low-pass filter of Fig. 4.3 for which $\omega_{-3} = 1$ rad/s, we reciprocate all elements and all values to obtain the filter shown in Fig. 4.9(a) for which $\omega_{-3} = 1$ rad/s.

Now we apply the scaling rules to raise all impedances to 600 Ω and the radian cut-off frequency to $2\pi 500$ rad/s as shown in Fig. 4.9(b).

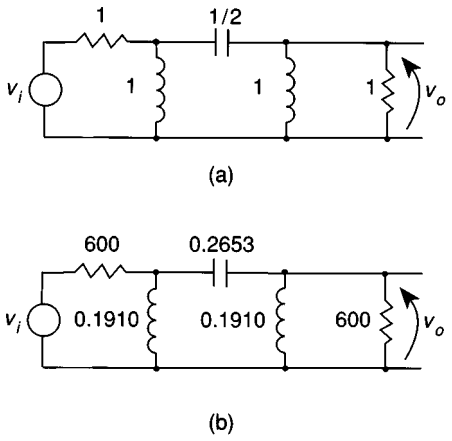


FIGURE 4.9 Third-order Butterworth high-pass filter: (a) normalized (in Ω, H, F); (b) scaled (in Ω, H, μF).

Low-Pass to Bandpass Transformation

To transform a low-pass filter to a bandpass filter we must first scale the low-pass filter so that the cut-off frequency is equal to the bandwidth of the normalized bandpass filter. The normalized center frequency of the bandpass filter is $\omega_0 = 1$ rad/s. Then we apply the transformation $s \rightarrow s + 1/s$. For an inductor

$$Z = Ls \text{ transforms to } Z = L\left(s + \frac{1}{s}\right)$$

For a capacitor

$$Y = Cs \text{ transforms to } Y = C\left(s + \frac{1}{s}\right)$$

The first step is then to determine the Q of the bandpass filter where

$$Q = \frac{f_0}{B} = \frac{\omega_0}{B_r}$$

(f_0 is the center frequency in Hz and B is the 3-dB bandwidth in Hz). Now we scale the low-pass filter to a cut-off frequency of $1/Q$ rad/s, then series tune every inductor, L, with a capacitor of value $1/L$ and parallel tune every capacitor, C, with an inductor of value $1/C$.

Example 4.8. Design a bandpass filter centered at 100 kHz having a 3-dB bandwidth of 10 kHz starting with a third-order Butterworth low-pass filter. The source and load resistors are each to be 600 Ω.

The Q required is

$$Q = \frac{100 \text{ kHz}}{10 \text{ kHz}} = 10, \text{ or } \frac{1}{Q} = 0.1$$

Scaling the normalized third-order low-pass filter of Fig. 4.10(a) to $\omega_{-3\text{dB}} = 1/Q = 0.1$ rad/s, we obtain the filter of Fig. 4.10(b).

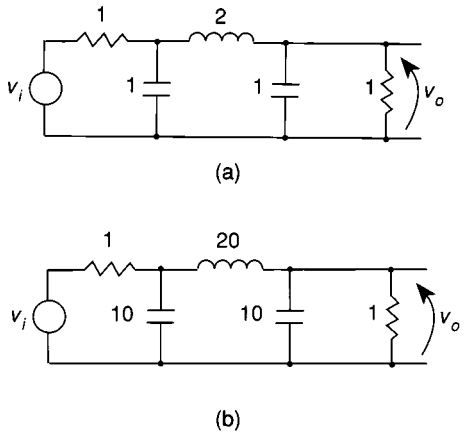


FIGURE 4.10 Third-order Butterworth low-pass filter: (a) normalized (in Ω, H, F); (b) scaled in (in Ω, H, F).

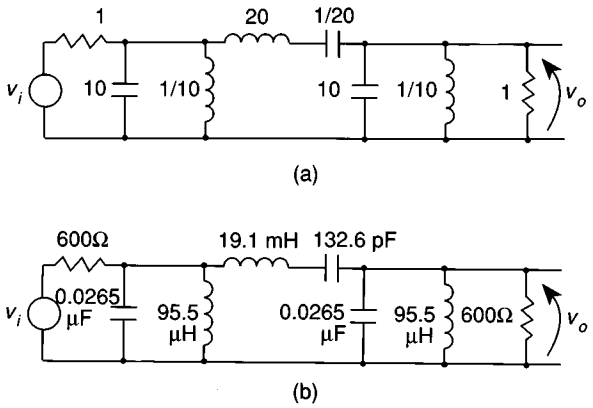


FIGURE 4.11 Sixth-order Butterworth bandpass filter ($Q = 10$): (a) normalized, $\omega_0 = 1$ rad/s (in Ω, H, F); (b) scaled.

Now converting to bandpass with $\omega_0 = 1$ rad/s, we obtain the normalized bandpass filter of Fig. 4.11(a). Next, scaling to an impedance of 600 Ω and to a center frequency of $f_0 = 100$ kHz ($\omega_0 = 2\pi 100$ k rad/s), we obtain the filter of Fig. 4.11(b).

Defining Terms

Bandpass filter: A filter whose passband extends from a finite lower cut-off frequency to a finite upper cut-off frequency.

Equal ripple: A frequency response function whose magnitude has equal maxima and equal minima in the passband.

Frequency scaling: The process of modifying a filter to change from a normalized set of element values to other usually more practical values by dividing all L, C elements by a constant equal to the ratio of the scaled (cut-off) frequency desired to the normalized cut-off frequency.

High-pass filter: A filter whose band extends from some finite cut-off frequency to infinity.

Impedance scaling: Modifying a filter circuit to change from a normalized set of element values to other usually more practical element values by multiplying all impedances by a constant equal to the ratio of the desired (scaled) impedance to the normalized impedance.

Low-pass filter: A filter whose passband extends from dc to some finite cut-off frequency.

Maximally flat magnitude (MFM) filter: A filter having a magnitude that is as flat as possible versus frequency while maintaining a monotonic characteristic.

Passband: A frequency region of signal transmission usually within 3 dB of the maximum transmission.

Stopband: The frequency response region in which the signal is attenuated, usually by more than 3 dB from the maximum transmission.

Transfer function: The Laplace transform of the response (output voltage) divided by the Laplace transform of the excitation (input voltage).

Transformation: The modification of a low-pass filter to convert it to an equivalent high-pass or bandpass filter.

Related Topics

6.1 Definitions and Properties • 10.3 The Ideal Linear-Phase Low-Pass Filter • 10.6 Butterworth Filters • 10.7 Chebyshev Filters

References

A. Budak, *Passive and Active Network Analysis and Synthesis,* Boston: Houghton Mifflin, 1974.

C. Nachtigal, Ed., *Instrumentation and Control: Fundamentals and Applications,* New York: John Wiley, 1990.

H.-J. Orchard, "Inductorless filters," *Electron. Lett.,* vol. 2, pp. 224–225, 1966.

P. Sydenham, Ed., *Handbook of Measurement Science,* Chichester, U.K.: John Wiley, 1982.

W. E. Thomson, "Maximally flat delay networks," *IRE Transactions,* vol. CT-6, p. 235, 1959.

L. Weinberg, *Network Analysis and Synthesis,* New York: McGraw-Hill, 1962.

L. Weinberg and P. Slepian, "Takahasi's results on Tchebycheff and Butterworth ladder networks," *IRE Transactions, Professional Group on Circuit Theory,* vol. CT-7, no. 2, pp. 88–101, 1960.

Hudgins, J.L., Bogart, Jr., T.F., Mayaram, K., Kennedy, M.P., Kolumbán, G. "Nonlinear Circuits"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

5

Nonlinear Circuits

Jerry L. Hudgins
University of South Carolina

Theodore F. Bogart, Jr.
University of Southern Mississippi

Kartikeya Mayaram
Washington State University

Michael Peter Kennedy
University College Dublin

Géza Kolumbán
Technical University of Budapest

5.1 Diodes and Rectifiers
Diodes • Rectifiers

5.2 Limiters
Limiting Circuits • Precision Rectifying Circuits

5.3 Distortion
Harmonic Distortion • Power-Series Method • Differential-Error Method • Three-Point Method • Five-Point Method • Intermodulation Distortion • Triple-Beat Distortion • Cross Modulation • Compression and Intercept Points • Crossover Distortion • Failure-to-Follow Distortion • Frequency Distortion • Phase Distortion • Computer Simulation of Distortion Components

5.4 Communicating with Chaos
Elements of Chaotic Digital Communications Systems • Chaotic Digital Modulation Schemes • Low-Pass Equivalent Models for Chaotic Communications Systems • Multipath Performance of FM-DCSK

5.1 Diodes and Rectifiers

Jerry L. Hudgins

A **diode** generally refers to a two-terminal solid-state semiconductor device that presents a low impedance to current flow in one direction and a high impedance to current flow in the opposite direction. These properties allow the diode to be used as a one-way current valve in electronic circuits. *Rectifiers* are a class of circuits whose purpose is to convert ac waveforms (usually sinusoidal and with zero average value) into a waveform that has a significant non-zero average value (dc component). Simply stated, rectifiers are ac-to-dc energy converter circuits. Most rectifier circuits employ diodes as the principal elements in the energy conversion process; thus the almost inseparable notions of diodes and rectifiers. The general electrical characteristics of common diodes and some simple rectifier topologies incorporating diodes are discussed.

Diodes

Most diodes are made from a host crystal of silicon (Si) with appropriate impurity elements introduced to modify, in a controlled manner, the electrical characteristics of the device. These diodes are the typical ***pn*-junction** (or **bipolar**) devices used in electronic circuits. Another type is the **Schottky diode** (unipolar), produced by placing a metal layer directly onto the semiconductor [Schottky, 1938; Mott, 1938]. The metal-semiconductor interface serves the same function as the *pn*-junction in the common diode structure. Other semiconductor materials such as gallium-arsenide (GaAs) and silicon-carbide (SiC) are also in use for new and specialized applications of diodes. Detailed discussion of diode structures and the physics of their operation can be found in later paragraphs of this section.

The electrical circuit symbol for a bipolar diode is shown in Fig. 5.1. The polarities associated with the forward voltage drop for forward current flow are also included. Current or voltage opposite to the polarities indicated in Fig. 5.1 are considered to be negative values with respect to the diode conventions shown.

The characteristic curve shown in Fig. 5.2 is representative of the current-voltage dependencies of typical diodes. The diode conducts forward current with a small forward voltage drop across the device, simulating a closed switch. The relationship between the forward current and forward voltage is approximately given by the Shockley diode equation [Shockley, 1949]:

$$i_D = I_s\left[\exp\left(\frac{qV_D}{nkT}\right) - 1\right] \qquad (5.1)$$

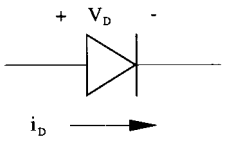


FIGURE 5.1 Circuit symbol for a bipolar diode indicating the polarity associated with the forward voltage and current directions.

where I_s is the leakage current through the diode, q is the electronic charge, n is a correction factor, k is Boltzmann's constant, and T is the temperature of the semiconductor. Around the knee of the curve in Fig. 5.2 is a positive voltage that is termed the turn-on or sometimes the threshold voltage for the diode. This value is an approximate voltage above which the diode is considered turned "on" and can be modeled to first degree as a closed switch with constant forward drop. Below the threshold voltage value the diode is considered weakly conducting and approximated as an open switch. The exponential relationship shown in Eq. (5.1) means that the diode forward current can change by orders of magnitude before there is a large change in diode voltage, thus providing the simple circuit model during conduction. The nonlinear relationship of Eq. (5.1) also provides a means of frequency mixing for applications in modulation circuits.

Reverse voltage applied to the diode causes a small leakage current (negative according to the sign convention) to flow that is typically orders of magnitude lower than current in the forward direction. The diode can withstand reverse voltages up to a limit determined by its physical construction and the semiconductor material used. Beyond this value the reverse voltage imparts enough energy to the charge carriers to cause large increases in current. The mechanisms by which this current increase occurs are impact ionization (avalanche) [McKay, 1954] and a tunneling phenomenon (Zener breakdown) [Moll, 1964]. Avalanche breakdown results in large power dissipation in the diode, is generally destructive, and should be avoided at all times. Both breakdown regions are superimposed in Fig. 5.2 for comparison of their effects on the shape of the diode characteristic curve. Avalanche breakdown occurs for reverse applied voltages in the range of volts to kilovolts depending on the exact design of the diode. Zener breakdown occurs at much lower voltages than the avalanche mechanism. Diodes specifically designed to operate in the Zener breakdown mode are used extensively as voltage regulators in regulator integrated circuits and as discrete components in large regulated power supplies.

During forward conduction the power loss in the diode can become excessive for large current flow. Schottky diodes have an inherently lower turn-on voltage than *pn*-junction diodes and are therefore more desirable in applications where the energy losses in the diodes are significant (such as output rectifiers in switching power supplies). Other considerations such as recovery characteristics from forward conduction to reverse blocking

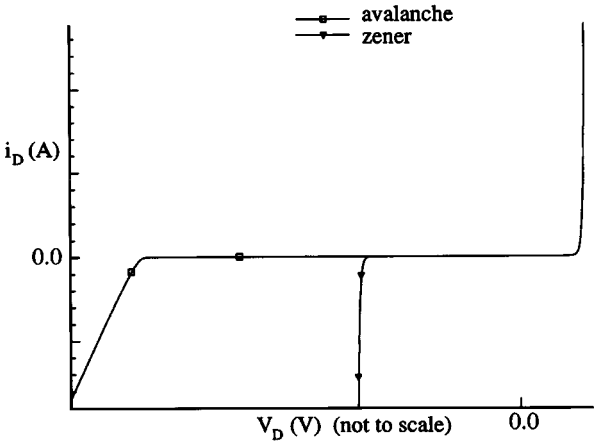


FIGURE 5.2 A typical diode dc characteristic curve showing the current dependence on voltage.

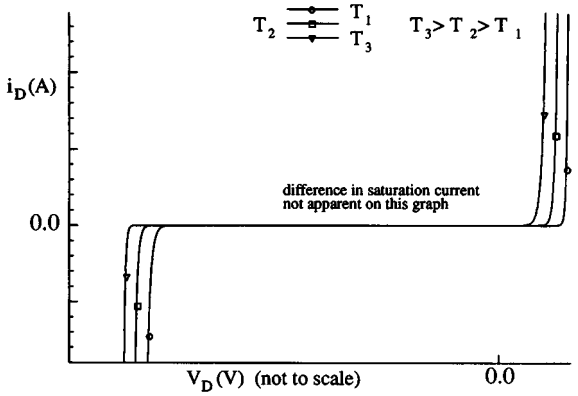


FIGURE 5.3 The effects of temperature variations on the forward voltage drop and the avalanche breakdown voltage in a bipolar diode.

may also make one diode type more desirable than another. Schottky diodes conduct current with one type of charge carrier and are therefore inherently faster to turn off than bipolar diodes. However, one of the limitations of Schottky diodes is their excessive forward voltage drop when designed to support reverse biases above about 200 V. Therefore, high-voltage diodes are the *pn*-junction type.

The effects due to an increase in the temperature in a bipolar diode are many. The forward voltage drop during conduction will decrease over a large current range, the reverse leakage current will increase, and the reverse avalanche breakdown voltage (V_{BD}) will increase as the device temperature climbs. A family of static characteristic curves highlighting these effects is shown in Fig. 5.3 where $T_3 > T_2 > T_1$. In addition, a major effect on the switching characteristic is the increase in the reverse recovery time during turn-off. Some of the key parameters to be aware of when choosing a diode are its repetitive peak inverse voltage rating, V_{RRM} (relates to the avalanche breakdown value), the peak forward surge current rating, I_{FSM} (relates to the maximum allowable transient heating in the device), the average or rms current rating, I_O (relates to the steady-state heating in the device), and the reverse recovery time, t_{rr} (relates to the switching speed of the device).

Rectifiers

This section discusses some simple **uncontrolled rectifier** circuits that are commonly encountered. The term *uncontrolled* refers to the absence of any control signal necessary to operate the primary switching elements (diodes) in the rectifier circuit. The discussion of controlled rectifier circuits, and the controlled switches themselves, is more appropriate in the context of power electronics applications [Hoft, 1986]. Rectifiers are the fundamental building block in dc power supplies of all types and in dc power transmission used by some electric utilities.

A single-phase full-wave rectifier circuit with the accompanying input and output voltage waveforms is shown in Fig. 5.4. This topology makes use of a center-tapped transformer with each diode conducting on opposite half-cycles of the input voltage. The forward drop across the diodes is ignored on the output graph, which is a valid approximation if the peak voltages of the input and output are large compared to 1 V. The circuit changes a sinusoidal waveform with no dc component (zero average value) to one with a dc component of $2V_{peak}/\pi$. The rms value of the output is $0.707V_{peak}$.

The dc value can be increased further by adding a low-pass filter in cascade with the output. The usual form of this filter is a shunt capacitor or an LC filter as shown in Fig. 5.5. The resonant frequency of the LC filter should be lower than the fundamental frequency of the rectifier output for effective performance. The ac portion of the output signal is reduced while the dc and rms values are increased by adding the filter. The remaining ac portion of the output is called the **ripple.** Though somewhat confusing, the transformer, diodes, and filter are often collectively called the rectifier circuit.

Another circuit topology commonly encountered is the bridge rectifier. Figure 5.6 illustrates single- and three-phase versions of the circuit. In the single-phase circuit diodes D1 and D4 conduct on the positive half-cycle of the input while D2 and D3 conduct on the negative half-cycle of the input. Alternate pairs of diodes conduct in the three-phase circuit depending on the relative amplitude of the source signals.

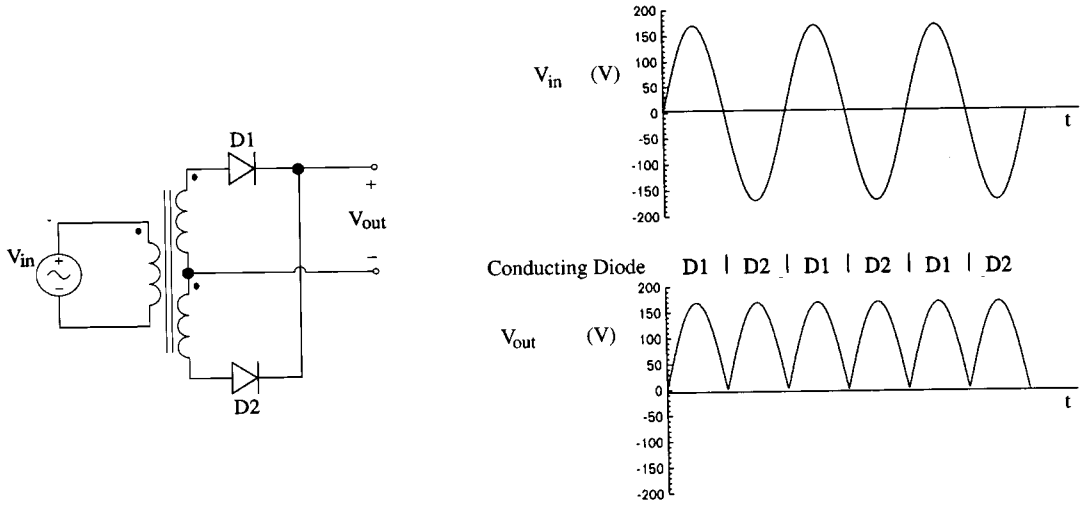


FIGURE 5.4 A single-phase full-wave rectifier circuit using a center-tapped transformer with the associated input and output waveforms.

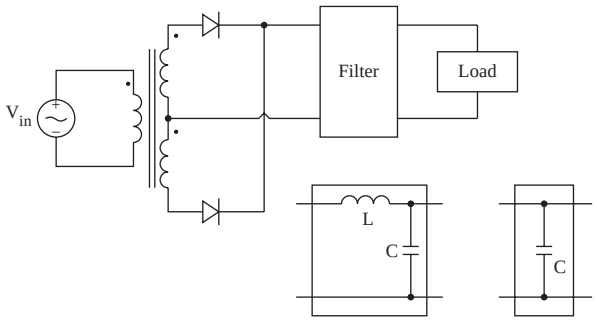


FIGURE 5.5 A single-phase full-wave rectifier with the addition of an output filter.

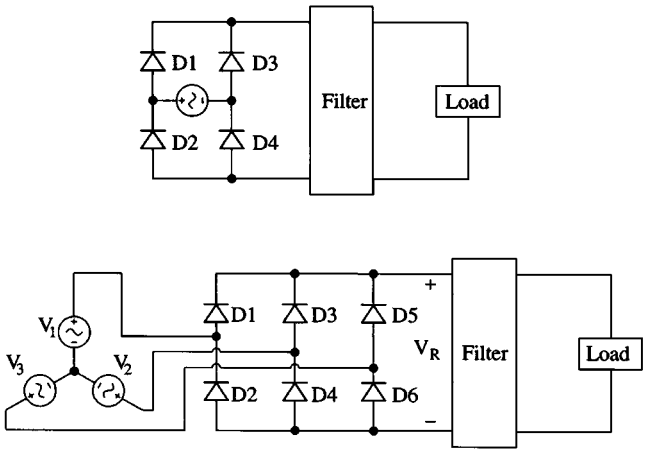


FIGURE 5.6 Single- and three-phase bridge rectifier circuits.

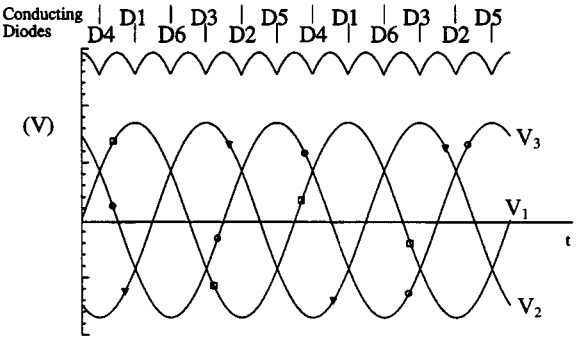


FIGURE 5.7 Three-phase rectifier output compared to the input signals. The input signals as well as the conducting diode labels are those referenced to Fig. 5.6.

The three-phase inputs with the associated rectifier output voltage are shown in Fig. 5.7 as they would appear without the low-pass filter section. The three-phase bridge rectifier has a reduced ripple content of 4% as compared to a ripple content of 47% in the single-phase bridge rectifier [Milnes, 1980]. The corresponding diodes that conduct are also shown at the top of the figure. This output waveform assumes a purely resistive load connected as shown in Fig. 5.6. Most loads (motors, transformers, etc.) and many sources (power grid) include some inductance, and in fact may be dominated by inductive properties. This causes phase shifts between the input and output waveforms. The rectifier output may thus vary in shape and phase considerably from that shown in Fig. 5.7 [Kassakian et al., 1991]. When other types of switches are used in these circuits the inductive elements can induce large voltages that may damage sensitive or expensive components. Diodes are used regularly in such circuits to shunt current and clamp induced voltages at low levels to protect expensive components such as electronic switches.

One variation of the typical rectifier is the Cockroft-Walton circuit used to obtain high voltages without the necessity of providing a high-voltage transformer. The circuit in Fig. 5.8 multiplies the peak secondary voltage by a factor of six. The steady-state voltage level at each filter capacitor node is shown in the figure. Adding additional stages increases the load voltage further. As in other rectifier circuits, the value of the capacitors will determine the amount of ripple in the output waveform for given load resistance values. In general, the capacitors in a lower voltage stage should be larger than in the next highest voltage stage.

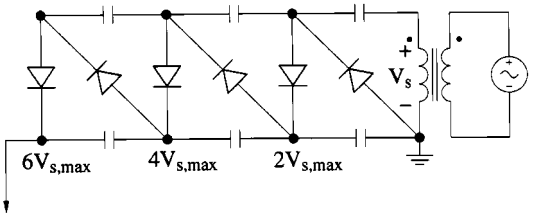


FIGURE 5.8 Cockroft-Walton circuit used for voltage multiplication.

Defining Terms

Bipolar device: Semiconductor electronic device that uses positive and negative charge carriers to conduct electric current.

Diode: Two-terminal solid-state semiconductor device that presents a low impedance to current flow in one direction and a high impedance to current flow in the opposite direction.

***pn*-junction:** Metallurgical interface of two regions in a semiconductor where one region contains impurity elements that create equivalent positive charge carriers (*p*-type) and the other semiconductor region contains impurities that create negative charge carriers (*n*-type).

Ripple: The ac (time-varying) portion of the output signal from a rectifier circuit.

Schottky diode: A diode formed by placing a metal layer directly onto a unipolar semiconductor substrate.

Uncontrolled rectifier: A rectifier circuit employing switches that do not require control signals to operate them in their "on" or "off" states.

Related Topics

22.2 Diodes • 30.1 Power Semiconductor Devices

References

R.G. Hoft, *Semiconductor Power Electronics,* New York: Van Nostrand Reinhold, 1986.
J.G. Kassakian, M.F. Schlecht, and G.C. Verghese, *Principles of Power Electronics,* Reading, Mass.: Addison-Wesley, 1991.
K.G. McKay, "Avalanche breakdown in silicon," *Physical Review,* vol. 94, p. 877, 1954.
A.G. Milnes, *Semiconductor Devices and Integrated Electronics,* New York: Van Nostrand Reinhold, 1980.
J.L. Moll, *Physics of Semiconductors,* New York: McGraw-Hill, 1964.
N.F. Mott, "Note on the contact between a metal and an insulator or semiconductor," *Proc. Cambridge Philos. Soc.,* vol. 34, p. 568, 1938.
W. Schottky, "Halbleitertheorie der Sperrschicht," *Naturwissenschaften,* vol. 26, p. 843, 1938.
W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," *Bell System Tech. J.,* vol. 28, p. 435, 1949.

Further Information

A good introduction to solid-state electronic devices with a minimum of mathematics and physics is *Solid State Electronic Devices,* 3rd edition, by B.G. Streetman, Prentice-Hall, 1989. A rigorous and more detailed discussion is provided in *Physics of Semiconductor Devices,* 2nd edition, by S.M. Sze, John Wiley & Sons, 1981. Both of these books discuss many specialized diode structures as well as other semiconductor devices. Advanced material on the most recent developments in semiconductor devices, including diodes, can be found in technical journals such as the *IEEE Transactions on Electron Devices, Solid State Electronics,* and *Journal of Applied Physics.* A good summary of advanced rectifier topologies and characteristics is given in *Basic Principles of Power Electronics* by K. Heumann, Springer-Verlag, 1986. Advanced material on rectifier designs as well as other power electronics circuits can be found in *IEEE Transactions on Power Electronics, IEEE Transactions on Industry Applications,* and the *EPE Journal.* Two good industry magazines that cover power devices such as diodes and power converter circuitry are *Power Control and Intelligent Motion* (PCIM) and *Power Technics.*

5.2 Limiters[1]

Theodore F. Bogart, Jr.

Limiters are named for their ability to limit voltage excursions at the output of a circuit whose input may undergo unrestricted variations. They are also called *clipping circuits* because waveforms having rounded peaks that exceed the limit(s) imposed by such circuits appear, after limiting, to have their peaks flattened, or "clipped" off. **Limiters** may be designed to clip positive voltages at a certain level, negative voltages at a different level, or to do both. The simplest types consist simply of diodes and dc voltage sources, while more elaborate designs incorporate operational amplifiers.

Limiting Circuits

Figure 5.9 shows how the transfer characteristics of limiting circuits reflect the fact that outputs are clipped at certain levels. In each of the examples shown, note that the characteristic becomes horizontal at the output level where clipping occurs. The horizontal line means that the output remains constant regardless of the input level in that region. Outside of the clipping region, the transfer characteristic is simply a line whose slope equals

[1]Excerpted from T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio:Macmillan/Merrill, 1993, pp. 689–697. With permission.

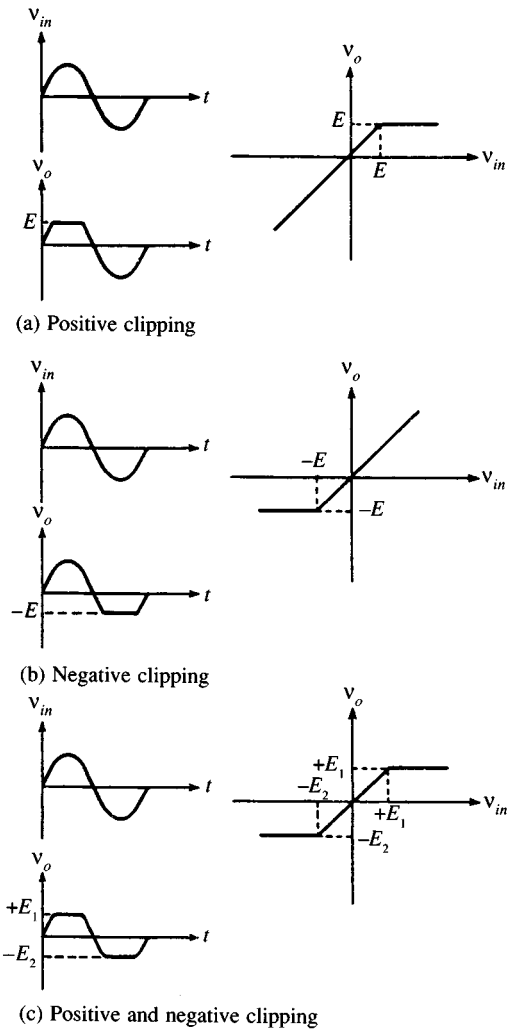


FIGURE 5.9 Waveforms and transfer characteristics of limiting circuits. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 676. With permission.)

the gain of the device. This is the region of linear operation. In these examples, the devices are assumed to have unity gain, so the slope of each line in the linear region is 1.

Figure 5.10 illustrates a somewhat different kind of limiting action. Instead of the positive or negative peaks being clipped, the output follows the input when the signal is above or below a certain level. The transfer characteristics show that linear operation occurs only when certain signal levels are reached and that the output remains constant below those levels. This form of limiting can also be thought of as a special case of that shown in Fig. 5.9. Imagine, for example, that the clipping level in Fig. 5.9(b) is raised to a positive value; then the result is the same as Fig. 5.10(a).

Limiting can be accomplished using **biased diodes.** Such circuits rely on the fact that diodes have very low impedances when they are forward biased and are essentially open circuits when reverse biased. If a certain point in a circuit, such as the output of an amplifier, is connected through a very small impedance to a *constant* voltage, then the voltage at the circuit point cannot differ significantly from the constant voltage. We say in this case that the point is *clamped* to the fixed voltage. An ideal, forward-biased diode is like a closed switch, so if it is connected between a point in a circuit and a fixed voltage source, the diode very effectively holds the point to the fixed voltage. Diodes can be connected in operational amplifier circuits, as well as other circuits,

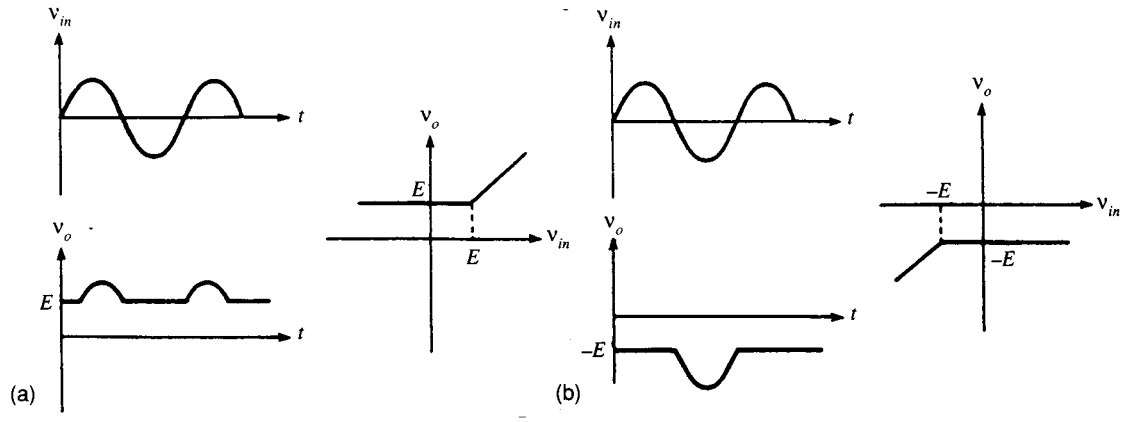


FIGURE 5.10 Another form of clipping. Compare with Fig. 5.9. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 690. With permission.)

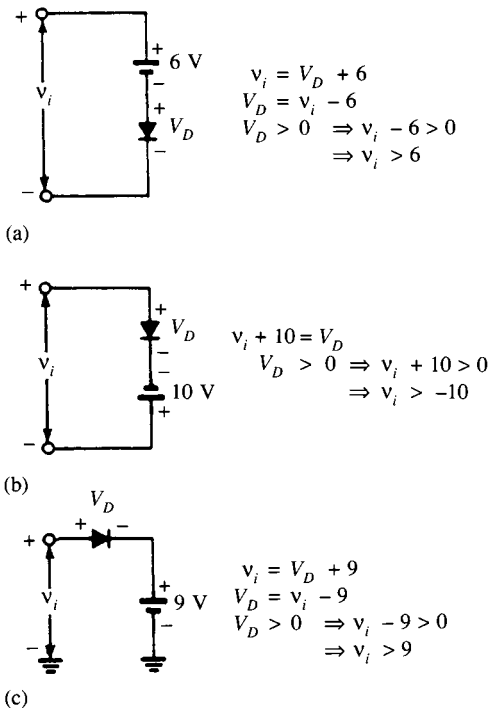


FIGURE 5.11 Examples of biased diodes and the signal voltages v_i required to forward bias them. (Ideal diodes are assumed.) In each case, we solve for the value of v_i that is necessary to make $V_D > 0$. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 691. With permission.)

in such a way that they become forward biased when a signal reaches a certain voltage. When the forward-biasing level is reached, the diode serves to hold the output to a fixed voltage and thereby establishes a clipping level.

A biased diode is simply a diode connected to a fixed voltage source. The value and polarity of the voltage source determine what value of total voltage across the combination is necessary to forward bias the diode. Figure 5.11 shows several examples. (In practice, a series resistor would be connected in each circuit to limit current flow when the diode is forward biased.) In each part of the figure, we can write Kirchhoff's voltage law

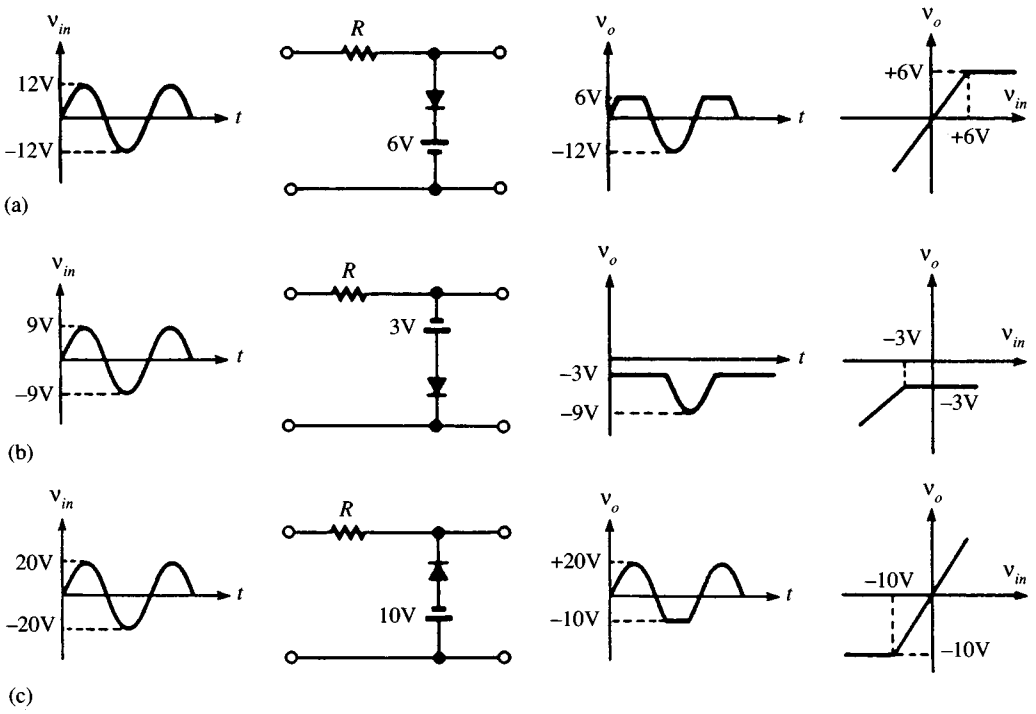


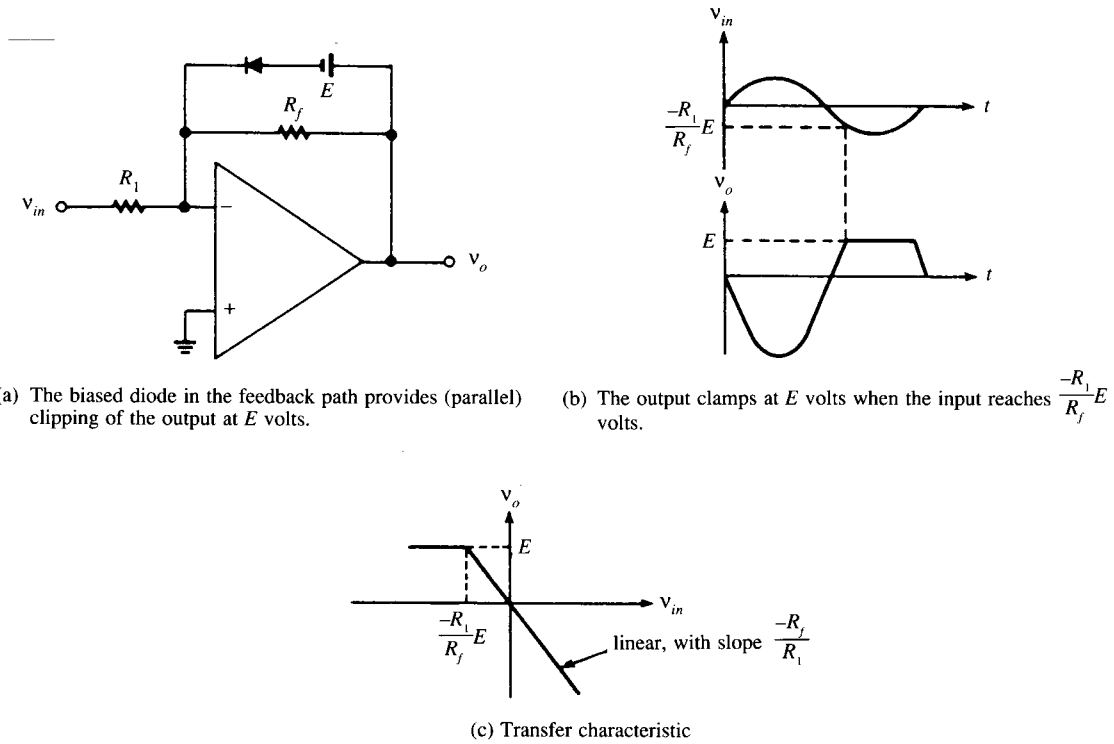
FIGURE 5.12 Examples of parallel clipping circuits. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 692. With permission.)

around the loop to determine the value of input voltage v_i that is necessary to forward bias the diode. Assuming that the diodes are ideal (neglecting their forward voltage drops), we determine the value v_i necessary to forward bias each diode by determining the value v_i necessary to make $v_D > 0$. When v_i reaches the voltage necessary to make $V_D > 0$, the diode becomes forward biased and the signal source is forced to, or held at, the dc source voltage. If the forward voltage drop across the diode is not neglected, the clipping level is found by determining the value of v_i necessary to make V_D greater than that forward drop (e.g., $V_D > 0.7$ V for a silicon diode).

Figure 5.12 shows three examples of clipping circuits using ideal biased diodes and the waveforms that result when each is driven by a sine-wave input. In each case, note that the output equals the dc source voltage when the input reaches the value necessary to forward bias the diode. Note also that the type of clipping we showed in Fig. 5.9 occurs when the fixed bias voltage tends to *reverse* bias the diode, and the type shown in Fig. 5.10 occurs when the fixed voltage tends to *forward* bias the diode. When the diode is reverse biased by the input signal, it is like an open circuit that disconnects the dc source, and the output follows the input. These circuits are called *parallel* clippers because the biased diode is in parallel with the output. Although the circuits behave the same way whether or not one side of the dc voltage source is connected to the common (low) side of the input and output, the connections shown in Fig. 5.12(a) and (c) are preferred to that in (b), because the latter uses a floating source.

Figure 5.13 shows a biased diode connected in the feedback path of an inverting operational amplifier. The diode is in parallel with the feedback resistor and forms a parallel clipping circuit like that shown in Fig. 5.12. In an operational amplifier circuit, $v^- \approx v^+$, and since $v^+ = 0$ V in this circuit, v^- is approximately 0 V (virtual ground). Thus, the voltage across R_f is the same as the output voltage v_o. Therefore, when the output voltage reaches the bias voltage E, the output is held at E volts. Figure 5.13(b) illustrates this fact for a sinusoidal input. So long as the diode is reverse biased, it acts like an open circuit and the amplifier behaves like a conventional inverting amplifier. Notice that output clipping occurs at *input* voltage $-(R_1/R_f)E$, since the amplifier inverts and has closed-loop gain magnitude R_f/R_1. The resulting transfer characteristic is shown in Fig. 5.13(c).

In practice, the biased diode shown in the feedback of Fig. 5.13(a) is often replaced by a *Zener* diode in series with a conventional diode. This arrangement eliminates the need for a floating voltage source. Zener diodes



(a) The biased diode in the feedback path provides (parallel) clipping of the output at E volts.

(b) The output clamps at E volts when the input reaches $\frac{-R_1}{R_f}E$ volts.

(c) Transfer characteristic

FIGURE 5.13 An operational amplifier limiting circuit. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio:Macmillan/Merrill, 1993, p. 693. With permission.)

are in many respects functionally equivalent to biased diodes. Figure 5.14 shows two operational amplifier clipping circuits using Zener diodes. The Zener diode conducts like a conventional diode when it is forward biased, so it is necessary to connect a reversed diode in series with it to prevent shorting of R_f. When the reverse voltage across the Zener diode reaches V_Z, the diode breaks down and conducts heavily, while maintaining an essentially constant voltage, V_Z, across it. Under those conditions, the total voltage across R_f, i.e., v_o, equals V_Z plus the forward drop, V_D, across the conventional diode.

Figure 5.15 shows *double-ended* limiting circuits, in which both positive and negative peaks of the output waveform are clipped. Figure 5.15(a) shows the conventional parallel clipping circuit and (b) shows how double-ended limiting is accomplished in an operational amplifier circuit. In each circuit, note that no more than one diode is forward biased at any given time and that both diodes are reverse biased for $-E_1 < v_o < E_2$, the linear region.

Figure 5.16 shows a double-ended limiting circuit using back-to-back Zener diodes. Operation is similar to that shown in Fig. 5.14, but no conventional diode is required. Note that diode D_1 is conducting in a forward direction when D_2 conducts in its reverse breakdown (Zener) region, while D_2 is forward biased when D_1 is conducting in its reverse breakdown region. Neither diode conducts when $-(V_{Z2} + 0.7) < v_o < (V_{Z1} + 0.7)$, which is the region of linear amplifier operation.

Precision Rectifying Circuits

A *rectifier* is a device that allows current to pass through it in one direction only. A diode can serve as a rectifier because it permits generous current flow in only one direction—the direction of forward bias. Rectification is the same as limiting at the 0-V level: all of the waveform below (or above) the zero-axis is eliminated. However, a diode rectifier has certain intervals of nonconduction and produces resulting "gaps" at the zero-crossing points of the output voltage, due to the fact that the input must overcome the diode drop (0.7 V for silicon) before

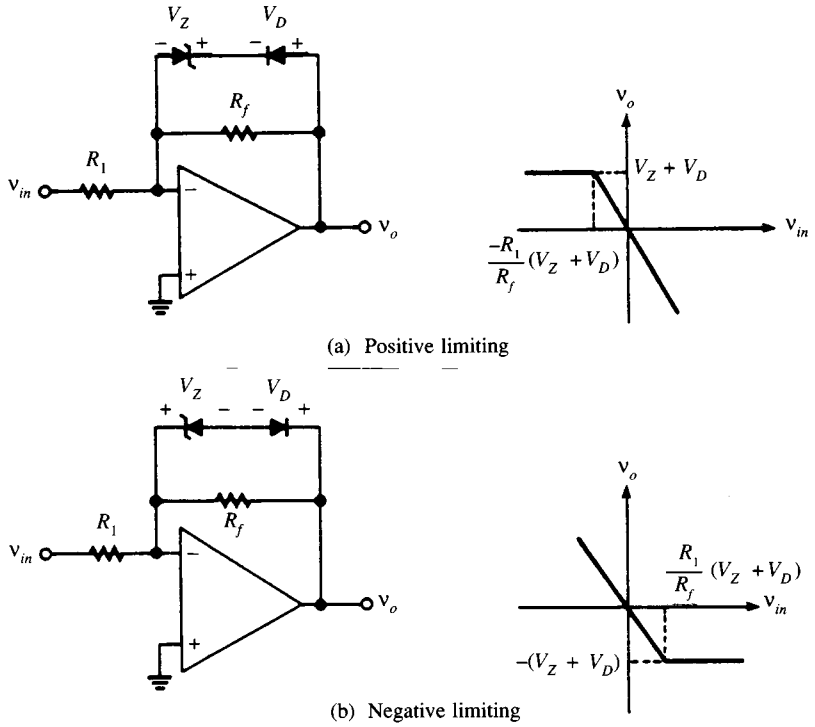


FIGURE 5.14 Operational amplifier limiting circuits using Zener diodes. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 694. With permission.)

conduction begins. In power-supply applications, where input voltages are quite large, these gaps are of no concern. However, in many other applications, especially in instrumentation, the 0.7-V drop can be a significant portion of the total input voltage swing and can seriously affect circuit performance. For example, most ac instruments rectify ac inputs so they can be measured by a device that responds to dc levels. It is obvious that small ac signals could not be measured if it were always necessary for them to reach 0.7 V before rectification could begin. For these applications, *precision* rectifiers are necessary.

Figure 5.17 shows one way to obtain precision rectification using an operational amplifier and a diode. The circuit is essentially a noninverting voltage follower (whose output follows, or duplicates, its input) when the diode is forward biased. When v_{in} is positive, the output of the amplifier, v_o, is positive, the diode is forward biased, and a low-resistance path is established between v_o and v^-, as necessary for a voltage follower. The load voltage, v_L, then follows the positive variations of $v_{in} = v^+$. Note that even a very small positive value of v_{in} will cause this result, because of the large differential gain of the amplifier. That is, the large gain and the action of the feedback cause the usual result that $v^+ \approx v^-$. Note also that the drop across the diode does not appear in v_L.

When the input goes negative, v_o becomes negative, and the diode is reverse biased. This effectively opens the feedback loop, so v_L no longer follows v_{in}. The amplifier itself, now operating open-loop, is quickly driven to its maximum negative output, thus holding the diode well into reverse bias.

Another precision rectifier circuit is shown in Fig. 5.18. In this circuit, the load voltage is an amplified and inverted version of the *negative* variations in the input signal, and is 0 when the input is positive. Also in contrast with the previous circuit, the amplifier in this rectifier is not driven to one of its output extremes. When v_{in} is negative, the amplifier output, v_o, is positive, so diode D_1 is reverse biased and diode D_2 is forward biased. D_1 is open and D_2 connects the amplifier output through R_f to v^-. Thus, the circuit behaves like an ordinary inverting amplifier with gain $-R_f/R_1$. The load voltage is an amplified and inverted (positive) version of the negative variations in v_{in}. When v_{in} becomes positive, v_o is negative, D_1 is forward biased, and D_2 is reverse biased. D_1 shorts the output v_o to v^-, which is held at virtual ground, so v_L is 0.

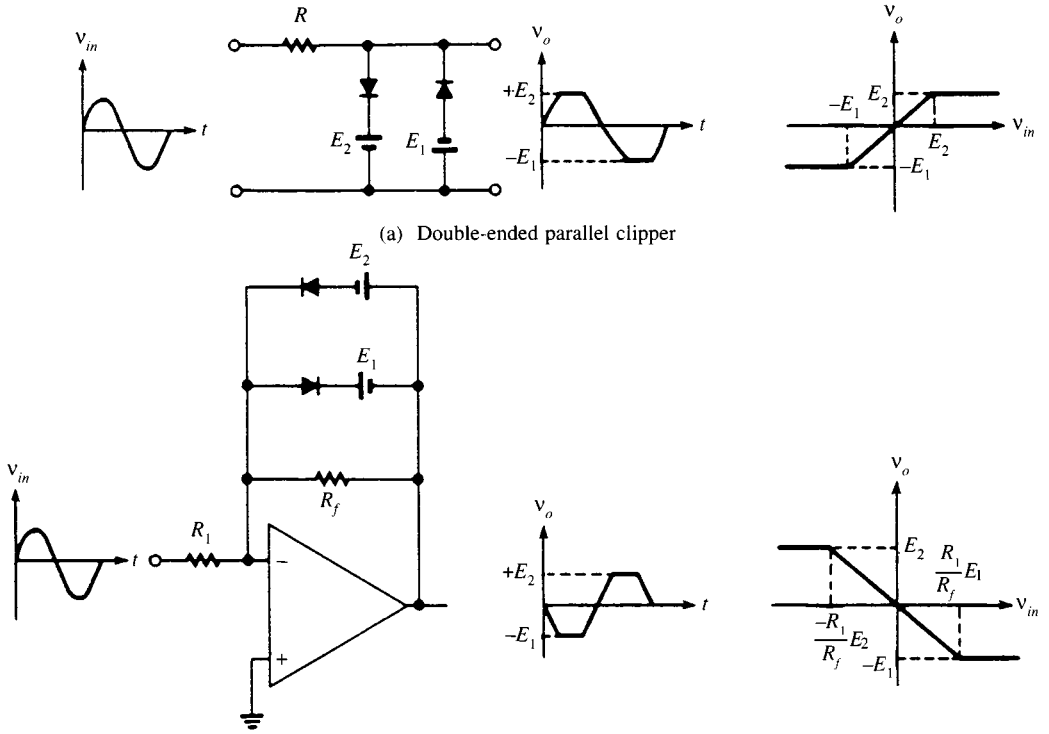


FIGURE 5.15 Double-ended clipping, or limiting. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits*, 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 695. With permission.)

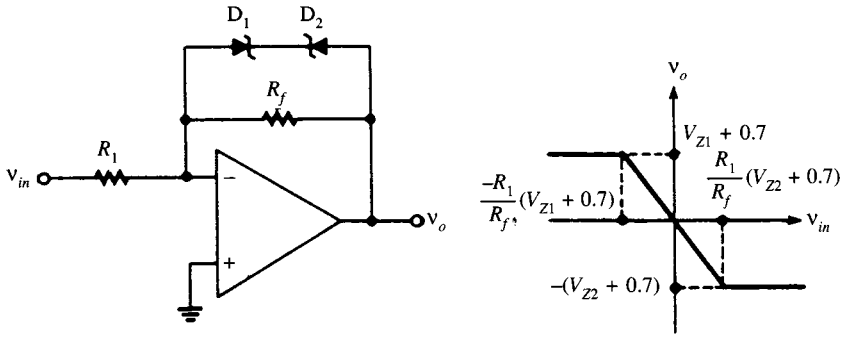


FIGURE 5.16 A double-ended limiting circuit using Zener diodes. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits*, 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 695. With permission.)

Defining Terms

Biased diode: A diode connected in series with a dc voltage source in order to establish a clipping level. Clipping occurs when the voltage across the combination is sufficient to forward bias the diode.

Limiter: A device or circuit that restricts voltage excursions to prescribed level(s). Also called a clipping circuit.

Related Topics

5.1 Diodes and Rectifiers • 27.1 Ideal and Practical Models

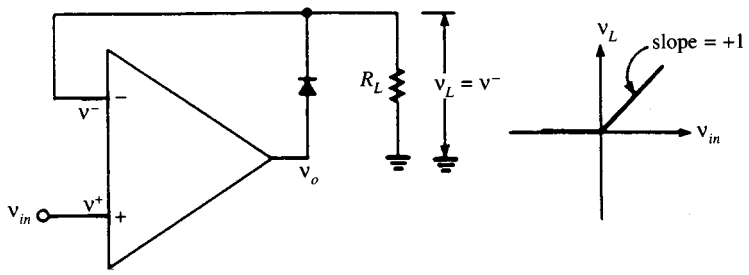


FIGURE 5.17 A precision rectifier. When v_{in} is positive, the diode is forward biased, and the amplifier behaves like a voltage follower, maintaining $v^+ \approx v^- = v_L$. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 696. With permission.)

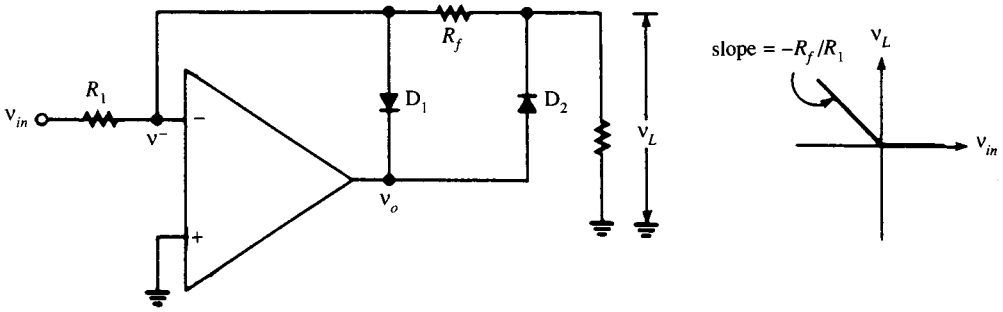


FIGURE 5.18 A precision rectifier circuit that amplifies and inverts the negative variations in the input voltage. (*Source:* T.F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993, p. 697. With permission.)

References

W.H. Baumgartner, *Pulse Fundamentals and Small-Scale Digital Circuits,* Reston, Va.: Reston Publishing, 1985.
T. F. Bogart, Jr., *Electronic Devices and Circuits,* 3rd ed., Columbus, Ohio: Macmillan/Merrill, 1993.
R.A. Gayakwad, *Op-Amps and Linear Integrated Circuit Technology,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.
A.S. Sedra and K.C. Smith, *Microelectronic Circuits,* New York: CBS College Publishing, 1982.
H. Zanger, *Semiconductor Devices and Circuits,* New York: John Wiley & Sons, 1984.

5.3 Distortion

Kartikeya Mayaram

The diode was introduced in the previous sections as a nonlinear device that is used in rectifiers and limiters. These are applications that depend on the nonlinear nature of the diode. Typical electronic systems are composed not only of diodes but also of other nonlinear devices such as transistors (Section III). In analog applications transistors are used to amplify weak signals (amplifiers) and to drive large loads (output stages). For such situations it is desirable that the output be an amplified true reproduction of the input signal; therefore, the transistors must operate as linear devices. However, the inherent nonlinearity of transistors results in an output which is a "distorted" version of the input.

The distortion due to a nonlinear device is illustrated in Fig. 5.19. For an input X the output is $Y = F(X)$ where F denotes the nonlinear transfer characteristics of the device; the dc operating point is given by X_0. Sinusoidal input signals of two different amplitudes are applied and the output responses corresponding to these inputs are also shown.

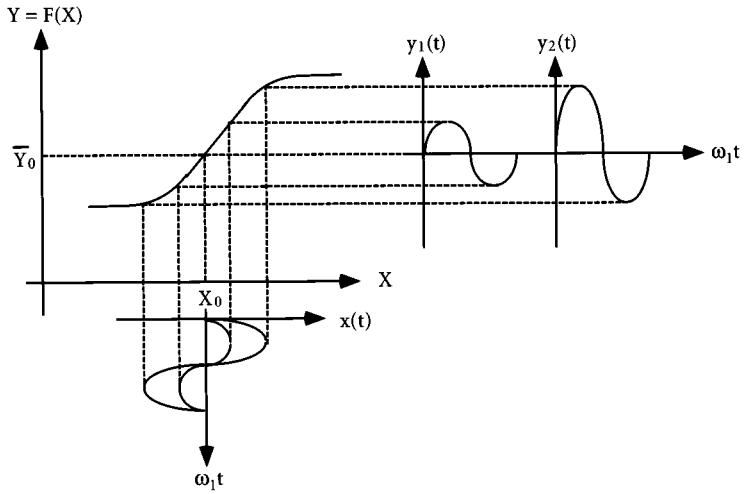


FIGURE 5.19 DC transfer characteristics of a nonlinear circuit and the input and output waveforms. For a large input amplitude the output is distorted.

For an input signal of small amplitude the output faithfully follows the input, whereas for large-amplitude signals the output is distorted; a flattening occurs at the negative peak value. The distortion in amplitude results in the output having frequency components that are integer multiples of the input frequency, *harmonics*, and this type of distortion is referred to as **harmonic distortion.**

The distortion level places a restriction on the amplitude of the input signal that can be applied to an electronic system. Therefore, it is essential to characterize the distortion in a circuit. In this section different types of distortion are defined and techniques for distortion calculation are presented. These techniques are applicable to simple circuit configurations. For larger circuits a circuit simulation program is invaluable.

Harmonic Distortion

When a sinusoidal signal of a single frequency is applied at the input of a nonlinear device or circuit, the resulting output contains frequency components that are integer multiples of the input signal. These harmonics are generated by the nonlinearity of the circuit and the *harmonic distortion* is measured by comparing the magnitudes of the harmonics with the fundamental component (input frequency) of the output.

Consider the input signal to be of the form:

$$x(t) = X_1 \cos \omega_1 t \tag{5.2}$$

where $f_1 = \omega_1/2\pi$ is the frequency and X_1 is the amplitude of the input signal. Let the output of the nonlinear circuit be

$$y(t) = Y_0 + Y_1 \cos \omega_1 t + Y_2 \cos 2\omega_1 t + Y_3 \cos 3\omega_1 t + \ldots \tag{5.3}$$

where Y_0 is the dc component of the output, Y_1 is the amplitude of the fundamental component, and Y_2, Y_3 are the amplitudes of the second and third harmonic components. The *second harmonic distortion factor* (HD_2), the *third harmonic distortion factor* (HD_3), and the *n*th *harmonic distortion factor* (HD_n) are defined as

$$HD_2 = \frac{|Y_2|}{|Y_1|} \tag{5.4}$$

$$\mathrm{HD}_3 = \frac{|Y_3|}{|Y_1|} \tag{5.5}$$

$$\mathrm{HD}_n = \frac{|Y_n|}{|Y_1|} \tag{5.6}$$

The **total harmonic distortion** (THD) of a waveform is defined to be the ratio of the rms (root-mean-square) value of the harmonics to the amplitude of the fundamental component.

$$\mathrm{THD} = \frac{\sqrt{Y_2^2 + Y_3^2 + \cdots + Y_n^2}}{|Y_1|} \tag{5.7}$$

THD can be expressed in terms of the individual **harmonic distortion factors**

$$\mathrm{THD} = \sqrt{\mathrm{HD}_2^2 + \mathrm{HD}_3^2 + \cdots + \mathrm{HD}_n^2} \tag{5.8}$$

Various methods for computing the harmonic distortion factors are described next.

Power-Series Method

In this method a truncated power-series expansion of the dc transfer characteristics of a nonlinear circuit is used. Therefore, the method is suitable only when energy storage effects in the nonlinear circuit are negligible and the input signal is small. In general, the input and output signals comprise both dc and time-varying components. For distortion calculation we are interested in the time-varying or incremental components around a quiescent[1] operating point. For the transfer characteristic of Fig. 5.19, denote the quiescent operating conditions by X_0 and $\bar{Y}_0$ and the incremental variables by $x(t)$ and $y(t)$, at the input and output, respectively. The output can be expressed as a function of the input using a series expansion

$$\bar{Y}_0 + y = F(X_0 + x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots \tag{5.9}$$

where $a_0 = \bar{Y}_0 = F(X_0)$ is the output at the dc operating point. The incremental output is

$$y = a_1 x + a_2 x^2 + a_3 x^3 + \ldots \tag{5.10}$$

Depending on the amplitude of the input signal, the series can be truncated at an appropriate term. Typically only the first few terms are used, which makes this technique applicable only to small input signals. For a pure sinusoidal input [Eq. (5.2)], the distortion in the output can be estimated by substituting for x in Eq. (5.10) and by use of trigonometric identities one can arrive at the form given by Eq. (5.3). For a series expansion that is truncated after the cubic term

[1]Defined as the operating condition when the input has no time-varying component.

$$
\begin{aligned}
Y_0 &= \frac{a_2 X_1^2}{2} \\
Y_1 &= a_1 X_1 + \frac{3a_3 X_1^3}{4} \cong a_1 X_1 \\
Y_2 &= \frac{a_2 X_1^2}{2} \\
Y_3 &= \frac{a_3 X_1^3}{4}
\end{aligned}
\tag{5.11}
$$

Notice that a dc term Y_0 is present in the output (produced by the even-powered terms) which results in a shift of the operating point of the circuit due to distortion. In addition, depending on the sign of a_3 there can be an *expansion* or *compression* of the fundamental component. The harmonic distortion factors (assuming $Y_1 = a_1 X_1$) are

$$
\begin{aligned}
\mathrm{HD}_2 &= \frac{|Y_2|}{|Y_1|} = \frac{1}{2}\left|\frac{a_2}{a_1} X_1\right| \\
\mathrm{HD}_3 &= \frac{|Y_3|}{|Y_1|} = \frac{1}{4}\left|\frac{a_3}{a_1} X_1^2\right|
\end{aligned}
\tag{5.12}
$$

As an example, choose as the transfer function $Y = F(X) = \exp(X)$; then, $a_1 = 1$, $a_2 = 1/2$, $a_3 = 1/6$. For an input signal amplitude of 0.1, $\mathrm{HD}_2 = 2.5\%$ and $\mathrm{HD}_3 = 0.04\%$.

Differential-Error Method

This technique is also applicable to nonlinear circuits in which energy storage effects can be neglected. The method is valuable for circuits that have small distortion levels and relies on one's ability to calculate the small-signal gain of the nonlinear function at the quiescent operating point and at the maximum and minimum excursions of the input signal. Again the power-series expansion provides the basis for developing this technique. The small-signal gain[1] at the quiescent state ($x = 0$) is a_1. At the extreme values of the input signal X_1 (positive peak) and $-X_1$ (negative peak) let the small-signal gains be a^+ and a^-, respectively. By defining two new parameters, the differential errors, E^+ and E^-, as

$$
E^+ = \frac{a^+ - a_1}{a_1} \qquad E^- = \frac{a^- - a_1}{a_1} \tag{5.13}
$$

the distortion factors are given by

$$
\begin{aligned}
\mathrm{HD}_2 &= \frac{E^+ - E^-}{8} \\
\mathrm{HD}_3 &= \frac{E^+ + E^-}{24}
\end{aligned}
\tag{5.14}
$$

[1]Small-signal gain $= dy/dx = a_1 + 2a_2 x + 3a_3 x^2 + \ldots$

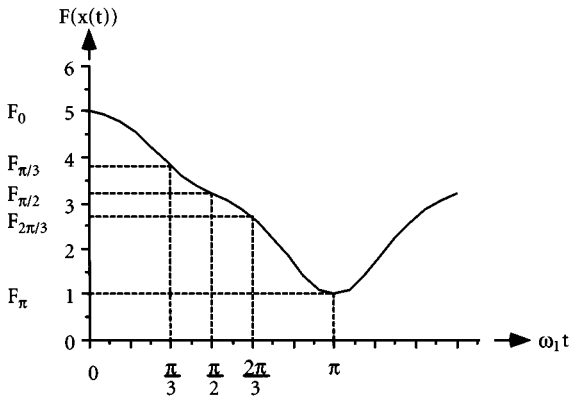


FIGURE 5.20 Output waveform from a nonlinear circuit.

The advantage of this method is that the transfer characteristics of a nonlinear circuit can be directly used; an explicit power-series expansion is not required. Both the power-series and the differential-error techniques cannot be applied when only the output waveform is known. In such a situation the distortion factors are calculated from the output signal waveform by a simplified Fourier analysis as described in the next section.

Three-Point Method

The three-point method is a simplified analysis applicable to small levels of distortion and can only be used to calculate HD_2. The output is written directly as a Fourier cosine series as in Eq. (5.3) where only terms up to the second harmonic are retained. The dc component includes the quiescent state and the contribution due to distortion that results in a shift of the dc operating point. The output waveform values at $\omega_1 t = 0$ (F_0), $\omega_1 t = \pi/2$ $(F_{\pi/2})$, $\omega_1 t = \pi$ (F_π), as shown in Fig. 5.20, are used to calculate Y_0, Y_1, and Y_2.

$$
\begin{aligned}
Y_0 &= \frac{F_0 + 2F_{\pi/2} + F_\pi}{4} \\
Y_1 &= \frac{F_0 - F_\pi}{2} \\
Y_2 &= \frac{F_0 - 2F_{\pi/2} + F_\pi}{4}
\end{aligned}
\tag{5.15}
$$

The second harmonic distortion is calculated from the definition. From Fig. 5.20, $F_0 = 5$, $F_{\pi/2} = 3.2$, $F_\pi = 1$, $Y_0 = 3.1$, $Y_1 = 2.0$, $Y_2 = -0.1$, and $HD_2 = 5.0\%$.

Five-Point Method

The five-point method is an extension of the above technique and allows calculation of third and fourth harmonic distortion factors. For distortion calculation the output is expressed as a Fourier cosine series with terms up to the fourth harmonic where the dc component includes the quiescent state and the shift due to distortion. The output waveform values at $\omega_1 t = 0$ (F_0), $\omega_1 t = \pi/3$ $(F_{\pi/3})$, $\omega_1 t = \pi/2$ $(F_{\pi/2})$, $\omega_1 t = 2\pi/3$ $(F_{2\pi/3})$, $\omega_1 t = \pi$ (F_π), as shown in Fig. 5.20, are used to calculate Y_0, Y_1, Y_2, Y_3, and Y_4.

$$
\begin{aligned}
Y_0 &= \frac{F_0 + 2F_{\pi/3} + 2F_{2\pi/3} + F_\pi}{6} \\
Y_1 &= \frac{F_0 + F_{\pi/3} - F_{2\pi/3} - F_\pi}{3} \\
Y_2 &= \frac{F_0 - 2F_{\pi/2} + F_\pi}{4} \\
Y_3 &= \frac{F_0 - 2F_{\pi/3} + 2F_{2\pi/3} - F_\pi}{6} \\
Y_4 &= \frac{F_0 - 4F_{\pi/3} + 6F_{\pi/2} - 4F_{2\pi/3} + F_\pi}{12}
\end{aligned} \tag{5.16}
$$

For $F_0 = 5$, $F_{\pi/3} = 3.8$, $F_{\pi/2} = 3.2$, $F_{2\pi/3} = 2.7$, $F_\pi = 1$, $Y_0 = 3.17$, $Y_1 = 1.7$, $Y_2 = -0.1$, $Y_3 = 0.3$, $Y_4 = -0.07$, and $HD_2 = 5.9\%$, $HD_3 = 17.6\%$. This particular method allows calculation of HD_3 and also gives a better estimate of HD_2. To obtain higher-order harmonics a detailed Fourier series analysis is required and for such applications a circuit simulator, such as SPICE, should be used.

Intermodulation Distortion

The previous sections have examined the effect of nonlinear device characteristics when a single-frequency sinusoidal signal is applied at the input. However, if there are two or more sinusoidal inputs, then the nonlinearity results in not only the fundamental and harmonics but also additional frequencies called the *beat frequencies* at the output. The distortion due to the components at the beat frequencies is called **intermodulation distortion.** To characterize this type of distortion consider the incremental output given by Eq. (5.10) and the input signal to be

$$x(t) = X_1 \cos \omega_1 t + X_2 \cos \omega_2 t \tag{5.17}$$

where $f_1 = \omega_1/2\pi$ and $f_2 = \omega_2/2\pi$ are the two input frequencies. The output frequency spectrum due to the quadratic term is shown in Table 5.1.

In addition to the dc term and the second harmonics of the two frequencies, there are additional terms at the sum and difference frequencies, $f_1 + f_2$, $f_1 - f_2$, which are the beat frequencies. The *second-order intermodulation distortion* (IM_2) is defined as the ratio of the amplitude at a beat frequency to the amplitude of the fundamental component.

$$IM_2 = \left| \frac{a_2 X_1 X_2}{a_1 X_1} \right| = \left| \frac{a_2 X_2}{a_1} \right| \tag{5.18}$$

where it has been assumed that the contribution to second-order intermodulation by higher-order terms is negligible. In defining IM_2 the input signals are assumed to be of equal amplitude and for this particular condition $IM_2 = 2\ HD_2$ [Eq. (5.12)].

TABLE 5.1 Output Frequency Spectrum Due to the Quadratic Term

Frequency	0	$2f_1$	$2f_2$	$f_1 \pm f_2$
Amplitude	$\frac{a_2}{2}[X_1^2 + X_2^2]$	$\frac{a_2}{2}X_1^2$	$\frac{a_2}{2}X_2^2$	$a_2 X_1 X_2$

TABLE 5.2 Output Frequency Spectrum Due to the Cubic Term

Frequency	f_1	f_2	$2f_1 \pm f_2$	$2f_2 \pm f_1$	$3f_1$	$3f_2$
Amplitude	$\frac{3a_3}{4}[X_1^3 + X_1X_2^2]$	$\frac{3a_3}{4}[X_2^3 + X_1^2X_2]$	$\frac{3}{4}a_3X_1^2X_2$	$\frac{3}{4}a_3X_1X_2^2$	$\frac{1}{4}a_3X_1^3$	$\frac{1}{4}a_3X_2^3$

The cubic term of the series expansion for the nonlinear circuit gives rise to components at frequencies $2f_1 + f_2$, $2f_2 + f_1$, $2f_1 - f_2$, $2f_2 - f_1$, and these terms result in *third-order intermodulation distortion* (IM_3). The frequency spectrum obtained from the cubic term is shown in Table 5.2.

For definition purposes the two input signals are assumed to be of equal amplitude and IM_3 is given by (assuming negligible contribution to the fundamental by the cubic term)

$$IM_3 = \frac{3}{4}\left|\frac{a_3X_1^3}{a_1X_1}\right| = \frac{3}{4}\left|\frac{a_3X_1^2}{a_1}\right| \tag{5.19}$$

Under these conditions $IM_3 = 3\ HD_3$ [Eq. (5.12)]. When f_1 and f_2 are close to one another, then the third-order intermodulation components, $2f_1 - f_2$, $2f_2 - f_1$, are close to the fundamental and are difficult to filter out.

Triple-Beat Distortion

When three sinusoidal signals are applied at the input then the output consists of components at the triple-beat frequencies. The cubic term in the nonlinearity results in the triple-beat terms

$$\frac{3}{2}a_3X_1X_2X_2\ \cos[\omega_1 \pm \omega_2 \pm \omega_3]t \tag{5.20}$$

and the *triple-beat distortion factor* (TB) is defined for equal amplitude input signals.

$$TB = \frac{3}{2}\left|\frac{a_3X_1^2}{a_1}\right| \tag{5.21}$$

From the above definition $TB = 2\ IM_3$. If all of the frequencies are close to one another, the triple beats will be close to the fundamental and cannot be easily removed.

Cross Modulation

Another form of distortion that occurs in amplitude-modulated (AM) systems (Chapter 63) due to the circuit nonlinearity is **cross modulation**. The modulation from an unwanted AM signal is transferred to the signal of interest and results in distortion. Consider an AM signal

$$x(t) = X_1\cos\omega_1t + X_2[1 + m\cos\omega_mt]\cos\omega_2t \tag{5.22}$$

where $m < 1$ is the modulation index. Due to the cubic term of the nonlinearity the modulation from the second signal is transferred to the first and the modulated component corresponding to the fundamental is

$$a_1X_1\left[1 + \frac{3a_3X_2^2m}{a_1}\cos\omega_mt\right]\cos\omega_1t \tag{5.23}$$

The *cross-modulation factor* (CM) is defined as the ratio of the transferred modulation index to the original modulation.

$$\mathrm{CM} = 3\left|\frac{a_3 X_2^2}{a_1}\right| \tag{5.24}$$

The cross modulation is a factor of four larger than IM_3 and twelve times as large as HD_3.

Compression and Intercept Points

For high-frequency circuits distortion is specified in terms of **compression and intercept points**. These quantities are derived from extrapolated small-signal output power levels. The 1 *dB compression point* is defined as the value of the fundamental output power for which the power is 1 dB below the extrapolated small-signal value.

The *nth-order intercept point* (IP_n), $n \geq 2$, is the output power at which the extrapolated small-signal powers of the fundamental and the nth harmonic intersect. Let P_{in} be an input power that is small enough to ensure small-signal operation. If P_1 is the output power of the fundamental, and P_n the output power of the nth harmonic, then the nth-order intercept point is given by $IP_n = \frac{nP_1 - P_n}{n-1}$, where power is measured in dB.

Crossover Distortion

This type of distortion occurs in circuits that use devices operating in a "push-pull" manner. The devices are used in pairs and each device operates only for half a cycle of the input signal (Class AB operation). One advantage of such an arrangement is the cancellation of even harmonic terms resulting in smaller total harmonic distortion. However, if the circuit is not designed to achieve a smooth crossover or transition from one device to another, then there is a region of the transfer characteristics when the output is zero. The resulting distortion is called **crossover distortion**.

Failure-to-Follow Distortion

When a properly designed peak detector circuit is used for AM demodulation the output follows the envelope of the input signal whereby the original modulation signal is recovered. A simple peak detector is a diode in series with a low-pass RC filter. The critical component of such a circuit is a linear element, the filter capacitance C. If C is large, then the output fails to follow the envelope of the input signal, resulting in **failure-to-follow distortion.**

Frequency Distortion

Ideally an amplifier circuit should provide the same amplification for all input frequencies. However, due to the presence of energy storage elements the gain of the amplifier is frequency dependent. Consequently different frequency components have different amplifications resulting in **frequency distortion.** The distortion is specified by a frequency response curve in which the amplifier output is plotted as a function of frequency. An ideal amplifier has a flat frequency response over the frequency range of interest.

Phase Distortion

When the phase shift (θ) in the output signal of an amplifier is not proportional to the frequency, the output does not preserve the form of the input signal, resulting in **phase distortion.** If the phase shift is proportional to frequency, different frequency components have a constant delay time (θ/ω) and no distortion is observed. In TV applications phase distortion can result in a smeared picture.

Computer Simulation of Distortion Components

Distortion characterization is important for nonlinear circuits. However, the techniques presented for distortion calculation can only be used for simple circuit configurations and at best to determine the second and third

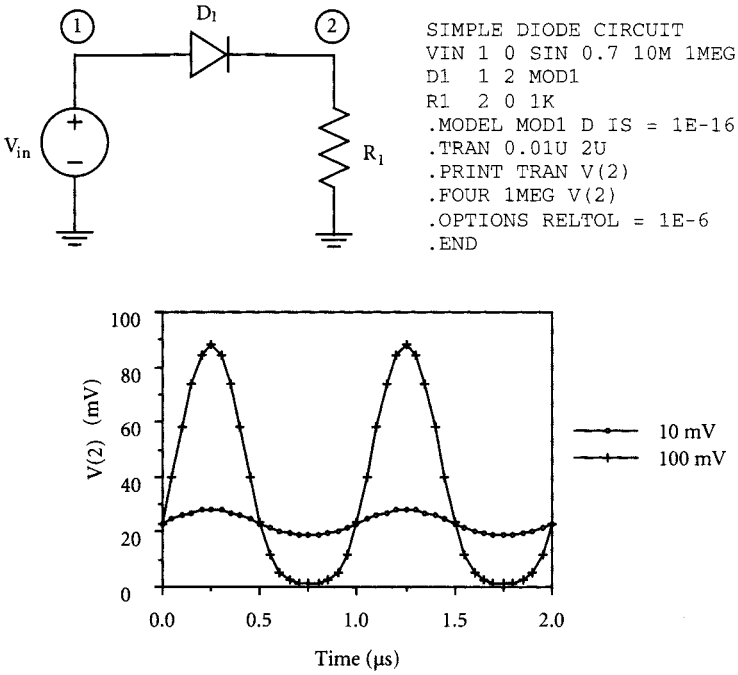


FIGURE 5.21 Simple diode circuit, SPICE input file, and output voltage waveforms.

harmonic distortion factors. In order to determine the distortion generation in actual circuits one must fabricate the circuit and then use a harmonic analyzer for sine curve inputs to determine the harmonics present in the output. An attractive alternative is the use of circuit simulation programs that allow one to investigate circuit performance before fabricating the circuit. In this section a brief overview of the techniques used in circuit simulators for distortion characterization is provided.

The simplest approach is to simulate the time-domain output for a circuit with a specified sinusoidal input signal and then perform a Fourier analysis of the output waveform. The simulation program SPICE2 provides a capability for computing the Fourier components of any waveform using a *.FOUR* command and specifying the voltage or current for which the analysis has to be performed. A simple diode circuit, the SPICE input file, and transient voltage waveforms for an input signal frequency of 1 MHz and amplitudes of 10 and 100 mV are shown in Fig. 5.21. The Fourier components of the resistor voltage are shown in Fig. 5.22; only the fundamental and first two significant harmonics are shown (SPICE provides information to the ninth harmonic).

In this particular example the input signal frequency is 1 MHz, and this is the frequency at which the Fourier analysis is requested. Since there are no energy storage elements in the circuit another frequency would have given identical results. To determine the Fourier components accurately a small value of the parameter RELTOL is used and a sufficient number of points for transient analysis are specified. From the output voltage waveforms and the Fourier analysis it is seen that the harmonic distortion increases significantly when the input voltage amplitude is increased from 10 mV to 100 mV.

The transient approach can be computationally expensive for circuits that reach their periodic steady state after a long simulation time. Results from the Fourier analysis are meaningful only in the periodic steady state, and although this approach works well for large levels of distortion it is inaccurate for small distortion levels.

For small distortion levels accurate distortion analysis can be performed by use of the Volterra series method. This technique is a generalization of the power-series method and is useful for analyzing harmonic and intermodulation distortion due to frequency-dependent nonlinearities. The SPICE3 program supports this analysis technique (in addition to the Fourier analysis of SPICE2) whereby the second and third harmonic and intermodulation components can be efficiently obtained by three small-signal analyses of the circuit.

FOURIER COMPONENTS OF TRANSIENT RESPONSE V(2) VIN = 10mV

DC COMPONENT = 2.330D-02

HARMONIC NO	FREQUENCY (HZ)	FOURIER COMPONENT	NORMALIZED COMPONENT	PHASE (DEG)	NORMALIZED PHASE (DEG)
1	1.000D+06	4.695D-03	1.000000	-0.001	0.000
2	2.000D+06	1.242D-04	0.026462	-89.989	-89.988
3	3.000D+06	1.705D-06	0.000363	-3.241	-3.239

TOTAL HARMONIC DISTORTION = 2.646409 PERCENT

FOURIER COMPONENTS OF TRANSIENT RESPONSE V(2) VIN = 100mV

DC COMPONENT = 3.445D-02

HARMONIC NO	FREQUENCY (HZ)	FOURIER COMPONENT	NORMALIZED COMPONENT	PHASE (DEG)	NORMALIZED PHASE (DEG)
1	1.000D+06	4.402D-02	1.000000	-0.011	0.000
2	2.000D+06	1.059D-02	0.240634	-89.993	-89.983
3	3.000D+06	1.658D-04	0.015127	-0.686	-0.675

TOTAL HARMONIC DISTORTION = 24.132679 PERCENT

FIGURE 5.22 Fourier components of the resistor voltage for input amplitudes of 10 and 100 mV, respectively.

An approach based on the *harmonic balance* technique available in the simulation program SPECTRE is applicable to both large and small levels of distortion. The program determines the periodic steady state of a circuit with a sinusoidal input. The unknowns are the magnitudes of the circuit variables at the fundamental frequency and at all the significant harmonics of the fundamental. The distortion levels can be simply calculated by taking the ratios of the magnitudes of the appropriate harmonics to the fundamental.

Defining Terms

Compression and Intercept Points: Characterize distortion in high-frequency circuits. These quantities are derived from extrapolated small-signal output power levels.

Cross modulation: Occurs in amplitude-modulated systems when the modulation of one signal is transferred to another by the nonlinearity of the system.

Crossover distortion: Present in circuits that use devices operating in a push-pull arrangement such that one device conducts when the other is off. Crossover distortion results if the transition or crossover from one device to the other is not smooth.

Failure-to-follow distortion: Can occur during demodulation of an amplitude-modulated signal by a peak detector circuit. If the capacitance of the low-pass RC filter of the peak detector is large, then the output fails to follow the envelope of the input signal, resulting in failure-to-follow distortion.

Frequency distortion: Caused by the presence of energy storage elements in an amplifier circuit. Different frequency components have different amplifications, resulting in frequency distortion and the distortion is specified by a frequency response curve.

Harmonic distortion: Caused by the nonlinear transfer characteristics of a device or circuit. When a sinusoidal signal of a single frequency (the *fundamental* frequency) is applied at the input of a nonlinear circuit, the output contains frequency components that are integer multiples of the fundamental frequency (*harmonics*). The resulting distortion is called harmonic distortion.

Harmonic distortion factors: A measure of the harmonic content of the output. The *n*th *harmonic distortion factor* is the ratio of the amplitude of the *n*th harmonic to the amplitude of the fundamental component of the output.

Intermodulation distortion: Distortion caused by the mixing or beating of two or more sinusoidal inputs due to the nonlinearity of a device. The output contains terms at the sum and difference frequencies called the *beat frequencies*.

Phase distortion: Occurs when the phase shift in the output signal of an amplifier is not proportional to the frequency.

Total harmonic distortion: The ratio of the root-mean-square value of the harmonics to the amplitude of the fundamental component of a waveform.

Related Topics

13.1 Analog Circuit Simulation • 47.5 Distortion and Second-Order Effects • 62.1 Power Quality Disturbances

References

K.K. Clarke and D.T. Hess, *Communication Circuits: Analysis and Design,* Reading, Mass.: Addison-Wesley, 1971.

P.R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits,* New York: John Wiley and Sons, 1992.

K.S. Kundert, *Spectre User's Guide: A Frequency Domain Simulator for Nonlinear Circuits,* EECS Industrial Liaison Program Office, University of California, Berkeley, 1987.

K.S. Kundert, *The Designer's Guide to SPICE and SPECTRE,* Mass.: Kluwer Academic Publishers, 1995.

L.W. Nagel, "SPICE2: A Computer Program to Simulate Semiconductor Circuits," Memo No. ERL-M520, Electronics Research Laboratory, University of California, Berkeley, 1975.

D.O. Pederson and K. Mayaram, *Analog Integrated Circuits for Communication: Principles, Simulation and Design,* Boston: Kluwer Academic Publishers, 1991.

T.L. Quarles, *SPICE3C.1 User's Guide,* EECS Industrial Liaison Program Office, University of California, Berkeley, 1989.

J.S. Roychowdhury, "SPICE 3 Distortion Analysis," Memo No. UCB/ERL M89/48, Electronics Research Laboratory, University of California, Berkeley, 1989.

D.D. Weiner and J.F. Spina, *Sinusoidal Analysis and Modeling of Weakly Nonlinear Circuits,* New York: Van Nostrand Reinhold Company, 1980.

Further Information

Characterization and simulation of distortion in a wide variety of electronic circuits (with and without feedback) is presented in detail in Pederson and Mayaram [1991]. Also derivations for the simple analysis techniques are provided and verified using SPICE2 simulations. Algorithms for computer-aided analysis of distortion are available in Weiner and Spina [1980], Nagel [1975], Roychowdhury [1989], and Kundert [1987]. Chapter 5 of Kundert [1995] gives valuable information on use of Fourier analysis in SPICE for distortion calculation in circuits. The software packages SPICE2, SPICE3 and SPECTRE are available from EECS Industrial Liaison Program Office, University of California, Berkeley, CA 94720.

5.4 Communicating with Chaos

Michael Peter Kennedy and Géza Kolumbán

The goal of a digital communications system is to deliver information represented by a sequence of binary symbols from a transmitter, through a physical channel, to a receiver. The mapping of these symbols into analog signals is called digital modulation.

In a conventional digital modulation scheme, the modulator represents each symbol to be transmitted as a weighted sum of a number of *periodic* basis functions. For example, two orthogonal signals, such as a sine and a cosine, can be used. Each symbol represents a certain bit sequence and is mapped to a corresponding set of weights. The objective of the receiver is to recover the weights associated with the received signal and thereby

to decide which symbol was transmitted [1]. The receiver's estimate of the transmitted symbol is mapped back to a bit sequence by a decoder.

When sinusoidal basis functions are used, the modulated signal consists of segments of periodic waveforms corresponding to the individual symbols. A unique segment of analog waveform corresponds to each symbol. If the spread spectrum technique is not used, the transmitted signal is narrow-band. Consequently, multipath propagation can cause high attenuation or even dropout of the transmitted narrow-band signal.

Chaotic signals are *nonperiodic* waveforms, generated by deterministic systems, which are characterized by a continuous "noise-like" broad power spectrum [2]. In the time domain, chaotic signals appear "random." Chaotic systems are characterized by "sensitive dependence on initial conditions"; an arbitrarily small perturbation eventually causes a large change in the state of the system. Equivalently, chaotic signals decorrelate rapidly with themselves. The autocorrelation function of a chaotic signal has a large peak at zero and decays rapidly.

Thus, while chaotic signals share many of the properties of stochastic processes, they also possess a deterministic structure that makes it possible to generate noise-like chaotic signals in a theoretically reproducible manner. In particular, a continuous-time chaotic system can be used to generate a wideband noise-like signal with robust and reproducible statistical properties [2].

Due to its wide-band nature, a signal comprising chaotic basis functions is potentially more resistant to multipath propagation than one constructed of sinusoids. Thus, **chaotic digital modulation,** where the digital information signal to be transmitted is mapped to chaotic waveforms, is potentially useful in propagation environments where multipath effects dominate.

In this chapter section, four chaotic digital modulation techniques are described in detail: Chaos Shift Keying (CSK), **Chaotic On-Off Keying** (COOK), **Differential Chaos Shift Keying** (DCSK), and FM-DCSK.

Elements of Chaotic Digital Communications Systems

In a digital communications system, the symbol to be transmitted is mapped by the modulator to an analog sample function and this analog signal passes through an analog channel. The analog signal in the channel is subject to a number of disturbing influences, including attenuation, bandpass filtering, and additive noise. The role of the demodulator is to decide, on the basis of the received corrupted sample function, which symbol was transmitted.

Transmitter

The sample function of duration T representing a symbol i is a weighted sum of analog basis functions $g_j(t)$:

$$s_i(t) = \sum_{j=1}^{N} s_{ij} g_j(t) \tag{5.25}$$

In a conventional digital modulation scheme, the analog sample function of duration T that represents a symbol is a linear combination of *periodic, orthogonal* basis functions (e.g., a sine and a cosine, or sinusoids at different frequencies), and the symbol duration T is an integer multiple of the period of the basis functions.

In a *chaotic* digital communications system, shown schematically in Fig. 5.23, the analog sample function of duration T that represents a symbol is a weighted sum of inherently nonperiodic *chaotic* basis function(s).

Channel Model

In any practical communications system, the signal $r_i(t)$ that is present at the input to the demodulator differs from that which was transmitted, due to the effects of the channel.

The simplest realistic model of the channel is a linear bandpass channel with additive white Gaussian noise (AWGN). A block diagram of the bandpass AWGN channel model that is considered throughout this section and the next is shown in Fig. 5.24. The additive noise is characterized by its power spectral density No.

Receiver

The role of the receiver in a digital communications system is to decide, on the basis of the received signal $r_i(t)$, which symbol was transmitted. This decision is made by estimating some property of the received sample

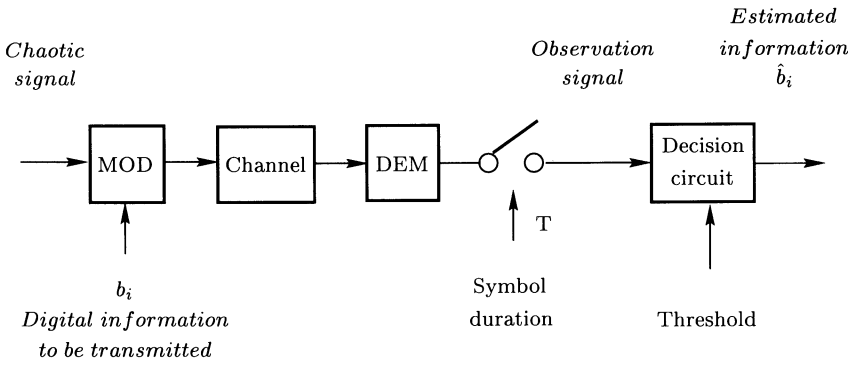


FIGURE 5.23 Block diagram of a chaotic communications scheme. The modulator and demodulator are labeled MOD and DEM, respectively.

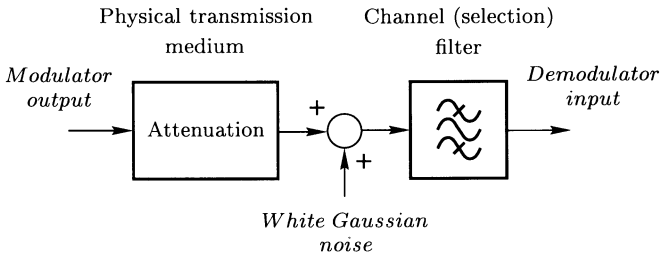


FIGURE 5.24 Model of an additive white Gaussian noise channel including the frequency selectivity of the receiver.

function. The property, for example, could be the weights of the coefficients of the basis functions, the energy of the received signal, or the correlation measured between different parts of the transmitted signal.

If the basis functions $g_j(t)$ are chosen such that they are periodic and orthogonal — that is:

$$\int_T g_i(t)\, g_j(t)\, dt = \begin{cases} K & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{5.26}$$

then the coefficients s_{ij} for symbol s_i can be recovered in the receiver by evaluating the observation signals

$$z_{ij} = \frac{1}{K} \int_T r_i(t)\, g_j(t)\, dt \tag{5.27}$$

Clearly, if $r_i(t) = s_i(t)$, then $z_{ij} = s_{ij}$ for every j, and the transmitted symbol can be identified.

In every physical implementation of a digital communications system, the received signal is corrupted by noise and the observation signal becomes a random process. The decision rule is very simple: decide in favor of the symbol to which the observation signal is closest.

Unlike periodic waveforms, chaotic basis functions are *inherently nonperiodic* and are different in each interval of duration T. Chaotic basis functions have the advantage that each transmitted symbol is represented by a unique analog sample function, and the correlation between chaotic sample functions is extremely low. However, it also produces a problem associated with estimating long-term statistics of a chaotic process from sample functions of finite duration.

This is the so-called *estimation problem*, discussed next [3]. It arises in all chaotic digital modulation schemes where the energy associated with a transmitted symbol is different every time that symbol is transmitted.

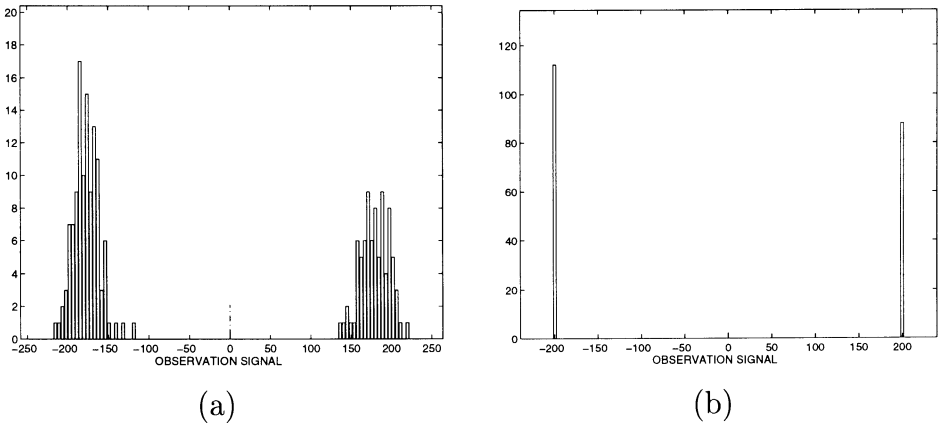


FIGURE 5.25 Histograms of the observation signal z_i for (a) non-constant and (b) constant energy per symbol.

The Estimation Problem

In modulation schemes that use periodic basis functions, $s_i(t)$ is periodic and the bit duration T is an integer multiple of the period of the basis function(s); hence, $\int_T s_i^2(t)dt$ is constant. By contrast, chaotic signals are inherently nonperiodic, so $\int_T s_i^2(t)dt$ varies from one sample function of length T to the next.

This effect is illustrated in Fig. 5.25(a), which shows a histogram of the observation signal in a noise-free binary **chaotic digital modulation** scheme where $s_{i1}(t) = g(t)$ and $s_{i2}(t) = -g(t)$. The observation signal is given by

$$z_i = \begin{cases} +\int_T g^2(t)dt & \text{when symbol } i \text{ is "1"} \\ -\int_T g^2(t)dt & \text{when symbol } i \text{ is "0"} \end{cases} \tag{5.28}$$

Because the basis function $g(\cdot)$ is *not periodic*, the value $\int_T g^2(t)dt$ varies from one symbol period of length T to the next. Consequently, the samples of the observation signal z_i corresponding to symbols "0" and "1" are clustered *with non-zero variance* about –180 and +180, respectively. Thus, the nonperiodic nature of the chaotic signal itself produces an effect that is indistinguishable at the receiver from the effect of additive channel noise.

By increasing the bit duration T, the variance of estimation can be reduced, but it also imposes a constraint on the maximum symbol rate. The estimation problem can be solved completely by keeping the energy per symbol constant. In this case, the variance of the samples of the observation signal is zero, as shown in Fig. 5.25(b).

Chaotic Digital Modulation Schemes

Chaos Shift Keying CSK

In Chaos Shift Keying (CSK), each symbol is represented by a weighted sum of chaotic basis functions $g_j(t)$. A binary CSK transmitter is shown in Fig. 5.26. The sample function $s_i(t)$ is $g_1(t)$ or $g_2(t)$, depending on whether symbol "1" or "0" is to be transmitted.

The required chaotic basis functions can be generated by different chaotic circuits (as shown in Fig. 5.26) or they can be produced by a single chaotic generator whose output is multiplied by two different constants. In both cases, the binary information to be transmitted is mapped to the bit energies of chaotic sample functions.

In chaotic digital communications systems, as in conventional communications schemes, the transmitted symbols can be recovered using either coherent or noncoherent demodulation techniques.

Coherent Demodulation of CSK

Coherent demodulation is accomplished by reproducing copies of the basis functions in the receiver, typically by means of a synchronization scheme [4]. When synchronization is exploited, the synchronization scheme must be able to recover the basis function(s) from the corrupted received signal.

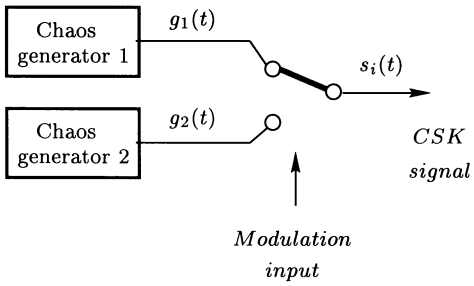


FIGURE 5.26 Block diagram of a CSK modulator.

If a single sinusoidal basis function is used, then a *narrow-band* phase-locked loop (PLL) can be used to recover it [1]. Noise corrupting the transmitted signal is suppressed because of the low-pass property of the PLL. When an inherently wideband chaotic basis function is used, the synchronized circuit must also be *wideband* in nature. Typically, both the "amplitude" and "phase" of the chaotic basis function must be recovered from the received signal. Because of the wideband property of the chaotic basis function, narrow-band linear filtering cannot be used to suppress the additive channel noise.

Figure 5.27 shows a coherent (synchronization-based) receiver using binary CSK modulation with two basis functions $g_1(t)$ and $g_2(t)$. Synchronization circuits at the receiver attempt to reproduce the basis functions, given the received noisy sample function $r_i(t) = s_i(t) + n(t)$.

An acquisition time T_S is allowed for the synchronization circuits to lock to the incoming signal. The recovered basis functions $\hat{g}_1(t)$ and $\hat{g}_2(t)$ are then correlated with $r_i(t)$ for the remainder of the bit duration T. A decision is made on the basis of the relative closeness of $r_i(t)$ to $\hat{g}_1(t)$ and $\hat{g}_2(t)$, as quantified by the observation variables z_{i1} and z_{i2}, respectively.

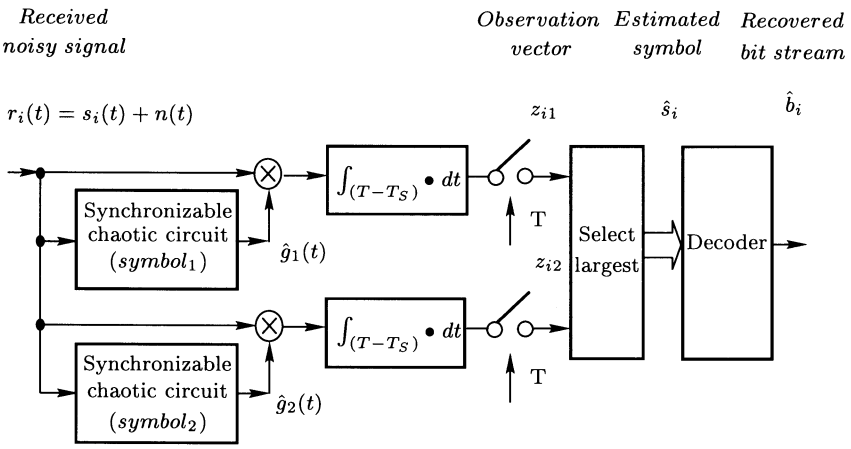


FIGURE 5.27 Block diagram of a coherent CSK receiver.

Studies of **chaotic synchronization,** where significant noise and filtering have been introduced in the channel, suggest that the performance of chaotic synchronization schemes is significantly worse at low *signal-to-noise ratio* (SNR) than that of the best synchronization schemes for sinusoids [4–6].

Noncoherent Demodulation of CSK

Synchronization (in the sense of carrier recovery) is not a necessary requirement for digital communications; demodulation can also be performed without synchronization. This is true for both periodic and chaotic sample functions.

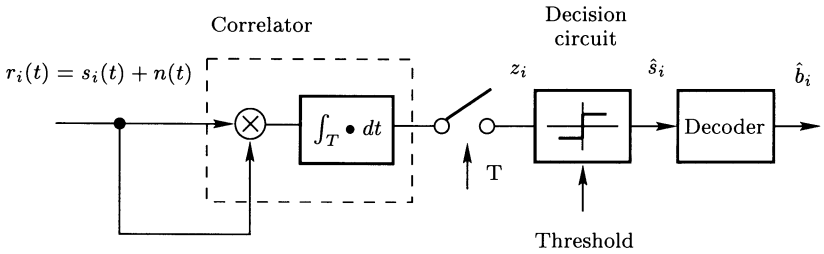


FIGURE 5.28 Block diagram of a non-coherent CSK receiver.

Due to the nonperiodic property of chaotic signals, the energy of chaotic sample functions varies from one sample function to the next, even if the same symbol is transmitted. If the mean bit energies $\int_T g_1^2(t)dt$ and $\int_T g_2^2(t)dt$ associated with symbols "1" and "0," respectively, are sufficiently different, then a CSK transmission can be demodulated without synchronization. In this case, the bit energy can be estimated by a correlator at the receiver, as shown in Fig. 5.28, without recovering the basis functions. The decision as to which symbol was transmitted is made by comparing this estimate against a threshold.

The observation signal z_i that is used by the decision circuit is defined by

$$z_i = \int_T r_i^2(t)dt \tag{5.29}$$

where $\int_T$ denotes integration over one bit period.

For a given noise level and chaotic signal, the best noise performance of CSK can be achieved if the distance between the mean bit energies of the two symbols is maximized; this requirement can be satisfied by the Chaotic On-Off Keying technique, described next.

Chaotic On-Off Keying (COOK)

In the Chaotic On-Off Keying (COOK) scheme, the chaotic signal is switched on and off to transmit symbols "1" and "0," respectively, as shown in Fig. 5.29.

If the average bit energy is E_b and both symbols are equiprobable, then the distance between the elements of the signal set is $2E_b$. It is well-known from the theory of communications systems that the greater the distance between the elements of the signal set, the better the noise performance of a modulation scheme [1]. The noise

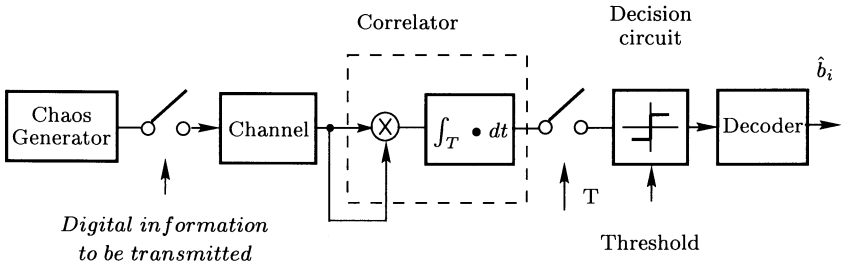


FIGURE 5.29 Block diagram of COOK modulation scheme with non-coherent demodulation.

performance of COOK represents the upper bound for CSK because the distance between the elements of the signal set is maximized.

Notice that the observation signal is determined by the energy per bit of the noisy received signal $r_i(t) = s_i(t) + n(t)$. This is why a significant drawback of the CSK system — namely that the threshold value of the decision circuit depends on the noise level — also applies to COOK. This means that using COOK, one can maximize the distance between the elements of the signal set, but the threshold level required by the decision

circuit depends on the SNR. The threshold can be kept constant by applying the Differential Chaos Shift Keying method.

Differential Chaos Shift Keying (DCSK)

In Differential Chaos Shift Keying (DCSK), every symbol to be transmitted is represented by two chaotic sample functions. The first sample function serves as a *reference*, while the second one carries the information. symbol "1" is sent by transmitting a reference signal provided by a chaos generator twice in succession; while for symbol "0," the reference chaotic signal is transmitted, followed by an inverted copy of the same integral. Thus,

$$s(t)=\begin{cases} x(t), & t_i \le t < t_i + T/2 \\ +x(t-T/2), & t_i + T/2 \le t < t_i + T \end{cases} \tag{5.30}$$

if symbol "1" is transmitted in $(t_i, t_i + T)$ and

$$s(t)=\begin{cases} x(t), & t_i \le t < t_i + T/2 \\ -x(t-T/2), & t_i + T/2 \le t < t_i + T \end{cases} \tag{5.31}$$

if symbol "0" is transmitted in $(t_i, t_i + T)$.

Figures 5.30 and 5.31 show a block diagram of a DCSK modulator and a typical DCSK signal corresponding to the binary sequence 1100. In this example, the chaotic signal is produced by an analog phase-locked loop (APLL) and the bit duration is 20 ms.

Since each bit is mapped to the correlation between successive segments of the transmitted signal of length $T/2$, the information signal can be recovered by a correlator. A block diagram of a DCSK demodulator is shown in Fig. 5.32.

The received noisy signal is delayed by half of the bit duration ($T/2$), and the correlation between the received signal and the delayed copy of itself is determined. The decision is made by a level comparator [7].

In contrast to the CSK and COOK schemes discussed above, DCSK is an antipodal modulation scheme. In addition to superior noise performance, the decision threshold is zero independently of the SNR [7].

A further advantage results from the fact that the reference- and information-bearing sample functions pass through the same channel, thereby rendering the modulation scheme insensitive to channel distortion. DCSK can also operate over a time-varying channel if the parameters of the channel remain constant for half the bit duration T.

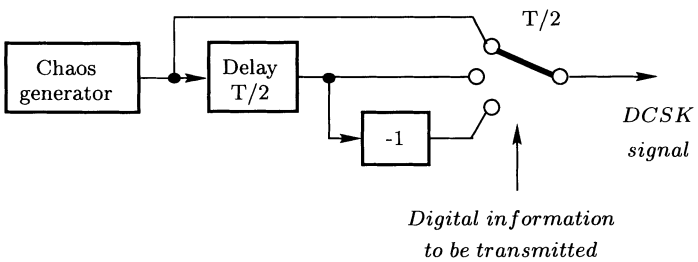


FIGURE 5.30 Block diagram of a DCSK modulator.

The principal drawback of DCSK arises from the fact that the correlation is performed over half the bit duration. Compared to conventional techniques where the elements of the signal set are available at the receiver, DCSK has half of the data rate, and only half the bit energy contributes to its noise performance [4,6].

In the CSK, COOK, and DCSK modulation schemes, the information signal to be transmitted is mapped to chaotic sample functions of finite length. The property required by the decision circuit at the receiver to

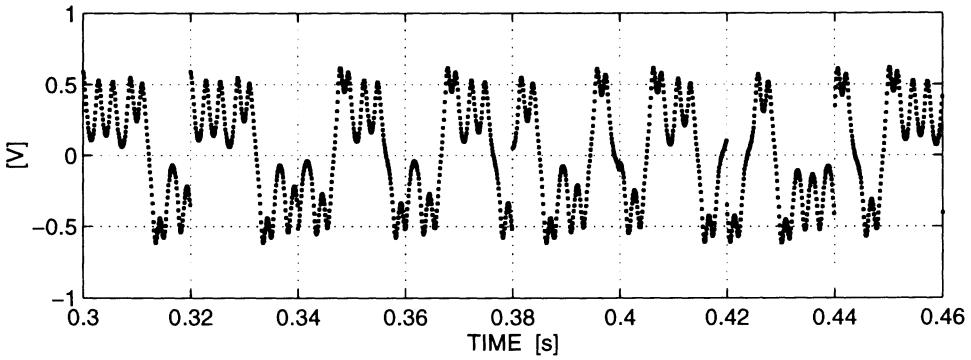


FIGURE 5.31 DCSK signal corresponding to binary sequence 1100.

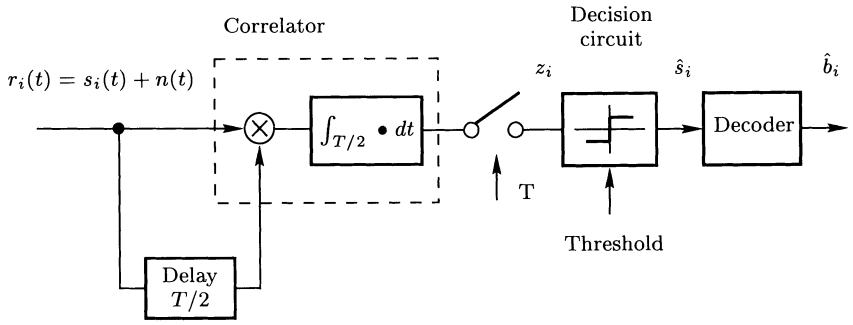


FIGURE 5.32 Block diagram of a DCSK receiver.

perform the demodulation can only be *estimated* because of the nonperiodic nature of chaotic signals. The estimation has a non-zero variance even in the noise-free case; this puts a lower bound on the bit duration and thereby limits the data rate.

One way to improve the data rate is to use a multilevel modulation scheme such as those described in [8]. Alternatively, one can solve the estimation problem directly by modifying the modulation scheme such that the transmitted energy for each symbol is kept constant. FM-DCSK is an example of the latter approach.

FM-DCSK

The power of a **frequency-modulated** (FM) signal is independent of the modulation. Therefore, if a chaotic signal is applied to the input of an FM modulator, and the output of the FM modulator is applied to the input of a DCSK modulator, then the resulting output of the DCSK modulator has constant energy per symbol. If this signal is applied directly to a DCSK correlation receiver, then the observation signal in the receiver has zero variance in the noise-free case and the estimation problem is solved.

As in the DCSK technique, every information bit is transmitted in two pieces: the first sample function serves as a reference, while the second one carries the information. The operation of the modulator shown in Fig. 5.33 is similar to DCSK, the difference being that the FM signal, rather than the chaotic signal itself, is applied to the input of the DCSK modulator. In this example, the chaotic signal is generated by an appropriately designed analog phase-locked loop (APLL).

The demodulator of an FM-DCSK system is a DCSK receiver. The only difference is that, instead of low-frequency chaotic signals, the noisy FM signals are correlated directly in the receiver, as shown in Fig. 5.34.

The noise performance of the FM-DCSK system is an attainable upper bound to that of DCSK. The main advantage of FM-DCSK modulation over CSK, COOK, and DCSK is that the data rate is not limited by the properties of the chaotic signal.

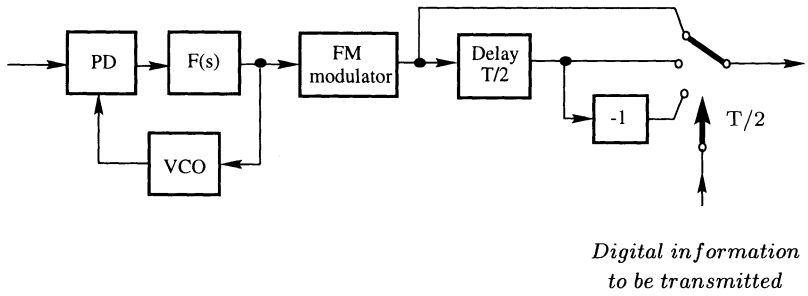


FIGURE 5.33 Block diagram of an FM-DCSK modulator.

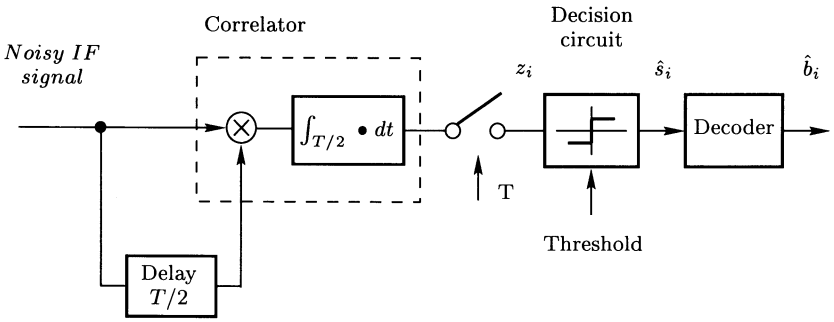


FIGURE 5.34 Block diagram of an FM-DCSK demodulator.

Performance Evaluation

The noise performance of a digital modulation scheme is characterized by plotting the *bit error rate* (BER) as a function of the ratio of the energy per bit to the noise spectial density. (E_b/N_0). The simulated noise performance of noncoherent CSK, COOK, and DCSK/FM-DCSK is summarized graphically in Fig. 5.35.

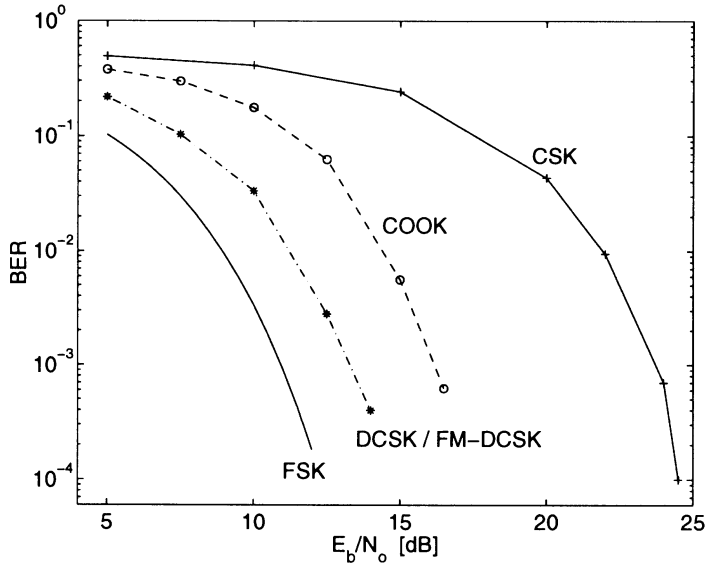


FIGURE 5.35 Noise performance of the CSK, COOK, and DCSK/FM-DCSK techniques. Non-coherent FSK is shown for comparison.

The upper bound on the data rate of DCSK can be increased by using multilevel modulation schemes or by keeping the transmitted energy constant for each symbol. The FM-DCSK technique, which is an antipodal modulation scheme with constant bit energy, represents an optimal solution in the sense that its noise performance is equal to that of DCSK but the data rate is not limited by the properties of the underlying chaotic signal.

Low-Pass Equivalent Models for Chaotic Communications Systems

The previous sections have described **chaotic digital modulation** schemes. The output of these modulations is generally a low-pass signal. Many telecommunications channels, such as a radio channel, can transmit only bandpass signals, so a second modulation scheme must be used to produce an RF output in these cases. An exception is the FM-DCSK modulation scheme, where the output of the FM modulator is already a bandpass RF signal and the DCSK modulation is applied directly to this signal.

The performance evaluation of communications systems can be done analytically only in the simplest cases; usually, computer simulation is required. However, if computer simulations of RF communications systems are performed directly in the RF domain, then the sampling frequency for the simulation depends on both the carrier frequency and the bandwidth of the transmitted signal. The high carrier frequency results in a high sampling frequency and consequently a long simulation time. On the other hand, the parameters of a bandpass system do not depend on the actual value of the carrier frequency.

It is well-known that a low-pass equivalent model can be developed for every bandpass system [1]. As a result, the carrier frequency can be removed from the model of an RF communications system and the sampling frequency is then determined solely by the bandwidth of the RF signal. This reduces significantly the computational effort required to characterize the performance of a chaotic communications system. This section illustrates the development of a low-pass equivalent model for the RF FM-DCSK system. For further details and models of other chaotic communications systems, see [9].

Theoretical Background

Representation of Bandpass Signals

A signal $x(t)$ is referred to as a bandpass signal if its energy is nonnegligible only in a frequency band of total extent $2BW$ centered about a carrier frequency f_c. Every bandpass signal can be expressed in terms of a slowly varying signal $\tilde{x}(t)$ and a complex exponential

$$x(t) = Re\left[\tilde{x}(t)e^{j\omega_c t}\right] \tag{5.32}$$

where $\tilde{x}(t)$ is called the complex envelope, and $\omega_c = 2\pi f_c$. In general, $\tilde{x}(t)$ is a complex-valued quantity; it can be expressed in terms of its *in-phase* and *quadrature* components, $x_I(t)$ and $x_Q(t)$, as follows:

$$\tilde{x}(t) = x_I(t) + jx_Q(t) \tag{5.33}$$

Both $x_I(t)$ and $x_Q(t)$ are low-pass signals limited to the frequency band $-BW \le f \le BW$.

The complex envelope $\tilde{x}(t)$ carries all of the information, except the carrier frequency, of the original bandpass signal $x(t)$. This means that if the complex envelope of a signal is given, then that signal is completely characterized. Knowing the carrier frequency, in addition, means that the original bandpass signal can be reconstructed.

The in-phase and quadrature components of the complex envelope can be generated from the bandpass signal $x(t)$ using the scheme shown in Fig. 5.36, where the ideal low-pass filters have bandwidth BW.

The original bandpass signal $x(t)$ can be reconstructed from the in-phase and quadrature components of $\tilde{x}(t)$ as shown in Fig. 5.37.

Representation of Bandpass Systems

Let the bandpass input signal $x(t)$ be applied to a linear time-invariant bandpass system with impulse response $h(t)$, and let the bandwidth of the bandpass system be equal to $2B$ and centered about the carrier frequency f_c. Then, by analogy with the representation of bandpass signals, the impulse response of the bandpass system can also be expressed in terms of a slowly varying complex impulse response $\tilde{h}(t)$ and a complex exponential:

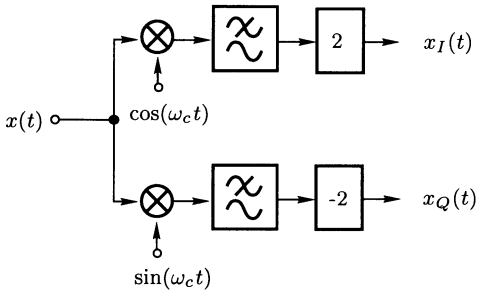


FIGURE 5.36 Generation of the in-phase and quadrature components of a bandpass signal.

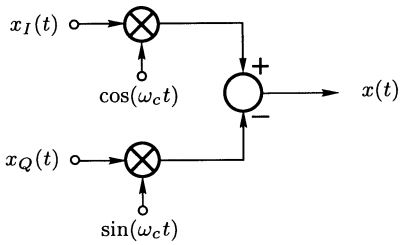


FIGURE 5.37 Reconstruction of the original bandpass signal from its in-phase and quadrature components.

$$h(t) = Re\left[\tilde{h}(t)e^{j\omega_c t}\right] \tag{5.34}$$

In general, the complex impulse response is a complex-valued quantity that can be expressed in terms of its in-phase and quadrature components

$$\tilde{h}(t) = h_I(t) + jh_Q(t) \tag{5.35}$$

where $\tilde{h}(t)$, $h_I(t)$, and $h_Q(t)$ are all low-pass functions limited to the frequency band $-B \le f \le B$.

Representation of Bandpass Gaussian Noise

If the channel noise $n(t)$ is a bandpass Gaussian random process and its spectrum is symmetric about the carrier frequency f_c, then $n(t)$ can also be represented by its complex envelope

$$\tilde{n}(t) = n_I(t) + jn_Q(t) \tag{5.36}$$

Low-Pass Equivalent of FM-DCSK System

The block diagram of a general chaotic communications system is given in Fig. 5.23. As shown in Fig. 5.34, the demodulator of an FM-DCSK system is a correlator, and the observation signal z_i is the correlator output sampled at the decision time instants. To derive the low-pass equivalent model of a chaotic communications scheme, the relationship between the analog input and output signals must be found; that is, the correlator output $z(t)$ must be determined for a given analog input signal.

The block diagram of the RF FM-DCSK system to be transformed is shown in Fig. 5.38, where $h(t)$ denotes

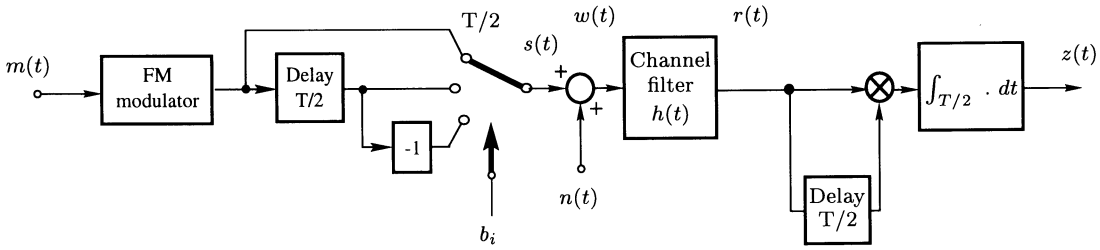


FIGURE 5.38 Block diagram of an RF FM-DCSK system.

the impulse response of channel filter, $n(t)$ is the channel noise, and $w(t)$ is the input to the channel filter.

Applying the theorems of the analytic signal approach [1], assuming a zero-phase channel filter and that half of the bit duration is equal to an entire multiple of the RF carrier period, the low-pass equivalent model of the RF FM-DCSK system can be developed as shown in Fig. 5.39 (for further details, see [9]).

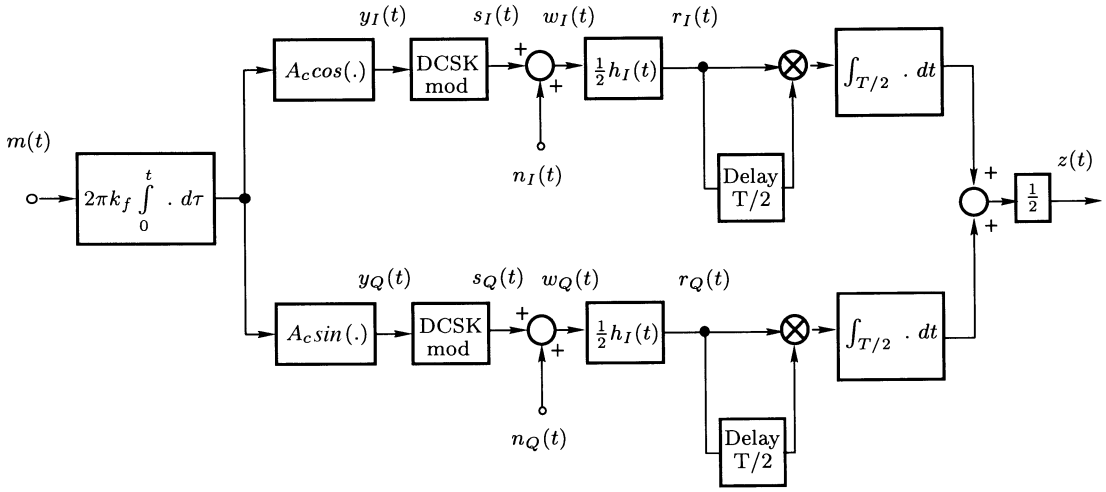


FIGURE 5.39 Low-pass equivalent model of the RF FM-DCSK chaotic communications system shown in Fig. 5.38.

Note that all RF signals and the carrier frequency have been removed from Fig. 5.39. Consequently, the sampling frequency required for computer simulations is determined exclusively by the slowly-varying low-pass signals. All noise performance curves shown in this chapter section have been determined using low-pass equivalent models derived in this way.

Multipath Performance of FM-DCSK

In many applications, such as mobile communications or indoor radio, the transmitted signal arrives at the receiver via multiple propagation paths with different delays, thus giving rise to multipath propagation. The components arriving via different propagation paths may add destructively, resulting in deep frequency-selective fading. Conventional narrow-band systems completely fail to operate if a *multipath-related null* (defined below) resulting from deep frequency-selective fading coincides with the carrier frequency. Because of the inherently broad-band nature of chaotic signals, chaotic modulation schemes have potentially better performance in multipath environments than narrow-band ones. In this section, the performance degradation of the FM-DCSK scheme resulting from multipath propagation is determined by computer simulation.

Multipath Model

A time-invariant multipath radio channel having two propagation paths can be modeled as shown in Fig. 5.40.

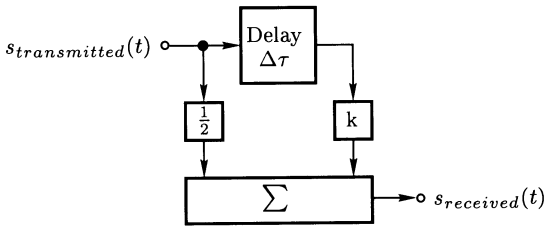


FIGURE 5.40 Tapped delay line model of a multipath radio channel.

In the worst case, the two received signals cancel each other completely at the carrier frequency ω_c; that is, $\Delta\tau\omega_c = (2n + 1)\,\pi$, $n = 0, \pm1, \pm2,\ldots$, and $k = -1/2$, where $\Delta\tau$ denotes the additional delay of the second path.

Let the multipath channel be characterized by its frequency response shown in Fig. 5.41. Note that the

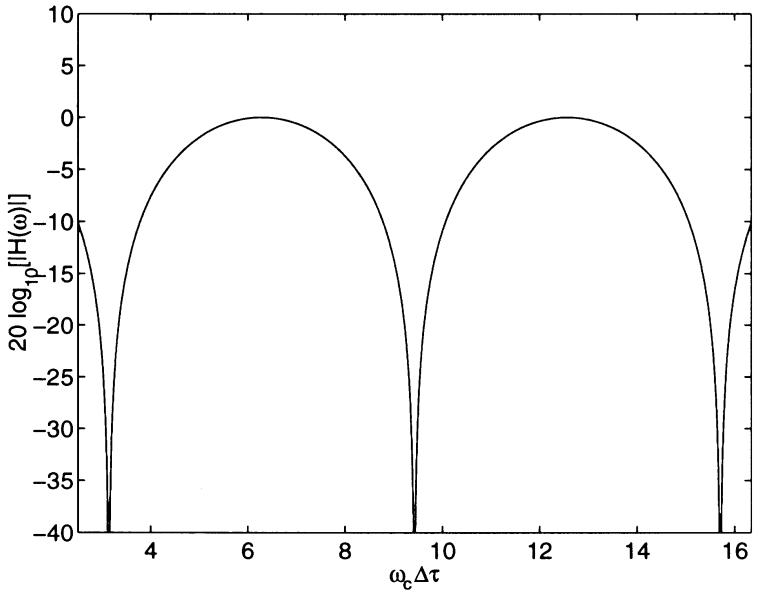


FIGURE 5.41 Magnitude of the frequency response of a multipath channel.

multipath-related nulls, where the attenuation becomes infinitely large, appear at

$$f_{\text{null}} = \frac{2n+1}{2\Delta\tau}, \quad n = 0, \pm1, \pm2, \ldots \tag{5.37}$$

Let the bandwidth of fading be defined as the frequency range over which the attenuation of the multipath channel is greater than 10 dB. Then the bandwidth of fading can be expressed as

$$\Delta f_{\text{null}} \approx \frac{0.1}{\Delta\tau} \tag{5.38}$$

Performance of FM-DCSK Modulation Scheme

Figure 5.41 shows why conventional narrow-band systems can fail to operate over a multipath channel. Due to high attenuation appearing about the multipath-related nulls, the SNR becomes extremely low at the input of the receiver. Consequently, the demodulator and the carrier recovery circuit, if used, cannot operate.

In a chaotic communications system, the power of the radiated signal is spread over a wide frequency range. The appearance of a multipath-related null means that part of the transmitted power is lost, but the system still operates. Of course, the lower SNR at the demodulator input results in a worse bit error rate.

The performance degradation of the FM-DCSK system due to multipath propagation is shown in Fig. 5.42,

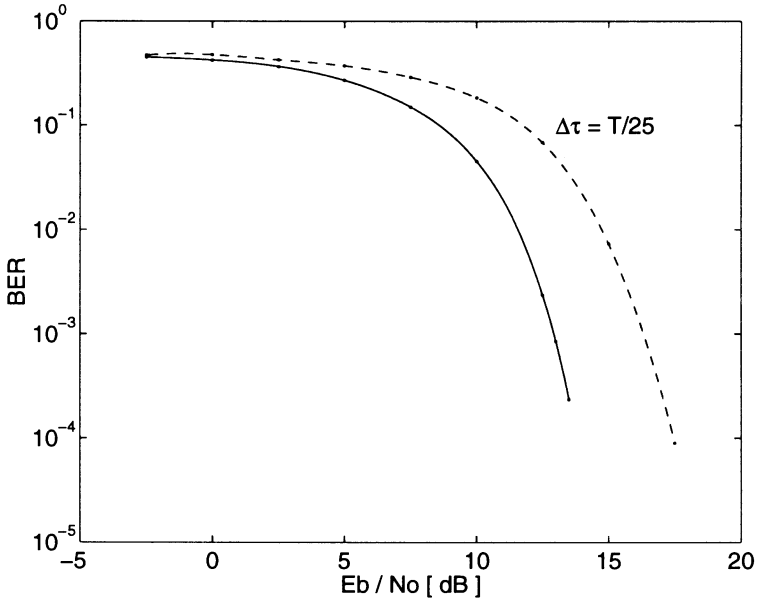


FIGURE 5.42 Noise performance of FM-DCSK with (dashed curve) and without (solid curve) multipath effects.

where $\Delta\tau = T/25$. The solid line shows the noise performance if multipath propagation is not present, while the system performance for $k = -1/2$ is given by the dashed curve. Note that FM-DCSK performs extremely well over a radio channel suffering from multipath effects; the performance degradation even in the worst case is less than a few dB. Note that conventional narrow-band systems cannot operate over this channel.

Defining Terms

Chaotic synchronization: The process by which a dynamical system is synchronized with a chaotic reference signal. In chaotic digital communications, chaotic (rather than periodic) basis functions must be recovered without distortion from the noisy received (reference) signal at the receiver. Noise corrupting the reference signal must be suppressed as much as possible.

Chaotic digital modulation: The mapping of information-source symbols into chaotic signals, which is performed to carry information through the analog transmission channel.

Chaos shift keying: A digital modulation scheme in which the source information is carried by the coefficients of a weighted sum of chaotic waveforms.

Chaotic on-off keying: A binary digital modulation scheme in which the chaotic carrier is switched on or off, depending on the binary information to be transmitted.

Differential chaos shift keying: A digital modulation scheme in which the source information is carried by the correlation between segments of a chaotic waveform that are separated in time.

Frequency-modulated differential chaos shift keying: A digital modulation scheme in which the source information is carried by the correlation between chaotic frequency-modulated waveforms.

References

1. S.S. Haykin. *Communication Systems*, 3rd edition, John Wiley & Sons, New York, 1994.
2. M.P. Kennedy. Bifurcation and chaos, in *The Circuits and Filters Handbook*, W.K. Chen, Editor, pages 1089–1164. CRC Press, 1995.
3. G. Kolumbán, M.P. Kennedy, and G. Kis. Determination of symbol duration in chaos-based communications, *Proc. NDES'97*, pages 217–222, Moscow, Russia, 26–27 June, 1997.
4. G. Kolumbán, M.P. Kennedy, and L.O. Chua. The role of synchronization in digital communications using chaos. Part I. Fundamentals of digital communications. *IEEE Trans. Circuits and Systems. Part I. Fundamental Theory and Applications*, 44(10):927–936, 1997.
5. G. Kolumbán, H. Dedieu, J. Schweizer, J. Ennitis, and B. Vizvári. Performance evaluation and comparison of chaos communication systems, in *Proc. NDES'96* pages 105–110, Sevilla, 27–28 June, 1996.
6. G. Kolumbán, M.P. Kennedy, and L.O. Chua. The role of synchronization in digital communications using chaos. Part II. Chaotic modulation and chaotic synchronization, *IEEE Trans. Circuits and Systems. Part I. Fundamental Theory and Applications*, 45(11):1129–1140, 1998.
7. G. Kolumbán, B. Vizvári, W. Schwarz, and A. Abel. Differential chaos shift keying: A robust coding for chaotic communications, *Proc. NDES'96*, pages 87–92, Sevilla, 27–28 June, 1996.
8. G. Kolumbán, M.P. Kennedy, and G. Kis. Multilevel differential chaos shift keying, *Proc. NDES'97*, pages 191–196, Moscow, Russia, 26–27 June, 1997.
9. G. Kolumbán. Performance evaluation of chaotic communications systems: determination of low-pass equivalent model, *Proc. NDES'98*, pages 41–51, Budapest, Hungary, 17–18 July, 1998.
10. R.C. Dixon. *Spread Spectrum Communication Systems with Commercial Applications*, 3rd edition, Wiley, New York, 1994.
11. L.M. Pecora and T.L. Carroll. Synchronization in chaotic systems, *Phys. Rev. Lett.*, 64(8):821–824, 1990.
12. M. Hasler. Engineering chaos for encryption and broadband communication, *Phil. Trans. R. Soc. Lond.*, 353(1701):115–126, 1995.
13. G. Kolumbán, G. Kis, Z. Jákó, and M.P. Kennedy. FM-DCSK: a robust modulation scheme for chaotic communications, *IEICE Transactions*, E81-A(9): 1798–1802, September 1998.
14. G. Heidari-Bateni and C.D. McGillem. A chaotic direct sequence spread spectrum communication system, *IEEE Trans. Commun.*, COM-42(2/3/4):1524–1527, 1994.
15. N.F. Rulkov, M.M. Sushchik, L.S. Tsimring, and H.D. Abarbanel. Generalized synchronization of chaos in directionally coupled chaotic systems, *Phys. Rev. E.*, 51(2):980–994, 1995.
16. S. Hayes, C. Grebogi, and E. Ott. Communicating with chaos, *Phys. Rev. Lett.*, 70(20):3031–3034, 1993.

Further Information

An introduction to chaos for electrical engineers can be found in [2].

Digital modulation theory and low-pass equivalent circuits of bandpass communications systems are described at an introductory level in [1]. The theory of spread spectrum communications can be found in [10].

The field of communicating with chaos has developed rapidly since the experiments by Pecora, Carroll, and others in the 1990s on chaotic synchronization [11]. Hasler [12] has written an overview of early work in this field.

The role of synchronization in chaotic digital modulation is explored in [4,6]. These papers also describe the state of the art in noncoherent receivers for chaotic digital communications. FM-DCSK is developed in [13].

Advances in the theory and practice of chaotic communications in electrical engineering are reported in *Electronics Letters*, the *IEEE Transactions on Circuits and Systems*, and the *IEEE Transactions on Communications*.

This section has focused exclusively on chaotic modulation techniques. Other applications of chaotic signals and synchronization schemes have been proposed but they are less close to practice: discrete-time chaotic sequences for spread spectrum systems were introduced in [14]; synchronization techniques for chaotic systems,

such as [15] and methods for transmitting or hiding information (e.g., [16]), are frequently reported in physics journals such as *Physical Review Letters* and the *Physical Review E*.

Dorf, R.C., Wan, Z., Johnson, D.E. "Laplace Transform"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

6

Laplace Transform

Richard C. Dorf
University of California, Davis

Zhen Wan
University of California, Davis

David E. Johnson
Birmingham-Southern College

6.1 Definitions and Properties
Laplace Transform Integral • Region of Absolute Convergence • Properties of Laplace Transform • Time-Convolution Property • Time-Correlation Property • Inverse Laplace Transform

6.2 Applications
Differentiation Theorems • Applications to Integrodifferential Equations • Applications to Electric Circuits • The Transformed Circuit • Thévenin's and Norton's Theorems • Network Functions • Step and Impulse Responses • Stability

6.1 Definitions and Properties

Richard C. Dorf and Zhen Wan

The **Laplace transform** is a useful analytical tool for converting time-domain signal descriptions into functions of a complex variable. This *complex domain* description of a signal provides new insight into the analysis of signals and systems. In addition, the Laplace transform method often simplifies the calculations involved in obtaining system response signals.

Laplace Transform Integral

The Laplace transform completely characterizes the exponential response of a time-invariant linear function. This transformation is formally generated through the process of multiplying the linear characteristic signal $x(t)$ by the signal e^{-st} and then integrating that product over the time interval $(-\infty, +\infty)$. This systematic procedure is more generally known as *taking the Laplace transform* of the signal $x(t)$.

Definition: The Laplace transform of the continuous-time signal $x(t)$ is

$$X(s) = \int_{-\infty}^{+\infty} x(t)e^{-st}dt$$

The variable s that appears in this integrand exponential is generally complex valued and is therefore often expressed in terms of its rectangular coordinates

$$s = \sigma + j\omega$$

where $\sigma = \mathrm{Re}(s)$ and $\omega = \mathrm{Im}(s)$ are referred to as the *real* and *imaginary* components of s, respectively.

The signal $x(t)$ and its associated Laplace transform $X(s)$ are said to form a *Laplace transform pair*. This reflects a form of equivalency between the two apparently different entities $x(t)$ and $X(s)$. We may symbolize this interrelationship in the following suggestive manner:

$$X(s) = \mathcal{L}[x(t)]$$

where the operator notation $\mathcal{L}$ means to multiply the signal $x(t)$ being operated upon by the complex exponential e^{-st} and then to integrate that product over the time interval $(-\infty, +\infty)$.

Region of Absolute Convergence

In evaluating the Laplace transform integral that corresponds to a given signal, it is generally found that this integral will exist (that is, the integral has finite magnitude) for only a restricted set of s values.

The definition of **region of absolute convergence** is as follows. The set of complex numbers s for which the magnitude of the Laplace transform integral is finite is said to constitute the region of absolute convergence for that integral transform. This region of convergence is always expressible as

$$\sigma_+ < \mathrm{Re}(s) < \sigma_-$$

where σ_+ and σ_- denote real parameters that are related to the causal and anticausal components, respectively, of the signal whose Laplace transform is being sought.

Laplace Transform Pair Tables

It is convenient to display the Laplace transforms of standard signals in one table. Table 6.1 displays the time signal $x(t)$ and its corresponding Laplace transform and region of absolute convergence and is sufficient for our needs.

Example. To find the Laplace transform of the first-order causal exponential signal

$$x_1(t) = e^{-at}u(t)$$

where the constant a can in general be a complex number.

The Laplace transform of this general exponential signal is determined upon evaluating the associated Laplace transform integral

$$\begin{aligned} X_1(s) &= \int_{-\infty}^{+\infty} e^{-at}u(t)e^{-st}dt = \int_0^{+\infty} e^{-(s+a)t}dt \\ &= \left.\frac{e^{-(s+a)t}}{-(s+a)}\right|_0^{+\infty} \end{aligned} \tag{6.1}$$

In order for $X_1(s)$ to exist, it must follow that the real part of the exponential argument be positive, that is,

$$\mathrm{Re}(s+a) = \mathrm{Re}(s) + \mathrm{Re}(a) > 0$$

If this were not the case, the evaluation of expression (6.1) at the upper limit $t = +\infty$ would either be unbounded if $\mathrm{Re}(s) + \mathrm{Re}(a) < 0$ or undefined when $\mathrm{Re}(s) + \mathrm{Re}(a) = 0$. On the other hand, the upper limit evaluation is zero when $\mathrm{Re}(s) + \mathrm{Re}(a) > 0$, as is already apparent. The lower limit evaluation at $t = 0$ is equal to $1/(s + a)$ for all choices of the variable s.

The Laplace transform of exponential signal $e^{-at}\,u(t)$ has therefore been found and is given by

$$\mathcal{L}\ [e^{-at}u(t)] = \frac{1}{s+a} \quad \text{for } \mathrm{Re}(s) > -\mathrm{Re}(a)$$

TABLE 6.1 Laplace Transform Pairs

	Time Signal $x(t)$	Laplace Transform $X(s)$	Region of Absolute Convergence
1.	$e^{-at}u(t)$	$\frac{1}{s+a}$	$\mathrm{Re}(s) > -\mathrm{Re}(a)$
2.	$t^k e^{-at}u(-t)$	$\frac{k!}{(s+a)^{k+1}}$	$\mathrm{Re}(s) > -\mathrm{Re}(a)$
3.	$-e^{-at}u(-t)$	$\frac{1}{(s+a)}$	$\mathrm{Re}(s) < -\mathrm{Re}(a)$
4.	$(-t)^k e^{-at}u(-t)$	$\frac{k!}{(s+a)^{k+1}}$	$\mathrm{Re}(s) < -\mathrm{Re}(a)$
5.	$u(t)$	$\frac{1}{s}$	$\mathrm{Re}(s) > 0$
6.	$\delta(t)$	1	all s
7.	$\frac{d^k\delta(t)}{dt^k}$	s^k	all s
8.	$t^k u(t)$	$\frac{k!}{s^{k+1}}$	$\mathrm{Re}(s) > 0$
9.	$\mathrm{sgn}\, t = \begin{cases} 1, t \geq 0 \\ -1, t < 0 \end{cases}$	$\frac{2}{s}$	$\mathrm{Re}(s) = 0$
10.	$\sin \omega_0 t\ u(t)$	$\frac{\omega_0}{s^2+\omega_0^2}$	$\mathrm{Re}(s) > 0$
11.	$\cos \omega_0 t\ u(t)$	$\frac{s}{s^2+\omega_0^2}$	$\mathrm{Re}(s) > 0$
12.	$e^{-at} \sin \omega_0 t\ u(t)$	$\frac{\omega}{(s+a)^2+\omega_0^2}$	$\mathrm{Re}(s) > -\mathrm{Re}(a)$
13.	$e^{-at} \cos \omega_0 t\ u(t)$	$\frac{s+a}{(s+a)^2+\omega_0^2}$	$\mathrm{Re}(s) > -\mathrm{Re}(a)$

Source: J.A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 133. With permission.

Properties of Laplace Transform

Linearity

Let us obtain the Laplace transform of a signal, $x(t)$, that is composed of a linear combination of two other signals,

$$x(t) = \alpha_1 x_1(t) + \alpha_2 x_2(t)$$

where α_1 and α_2 are constants.

The linearity property indicates that

$$\mathcal{L}[\alpha_1 x_1(t) + \alpha_2 x_2(t)] = \alpha_1 X_1(s) + \alpha_2 X_2(s)$$

and the region of absolute convergence is *at least as large* as that given by the expression

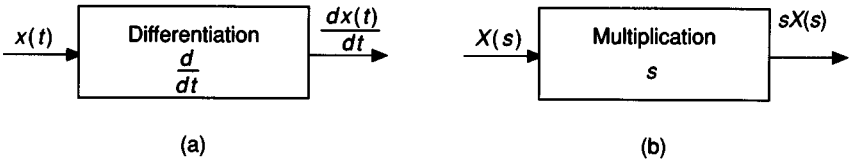


FIGURE 6.1 Equivalent operations in the (a) time-domain operation and (b) Laplace transform-domain operation. (*Source:* J.A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 138. With permission.)

$$\max(\sigma_+^1; \sigma_+^2) < \mathrm{Re}(s) < \min(\sigma_-^1; \sigma_-^2)$$

where the pairs $(\sigma_+^1; \sigma_+^2) < \mathrm{Re}(s) < \min(\sigma_-^1; \sigma_-^2)$ identify the regions of convergence for the Laplace transforms $X_1(s)$ and $X_2(s)$, respectively.

Time-Domain Differentiation

The operation of time-domain differentiation has then been found to correspond to a multiplication by s in the Laplace variable s domain.

The Laplace transform of differentiated signal $dx(t)/dt$ is

$$\mathcal{L}\left[\frac{dx(t)}{dt}\right] = sX(s)$$

Furthermore, it is clear that the region of absolute convergence of $dx(t)/dt$ is at least as large as that of $x(t)$. This property may be envisioned as shown in Fig. 6.1.

Time Shift

The signal $x(t - t_0)$ is said to be a version of the signal $x(t)$ right shifted (or delayed) by t_0 seconds. Right shifting (delaying) a signal by a t_0 second duration in the time domain is seen to correspond to a multiplication by e^{-st0} in the Laplace transform domain. The desired Laplace transform relationship is

$$\mathcal{L}[x(t - t_0)] = e^{-st_0}X(s)$$

where $X(s)$ denotes the Laplace transform of the unshifted signal $x(t)$. As a general rule, any time a term of the form e^{-st0} appears in $X(s)$, this implies some form of time shift in the time domain. This most important property is depicted in Fig. 6.2. It should be further noted that the regions of absolute convergence for the signals $x(t)$ and $x(t - t_0)$ are identical.

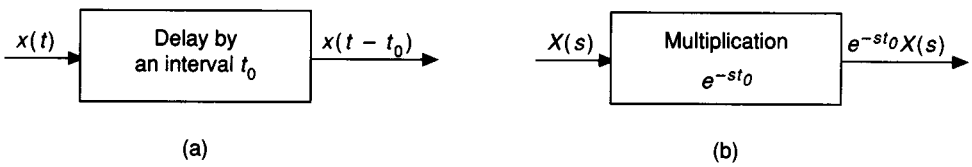


FIGURE 6.2 Equivalent operations in (a) the time domain and (b) the Laplace transform domain. (*Source:* J.A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 140. With permission.)

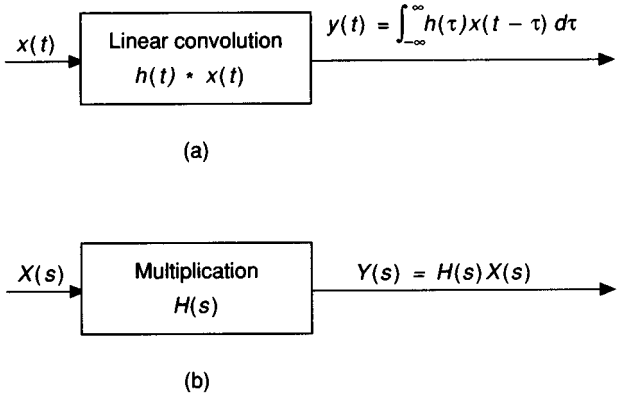


FIGURE 6.3 Representation of a time-invariant linear operator in (a) the time domain and (b) the *s*-domain. *(Source:* J. A. Cadzow and H. F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 144. With permission.)

Time-Convolution Property

The convolution integral signal $y(t)$ can be expressed as

$$y(t) = \int_{-\infty}^{\infty} h(\tau)x(t - \tau)d\tau$$

where $x(t)$ denotes the input signal, the $h(t)$ characteristic signal identifying the operation process.

The Laplace transform of the response signal is simply given by

$$Y(s) = H(s)X(s)$$

where $H(s) = \mathcal{L}[h(t)]$ and $X(s) = \mathcal{L}[x(t)]$. Thus, the convolution of two time-domain signals is seen to correspond to the multiplication of their respective Laplace transforms in the *s*-domain. This property may be envisioned as shown in Fig. 6.3.

Time-Correlation Property

The operation of correlating two signals $x(t)$ and $y(t)$ is formally defined by the integral relationship

$$\phi_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t + \tau)dt$$

The Laplace transform property of the correlation function $\phi_{xy}(\tau)$ is

$$\Phi_{xy}(s) = X(-s)Y(s)$$

in which the region of absolute convergence is given by

$$\max(-\sigma_{x^-}, \sigma_{y^+}) < \mathrm{Re}(s) < \min(-\sigma_{x^+}, \sigma_{y^-})$$

Autocorrelation Function

The autocorrelation function of the signal $x(t)$ is formally defined by

$$\phi_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt$$

The Laplace transform of the autocorrelation function is

$$\Phi_{xx}(s) = X(-s)X(s)$$

and the corresponding region of absolute convergence is

$$\max(-\sigma_{x^-}, \sigma_{y^+}) < \mathrm{Re}(s) < \min(-\sigma_{x^+}, \sigma_{y^-})$$

Other Properties

A number of properties that characterize the Laplace transform are listed in Table 6.2. Application of these properties often enables one to efficiently determine the Laplace transform of seemingly complex time functions.

TABLE 6.2 Laplace Transform Properties

Property	Signal $x(t)$ Time Domain	Laplace Transform $X(s)$ s Domain	Region of Convergence of $X(s)$ $\sigma_+ < \mathrm{Re}(s) < \sigma_-$
Linearity	$\alpha_1 x_1(t) + \alpha_2 x_2(t)$	$\alpha_1 X_1(s) + \alpha_2 X_2(s)$	At least the intersection of the region of convergence of $X_1(s)$ and $X_2(s)$
Time differentiation	$\frac{dx(t)}{dt}$	sX(s)	At least $\sigma_+ < \mathrm{Re(s)}$ and $\mathrm{X_2(s)}$
Time shift	$x(t-t_0)$	$e^{-st0}X(s)$	$\sigma_+ < \mathrm{Re}(s) < \sigma_-$
Time convolution	$\int_{-\infty}^{\infty} h(\tau)\,x(t-\tau)\,dt$	$H(s)X(s)$	At least the intersection of the region of convergence of $H(s)$ and $X(s)$
Time scaling	$x(at)$	$\frac{1}{\lvert a\rvert} X\left(\frac{s}{a}\right)$	$\sigma_+ < \mathrm{Re}\left(\frac{s}{a}\right) < \sigma_-$
Frequency shift	$e^{-at}x(t)$	$X(s+a)$	$\sigma_+ - \mathrm{Re}(a) < \mathrm{Re}(s) < \sigma_- - \mathrm{Re}(a)$
Multiplication (frequency convolution)	$x_1(t)x_2(t)$	$\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} X_1(u)X_2(s-u)\,d$	$\sigma_+^{(1)} + \sigma_+^{(2)} < \mathrm{Re}(s) < \sigma_-^{(1)} + \sigma_-^{(2)}$ $\sigma_+^{(1)} + \sigma_+^{(2)} < c < \sigma_-^{(1)} + \sigma_-^{(2)}$
Time integration	$\int_{-\infty}^{t} x(\tau)\,d\tau$	$\frac{1}{s}X(s)$ for $X(0) =$	At least $\sigma_+ < \mathrm{Re}(s) < \sigma_-$
Frequency differentiation	$(-t)^k x(t)$	$\frac{d^k X(s)}{ds^k}$	At least $\sigma_+ < \mathrm{Re}(s) < \sigma_-$
Time correlation	$\int_{-\infty}^{+\infty} x(t)\,y(t+z)dt$	$X(-s)\,Y(s)$	$\max(-\sigma_{x-}, \sigma_{y+}) < \mathrm{Re}(s) < \min(-\sigma_{x+}, \sigma_{y-})$
Autocorrelation function	$\int_{-\infty}^{+\infty} x(t)\,x(t+z)dt$	$X(-s)\,X(s)$	$\max(-\sigma_{x-}, \sigma_{x+}) < \mathrm{Re}(s) < \min(-\sigma_{x+}, \sigma_{x-})$

Source: J. A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.

Inverse Laplace Transform

Given a transform function $X(s)$ and its region of convergence, the procedure for finding the signal $x(t)$ that generated that transform is called *finding the inverse Laplace transform* and is symbolically denoted as

$$x(t) = \mathcal{L}^{\pm 1}[X(s)]$$

The signal $x(t)$ can be recovered by means of the relationship

$$x(t) = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} X(s)e^{st}ds$$

In this integral, the real number c is to be selected so that the complex number $c + j\omega$ lies entirely within the region of convergence of $X(s)$ for all values of the imaginary component ω. For the important class of rational Laplace transform functions, there exists an effective alternate procedure that does not necessitate directly evaluating this integral. This procedure is generally known as the *partial-fraction expansion method.*

Partial Fraction Expansion Method

As just indicated, the partial fraction expansion method provides a convenient technique for reacquiring the signal that generates a given rational Laplace transform. Recall that a transform function is said to be rational if it is expressible as a ratio of polynomial in s, that is,

$$X(s) = \frac{B(s)}{A(s)} = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0}$$

The partial fraction expansion method is based on the appealing notion of equivalently expressing this rational transform as a sum of n elementary transforms whose corresponding inverse Laplace transforms (i.e., generating signals) are readily found in standard Laplace transform pair tables. This method entails the simple five-step process as outlined in Table 6.3. A description of each of these steps and their implementation is now given.

I. Proper Form for Rational Transform. This division process yields an expression in the proper form as given by

$$\begin{aligned} X(s) &= \frac{B(s)}{A(s)} \\ &= Q(s) + \frac{R(s)}{A(s)} \end{aligned}$$

TABLE 6.3 Partial Fraction Expansion Method for Determining the Inverse Laplace Transform

I.	Put rational transform into proper form whereby the degree of the numerator polynomial is less than or equal to that of the denominator polynomial.
II.	Factor the denominator polynomial.
III.	Perform a partial fraction expansion.
IV.	Separate partial fraction expansion terms into causal and anticausal components using the associated region of absolute convergence for this purpose.
V.	Using a Laplace transform pair table, obtain the inverse Laplace transform.

Source: J. A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 153. With permission.

in which $Q(s)$ and $R(s)$ are the quotient and remainder polynomials, respectively, with the division made so that the degree of $R(s)$ is less than or equal to that of $A(s)$.

II. Factorization of Denominator Polynomial. The next step of the partial fraction expansion method entails the factorizing of the nth-order denominator polynomial $A(s)$ into a product of n first-order factors. This factorization is always possible and results in the equivalent representation of $A(s)$ as given by

$$A(s) = (s - p_1)(s - p_2)\ldots(s - p_n)$$

The terms $p_1, p_2, \ldots, p_n$ constituting this factorization are called the *roots of polynomial* $A(s)$, or the *poles of* $X(s)$.

III. Partial Fraction Expansion. With this factorization of the denominator polynomial accomplished, the rational Laplace transform $X(s)$ can be expressed as

$$X(s) = \frac{B(s)}{A(s)} = \frac{b_n s^n + b_{n-1}s^{n-1} + \cdots + b_0}{(s - p_1)(s - p_2)\cdots(s - p_n)} \tag{6.2}$$

We shall now *equivalently represent* this transform function as a linear combination of elementary transform functions.

Case 1: $A(s)$ Has Distinct Roots.

$$X(s) = \alpha_0 + \frac{\alpha_1}{s - p_1} + \frac{\alpha_2}{s - p_2} + \cdots + \frac{\alpha_n}{s - p_n}$$

where the α_k are constants that identify the expansion and must be properly chosen for a valid representation.

$$\alpha_k = (s - p_k)X(s)\Big|_{s=p_k} \quad \text{for } k = 1, 2, \ldots, n$$

and

$$\alpha_0 = b_n$$

The expression for parameter α_0 is obtained by letting s become unbounded (i.e., $s = +\infty$) in expansion (6.2).

Case 2: $A(s)$ Has Multiple Roots.

$$X(s) = \frac{B(s)}{A(s)} = \frac{B(s)}{(s - p_1)^q A_1(s)}$$

The appropriate partial fraction expansion of this rational function is then given by

$$X(s) = \alpha_0 + \frac{\alpha_1}{(s - p_1)^1} + \cdots + \frac{\alpha_q}{(s - p_1)^q} + (n - q)\ \text{other elementary terms due to the roots of } A_1(s)$$

The coefficient α_0 may be expediently evaluated by letting s approach infinity, whereby each term on the right side goes to zero except α_0. Thus,

$$\alpha_0 = \lim_{s \to +\infty} X(s) = 0$$

The α_q coefficient is given by the convenient expression

$$\begin{aligned} \alpha_q &= (s - p_1)^q X(s) \Big|_{s=p_1} \\ &= \frac{B(p_1)}{A_1(p_1)} \end{aligned} \tag{6.3}$$

The remaining coefficients $\alpha_1, \alpha_2, \ldots, \alpha_{q-1}$ associated with the multiple root p_1 may be evaluated by solving Eq. (6.3) by setting s to a specific value.

IV. Causal and Anticausal Components. In a partial fraction expansion of a rational Laplace transform $X(s)$ whose region of absolute convergence is given by

$$\sigma_+ < \mathrm{Re}(s) < \sigma_-$$

it is possible to decompose the expansion's elementary transform functions into causal and anticausal functions (and possibly impulse-generated terms). Any elementary function is interpreted as being (1) *causal* if the real component of its pole is less than or equal to σ_+ and (2) *anticausal* if the real component of its pole is greater than or equal to σ_-.

The poles of the rational transform that lie to the left (right) of the associated region of absolute convergence correspond to the causal (anticausal) component of that transform. Figure 6.4 shows the location of causal and anticausal poles of rational transform.

V. Table Look-Up of Inverse Laplace Transform. To complete the inverse Laplace transform procedure, one need simply refer to a standard Laplace transform function table to determine the time signals that generate each of the elementary transform functions. The required time signal is then equal to the same linear combination of the inverse Laplace transforms of these elementary transform functions.

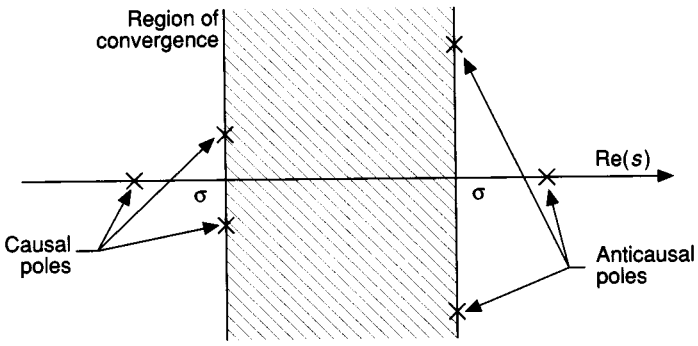


FIGURE 6.4 Location of causal and anticausal poles of a rational transform. (*Source:* J.A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 161. With permission.)

Defining Terms

Laplace transform: A transformation of a function $f(t)$ from the time domain into the complex frequency domain yielding $F(s)$.

$$F(s) = \int_{-\infty}^{\infty} f(t)e^{-st}dt$$

where $s = \sigma + j\omega$.

Region of absolute convergence: The set of complex numbers s for which the magnitude of the Laplace transform integral is finite. The region can be expressed as

$$\sigma_+ < \mathrm{Re}(s) < \sigma_-$$

where σ_+ and σ_- denote real parameters that are related to the causal and anticausal components, respectively, of the signal whose Laplace transform is being sought.

Related Topic

4.1 Introduction

References

J.A. Cadzow and H.F. Van Landingham, *Signals, Systems, and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985.

E. Kamen, *Introduction to Signals and Systems,* 2nd Ed., Englewood Cliffs, N.J.: Prentice-Hall, 1990.

B.P. Lathi, *Signals and Systems,* Carmichael, Calif.: Berkeley-Cambridge Press, 1987.

6.2 Applications[1]

David E. Johnson

In applications such as electric circuits, we start counting time at $t = 0$, so that a typical function $f(t)$ has the property $f(t) = 0$, $t < 0$. Its transform is given therefore by

$$F(s) = \int_{0}^{\infty} f(t)e^{-st}\ dt$$

which is sometimes called the *one-sided Laplace transform.* Since $f(t)$ is like $x(t)u(t)$ we may still use Table 6.1 of the previous section to look up the transforms, but for simplicity we will omit the factor $u(t)$, which is understood to be present.

Differentiation Theorems

Time-Domain Differentiation

If we replace $f(t)$ in the one-sided transform by its derivative $f'(t)$ and integrate by parts, we have the transform of the derivative,

[1]Based on D.E. Johnson, J.R. Johnson, and J.L. Hilburn, *Electric Circuit Analysis,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1992, chapters 19 and 20. With permission.

$$\mathcal{L}[f'(t)] = sF(s) - f(0) \tag{6.4}$$

We may formally replace f by f' to obtain

$$\mathcal{L}[f''(t)] = s\,\mathcal{L}\,[f'(t)] - f'(0)$$

or by (6.4),

$$\mathcal{L}[f''(t)] = s^2F(s) - sf(0) - f'(0) \tag{6.5}$$

We may replace f by f' again in (6.5) to obtain $\mathcal{L}\,[f'''(t)]$, and so forth, obtaining the general result,

$$\mathcal{L}[f^{(n)}(t)] = s^nF(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0) \tag{6.6}$$

where $f^{(n)}$ is the nth derivative. The functions $f, f', \ldots, f^{(n-1)}$ are assumed to be continuous on $(0,\infty)$, and $f^{(n)}$ is continuous except possibly for a finite number of finite discontinuities.

Example 6.2.1 As an example, let $f(t) = t^n$, for n a nonnegative integer. Then $f^{(n)}(t) = n!$ and $f(0) = f'(0) = \ldots = f^{(n-1)}(0) = 0$. Therefore, we have

$$\mathcal{L}[n!\,] = s^n\mathcal{L}\,[t^n]$$

or

$$\mathcal{L}[t^n] = \frac{1}{s^n}\,\mathcal{L}\,[n!\,] = \frac{n!}{s^{n+1}}; \quad n = 0, 1, 2, \ldots \tag{6.7}$$

■

Example 6.2.2 As another example, let us invert the transform

$$F(s) = \frac{8}{s^3(s+2)}$$

which has the partial fraction expansion

$$F(s) = \frac{A}{s^3} + \frac{B}{s^2} + \frac{C}{s} + \frac{D}{s+2}$$

where

$$A = s^3F(s)\Big|_{s=0} = 4$$

and

$$D = (s+2)F(s)\Big|_{s=-2} = -1$$

To obtain B and C, we clear $F(s)$ of fractions, resulting in

$$8 = 4(s + 2) + Bs(s + 2) + Cs^2(s + 2) - s^3$$

Equating coefficients of s^3 yields $C = 1$, and equating those of s^2 yields $B = -2$. The transform is therefore

$$F(s) = 2\frac{2!}{s^3} - 2\frac{1!}{s^2} + \frac{1}{s} - \frac{1}{s + 2}$$

so that

$$f(t) = 2t^2 - 2t + 1 - e^{-2t}$$ ■

Frequency-Domain Differentiation

Frequency-domain differentiation formulas may be obtained by differentiating the Laplace transform with respect to s. That is, if $F(s) = \mathcal{L}[f(t)]$,

$$\frac{dF(s)}{ds} = \frac{d}{ds}\int_0^\infty f(t)e^{-st}dt$$

Assuming that the operations of differentiation and integration may be interchanged, we have

$$\frac{dF(s)}{ds} = \int_0^\infty \frac{d}{ds}[f(t)e^{-st}]\,dt$$

$$= \int_0^\infty [-tf(t)]\,e^{-st}dt$$

From the last integral it follows by definition of the transform that

$$\mathcal{L}[tf(t)] = -\frac{dF(s)}{ds} \tag{6.8}$$

Example 6.2.3 As an example, if $f(t) = \cos kt$, then $F(s) = s/(s^2 + k^2)$, and we have

$$\mathcal{L}[t\cos kt] = -\frac{d}{ds}\left(\frac{s}{s^2 + k^2}\right) = \frac{s^2 - k^2}{(s^2 + k^2)^2}$$ ■

We may repeatedly differentiate the transform to obtain the general case

$$\frac{d^nF(s)}{ds^n} = \int_0^\infty [(-t)^n f(t)]\,e^{-st}dt$$

from which we conclude that

TABLE 6.4 One-Sided Laplace Transform Properties

	$f(t)$	$F(s)$
1.	$cf(t)$	$cF(s)$
2.	$f_1(t) + f_2(t)$	$F_1(s) + F_2(s)$
3.	$\dfrac{df(t)}{dt}$	$sF(s) - f(0)$
4.	$\dfrac{d^n f(t)}{dt^n}$	$s^n F(s) - s^{n-1} f(0) - s^{n-2} f'(0) - s^{n-1} f''(0) - \cdots - f^{(n-1)}(0)$
5.	$\displaystyle\int_0^t f(\tau)\, d\tau$	$\dfrac{F(s)}{s}$
6.	$e^{-at} f(t)$	$F(s + a)$
7.	$f(t - \tau)u(t - \tau)$	$e^{-s\tau} F(s)$
8.	$f * g = \displaystyle\int_0^t f(\tau) g(t - \tau)\, d\tau$	$F(s)G(s)$
9.	$f(ct), c > 0$	$\dfrac{1}{c} F\left(\dfrac{s}{c}\right)$
10.	$t^n f(t), n = 0,1,2, \ldots$	$(-1)^n F^{(n)}(s)$

$$\mathcal{L}[t^n f(t)] = (-1)^n \frac{d^n F(s)}{ds^n}; \quad n = 0, 1, 2, \ldots \tag{6.9}$$

Properties of the Laplace transform obtained in this and the previous section are listed in Table 6.4.

Applications to Integrodifferential Equations

If we transform both members of a linear differential equation with constant coefficients, the result will be an algebraic equation in the transform of the unknown variable. This follows from Eq. (6.6), which also shows that the initial conditions are automatically taken into account. The transformed equation may then be solved for the transform of the unknown and inverted to obtain the time-domain answer.

Thus, if the differential equation is

$$a_n x^{(n)} + a_{n-1} x^{(n-1)} + \ldots + a_0 x = f(t)$$

the transformed equation is

$$\begin{aligned} &a_n\left[s^n X(s) - s^{n-1} x(0) - \ldots - x^{(n-1)}(0)\right] \\ &+ a_{n-1}\left[s^{n-1} X(s) - s^{n-2} x(0) - \ldots - x^{(n-2)}(0)\right] \\ &+ \ldots + a_0 X(s) = F(s) \end{aligned}$$

The transform $X(s)$ may then be found and inverted to give $x(t)$.

Example 6.2.4 As an example, let us find the solution $x(t)$, for $t > 0$, of the system of equations

$$x'' + 4x' + 3x = e^{-2t}$$
$$x(0) = 1, \quad x'(0) = 2$$

Transforming, we have

$$s^2X(s) - s - 2 + 4[sX(s) - 1] + 3X(s) = \frac{1}{s+2}$$

from which

$$X(s) = \frac{s^2 + 8s + 13}{(s+1)(s+2)(s+3)}$$

The partial fraction expansion is

$$X(s) = \frac{3}{s+1} - \frac{1}{s+2} - \frac{1}{s+3}$$

from which

$$x(t) = 3e^{-t} - e^{-2t} - e^{-3t}$$

■

Integration Property

Certain integrodifferential equations may be transformed directly without first differentiating to remove the integrals. We need only transform the integrals by means of

$$\mathcal{L}\left[\int_0^t f(\tau)\, d\tau\right] = \frac{F(s)}{s}$$

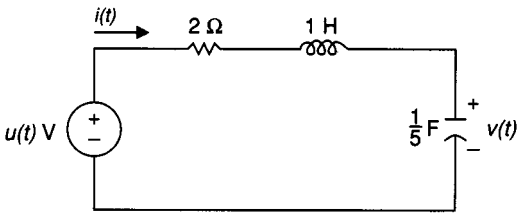


FIGURE 6.5 An *RLC* circuit.

Example 6.2.5 As an example, the current $i(t)$ in Fig. 6.5, with no initial stored energy, satisfies the system of equations,

$$\frac{di}{dt} + 2i + 5\int_0^t i\, dt = u(t)$$
$$i(0) = 0$$

Transforming yields

$$sI(s) + 2I(s) + \frac{5}{s} I(s) = \frac{1}{s}$$

or

$$I(s) = \frac{1}{s^2 + 2s + 5} = \frac{1}{2}\left[\frac{2}{(s+1)^2 + 4}\right]$$

Therefore the current is

$$i(t) = 0.5e^{-t} \sin 2t \quad \text{A}$$ ■

Applications to Electric Circuits

As the foregoing example shows, the Laplace transform method is an elegant procedure than can be used for solving electric circuits by transforming their describing integrodifferential equations into algebraic equations and applying the rules of algebra. If there is more than one loop or nodal equation, their transformed equations are solved simultaneously for the desired circuit current or voltage transforms, which are then inverted to obtain the time-domain answers. Superposition is not necessary because the various source functions appearing in the equations are simply transformed into algebraic quantities.

The Transformed Circuit

Instead of writing the describing circuit equations, transforming the results, and solving for the transform of the circuit current or voltage, we may go directly to a **transformed circuit**, which is the original circuit with the currents, voltages, sources, and passive elements replaced by transformed equivalents. The current or voltage transforms are then found using ordinary circuit theory and the results inverted to the time-domain answers.

Voltage Law Transformation

First, let us note that if we transform Kirchhoff's voltage law,

$$v_1(t) + v_2(t) + \cdots + v_n(t) = 0$$

we have

$$V_1(s) + V_2(s) + \cdots + V_n(s) = 0$$

where $V_i(s)$ is the transform of $v_i(t)$.The transformed voltages thus satisfy Kirchhoff's voltage law. A similar procedure will show that transformed currents satisfy Kirchhoff's current law, as well. Next, let us consider the passive elements. For a resistance R, with current i_R and voltage v_R, for which

$$v_R = Ri_R$$

the transformed equation is

$$V_R(s) = RI_R(s) \tag{6.10}$$

This result may be represented by the transformed resistor element of Fig. 6.6(a).

Inductor Transformation

For an inductance L, the voltage is

$$v_L = L\, di_L/dt$$

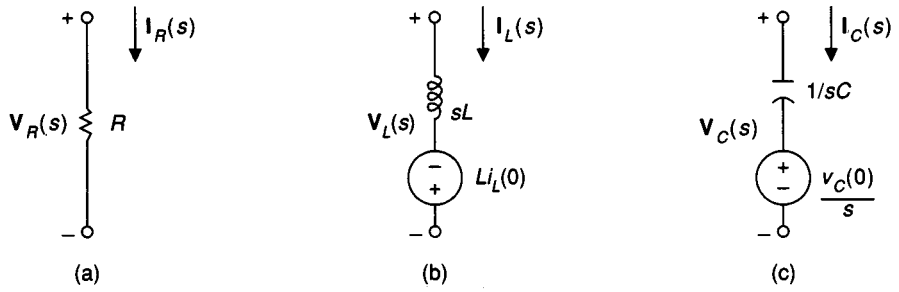


FIGURE 6.6 Transformed circuit elements.

Transforming, we have

$$V_L(s) = sLI_L(s) - Li_L(0) \tag{6.11}$$

which may be represented by an inductor with impedance sL in series with a source, $Li_L(0)$, with the proper polarity, as shown in Fig. 6.6(b). The included voltage source takes into account the initial condition $i_L(0)$.

Capacitor Transformation

In the case of a capacitance C we have

$$v_C = \frac{1}{C}\int_0^t i_C \, dt + v_C(0)$$

which transforms to

$$V_C(s) = \frac{1}{sC} I_C(s) + \frac{1}{s} v_C(0) \tag{6.12}$$

This is represented in Fig. 6.6(c) as a capacitor with impedance $1/sC$ in series with a source, $v_C(0)/s$, accounting for the initial condition.

We may solve Eqs. (6.10), (6.11), and (6.12) for the transformed currents and use the results to obtain alternate transformed elements useful for nodal analysis, as opposed to those of Fig. 6.6, which are ideal for loop analysis. The alternate elements are shown in Fig. 6.7.

Source Transformation

Independent sources are simply labeled with their transforms in the transformed circuit. Dependent sources are transformed in the same way as passive elements. For example, a controlled voltage source defined by

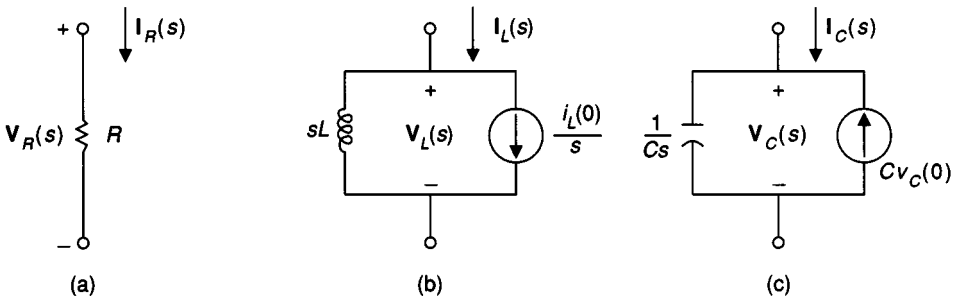


FIGURE 6.7 Transformed elements useful for nodal analysis.

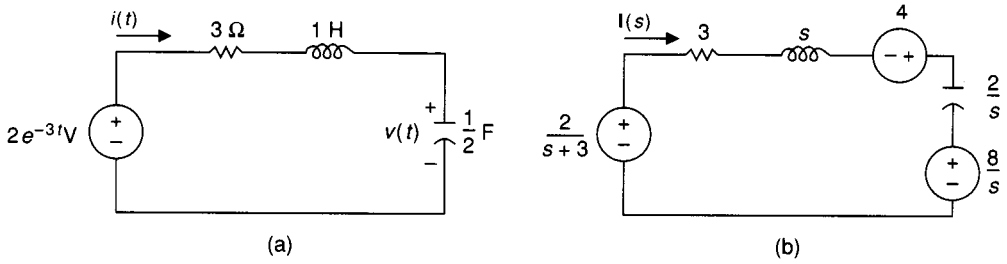


FIGURE 6.8 (a) A circuit and (b) its transformed counterpart.

$$v_1(t) = Kv_2(t)$$

transforms to

$$V_1(s) = KV_2(s)$$

which in the transformed circuit is the transformed source controlled by a transformed variable. Since Kirchhoff's laws hold and the rules for impedance hold, the transformed circuit may be analyzed exactly as we would an ordinary resistive circuit.

Example 6.2.6 To illustrate, let us find $i(t)$ in Fig. 6.8(a), given that $i(0) = 4$ A and $v(0) = 8$ V. The transformed circuit is shown in Fig. 6.8(b), from which we have

$$I(s) = \frac{[2/(s+3)] + 4 - (8/s)}{3 + s + (2/s)}$$

This may be written

$$I(s) = -\frac{13}{s+1} + \frac{20}{s+2} - \frac{3}{s+3}$$

so that

$$i(t) = -13e^{-t} + 20e^{-2t} - 3e^{-3t} \quad \text{A}$$ ■

Thévenin's and Norton's Theorems

Since the procedure using transformed circuits is identical to that using the phasor equivalent circuits in the ac steady-state case, we may obtain transformed Thévenin and Norton equivalent circuits exactly as in the phasor case. That is, the Thévenin impedance will be $Z_{th}(s)$ seen at the terminals of the transformed circuit with the sources made zero, and the open-circuit voltage and the short-circuit current will be $V_{oc}(s)$ and $I_{sc}(s)$, respectively, at the circuit terminals. The procedure is exactly like that for resistive circuits, except that in the transformed circuit the quantities involved are functions of s. Also, as in the resistor and phasor cases, the open-circuit voltage and short-circuit current are related by

$$V_{oc}(s) = Z_{th}(s)I_{sc}(s) \tag{6.13}$$

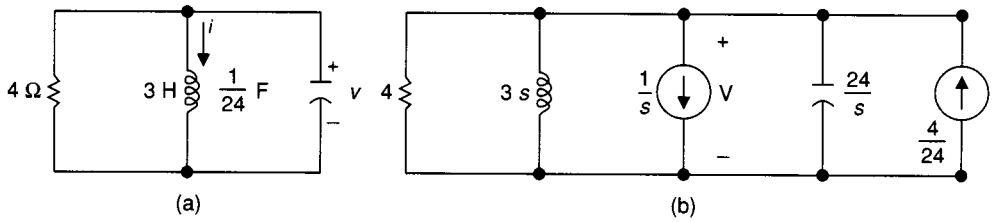


FIGURE 6.9 (a) An *RLC* parallel circuit and (b) its transformed circuit.

Example 6.2.7 As an example, let us consider the circuit of Fig. 6.9(a) with the transformed circuit shown in Fig. 6.9(b). The initial conditions are $i(0) = 1$ A and $v(0) = 4$ V. Let us find $v(t)$ for $t > 0$ by replacing everything to the right of the 4-Ω resistor in Fig. 6.9(b) by its Thévenin equivalent circuit. We may find $Z_{th}(s)$ directly from Fig. 6.9(b) as the impedance to the right of the resistor with the two current sources made zero (open circuited). For illustrative purposes we choose, however, to find the open-circuit voltage and short-circuit current shown in Figs. 6.10(a) and (b), respectively, and use Eq. (6.13) to get the Thévenin impedance.

The nodal equation in Fig. 6.10(a) is

$$\frac{V_{oc}(s)}{3s} + \frac{1}{s} + \frac{s}{24} V_{oc}(s) = \frac{1}{6}$$

from which we have

$$V_{oc}(s) = \frac{4(s-6)}{s^2 + 8}$$

From Fig. 6.10(b)

$$I_{sc}(s) = \frac{s-6}{6s}$$

The Thévenin impedance is therefore

$$Z_{th}(s) = \frac{V_{oc}(s)}{I_{sc}(s)} = \frac{\left[\dfrac{4(s-6)}{s^2+8}\right]}{\left[\dfrac{s-6}{6s}\right]} = \frac{24s}{s^2+8}$$

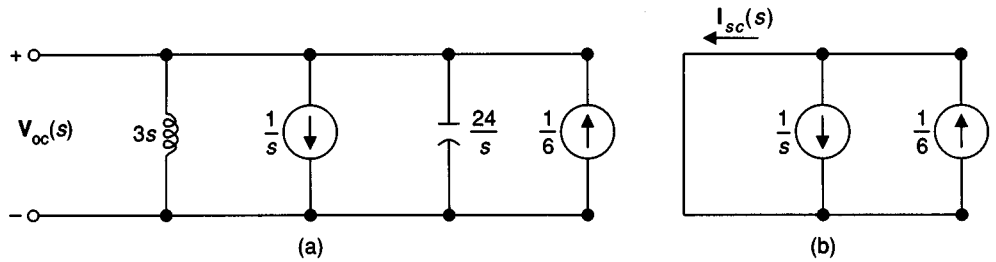


FIGURE 6.10 Circuit for obtaining (a) $V_{oc}(s)$ and (b) $I_{sc}(s)$.

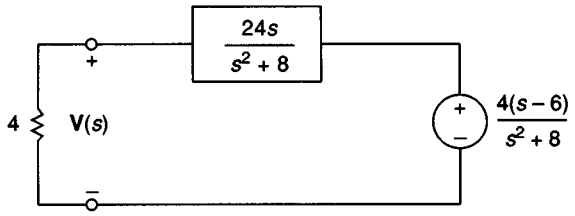


FIGURE 6.11 Thévenin equivalent circuit terminated in a resistor.

and the Thévenin equivalent circuit, with the 4 Ω connected, is shown in Fig. 6.11. From this circuit we find the transform

$$V(s) = \frac{4(s-6)}{(s+2)(s+4)} = \frac{-16}{s+2} + \frac{20}{s+4}$$

from which

$$v(t) = -16e^{-2t} + 20e^{-4t} \quad \text{V}$$ ■

Network Functions

A **network function** or **transfer function** is the ratio $H(s)$ of the Laplace transform of the output function, say $v_o(t)$, to the Laplace transform of the input, say $v_i(t)$, assuming that there is only one input. (If there are multiple inputs, the transfer function is based on one of them with the others made zero.) Suppose that in the general case the input and output are related by the differential equation

$$a_n \frac{d^n v_o}{dt^n} + a_{n-1} \frac{d^{n-1} v_o}{dt^{n-1}} + \cdots + a_1 \frac{dv_o}{dt} + a_o v_o = b_m \frac{d^m v_i}{dt^m} + b_{m-1} \frac{d^{m-1} v_i}{dt^{m-1}} + \cdots + b_1 \frac{dv_i}{dt} + b_o v_i$$

and that the initial conditions are all zero; that is,

$$v_o(0) = \frac{dv_o(0)}{dt} = \cdots = \frac{d^{n-1} v_o(0)}{dt^{n-1}} = v_i(0) = \frac{dv_i(0)}{dt} = \cdots = \frac{d^{m-1} v_i(0)}{dt^{m-1}} = 0$$

Then, transforming the differential equation results in

$$(a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0) V_o(s) = (b_m s^m + b_{m-1} s^{m-1} + \cdots + b_1 s + b_0) V_i(s)$$

from which the network function, or transfer function, is given by

$$H(s) = \frac{V_o(s)}{V_i(s)} = \frac{b_m s^m + b_{m-1} s^{m-1} + \cdots + b_1 s + b_0}{a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0} \tag{6.14}$$

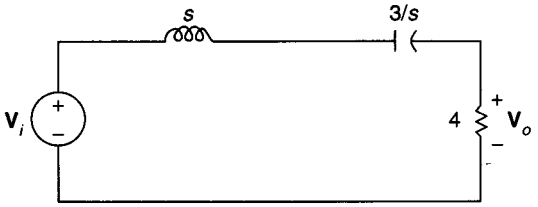


FIGURE 6.12 An *RLC* circuit.

Example 6.2.8 As an example, let us find the transfer function for the transformed circuit of Fig. 6.12, where the transfer function is $V_o(s)/V_i(s)$. By voltage division we have

$$H(s) = \frac{V_o(s)}{V_i(s)} = \frac{4}{s + 4 + (3/s)} = \frac{4s}{(s + 1)(s + 3)} \tag{6.15}$$

■

Step and Impulse Responses

In general, if $Y(s)$ and $X(s)$ are the transformed output and input, respectively, then the network function is $H(s) = Y(s)/X(s)$ and the output is

$$Y(s) = H(s)X(s) \tag{6.16}$$

The **step response** $r(t)$ is the output of a circuit when the input is the unit step function $u(t)$, with transform $1/s$. Therefore, the transform of the step response $R(s)$ is given by

$$R(s) = H(s)/s \tag{6.17}$$

The **impulse response** $h(t)$ is the output when the input is the unit impulse $\delta(t)$. Since $\mathcal{L}[\delta(t)] = 1$, we have from Eq. (6.16),

$$h(t) = \mathcal{L}^{-1}[H(s)/1] = \mathcal{L}^{-1}[H(s)] \tag{6.18}$$

Example 6.2.9 As an example, for the circuit of Fig. 6.12, $H(s)$, given in Eq. (6.15), has the partial fraction expansion,

$$H(s) = \frac{-2}{s + 1} + \frac{6}{s + 3}$$

so that

$$h(t) = -2e^{-t} + 6e^{-3t} \quad \text{V}$$

■

If we know the impulse response, we can find the transfer function,

$$H(s) = \mathcal{L}[h(t)]$$

from which we can find the response to *any* input. In the case of the step and impulse responses, it is understood that there are no other inputs except the step or the impulse. Otherwise, the transfer function would not be defined.

Stability

An important concern in circuit theory is whether the output signal remains bounded or increases indefinitely following the application of an input signal. An unbounded output could damage or even destroy the circuit, and thus it is important to know before applying the input if the circuit can accommodate the expected output. This question can be answered by determining the *stability* of the circuit.

A circuit is defined to have **bounded input–bounded output** (BIBO) stability if any bounded input results in a bounded output. The circuit in this case is said to be **absolutely stable** or *unconditionally stable*. BIBO stability can be determined by examining the *poles* of the network function (6.14).

If the denominator of $H(s)$ in Eq. (6.14) contains a factor $(s - p)^n$, then p is said to be a pole of $H(s)$ of *order n*. The output $V_o(s)$ would also contain this factor, and its partial fraction expansion would contain the term $K/(s - p)^n$. Thus, the inverse transform $v_o(t)$ is of the form

$$v_o(t) = A_n t^{n-1} e^{pt} + A_{n-1} t^{n-2} e^{pt} + \cdots + A_1 e^{pt} + v_1(t) \tag{6.19}$$

where $v_1(t)$ results from other poles of $V_o(s)$. If p is a real positive number or a complex number with a positive real part, $v_o(t)$ is unbounded because e^{pt} is a growing exponential. Therefore, for absolute stability there can be no pole of $V_o(s)$ that is positive or has a positive real part. This is equivalent to saying that $V_o(s)$ has no poles in the right half of the s-plane. Since $v_i(t)$ is bounded, $V_i(s)$ has no poles in the right half-plane. Therefore, since the only poles of $V_o(s)$ are those of $H(s)$ and $V_i(s)$, no pole of $H(s)$ for an absolutely stable circuit can be in the right-half of the s-plane.

From Eq. (6.19) we see that $v_i(t)$ is bounded, as far as pole p is concerned, if p is a *simple* pole (of order 1) and is purely imaginary. That is, $p = j\omega$, for which

$$e^{pt} = \cos \omega t + j \sin \omega t$$

which has a bounded magnitude. Unless $V_i(s)$ contributes an identical pole $j\omega$, $v_o(t)$ is bounded. Thus, $v_o(t)$ is bounded on the *condition* that any $j\omega$ pole of $H(s)$ is simple.

In summary, a network is *absolutely stable* if its network function $H(s)$ has only left half-plane poles. It is **conditionally stable** if $H(s)$ has only simple $j\omega$-axis poles and possibly left half-plane poles. It is *unstable* otherwise (right half-plane or multiple $j\omega$-axis poles).

Example 6.2.10 As an example, the circuit of Fig. 6.12 is absolutely stable, since from Eq. (6.15) the only poles of its transfer function are $s = -1, -3$, which are both in the left half-plane. There are countless examples of conditionally stable circuits that are extremely useful, for example, a network consisting of a single capacitor with $C = 1$ F with input current $I(s)$ and output voltage $V(s)$. The transfer function is $H(s) = Z(s) = 1/Cs = 1/s$, which has the simple pole $s = 0$ on the $j\omega$-axis. Figure 6.13 illustrates a circuit which is unstable. The transfer function is

$$H(s) = I(s)/V_i(s) = 1/(s - 2)$$

which has the right half-plane pole $s = 2$. ■

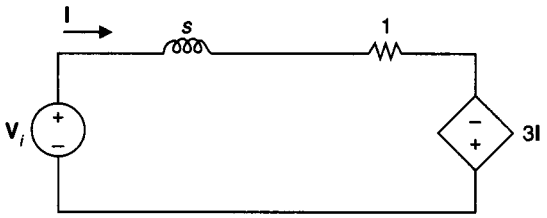


FIGURE 6.13 Unstable circuit.

Defining Terms

Absolute stability: When the network function $H(s)$ has only left half-plane poles.
Bounded input–bounded output stability: When any bounded input results in a bounded output.
Conditional stability: When the network function $H(s)$ has only simple $j\omega$-axis poles and possibly left half-plane poles.
Impulse response, $h(t)$: The output when the input is the unit impulse $\delta(t)$.
Network or transfer function: The ratio $H(s)$ of the Laplace transform of the output function to the Laplace transform of the input function.
Step response, $r(t)$: The output of a circuit when the input is the unit step function $u(t)$, with transform $1/s$.
Transformed circuit: An original circuit with the currents, voltages, sources, and passive elements replaced by transformed equivalents.

Related Topics

3.1 Voltage and Current Laws • 3.3 Network Theorems • 12.1 Introduction

References

R.C. Dorf, *Introduction to Electric Circuits,* 2nd ed., New York: John Wiley, 1993.
J.D. Irwin, *Basic Engineering Circuit Analysis,* 3rd ed., New York: Macmillan, 1989.
D.E. Johnson, J.R. Johnson, J.L. Hilburn, and P.D. Scott, *Electric Circuit Analysis,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1997.
J.W. Nilsson, *Electric Circuits,* 5th ed., Reading, Mass.: Addison-Wesley, 1996.

Chen, W.K. "State Variables: Concept and Formulation"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

7

State Variables: Concept and Formulation

7.1 Introduction
7.2 State Equations in Normal Form
7.3 The Concept of State and State Variables and Normal Tree
7.4 Systematic Procedure in Writing State Equations
7.5 State Equations for Networks Described by Scalar Differential Equations
7.6 Extension to Time-Varying and Nonlinear Networks

Wai-Kai Chen
University of Illinois, Chicago

7.1 Introduction

An electrical network is describable by a system of algebraic and differential equations known as the primary system of equations obtained by applying the Kirchhoff's current and voltage laws and the element *v-i* relations. In the case of linear networks, these equations can be transformed into a system of linear algebraic equations by means of the Laplace transformation, which is relatively simple to manipulate. The main drawback is that it contains a large number equations. To reduce this number, three secondary systems of equations are available: the nodal system, the cutset system, and the loop system. If a network has n nodes, b branches, and c components, there are $n - c$ linearly independent equations in nodal or cutset analysis and $b - n + c$ linearly independent equations in loop analysis. These equations can then be solved to yield the Laplace transformed solution. To obtain the final time-domain solution, we must take the inverse Laplace transformation. For most practical networks, the procedure is usually long and complicated and requires an excessive amount of computer time.

As an alternative we can formulate the network equations in the time domain as a system of first-order differential equations, which describe the dynamic behavior of the network. Some advantages of representing the network equations in this form are the following. First, such a system has been widely studied in mathematics, and its solution, both analytic and numerical, is known and readily available. Second, the representation can easily and naturally be extended to time-varying and nonlinear networks. In fact, computer-aided solution of time-varying, nonlinear network problems is almost always accomplished using the state-variable approach. Finally, the first-order differential equations can easily be programmed for a digital computer or simulated on an analog computer. Even if it were not for the above reasons, the approach provides an alternative view of the physical behavior of the network.

The term **state** is an abstract concept that may be represented in many ways. If we call the set of instantaneous values of all the branch currents and voltages as the *state* of the network, then the knowledge of the instantaneous values of all these variables determines this instantaneous state. Not all of these instantaneous values are required in order to determine the instantaneous state, however, because some can be calculated from the others. A set of data qualifies to be called the *state* of a system if it fulfills the following two requirements:

1. The state of any time, say, t_0, and the input to the system from t_0 on determine uniquely the state at any time $t > t_0$.

2. The state at time t and the inputs together with some of their derivatives at time t determine uniquely the value of any system variable at the time t.

The state may be regarded as a vector, the components of which are *state variables*. Network variables that are candidates for the state variables are the branch currents and voltages. Our problem is to choose state variables in order to formulate the *state equations*. Like the nodal, cutset, or loop system of equations, the state equations are formulated from the primary system of equations. For our purposes, we shall focus our attention on how to obtain state equations for linear systems.

7.2 State Equations in Normal Form

For a linear network containing k energy storage elements and h independent sources, our objective is to write a system of k first-order differential equations from the primary system of equations, as follows:

$$\dot{x}_i(t) = \sum_{j=1}^{k} a_{ij}x_j(t) + \sum_{j=1}^{h} b_{ij}u_j(t), \qquad (i = 1, 2, \ldots, k) \tag{7.1}$$

In matrix notation, Eq. (7.1) becomes

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_k(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \cdot \\ \cdot \\ \cdot \\ x_k(t) \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1h} \\ b_{21} & b_{22} & \cdots & b_{2h} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ b_{k1} & b_{k2} & \cdots & b_{kh} \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \\ \cdot \\ \cdot \\ \cdot \\ u_h(t) \end{bmatrix} \tag{7.2}$$

or, more compactly,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{7.3}$$

The real functions $x_1(t)$, $x_2(t)$, ..., $x_k(t)$ of the time t are called the **state variables**, and the k-vector $\mathbf{x}(t)$ formed by the state variables is known as the **state vector**. The h-vector $\mathbf{u}(t)$ formed by the h known forcing functions or excitations $u_j(t)$ is referred to as the **input vector**. The coefficient matrices $\mathbf{A}$ and $\mathbf{B}$, depending only upon the network parameters, are of orders $k \times k$ and $k \times h$, respectively. Equation (7.3) is usually called the **state equation in normal form**.

The state variables x_j may or may not be the desired output variables. We therefore must express the desired output variables in terms of the state variables and excitations. In general, if there are q output variables $y_j(t)$ ($j = 1, 2, \ldots, q$) and h input excitations, the **output vector** $\mathbf{y}(t)$ formed by the q output variables $y_j(t)$ can be expressed in terms of the state vector $\mathbf{x}(t)$ and the input vector $\mathbf{u}(t)$ by the matrix equation

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{7.4}$$

where the known coefficient matrices **C** and **D**, depending only on the network parameters, are of orders $q \times k$ and $q \times h$, respectively. Equation (7.4) is called the **output equation**. The state equation, Eq. (7.3), and the output equation, Eq. (7.4), together are known as the **state equations**.

7.3 The Concept of State and State Variables and Normal Tree

Our immediate problem is to choose the network variables as the state variables in order to formulate the state equations. If we call the set of instantaneous values of all the branch currents and voltages the *state* of the network, then the knowledge of the instantaneous values of all these variables determines this instantaneous state. Not all of these instantaneous values are required in order to determine the instantaneous state, however, because some can be calculated from the others. For example, the instantaneous voltage of a resistor can be obtained from its instantaneous current through Ohm's law. The question arises as to the minimum number of instantaneous values of branch voltages and currents that are sufficient to determine completely the instantaneous state of the network.

In a given network, a minimal set of its branch variables is said to be a **complete set of state variables** if their instantaneous values are sufficient to determine completely the instantaneous values of all the branch variables. For a linear time-invariant nondegenerate network, it is convenient to choose the capacitor voltages and inductor currents as the state variables. A **nondegenerate network** is one that contains neither a circuit composed only of capacitors and/or independent or dependent voltage sources nor a cutset composed only of inductors and/or independent or dependent current sources, where a cutset is a minimal subnetwork the removal of which cuts the original network into two connected pieces. Thus, not all the capacitor voltages and inductor currents of a degenerate network can be state variables. To help systematically select the state variables, we introduce the notion of normal tree.

A **tree** of a connected network is a connected subnetwork that contains all the nodes but does not contain any circuit. A **normal tree** of a connected network is a tree that contains all the independent voltage sources, the maximum number of capacitors, the minimum number of inductors, and none of the independent current sources. This definition excludes the possibility of having unconnected networks. In the case of unconnected networks, we can consider the normal trees of the individual components. We remark that the representation of the state of a network is generally not unique, but the state of a network itself is.

7.4 Systematic Procedure in Writing State Equations

In the following we present a systematic step-by-step procedure for writing the state equation for a network. They are a systematic way to eliminate the unwanted variables in the primary system of equations.

1. In a given network N, assign the voltage and current references of its branches.
2. In N select a normal tree T and choose as the state variables the capacitor voltages of T and the inductor currents of the **cotree** $\bar{T}$, the complement of T in N.
3. Assign each branch of T a voltage symbol, and assign each element of $\bar{T}$, called the **link**, a current symbol.
4. Using Kirchhoff's current law, express each tree-branch current as a sum of cotree-link currents, and indicate it in N if necessary.
5. Using Kirchhoff's voltage law, express each cotree-link voltage as a sum of tree-branch voltages, and indicate it in N if necessary.
6. Write the element v-i equations for the passive elements and separate these equations into two groups:
   a. Those element v-i equations for the tree-branch capacitors and the cotree-link inductors
   b. Those element v-i equations for all other passive elements
7. Eliminate the nonstate variables among the equations obtained in the preceding step. **Nonstate variables** are defined as those variables that are neither state variables nor known independent sources.
8. Rearrange the terms and write the resulting equations in normal form.

We illustrate the preceding steps by the following examples.

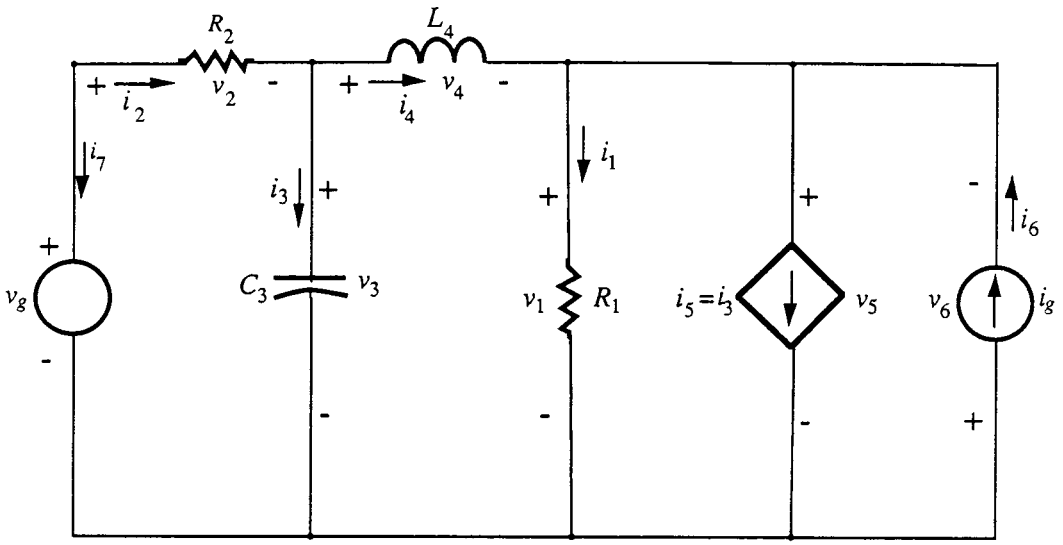


FIGURE 7.1 An active network used to illustrate the procedure for writing the state equations in normal form.

Example 1

We write the state equations for the network N of Fig. 7.1 by following the eight steps outlined above.

Step 1

The voltage and current references of the branches of the active network N are as indicated in Fig. 7.1.

Step 2

Select a normal tree T consisting of the branches R_1, C_3, and v_g. The subnetwork $C_3 i_5 v_g$ is another example of a normal tree.

Step 3

The tree branches R_1, C_3, and v_g are assigned the voltage symbols v_1, v_3, and v_g; and the cotree-links R_2, L_4, i_5, and i_g are assigned the current symbols i_2, i_4, i_3, and i_g, respectively. The controlled current source i_5 is given the current symbol i_3 because its current is controlled by the current of the branch C_3, which is i_3.

Step 4

Applying Kirchhoff's current law, the branch currents i_1, i_3, and i_7 can each be expressed as the sums of cotree-link currents:

$$i_1 = i_4 + i_g - i_3 \tag{7.5a}$$

$$i_3 = i_2 - i_4 \tag{7.5b}$$

$$i_7 = -i_2 \tag{7.5c}$$

Step 5

Applying Kirchhoff's voltage law, the cotree-link voltages v_2, v_4, v_5, and v_6 can each be expressed as the sums of tree-branch voltages:

$$v_2 = v_g - v_3 \tag{7.6a}$$

$$v_4 = v_3 - v_1 \tag{7.6b}$$

$$v_5 = v_1 \tag{7.6c}$$

$$v_6 = -v_1 \tag{7.6d}$$

Step 6

The element *v-i* equations for the tree-branch capacitor and the cotree-link inductor are found to be

$$C_3\dot{v}_3 = i_3 = i_2 - i_4 \tag{7.7a}$$

$$L_4\dot{i}_4 = v_4 = v_3 - v_1 \tag{7.7b}$$

Likewise, the element *v-i* equations for other passive elements are obtained as

$$v_1 = R_1 i_1 = R_1(i_4 + i_g - i_3) \tag{7.8a}$$

$$i_2 = \frac{v_2}{R_2} = \frac{v_g - v_3}{R_2} \tag{7.8b}$$

Step 7

The state variables are the capacitor voltage v_3 and inductor current i_4, and the known independent sources are i_g and v_g. To obtain the state equation, we must eliminate the nonstate variables v_1 and i_2 in Eq. (7.7). From Eqs. (7.5b) and (7.8) we express v_1 and i_2 in terms of the state variables and obtain

$$v_1 = R_1\left(2i_4 + i_g + \frac{v_3}{R_2} - \frac{v_g}{R_2}\right) \tag{7.9a}$$

$$i_2 = \frac{v_g - v_3}{R_2} \tag{7.9b}$$

Substituting these in Eq. (7.7) yields

$$C_3\dot{v}_3 = \frac{v_g - v_3}{R_2} - i_4 \tag{7.10a}$$

$$L_4\dot{i}_4 = \left(1 - \frac{R_1}{R_2}\right)v_3 - 2R_1 i_4 - R_1 i_g + \frac{R_1 v_g}{R_2} \tag{7.10b}$$

Step 8

Equations (7.10a) and (7.10b) are written in matrix form as

$$\begin{bmatrix} \dot{v}_3 \\ \dot{i}_4 \end{bmatrix} = \begin{bmatrix} -\dfrac{1}{R_2 C_3} & -\dfrac{1}{C_3} \\ \dfrac{1}{L_4} & -\dfrac{R_1}{R_2 L_4} & -\dfrac{2R_1}{L_4} \end{bmatrix} \begin{bmatrix} v_3 \\ i_4 \end{bmatrix} + \begin{bmatrix} \dfrac{1}{R_2 C_3} & 0 \\ \dfrac{R_1}{R_2 L_4} & -\dfrac{R_1}{L_4} \end{bmatrix} \begin{bmatrix} v_g \\ i_g \end{bmatrix} \tag{7.11}$$

This is the state equation in normal form for the active network N of Fig. 7.1.

Suppose that resistor voltage v_1 and capacitor current i_3 are the output variables. Then from Eqs. (7.5b) and (7.9) we obtain

$$v_1 = \frac{R_1}{R_2} v_3 + 2R_1 i_4 + R_1 \left(i_g - \frac{v_g}{R_2} \right) \tag{7.12a}$$

$$i_3 = -\frac{v_3}{R_2} - i_4 + \frac{v_g}{R_2} \tag{7.12b}$$

In matrix form, the output equation of the network becomes

$$\begin{bmatrix} v_1 \\ i_3 \end{bmatrix} = \begin{bmatrix} \dfrac{R_1}{R_2} & 2R_1 \\ -\dfrac{1}{R_2} & -1 \end{bmatrix} \begin{bmatrix} v_3 \\ i_4 \end{bmatrix} + \begin{bmatrix} -\dfrac{R_1}{R_2} & R_1 \\ \dfrac{1}{R_2} & 0 \end{bmatrix} \begin{bmatrix} v_g \\ i_g \end{bmatrix} \tag{7.13}$$

Equations (7.11) and (7.13) together are the state equations of the active network of Fig. 7.1.

7.5 State Equations for Networks Described by Scalar Differential Equations

In many situations we are faced with networks that are described by scalar differential equations of order higher than one. Our purpose here is to show that these networks can also be represented by the state equations in normal.

Consider a network that can be described by the nth-order linear differential equation

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \cdots + a_{n-1} \frac{dy}{dt} + a_n y = bu \tag{7.14}$$

Then its state equation can be obtained by defining

$$\begin{aligned} x_1 &= y \\ x_2 &= \dot{x}_1 \\ &\cdot \\ &\cdot \\ &\cdot \\ x_n &= \dot{x}_{n-1} \end{aligned} \tag{7.15}$$

showing that the nth-order linear differential Eq. (7.14) is equivalent to

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= x_3 \\
&\cdot \\
&\cdot \\
&\cdot \\
\dot{x}_{n-1} &= x_n \\
\dot{x}_n &= -a_n x_1 - a_{n-1} x_2 - \quad . \quad . \quad . - a_2 x_{n-1} - a_1 x_n + bu
\end{aligned} \tag{7.16}$$

or, in matrix form,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ b \end{bmatrix} \begin{bmatrix} u \end{bmatrix} \tag{7.17}$$

More compactly, Eq. (7.17) can be written as

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{7.18}$$

The coefficient matrix **A** is called the **companion matrix** of Eq. (7.14), and Eq. (7.17) is the state-equation representation of the network describable by the linear differential equation (7.14).

Let us now consider the more general situation where the right-hand side of (7.14) includes derivatives of the input excitation u. In this case, the different equation takes the general form

$$\begin{aligned}
&\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \ldots + a_{n-1} \frac{dy}{dt} + a_n y \\
&\qquad = b_0 \frac{d^n u}{dt^n} + b_1 \frac{d^{n-1} u}{dt^{n-1}} + \ldots + b_{n-1} \frac{du}{dt} + b_n u
\end{aligned} \tag{7.19}$$

Its state equation can be obtained by defining

$$\begin{aligned}
x_1 &= y - c_0 u \\
x_2 &= \dot{x}_1 - c_1 u \\
&\vdots \\
x_n &= \dot{x}_{n-1} - c_{n-1} u
\end{aligned} \tag{7.20}$$

The general state equation becomes

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix} \begin{bmatrix} u \end{bmatrix} \tag{7.21}$$

where $n > 1$,

$$\begin{aligned} c_1 &= b_1 - a_1 b_0 \\ c_2 &= \left(b_2 - a_2 b_0\right) - a_1 c_1 \\ c_3 &= \left(b_3 - a_3 b_0\right) - a_2 c_1 - a_1 c_2 \\ &\vdots \\ c_n &= \left(b_n - a_3 b_0\right) - a_{n-1} c_1 - a_{n-2} c_2 - \cdots - a_2 c_{n-2} - a_1 c_{n-1} \end{aligned} \tag{7.22}$$

and

$$x_1 = y - b_0 u \tag{7.23}$$

Finally, if y is the output variable, the output equation becomes

$$y(t) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_0 \end{bmatrix} \begin{bmatrix} u \end{bmatrix} \tag{7.24}$$

7.6 Extension to Time-Varying and Nonlinear Networks

A great advantage in the state-variable approach to network analysis is that it can easily be extended to time-varying and nonlinear networks, which are often not readily amenable to the conventional methods of analysis. In these cases, it is more convenient to choose the capacitor charges and inductor flux as the the state variables instead of capacitor voltages and inductor currents.

In the case of a linear time-varying network, its state equations can be written the same as before except that now the coefficient matrices are time-dependent:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{7.25a}$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \tag{7.25b}$$

Thus, with the state-variable approach, it is no more difficult to write the governing equations for a linear time-varying network than it is for a linear time-invariant network. Their solutions are, of course, a different matter.

For a nonlinear network, its state equation in normal form is describable by a coupled set of first-order differential equations:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{7.26}$$

If the function **f** satisfies the familiar Lipshitz condition with respect to **x** in a given domain, then for every set of initial conditions $\mathbf{x}_0(t_0)$ and every input **u** there exists a unique solution $\mathbf{x}(t)$, the components of which are the state variables of the network.

Defining Terms

Companion matrix: The coefficient matrix in the state-equation representation of the network describable by a linear differential equation.

Complete set of state variables: A minimal set of network variables, the instantaneous values of which are sufficient to determine completely the instantaneous values of all the network variables.

Cotree: The complement of a tree in a network.

Cutset: A minimal subnetwork, the removal of which cuts the original network into two connected pieces.

Cutset system: A secondary system of equations using cutset voltages as variables.

Input vector: A vector formed by the input variables to a network.

Link: An element of a cotree.

Loop system: A secondary system of equations using loop currents as variables.

Nodal system: A secondary system of equations using nodal voltages as variables.

Nondegenerate network: A network that contains neither a circuit composed only of capacitors and/or independent or dependent voltage sources nor a cutset composed only of inductors and/or independent or dependent current sources.

Nonstate variables: Network variables that are neither state variables nor known independent sources.

Normal tree: A tree that contains all the independent voltage sources, the maximum number of capacitors, the minimum number of inductors, and none of the independent current sources.

Output equation: An equation expressing the output vector in terms of the state vector and the input vector.

Output vector: A vector formed by the output variables of a network.

Primary system of equations: A system of algebraic and differential equations obtained by applying the Kirchhoff's current and voltage laws and the element *v-i* relations.

Secondary system of equations: A system of algebraic and differential equations obtained from the primary system of equations by transformation of network variables.

State: A set of data, the values of which at any time *t*, together with the input to the system at the time, determine uniquely the value of any network variable at the time *t*.

State equation in normal form: A system of first-order differential equations that describes the dynamic behavior of a network and that is put into a standard form.

State equations: Equations formed by the state equation and the output equation.

State variables: Network variables used to describe the state.

State vector: A vector formed by the state variables.

Tree: A connected subnetwork that contains all the nodes of the original network but does not contain any circuit.

Related Topics

3.1 Voltage and Current Laws • 3.2 Node and Mesh Analysis • 3.7 Two-Port Parameters and Transformations • 5.1 Diodes and Rectifiers • 100.2 Dynamic Response

References

W. K. Chen, *Linear Networks and Systems: Algorithms and Computer-Aided Implementations,* Singapore: World Scientific Publishing, 1990.

W. K. Chen, *Active Network Analysis,* Singapore: World Scientific Publishing, 1991.

L. O. Chua and P. M. Lin, *Computer-Aided Analysis of Electronics Circuits: Algorithms & Computational Techniques,* Englewood Cliffs, N.J.: Prentice-Hall, 1975.

E. S. Kuh and R. A. Rohrer, "State-variables approach to network analysis," *Proc. IEEE,* vol. 53, pp. 672–686, July 1965.

Further Information

An expository paper on the application of the state-variables technique to network analysis was originally written by E. S. Kuh and R. A. Rohrer ("State-variables approach to network analysis," *Proc. IEEE,* vol. 53, pp. 672–686, July 1965). A computer-aided network analysis based on state-variables approach is extensively discussed in the book by Wai-Kai Chen, *Linear Networks and Systems: Algorithms and Computer-Aided Implementations* (World Scientific Publishing Co., Singapore, 1990). The use of state variables in the analysis of electronics circuits and nonlinear networks is treated in the book by L. O. Chua and P. M. Lin, *Computer-Aided Analysis of Electronics Circuits: Algorithms & Computational Techniques* (Prentice-Hall, Englewood Cliffs, N.J., 1975). The application of state-variables technique to active network analysis is contained in the book by Wai-Kai Chen, *Active Network Analysis* (World Scientific Publishing Co., Singapore, 1991).

Dorf, R.C., Wan, Z. "The z-Transfrom"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

8

The *z*-Transform

Richard C. Dorf
University of California, Davis

Zhen Wan
University of California, Davis

8.1 Introduction
8.2 Properties of the *z*-Transform
Linearity • Translation • Convolution • Multiplication by a^n • Time Reversal
8.3 Unilateral *z*-Transform
Time Advance • Initial Signal Value • Final Value
8.4 *z*-Transform Inversion
Method 1 • Method 2 • Inverse Transform Formula (Method 2)
8.5 Sampled Data

8.1 Introduction

Discrete-time signals can be represented as sequences of numbers. Thus, if x is a discrete-time signal, its values can, in general, be indexed by n as follows:

$$x = \{\ldots, x(-2), x(-1), x(0), x(1), x(2), \ldots, x(n), \ldots\}$$

In order to work within a transform domain for discrete-time signals, we define the z-transform as follows. The z-transform of the sequence x in the previous equation is

$$\mathcal{Z}\{x(n)\} = X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

in which the variable z can be interpreted as being either a time-position marker or a complex-valued variable, and the script $\mathcal{Z}$ is the z-transform operator. If the former interpretation is employed, the number multiplying the marker z^{-n} is identified as being the nth element of the x sequence, i.e., $x(n)$. It will be generally beneficial to take z to be a complex-valued variable.

The z-transforms of some useful sequences are listed in Table 8.1.

8.2 Properties of the *z*-Transform

Linearity

Both the direct and inverse z-transform obey the property of linearity. Thus, if $\mathcal{Z}\{f(n)\}$ and $\mathcal{Z}\{g(n)\}$ are denoted by $F(z)$ and $G(z)$, respectively, then

$$\mathcal{Z}\{af(n) + bg(n)\} = aF(z) + bG(z)$$

where a and b are constant multipliers.

Table 8.1 Partial-Fraction Equivalents Listing Causal and Anticausal z-Transform Pairs

z-Domain: $F(z)$	Sequence Domain: $f(n)$
1a. $\dfrac{1}{z-a}$, for $\lvert z\rvert > \lvert a\rvert$	$a^{n-1}u(n-1) = \left\{0,\ 1,\ a,\ a^2,\ \ldots\right\}$
1b. $\dfrac{1}{z-a}$, for $\lvert z\rvert < \lvert a\rvert$	$-a^{n-1}u(-n) = \left\{\ldots,\ \dfrac{-1}{a^3},\ \dfrac{-1}{a^2},\ \dfrac{-1}{a}\right\}$
2a. $\dfrac{1}{(z-a)^2}$, for $\lvert z\rvert > \lvert a\rvert$	$(n-1)a^{n-2}u(n-1) = \left\{0,\ 1,\ 2a,\ 3a^2,\ \ldots\right\}$
2b. $\dfrac{1}{(z-a)^2}$, for $\lvert z\rvert < \lvert a\rvert$	$-(n-1)a^{n-2}u(-n) = \left\{\ldots,\ \dfrac{3}{a^4},\ \dfrac{2}{a^3},\ \dfrac{1}{a^2}\right\}$
3a. $\dfrac{1}{(z-a)^3}$, for $\lvert z\rvert > \lvert a\rvert$	$\dfrac{1}{2}(n-1)(n-2)a^{n-3}u(n-1) = \left\{0,\ 0,\ 1,\ 3a,\ 6a^2,\ \ldots\right\}$
3b. $\dfrac{1}{(z-a)^3}$, for $\lvert z\rvert < \lvert a\rvert$	$\dfrac{-1}{2}(n-1)(n-2)a^{n-3}u(-n) = \left\{\ldots,\ \dfrac{-6}{a^5},\ \dfrac{-3}{a^4},\ \dfrac{-1}{a^3}\right\}$
4a. $\dfrac{1}{(z-a)^m}$, for $\lvert z\rvert > \lvert a\rvert$	$\dfrac{1}{(m-1)!}\displaystyle\prod_{k=1}^{m-1}(n-k)a^{n-m}u(n-1)$
4b. $\dfrac{1}{(z-a)^m}$, for $\lvert z\rvert < \lvert a\rvert$	$\dfrac{-1}{(m-1)!}\displaystyle\prod_{k=1}^{m-1}(n-k)a^{n-m}u(-n)$
5a. z^{-m}, for $z \neq 0,\ m \geq 0$	$\delta(n-m) = \left\{\ldots,\ 0,\ 0,\ \ldots, 1,\ 0,\ \ldots,\ 0,\ \ldots\right\}$
5b. z^{+m}, for $\lvert z\rvert < \infty,\ m \geq 0$	$\delta(n+m) = \left\{\ldots,\ 0,\ 0,\ \ldots, 1,\ \ldots,\ 0,\ \ldots,\ 0,\ \ldots\right\}$

Source: J.A. Cadzow and H.F. Van Landingham, *Signals, Systems and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 191. With permission.

Translation

An important property when transforming terms of a difference equation is the *z-transform* of a sequence shifted in time. For a constant shift, we have

$$\mathcal{Z}\{f(n+k)\} = z^k F(z)$$

for positive or negative integer k. The region of convergence of $z^k F(z)$ is the same as for $F(z)$ for positive k; only the point $z = 0$ need be eliminated from the convergence region of $F(z)$ for negative k.

Convolution

In the z-domain, the time-domain convolution operation becomes a simple product of the corresponding transforms, that is,

$$\mathcal{Z}\{f(n) * g(n)\} = F(z)G(z)$$

Multiplication by a^n

This operation corresponds to a rescaling of the z-plane. For $a > 0$,

$$\mathcal{Z}\{a^n f(n)\} = F\left(\frac{z}{a}\right) \quad \text{for } aR_1 < |z| < aR_2$$

where $F(z)$ is defined for $R_1 < |z| < R_2$.

Time Reversal

$$\mathcal{Z}\{f(\pm n)\} = F(z^{-1}) \quad \text{for } R_2^{-1} < |z| < R_1^{-1}$$

where $F(z)$ is defined for $R_1 < |z| < R_2$.

8.3 Unilateral z-Transform

The unilateral z-transform is defined as

$$\mathcal{Z}_+\{x(n)\} = X(z) = \sum_{n=0}^{\infty} x(n)z^{-n} \quad \text{for } |z| > R$$

where it is called single-sided since $n \geq 0$, just as if the sequence $x(n)$ was in fact single-sided. If there is no ambiguity in the sequel, the subscript plus is omitted and we use the expression *z-transform* to mean either the double- or the single-sided transform. It is usually clear from the context which is meant. By restricting signals to be single-sided, the following useful properties can be proved.

Time Advance

For a single-sided signal $f(n)$,

$$\mathcal{Z}_+\{f(n+1)\} = zF(z) - zf(0)$$

More generally,

$$\mathcal{Z}_+\{f(n+k)\} = z^k F(z) - z^k f(0) - z^{k-1} f(1) - \cdots - zf(k-1)$$

This result can be used to solve linear constant-coefficient difference equations. Occasionally, it is desirable to calculate the initial or final value of a single-sided sequence without a complete inversion. The following two properties present these results.

Initial Signal Value

If $f(n) = 0$ for $n < 0$,

$$f(0) = \lim_{z \Rightarrow \infty} F(z)$$

where $F(z) = Z\{f(n)\}$ for $|z| > R$.

Final Value

If $f(n) = 0$ for $n < 0$ and $Z\{f(n)\} = F(z)$ is a rational function with all its denominator roots (poles) strictly inside the unit circle except possibly for a first-order pole at $z = 1$,

$$f(\infty) = \lim_{n \Rightarrow \infty} f(n) = \lim_{z \Rightarrow \infty} (1 - z^{-1}) F(z)$$

8.4 *z*-Transform Inversion

We operationally denote the inverse transform of $F(z)$ in the form

$$f(n) = Z^{-1}\{F(z)\}$$

There are three useful methods for inverting a transformed signal. They are:

1. Expansion into a series of terms in the variables z and z^{-1}
2. Complex integration by the method of residues
3. Partial-fraction expansion and table look-up

We discuss two of these methods in turn.

Method 1

For the expansion of $F(z)$ into a series, the theory of functions of a complex variable provides a practical basis for developing our inverse transform techniques. As we have seen, the general region of convergence for a transform function $F(z)$ is of the form $a < |z| < b$, i.e., an annulus centered at the origin of the z-plane. This first method is to obtain a series expression of the form

$$F(z) = \sum_{n=-\infty}^{\infty} c_n z^{-n}$$

which is valid in the annulus of convergence. When $F(z)$ has been expanded as in the previous equation, that is, when the coefficients c_n, $n = 0, \pm 1, \pm 2, \ldots$ have been found, the corresponding sequence is specified by $f(n) = c_n$ by uniqueness of the transform.

Method 2

We evaluate the inverse transform of $F(z)$ by the method of residues. The method involves the calculation of residues of a function both inside and outside of a simple closed path that lies inside the region of convergence. A number of key concepts are necessary in order to describe the required procedure.

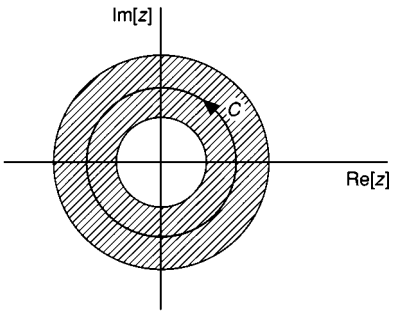


FIGURE 8.1 Typical convergence region for a transformed discrete-time signal (*Source:* J. A. Cadzow and H. F. Van Landingham, *Signals, Systems and Transforms*, Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 191. With permission.)

A complex-valued function $G(z)$ has a pole of order k at $z = z_0$ if it can be expressed as

$$G(z) = \frac{G_1(z_0)}{(z - z_0)^k}$$

where $G_1(z_0)$ is finite.

The residue of a complex function $G(z)$ at a pole of order k at $z = z_0$ is defined by

$$\text{Res}[G(z)]\Big|_{z=z_0} = \frac{1}{(k-1)!}\frac{d^{k-1}}{dz^{k-1}}[(z - z_0)^k G(z)]\Big|_{z=z_0}$$

Inverse Transform Formula (Method 2)

If $F(z)$ is convergent in the annulus $0 < a < |z| < b$ as shown in Fig. 8.1 and C is the closed path shown (the path C must lie entirely within the annulus of convergence), then

$$f(n)\begin{cases}\text{sum of residues of } F(z)z^{n-1} \text{ at poles of } F(z) \text{ inside } C, & m \geq 0\\ -(\text{sum of residues of } F(z)z^{n-1} \text{ at poles of } F(z) \text{ outside } C), & m < 0\end{cases}$$

where m is the least power of z in the numerator of $F(z)z^{n-1}$, e.g., m might equal $n - 1$. Figure 8.1 illustrates the previous equation.

8.5 Sampled Data

Data obtained for a signal only at discrete intervals (sampling period) is called sampled data. One advantage of working with sampled data is the ability to represent sequences as combinations of sampled time signals. Table 8.2 provides some key z-transform pairs. So that the table can serve a multiple purpose, there are three items per line: the first is an indicated sampled continuous-time signal, the second is the Laplace transform of the continuous-time signal, and the third is the z-transform of the uniformly sampled continous-time signal. To illustrate the interrelation of these entries, consider Fig. 8.2. For simplicity, only single-sided signals have been used in Table 8.2. Consequently, the convergence regions are understood in this context to be $\text{Re}[s] < \sigma_0$

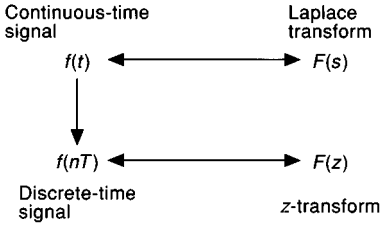


FIGURE 8.2 Signal and transform relationships for Table 8.2.

Table 8.2 *z*-Transforms for Sampled Data

$f(t)$, $t = nT$, $n = 0, 1, 2, \ldots$	$F(s)$, $\mathrm{Re}[s] > \sigma_0$	$F(z)$, $\lvert z \rvert > \rho_0$
1. 1 (unit step)	$\frac{1}{s}$	$\frac{z}{z-1}$
2. t (unit ramp)	$\frac{1}{s^2}$	$\frac{Tz}{(z-1)^2}$
3. t^2	$\frac{2}{s^3}$	$\frac{T^2 z(z+1)}{(z-1)^3}$
4. e^{-at}	$\frac{1}{s+a}$	$\frac{z}{z-e^{-aT}}$
5. te^{-at}	$\frac{1}{(s+a)^2}$	$\frac{Tze^{-aT}}{(z-e^{-aT})^2}$
6. $\sin \omega t$	$\frac{\omega}{s^2+\omega^2}$	$\frac{z \sin \omega T}{z^2 - 2z \cos \omega T + 1}$
7. $\cos \omega t$	$\frac{s}{s^2+\omega^2}$	$\frac{z(z - \cos \omega T)}{z^2 - 2z \cos \omega T + 1}$
8. $e^{-at} \sin \omega t$	$\frac{\omega}{(s+a)^2+\omega^2}$	$\frac{ze^{-aT} \sin \omega T}{z^2 - 2ze^{-aT} \cos \omega T + e^{-2aT}}$
9. $e^{-at} \cos \omega t$	$\frac{s+a}{(s+a)^2+\omega^2}$	$\frac{z(z - e^{-aT} \cos \omega T)}{z^2 - 2ze^{-aT} \cos \omega T + e^{-2aT}}$

Source: J. A. Cadzow and H. F. Landingham, *Signals, Systems and Transforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 191. With permission.

and $|z| > \rho_0$ for the Laplace and z-transforms, respectively. The parameters σ_0 and ρ_0 depend on the actual transformed functions; in factor z, the inverse sequence would begin at $n = 0$. Thus, we use a modified partial-fraction expansion whose terms have this extra z-factor.

Defining Terms

Sampled data: Data obtained for a variable only at discrete intervals. Data is obtained once every sampling period.

Sampling period: The period for which the sampled variable is held constant.

***z*-transform:** A transform from the *s*-domain to the *z*-domain by $z = e^{sT}$.

Related Topics

17.2 Video Signal Processing • 100.6 Digital Control Systems

References

J. A. Cadzow and H. F. Van Landingham, Signals, Systems and Transforms, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

R. C. Dorf, *Modern Control Systems*, 7th ed. Reading, Mass.: Addison-Wesley, 1995.

R. E. Ziemer, *Signals and Systems*, 2nd ed., New York: MacMillan, 1989.

Further Information

IEEE Transactions on Education
IEEE Transactions on Automatic Control
IEEE Transactions on Signal Processing
Contact IEEE, Piscataway, N.J. 08855-1313

Dorf, R.C., Wan, Z. "T-∏ Equivalent Networks"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

9

T–Π Equivalent Networks

Zhen Wan
University of California, Davis

Richard C. Dorf
University of California, Davis

9.1 Introduction
9.2 Three-Phase Connections
9.3 Wye ⇔ Delta Transformations

9.1 Introduction

Two very important two-ports are the T and Π networks shown in Fig. 9.1. Because we encounter these two geometrical forms often in two-port analyses, it is useful to determine the conditions under which these two networks are equivalent. In order to determine the equivalence relationship, we will examine *Z*-parameter equations for the T network and the *Y*-parameter equations for the Π network.

For the T network the equations are

$$\mathbf{V}_1 = (\mathbf{Z}_1 + \mathbf{Z}_3)\mathbf{I}_1 + \mathbf{Z}_3\mathbf{I}_2$$

$$\mathbf{V}_2 = \mathbf{Z}_3\mathbf{I}_1 + (\mathbf{Z}_2 + \mathbf{Z}_3)\mathbf{I}_2$$

and for the Π network the equations are

$$\mathbf{I}_1 = (\mathbf{Y}_a + \mathbf{Y}_b)\mathbf{V}_1 - \mathbf{Y}_b\mathbf{V}_2$$

$$\mathbf{I}_2 = -\mathbf{Y}_b\mathbf{V}_1 + (\mathbf{Y}_b + \mathbf{Y}_c)\mathbf{V}_2$$

Solving the equations for the T network in terms of $\mathbf{I}_1$ and $\mathbf{I}_2$, we obtain

$$\mathbf{I}_1 = \left(\frac{\mathbf{Z}_2 + \mathbf{Z}_3}{\mathbf{D}_1}\right)\mathbf{V}_1 - \frac{\mathbf{Z}_3\mathbf{V}_2}{\mathbf{D}_1}$$

$$\mathbf{I}_2 = -\frac{\mathbf{Z}_3\mathbf{V}_1}{\mathbf{D}_1} + \left(\frac{\mathbf{Z}_1 + \mathbf{Z}_3}{\mathbf{D}_1}\right)\mathbf{V}_2$$

where $\mathbf{D}_1 = \mathbf{Z}_1\mathbf{Z}_2 + \mathbf{Z}_2\mathbf{Z}_3 + \mathbf{Z}_1\mathbf{Z}_3$. Comparing these equations with those for the Π network, we find that

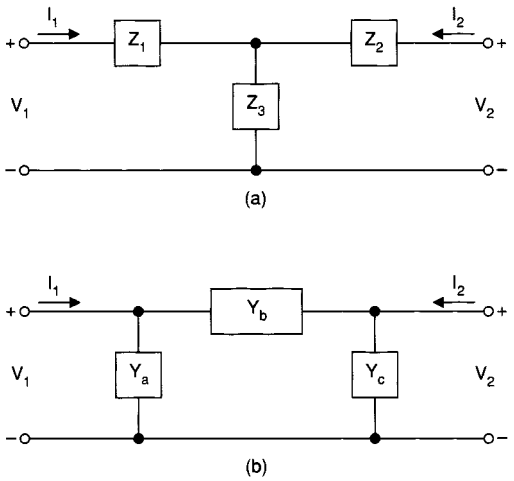


FIGURE 9.1 T and Π two-port networks.

$$\mathbf{Y}_a = \frac{\mathbf{Z}_2}{\mathbf{D}_1}$$
$$\mathbf{Y}_b = \frac{\mathbf{Z}_3}{\mathbf{D}_1}$$
$$\mathbf{Y}_c = \frac{\mathbf{Z}_1}{\mathbf{D}_1}$$

or in terms of the impedances of the Π network

$$\mathbf{Z}_a = \frac{\mathbf{D}_1}{\mathbf{Z}_2}$$
$$\mathbf{Z}_b = \frac{\mathbf{D}_1}{\mathbf{Z}_3}$$
$$\mathbf{Z}_c = \frac{\mathbf{D}_1}{\mathbf{Z}_1}$$

If we reverse this procedure and solve the equations for the Π network in terms of $\mathbf{V}_1$ and $\mathbf{V}_2$ and then compare the resultant equations with those for the T network, we find that

$$\mathbf{Z}_1 = \frac{\mathbf{Y}_c}{\mathbf{D}_2}$$
$$\mathbf{Z}_2 = \frac{\mathbf{Y}_a}{\mathbf{D}_2} \tag{9.1}$$
$$\mathbf{Z}_3 = \frac{\mathbf{Y}_b}{\mathbf{D}_2}$$

where $\mathbf{D}_2 = \mathbf{Y}_a\mathbf{Y}_b + \mathbf{Y}_b\mathbf{Y}_c + \mathbf{Y}_a\mathbf{Y}_c$. Equation (9.1) can also be written in the form

$$\mathbf{Z}_1 = \frac{\mathbf{Z}_a\mathbf{Z}_b}{\mathbf{Z}_a + \mathbf{Z}_b + \mathbf{Z}_c}$$

$$\mathbf{Z}_2 = \frac{\mathbf{Z}_b\mathbf{Z}_c}{\mathbf{Z}_a + \mathbf{Z}_b + \mathbf{Z}_c}$$

$$\mathbf{Z}_3 = \frac{\mathbf{Z}_a\mathbf{Z}_c}{\mathbf{Z}_a + \mathbf{Z}_b + \mathbf{Z}_c}$$

The T is a wye-connected network and the Π is a delta-connected network, as we discuss in the next section.

9.2 Three-Phase Connections

By far the most important polyphase voltage source is the balanced three-phase source. This source, as illustrated by Fig. 9.2, has the following properties. The phase voltages, that is, the voltage from each line *a*, *b*, and *c* to the neutral *n*, are given by

$$\mathbf{V}_{an} = \mathbf{V}_p \angle 0°$$

$$\mathbf{V}_{bn} = \mathbf{V}_p \angle -120° \qquad (9.2)$$

$$\mathbf{V}_{cn} = \mathbf{V}_p \angle +120°$$

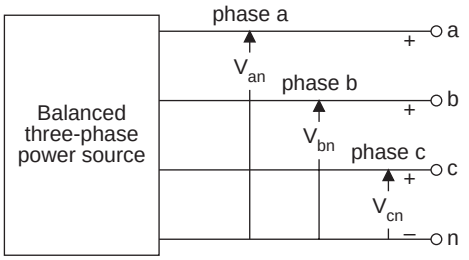


FIGURE 9.2 Balanced three-phase voltage source.

An important property of the balanced voltage set is that

$$\mathbf{V}_{an} + \mathbf{V}_{bn} + \mathbf{V}_{cn} = 0 \qquad (9.3)$$

From the standpoint of the user who connects a load to the balanced three-phase voltage source, it is not important how the voltages are generated. It is important to note, however, that if the load currents generated by connecting a load to the power source shown in Fig. 9.2 are also *balanced*, there are two possible equivalent configurations for the load. The equivalent load can be considered as being connected in either a *wye* (*Y*) or a *delta* (Δ) configuration. The balanced wye configuration is shown in Fig. 9.3. The delta configuration is shown in Fig. 9.4. Note that in the case of the delta connection, there is no neutral line. The actual function of the

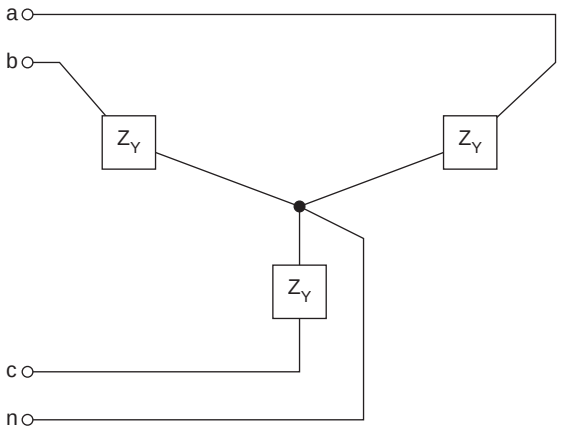


FIGURE 9.3 Wye (*Y*)-connected loads.

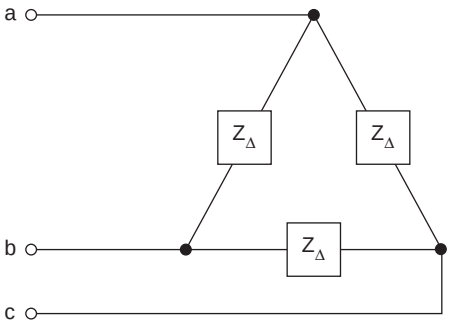


FIGURE 9.4 Delta (Δ)-connected loads.

neutral line in the wye connection will be examined and it will be shown that in a balanced system the neutral line carries no current and therefore may be omitted.

9.3 Wye ⇔ Delta Transformations

For a balanced system, the equivalent load configuration may be either wye or delta. If both of these configurations are connected at only three terminals, it would be very advantageous if an equivalence could be established between them. It is, in fact, possible characteristics are the same. Consider, for example, the two networks shown in Fig. 9.5. For these two networks to be equivalent at each corresponding pair of terminals it is necessary that the input impedances at the corresponding terminals be equal, for example, if at terminals *a* and *b*, with *c* open-circuited, the impedance is the same for both configurations. Equating the impedances at each port yields

TABLE 9.1 Current-Voltage Relationships for the Wye and Delta Load Configurations

Parameter	Wye Configuration	Delta Configuration
Voltage	$\mathbf{V}_{\text{line to line}} = \sqrt{3}\mathbf{V}_Y$	$\mathbf{V}_{\text{line to line}} = \mathbf{V}_\Delta$
Current	$\mathbf{I}_{\text{line}} = \mathbf{I}_Y$	$\mathbf{I}_{\text{line}} = \sqrt{3}\mathbf{I}_\Delta$

$$\begin{aligned}
\mathbf{Z}_{ab} &= \mathbf{Z}_a + \mathbf{Z}_b = \frac{\mathbf{Z}_1(\mathbf{Z}_2 + \mathbf{Z}_3)}{\mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3} \\
\mathbf{Z}_{bc} &= \mathbf{Z}_b + \mathbf{Z}_c = \frac{\mathbf{Z}_3(\mathbf{Z}_1 + \mathbf{Z}_2)}{\mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3} \\
\mathbf{Z}_{ca} &= \mathbf{Z}_c + \mathbf{Z}_a = \frac{\mathbf{Z}_2(\mathbf{Z}_1 + \mathbf{Z}_3)}{\mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3}
\end{aligned} \tag{9.4}$$

Solving this set of equations for $\mathbf{Z}_a$, $\mathbf{Z}_b$, and $\mathbf{Z}_c$ yields

$$\begin{aligned}
\mathbf{Z}_a &= \frac{\mathbf{Z}_1\mathbf{Z}_2}{\mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3} \\
\mathbf{Z}_b &= \frac{\mathbf{Z}_1\mathbf{Z}_3}{\mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3} \\
\mathbf{Z}_c &= \frac{\mathbf{Z}_2\mathbf{Z}_3}{\mathbf{Z}_1 + \mathbf{Z}_2 + \mathbf{Z}_3}
\end{aligned} \tag{9.5}$$

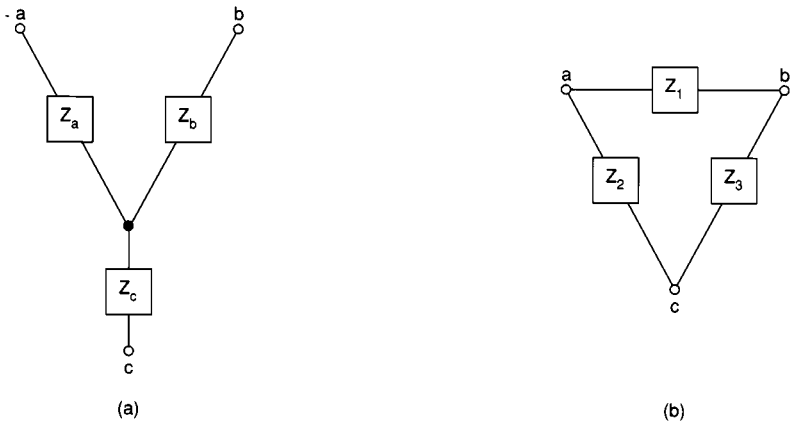


FIGURE 9.5 General wye- and delta-connected loads.

Similary, if we solve Eq. (9.4) for $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$, we obtain

$$\begin{aligned}\mathbf{Z}_1 &= \frac{\mathbf{Z}_a\mathbf{Z}_b + \mathbf{Z}_b\mathbf{Z}_c + \mathbf{Z}_c\mathbf{Z}_a}{\mathbf{Z}_c}\\ \mathbf{Z}_2 &= \frac{\mathbf{Z}_a\mathbf{Z}_b + \mathbf{Z}_b\mathbf{Z}_c + \mathbf{Z}_c\mathbf{Z}_a}{\mathbf{Z}_b}\\ \mathbf{Z}_3 &= \frac{\mathbf{Z}_a\mathbf{Z}_b + \mathbf{Z}_b\mathbf{Z}_c + \mathbf{Z}_c\mathbf{Z}_a}{\mathbf{Z}_a}\end{aligned} \tag{9.6}$$

Equations (9.5) and (9.6) are general relationships and apply to any set of impedances connected in a wye or delta configuration. For the balanced case where $\mathbf{Z}_a = \mathbf{Z}_b = \mathbf{Z}_c$ and $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{Z}_3$, the equations above reduce to

$$\mathbf{Z}_Y = \frac{1}{3}\mathbf{Z} \tag{9.7}$$

and

$$\mathbf{Z}_\Delta = 3\mathbf{Z}_Y \tag{9.8}$$

Defining Terms

Balanced voltages of the three-phase connection: The three voltages satisfy

$$\mathbf{V}_{an} + \mathbf{V}_{bn} + \mathbf{V}_{cn} = 0$$

where

$$\mathbf{V}_{an} = \mathbf{V}_p \angle 0°$$

$$\mathbf{V}_{bn} = \mathbf{V}_p \angle -120°$$

$$\mathbf{V}_{cn} = \mathbf{V}_p \angle +120°$$

T network: The equations of the T network are

$$\mathbf{V}_1 = (\mathbf{Z}_1 + \mathbf{Z}_3)\mathbf{I}_1 + \mathbf{Z}_3\mathbf{I}_2$$

$$\mathbf{V}_2 = \mathbf{Z}_3\mathbf{I}_1 + (\mathbf{Z}_2 + \mathbf{Z}_3)\mathbf{I}_2$$

Π network: The equations of Π network are

$$\mathbf{I}_1 = (\mathbf{Y}_a + \mathbf{Y}_b)\mathbf{V}_1 - \mathbf{Y}_b\mathbf{V}_2$$

$$\mathbf{I}_2 = -\mathbf{Y}_b\mathbf{V}_1 + (\mathbf{Y}_b + \mathbf{Y}_c)\mathbf{V}_2$$

T and Π can be transferred to each other.

Related Topic

3.5 Three-Phase Circuits

References

J.D. Irwin, *Basic Engineering Circuit Analysis,* 4th ed., New York: MacMillan, 1995.
R.C. Dorf, *Introduction to Electric Circuits,* 3rd ed., New York: John Wiley and Sons, 1996.

Further Information

IEEE Transactions on Power Systems
IEEE Transactions on Circuits and Systems, Part II: Analog and Digital Signal Processing

Dorf, R.C., Wan, Z. "Transfer Functions of Filters"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

10

Transfer Functions of Filters

Richard C. Dorf
University of California, Davis

Zhen Wan
University of California, Davis

10.1 Introduction
10.2 Ideal Filters
10.3 The Ideal Linear-Phase Low-Pass Filter
10.4 Ideal Linear-Phase Bandpass Filters
10.5 Causal Filters
10.6 Butterworth Filters
10.7 Chebyshev Filters

10.1 Introduction

Filters are widely used to pass signals at selected frequencies and reject signals at other frequencies. An *electrical filter* is a circuit that is designed to introduce gain or loss over a prescribed range of frequencies. In this section, we will describe ideal filters and then a selected set of practical filters.

10.2 Ideal Filters

An **ideal filter** is a system that completely rejects sinusoidal inputs of the form $x(t) = A \cos \omega t$, $-\infty < t < \infty$, for ω in certain frequency ranges and does not attenuate sinusoidal inputs whose frequencies are outside these ranges. There are four basic types of ideal filters: low-pass, high-pass, bandpass, and bandstop. The magnitude functions of these four types of filters are displayed in Fig. 10.1. Mathematical expressions for these magnitude functions are as follows:

$$\text{Ideal low-pass:} \quad |H(\omega)| = \begin{cases} 1, & -B \le \omega \le B \\ 0, & |\omega| > B \end{cases} \tag{10.1}$$

$$\text{Ideal high-pass:} \quad |H(\omega)| = \begin{cases} 0, & -B < \omega < B \\ 1, & |\omega| \ge B \end{cases} \tag{10.2}$$

$$\text{Ideal bandpass:} \quad |H(\omega)| = \begin{cases} 1, & B_1 \le |\omega| \le B_2 \\ 0, & \text{all other } \omega \end{cases} \tag{10.3}$$

$$\text{Ideal bandstop:} \quad |H(\omega)| = \begin{cases} 0, & B_1 \le |\omega| \le B_2 \\ 1, & \text{all other } \omega \end{cases} \tag{10.4}$$

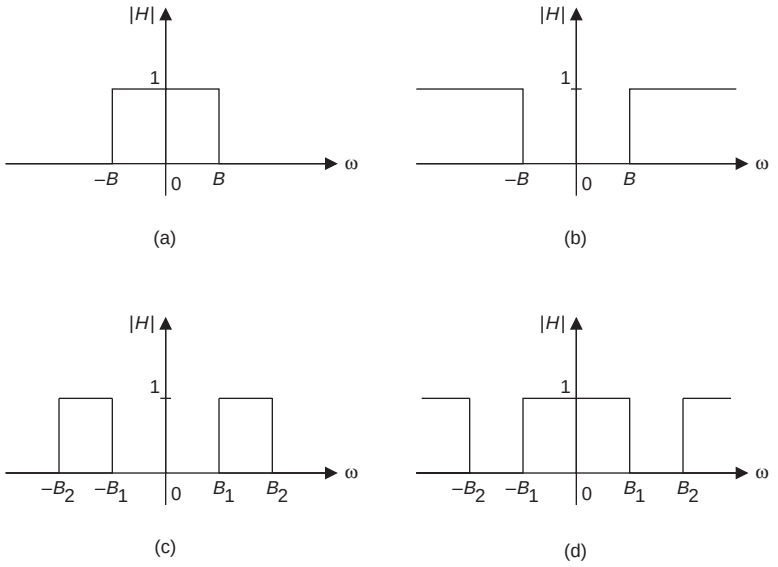


FIGURE 10.1 Magnitude functions of ideal filters:(a) low-pass; (b) high-pass; (c) bandpass; (d) bandstop.

The **stopband** of an ideal filter is defined to be the set of all frequencies ω for which the filter completely stops the sinusoidal input $x(t) = A\cos\omega t$, $-\infty < t < \infty$. The **passband** of the filter is the set of all frequencies ω for which the input $x(t)$ is passed without attenuation.

More complicated examples of ideal filters can be constructed by cascading ideal low-pass, high-pass, bandpass, and bandstop filters. For instance, by cascading bandstop filters with different values of B_1 and B_2, we can construct an ideal comb filter, whose magnitude function is illustrated in Fig. 10.2.

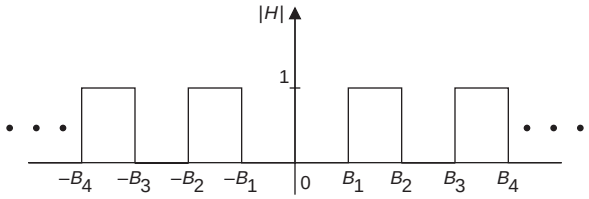


FIGURE 10.2 Magnitude function of an ideal comb filter.

10.3 The Ideal Linear-Phase Low-Pass Filter

Consider the ideal low-pass filter with the frequency function

$$H(\omega) = \begin{cases} e^{-j\omega t_d}, & -B \le \omega \le B \\ 0, & \omega < -B, \quad \omega > B \end{cases} \tag{10.5}$$

where t_d is a positive real number. Equation (10.5) is the polar-form representation of $H(\omega)$. From Eq. (10.5) we have

$$|H(\omega)| = \begin{cases} 1, & -B \le \omega \le B \\ 0, & \omega < -B, \quad \omega > B \end{cases}$$

and

$$\underline{/H(\omega)} = \begin{cases} -\omega t_d, & -B \le \omega \le B \\ 0, & \omega < -B, \quad \omega > B \end{cases}$$

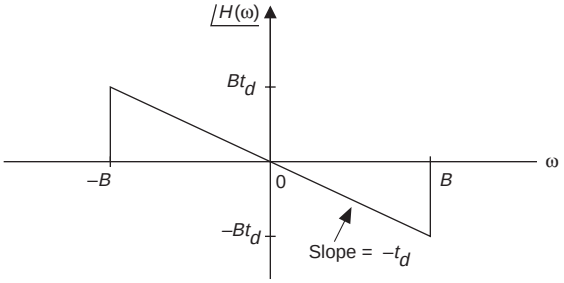


FIGURE 10.3 Phase function of ideal low-pass filter defined by Eq. (10.5).

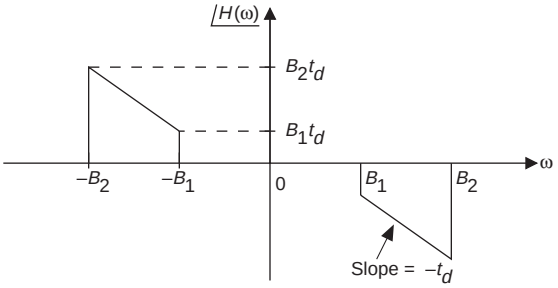


FIGURE 10.4 Phase function of ideal linear-phase bandpass filter.

The phase function $\underline{/H(\omega)}$ of the filter is plotted in Fig. 10.3. Note that over the frequency range 0 to B, the phase function of the system is linear with slope equal to $-t_d$.

The impulse response of the low-pass filter defined by Eq. (10.5) can be computed by taking the inverse Fourier transform of the frequency function $H(\omega)$. The impulse response of the ideal lowpass filter is

$$h(t) = \frac{B}{\pi} Sa[B(t - t_d)], \qquad -\infty < t < \infty \tag{10.6}$$

where $Sa(x) = (\sin x)/x$. The impulse response $h(t)$ of the ideal low-pass filter is not zero for $t < 0$. Thus, the filter has a response before the impulse at $t = 0$ and is said to be noncausal. As a result, it is not possible to build an ideal low-pass filter.

10.4 Ideal Linear-Phase Bandpass Filters

One can extend the analysis to ideal linear-phase bandpass filters. The frequency function of an ideal linear-phase bandpass filter is given by

$$H(\omega) = \begin{cases} e^{-j\omega t_d}, & B_1 \le |\omega| \le B_2 \\ 0, & \text{all other } \omega \end{cases}$$

where t_d, B_1, and B_2 are positive real numbers. The magnitude function is plotted in Fig. 10.1(c) and the phase function is plotted in Fig. 10.4. The passband of the filter is from B_1 to B_2. The filter will pass the signal within the band with no distortion, although there will be a time delay of t_d seconds.

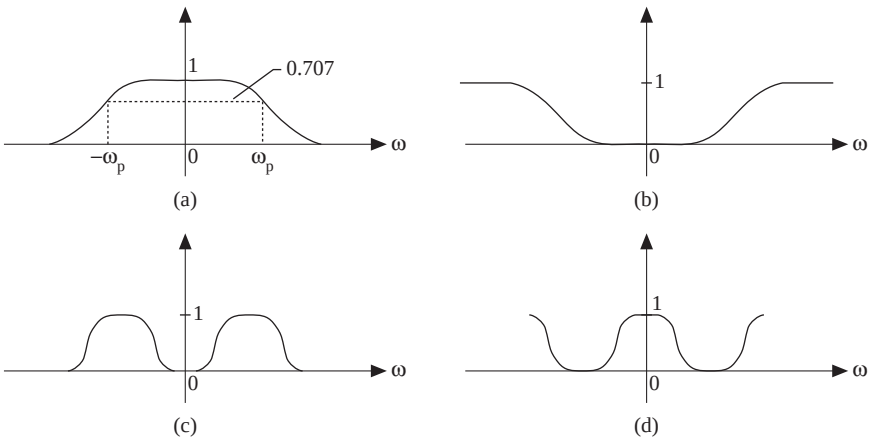


FIGURE 10.5 Causal filter magnitude functions: (a) low-pass; (b) high-pass; (c) bandpass; (d) bandstop.

10.5 Causal Filters

As observed in the preceding section, ideal filters cannot be utilized in real-time filtering applications, since they are noncausal. In such applications, one must use **causal filters**, which are necessarily nonideal; that is, the transition from the passband to the stopband (and vice versa) is gradual. In particular, the magnitude functions of causal versions of low-pass, high-pass, bandpass, and bandstop filters have gradual transitions from the passband to the stopband. Examples of magnitude functions for the basic filter types are shown in Fig. 10.5.

For a causal filter with frequency function $H(\omega)$, the passband is defined as the set of all frequencies ω for which

$$|H(\omega)| \geq \frac{1}{\sqrt{2}} |H(\omega_p)| \simeq 0.707 |H(\omega_p)| \tag{10.7}$$

where ω_p is the value of ω for which $|H(\omega)|$ is maximum. Note that Eq. (10.7) is equivalent to the condition that $|H(\omega)|_{dB}$ is less than 3 dB down from the peak value $|H(\omega_p)|_{dB}$. For low-pass or bandpass filters, the width of the passband is called the **3-dB bandwidth**.

A stopband in a causal filter is a set of frequencies ω for which $|H(\omega)|_{dB}$ is down some desired amount (e.g., 40 or 50 dB) from the peak value $|H(\omega_p)|_{dB}$. The range of frequencies between a passband and a stopband is called a **transition region**. In causal filter design, a key objective is to have the transition regions be suitably small in extent.

10.6 Butterworth Filters

The transfer function of the two-pole Butterworth filter is

$$H(s) = \frac{\omega_n^2}{s^2 + \sqrt{2}\ \omega_n s + \omega_n^2}$$

Factoring the denominator of $H(s)$, we see that the poles are located at

$$s = -\frac{\omega_n}{\sqrt{2}} \pm j\frac{\omega_n}{\sqrt{2}}$$

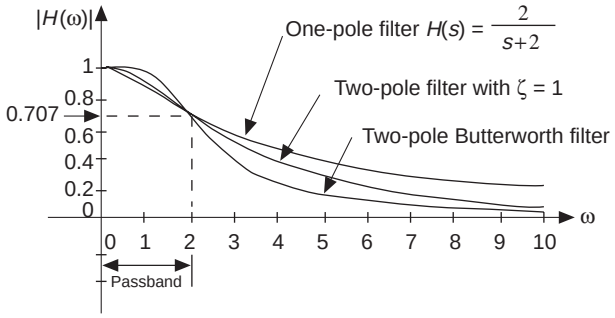


FIGURE 10.6 Magnitude curves of one- and two-pole low-pass filters.

Note that the magnitude of each of the poles is equal to ω_n.

Setting $s = j\omega$ in $H(s)$, we have that the magnitude function of the two-pole Butterworth filter is

$$|H(\omega)| = \frac{1}{\sqrt{1 + (\omega / \omega_n)^4}} \tag{10.8}$$

From Eq. (10.8) we see that the 3-dB bandwidth of the Butterworth filter is equal to ω_n. For the case $\omega_n = 2$ rad/s, the frequency response curves of the Butterworth filter are plotted in Fig. 10.6. Also displayed are the frequency response curves for the one-pole low-pass filter with transfer function $H(s) = 2/(s + 2)$, and the two-pole low-pass filter with $\zeta = 1$ and with 3-dB bandwidth equal to 2 rad/s. Note that the Butterworth filter has the sharpest cutoff of all three filters.

10.7 Chebyshev Filters

The magnitude function of the n-pole Butterworth filter has a monotone characteristic in both the passband and stopband of the filter. Here *monotone* means that the magnitude curve is gradually decreasing over the passband and stopband. In contrast to the Butterworth filter, the magnitude function of a type 1 Chebyshev filter has ripple in the passband and is monotone decreasing in the stopband (a type 2 Chebyshev filter has the opposite characteristic). By allowing ripple in the passband or stopband, we are able to achieve a sharper transition between the passband and stopband in comparison with the Butterworth filter.

The n-pole type 1 Chebyshev filter is given by the frequency function

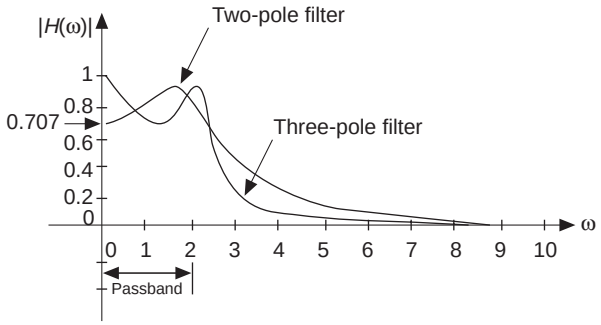
$$|H(\omega)| = \frac{1}{\sqrt{1 + \epsilon^2 T_n^2 (\omega/\omega_1)}} \tag{10.9}$$

where $T_n(\omega/\omega_1)$ is the nth-order Chebyshev polynomial. Note that ϵ is a numerical parameter related to the level of ripple in the passband. The Chebyshev polynomials can be generated from the recursion

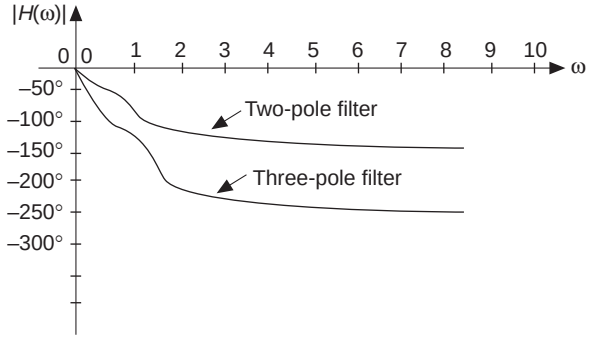
$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$$

where $T_0(x) = 1$ and $T_1(x) = x$. The polynomials for $n = 2, 3, 4, 5$ are

$$\begin{aligned} T_2(x) &= 2x(x) - 1 = 2x^2 - 1 \\ T_3(x) &= 2x(2x^2 - 1) - x = 4x^3 - 3x \\ T_4(x) &= 2x(4x^3 - 3x) - (2x^2 - 1) = 8x^4 - 8x^2 + 1 \\ T_5(x) &= 2x(8x^4 - 8x^2 + 1) - (4x^3 - 3x) = 16x^5 - 20x^3 + 5x \end{aligned} \tag{10.10}$$



(a)



(b)

FIGURE 10.7 Frequency curves of two- and three-pole Chebyshev filters with ω_c = 2.5 rad/s: (a) magnitude curves; (b) phase curves.

Using Eq. (10.10), the two-pole type 1 Chebyshev filter has the following frequency function

$$|H(\omega)| = \frac{1}{\sqrt{1 + \epsilon^2[2(\omega / \omega_1)^2 - 1]^2}}$$

For the case of a 3-dB ripple (ϵ = 1), the transfer functions of the two-pole and three-pole type 1 Chebyshev filters are

$$H(s) = \frac{0.50\omega_c^2}{s^2 + 0.645\omega_c s + 0.708\omega_c^2}$$

$$H(s) = \frac{0.251\omega_c^3}{s^3 + 0.597\omega_c s^2 + 0.928\omega_c^2 s + 0.251\omega_c^3}$$

where ω_c = 3-dB bandwidth. The frequency curves for these two filters are plotted in Fig. 10.7 for the case ω_c = 2.5 rad.

The magnitude response functions of the three-pole Butterworth filter and the three-pole type 1 Chebyshev filter are compared in Fig. 10.8 with the 3-dB bandwidth of both filters equal to 2 rad. Note that the transition from passband to stopband is sharper in the Chebyshev filter; however, the Chebyshev filter does have the 3-dB ripple over the passband.

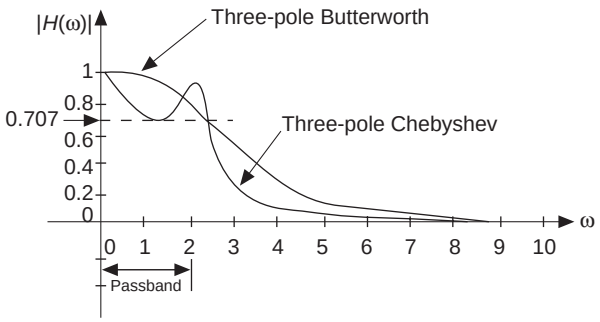


FIGURE 10.8 Magnitude curves of three-pole Butterworth and three-pole Chebyshev filters with 3-dB bandwidth equal to 2.5 rad/s.

Defining Terms

Causal filter: A filter of which the transition from the passband to the stopband is gradual, not ideal. This filter is realizable.

3-dB bandwidth: For a causal low-pass or bandpass filter with a frequency function $H(j\omega)$: the frequency at which $|H(\omega)|_{\text{dB}}$ is less than 3 dB down from the peak value $|H(\omega_p)|_{\text{dB}}$.

Ideal filter: An ideal filter is a system that completely rejects sinusoidal inputs of the form $x(t) = A \cos \omega t$, $-\infty < t < \infty$, for ω within a certain frequency range, and does not attenuate sinusoidal inputs whose frequencies are outside this range. There are four basic types of ideal filters:low-pass, high-pass, bandpass, and bandstop.

Passband: Range of frequencies ω for which the input is passed without attenuation.

Stopband: Range of frequencies ω for which the filter completely stops the input signal.

Transition region: The range of frequencies of a filter between a passband and a stopband.

Related Topics

4.2 Low-Pass Filter Functions • 4.3 Low Pass Filters • 11.1 Introduction • 29.1 Synthesis of Low-Pass Forms

References

R.C. Dorf, *Introduction to Electrical Circuits*, 3rd ed., New York: Wiley, 1996.
E.W. Kamen, *Introduction to Signals and Systems*, 2nd ed., New York: Macmillan, 1990.
G.R. Cooper and C.D. McGillem, *Modern Communications and Spread Spectrum*, New York: McGraw-Hill, 1986.

Further Information

IEEE Transactions on Circuits and Systems, Part I: Fundamental Theory and Applications.
IEEE Transactions on Circuits and Systems, Part II: Analog and Digital Signal Processing.
Available from IEEE.

Neudorfer, P. “Frequency Response”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

11

Frequency Response

11.1 Introduction
11.2 Linear Frequency Response Plotting
11.3 Bode Diagrams
11.4 A Comparison of Methods

Paul Neudorfer
Seattle University

11.1 Introduction

The IEEE Standard Dictionary of Electrical and Electronics Terms defines **frequency response** in stable, linear systems to be "the frequency-dependent relation in both gain and phase difference between steady-state sinusoidal inputs and the resultant steady-state sinusoidal outputs" [IEEE, 1988]. In certain specialized applications, the term *frequency response* may be used with more restrictive meanings. However, all such uses can be related back to the fundamental definition. The frequency response characteristics of a system can be found directly from its transfer function. A single-input/single-output linear time-invariant system is shown in Fig. 11.1.

For dynamic linear systems with no time delay, the transfer function $H(s)$ is in the form of a ratio of polynomials in the complex frequency s,

$$H(s) = K\frac{N(s)}{D(s)}$$

where K is a frequency-independent constant. For a system in the sinusoidal steady state, s is replaced by the sinusoidal frequency $j\omega$ ($j = \sqrt{-1}$) and the system function becomes

$$H(j\omega) = K\frac{N(j\omega)}{D(j\omega)} = |H(j\omega)|e^{j\arg H(j\omega)}$$

$H(j\omega)$ is a complex quantity. Its magnitude, $|H(j\omega)|$, and its argument or phase angle, $\arg H(j\omega)$, relate, respectively, the amplitudes and phase angles of sinusoidal steady-state input and output signals. Using the terminology of Fig. 11.1, if the input and output signals are

$$x(t) = X\cos(\omega t + \Theta_x)$$

$$y(t) = Y\cos(\omega t + \Theta_y)$$

then the output's amplitude Y and phase angle Θ_y are related to those of the input by the two equations

$$Y = |H(j\omega)|X$$

$$\Theta_y = \arg H(j\omega) + \Theta_x$$

The phrase *frequency response characteristics* usually implies a complete description of a system's sinusoidal steady-state behavior as a function of frequency. Because $H(j\omega)$ is complex and, therefore, two dimensional in nature, frequency response characteristics cannot be graphically displayed as a single curve plotted with respect to frequency. Instead, the magnitude and argument of $H(j\omega)$ can be separately plotted as functions of frequency. Often, only the magnitude curve is presented as a concise way of characterizing the system's behavior, but this must be viewed as an incomplete description. The most common form for such plots is the **Bode diagram** (developed by H.W. Bode of Bell Laboratories), which uses a logarithmic scale for frequency. Other forms of frequency response plots have also been developed. In the **Nyquist plot** (Harry Nyquist, also of Bell Labs), $H(j\omega)$ is displayed on the complex plane, $\mathrm{Re}[H(j\omega)]$ on the horizontal axis, and $\mathrm{Im}[H(j\omega)]$ on the vertical. Frequency is a parameter of such curves. It is sometimes numerically identified at selected points of the curve and sometimes omitted. The **Nichols chart** (N.B. Nichols) graphs magnitude versus phase for the system function. Frequency again is a parameter of the resultant curve, sometimes shown and sometimes not.

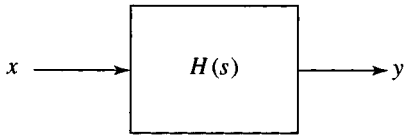


FIGURE 11.1 A single-input/single-output linear system.

Frequency response techniques are used in many areas of engineering. They are most obviously applicable to such topics as communications and filters, where the frequency response behaviors of systems are central to an understanding of their operations. It is, however, in the area of control systems where frequency response techniques are most fully developed as analytical and design tools. The Nichols chart, for instance, is used exclusively in the analysis and design of feedback control systems.

The remaining sections of this chapter describe several frequency response plotting methods. Applications of the methods can be found in other chapters throughout the *Handbook*.

11.2 Linear Frequency Response Plotting

Linear frequency response plots are prepared most directly by computing the magnitude and phase of $H(j\omega)$ and graphing each as a function of frequency (either f or ω), the frequency axis being scaled linearly. As an example, consider the transfer function

$$H(s) = \frac{160{,}000}{s^2 + 220s + 160{,}000}$$

Formally, the complex frequency variable s is replaced by the sinusoidal frequency $j\omega$ and the magnitude and phase found.

$$H(j\omega) = \frac{160{,}000}{(j\omega)^2 + 220(j\omega) + 160{,}000}$$

$$|H(j\omega)| = \frac{160{,}000}{\sqrt{(160{,}000 - \omega^2)^2 + (220\omega)^2}}$$

$$\arg H(j\omega) = -\tan^{-1}\frac{220\omega}{160{,}000 - \omega^2}$$

The plots of magnitude and phase are shown in Fig. 11.2.

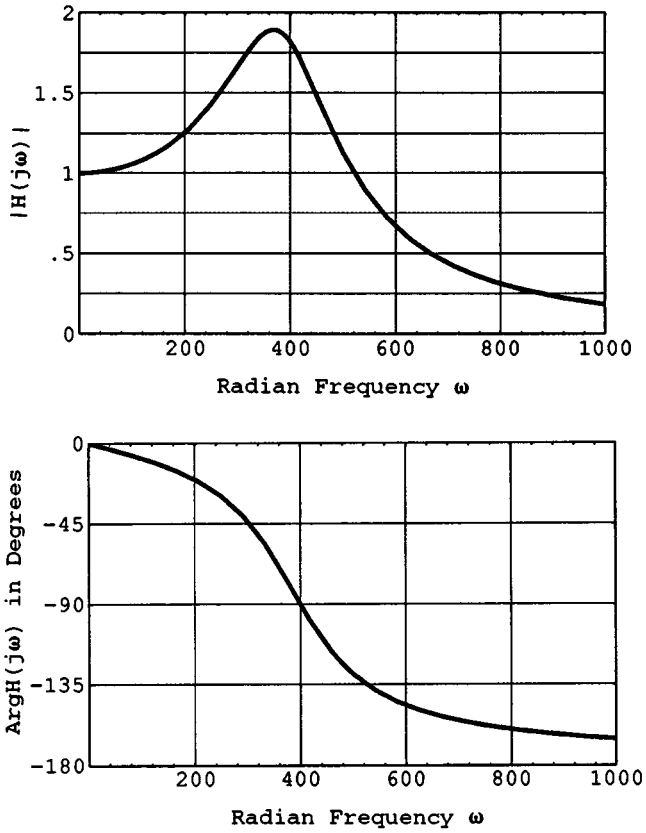


Figure 11.2 Linear frequency response curves of $H(j\omega)$.

11.3 Bode Diagrams

A Bode diagram consists of plots of the gain and phase of a transfer function, each with respect to logarithmically scaled frequency axes. In addition, the gain of the transfer function is scaled in **decibels** according to the definition

$$|H|_{\text{dB}} = H_{\text{dB}} = 20 \log_{10} |H(j\omega)|$$

This definition relates to transfer functions which are ratios of voltages and/or currents. The decibel gain between two *powers* has a multiplying factor of 10 rather than 20. This method of plotting frequency response information was popularized by H.W. Bode in the 1930s. There are two main advantages of the Bode approach. The first is that, with it, the gain and phase curves can be easily and accurately drawn. Second, once drawn, features of the curves can be identified both qualitatively and quantitatively with relative ease, even when those features occur over a wide dynamic range. Digital computers have rendered the first advantage obsolete. Ease of interpretation, however, remains a powerful advantage, and the Bode diagram is today the most common method chosen for the display of frequency response data.

A Bode diagram is drawn by applying a set of simple rules or procedures to a transfer function. The rules relate directly to the set of poles and zeros and/or time constants of the function. Before constructing a Bode diagram, the transfer function is normalized so that each pole or zero term (except those at $s = 0$) has a dc gain of one. For instance:

$$H(s) = K \frac{s + \omega_z}{s(s + \omega_p)} = \frac{K\omega_z}{\omega_p} \frac{s/\omega_z + 1}{s(s/\omega_p + 1)} = K' \frac{s\tau_z + 1}{s(s\tau_p + 1)}$$

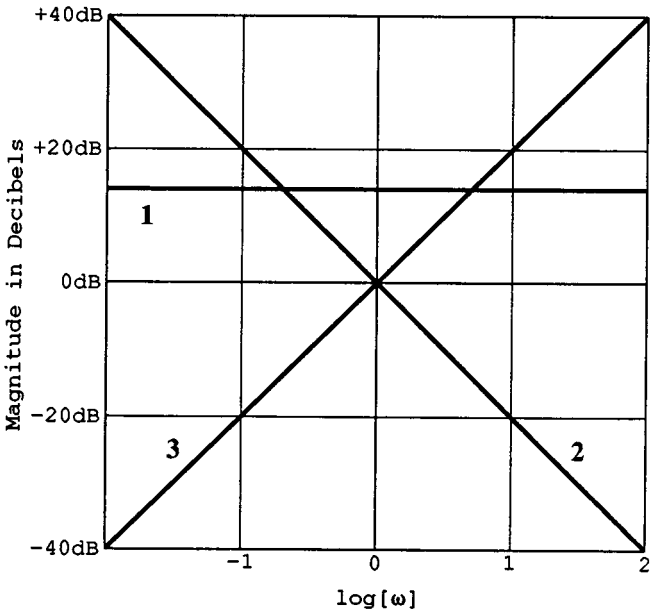


Figure 11.3 Bode magnitude functions for (1) $K = 5$, (2) $1/s$, and (3) s.

In the last form of the expression, $\tau_z = 1/\omega_z$ and $\tau_p = 1/\omega_p$. τ_p is a time constant of the system and $s = -\omega_p$ is the corresponding natural frequency. Because it is understood that Bode diagrams are limited to sinusoidal steady-state frequency response analysis, one can work directly from the transfer function $H(s)$ rather than resorting to the formalism of making the substitution $s = j\omega$.

Bode frequency response curves (gain and phase) for $H(s)$ are generated from the individual contributions of the four terms K', $s\tau_z + 1$, $1/s$, and $1/(s\tau_p + 1)$. As described in the following paragraph, the frequency response effects of these individual terms are easily drawn. To obtain the overall frequency response curves for the transfer function, the curves for the individual terms are added together.

The terms used as the basis for drawing Bode diagrams are found from factoring $N(s)$ and $D(s)$, the numerator and denominator polynomials of the transfer function. The factorization results in four standard forms. These are (1) a constant K; (2) a simple s term corresponding to either a zero (if in the numerator) or a pole (if in the denominator) at the origin; (3) a term such as $(s\tau + 1)$ corresponding to a real valued (nonzero) pole or zero; and (4) a quadratic term with a possible standard form of $[(s/\omega_n)^2 + (2\zeta/\omega_n)s + 1]$ corresponding to a pair of complex conjugate poles or zeros. The Bode magnitude and phase curves for these possibilities are displayed in Figs. 11.3–11.5. Note that both decibel magnitude and phase are plotted semilogarithmically. The frequency axis is logarithmically scaled so that every tenfold, or **decade**, change in frequency occurs over an equal distance. The magnitude axis is given in decibels. Customarily, this axis is marked in 20-dB increments. Positive decibel magnitudes correspond to amplifications between input and output that are greater than one (output amplitude larger than input). Negative decibel gains correspond to attenuation between input and output.

Figure 11.3 shows three separate magnitude functions. Curve 1 is trivial; the Bode magnitude of a constant K is simply the decibel-scaled constant $20 \log_{10} K$, shown for an arbitrary value of $K = 5$ ($20 \log_{10} 5 = 13.98$). Phase is not shown. However, a constant of $K > 0$ has a phase contribution of 0° for all frequencies. For $K < 0$, the contribution would be ±180° (Recall that $-\cos\theta = \cos(\theta \pm 180°)$. Curve 2 shows the magnitude frequency response curve for a pole at the origin ($1/s$). It is a straight line with a slope of –20 dB/decade. The line passes through 0 dB at $\omega = 0$ rad/s. The phase contribution of a simple pole at the origin is a constant –90°, independent of frequency. The effect of a zero at the origin (s) is shown in Curve 3. It is again a straight line that passes through 0 dB at $\omega = 0$ rad/s; however, the slope is +20 dB/decade. The phase contribution of a simple zero at $s = 0$ is +90°, independent of frequency.

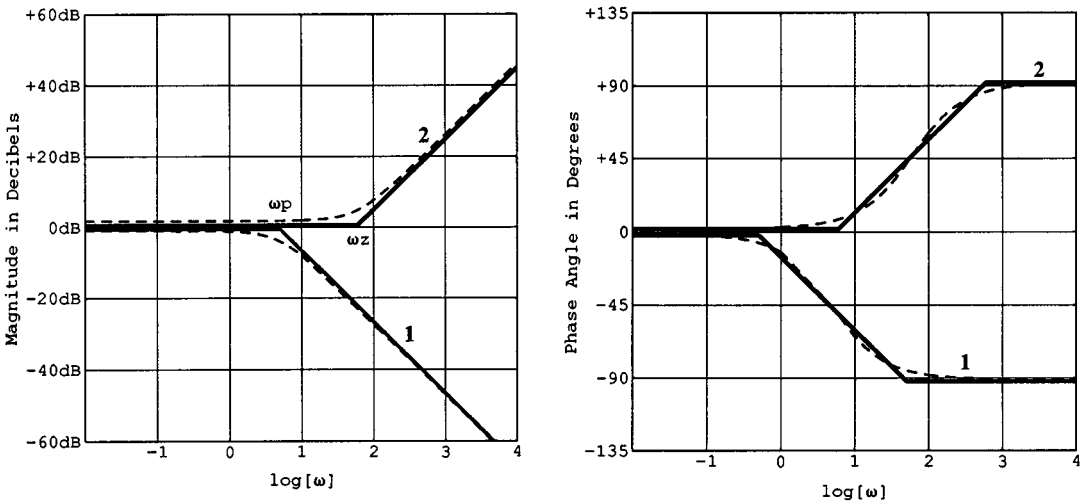


Figure 11.4 Bode curves for (1) a simple pole at $s = -\omega_p$ and (2) a simple zero at $s = -\omega_z$.

Note from Fig. 11.3 and the foregoing discussion that in Bode diagrams the effect of a pole term at a given location is simply the negative of that of a zero term at the same location. This is true for both magnitude and phase curves.

Figure 11.4 shows the magnitude and phase curves for a zero term of the form $(s/\omega_z + 1)$ and a pole term of the form $1/(s/\omega_p + 1)$. Exact plots of the magnitude and phase curves are shown as dashed lines. *Straight line approximations* to these curves are shown as solid lines. Note that the straight line approximations are so good that they obscure the exact curves at most frequencies. For this reason, some of the curves in this and later figures have been displaced slightly to enhance clarity. The greatest error between the exact and approximate magnitude curves is ±3 dB. The approximation for phase is always within 7° of the exact curve and usually much closer. The approximations for magnitude consist of two straight lines. The points of intersection between these two lines ($\omega = \omega_z$ for the zero term and $\omega = \omega_p$ for the pole) are **breakpoints** of the curves. Breakpoints of Bode gain curves always correspond to locations of poles or zeros in the transfer function.

In Bode analysis complex conjugate poles or zeros are always treated as pairs in the corresponding quadratic form $[(s/\omega_n)^2 + (2\zeta/\omega_n)s + 1]$.[1] For quadratic terms in stable, minimum phase systems, the **damping ratio** ζ (Greek letter zeta) is within the range $0 < \zeta < 1$. Quadratic terms cannot always be adequately represented by straight line approximations. This is especially true for lightly damped systems (small ζ). The traditional approach was to draw a preliminary representation of the contribution. This consists of a straight line of 0 dB from dc up to the breakpoint at ω_n followed by a straight line of slope ±40 dB/decade beyond the breakpoint, depending on whether the plot refers to a pair of poles or a pair of zeros. Then, referring to a family of curves as shown in Fig. 11.5, the preliminary representation was improved based on the value of ζ. The phase contribution of the quadratic term was similarly constructed. Note that Fig. 11.5 presents frequency response contributions for a quadratic pair of poles. For zeros in the corresponding locations, both the magnitude and phase curves would be negated. Digital computer applications programs render this procedure unnecessary for purposes of constructing frequency response curves. Knowledge of the technique is still valuable, however, in the qualitative and quantitative interpretation of frequency response curves. Localized peaking in the gain curve is a reflection of the existence of **resonance** in a system. The height of such a peak (and the corresponding value of ζ) is a direct indication of the degree of resonance.

Bode diagrams are easily constructed because, with the exception of lightly damped quadratic terms, each contribution can be reasonably approximated with straight lines. Also, the overall frequency response curve is found by adding the individual contributions. Two examples follow.

[1]Several such standard forms are used. This is the one most commonly encountered in controls applications.

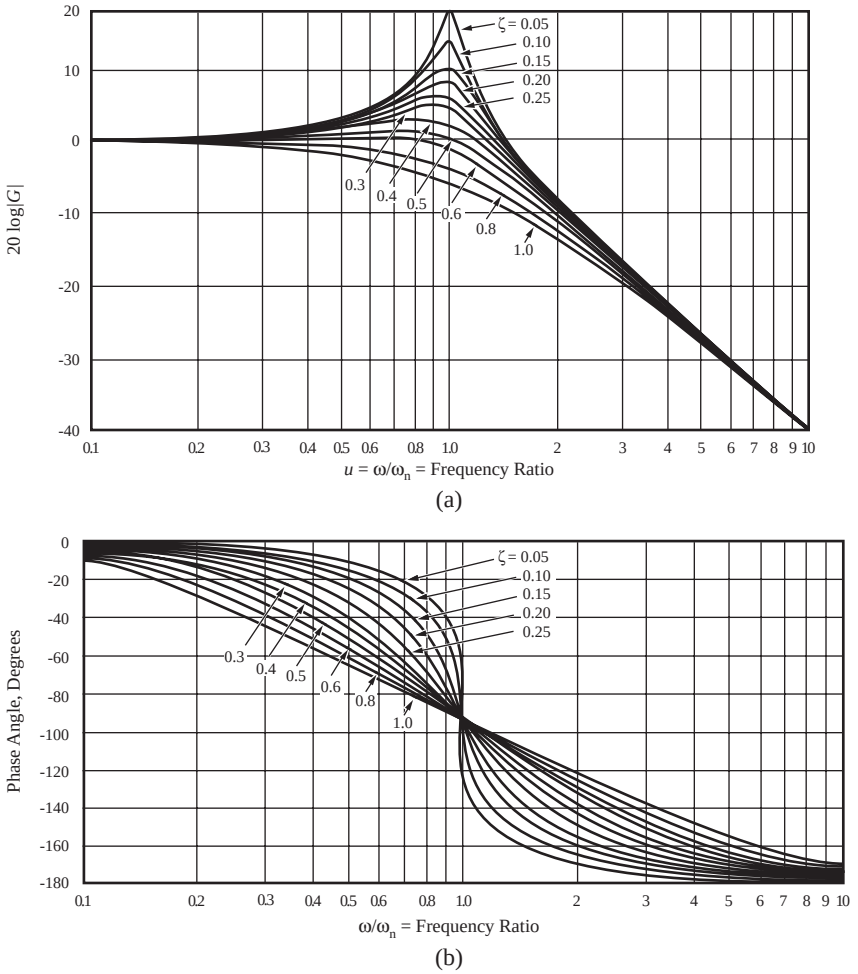


Figure 11.5 Bode diagram of $1/[(s/\omega_n)^2 + (2\zeta/\omega_n)s + 1]$.

Example 1

$$A(s) = \frac{10^4 s}{s^2 + 1100s + 10^5} = \frac{10^4 s}{(s + 100)(s + 1000)} = 10^{-1}\,\frac{s}{(s/100 + 1)(s/1000 + 1)}$$

In Fig. 11.6, the individual contributions of the four factored terms of $A(s)$ are shown as long dashed lines. The straight line approximations for gain and phase are shown with solid lines. The exact curves are presented with short dashed lines.

Example 2

$$G(s) = \frac{1000(s + 500)}{s^2 + 70s + 10{,}000} = \frac{50(s/500 + 1)}{(s/100)^2 + 2(0.35)(s/100) + 1}$$

Note that the damping factor for the quadratic term in the denominator is $\zeta = 0.35$. If drawing the response curves by hand, the resonance peak near the breakpoint at $\omega = 100$ would be estimated from Fig. 11.5. Figure 11.7 shows the exact gain and phase frequency response curves for $G(s)$.

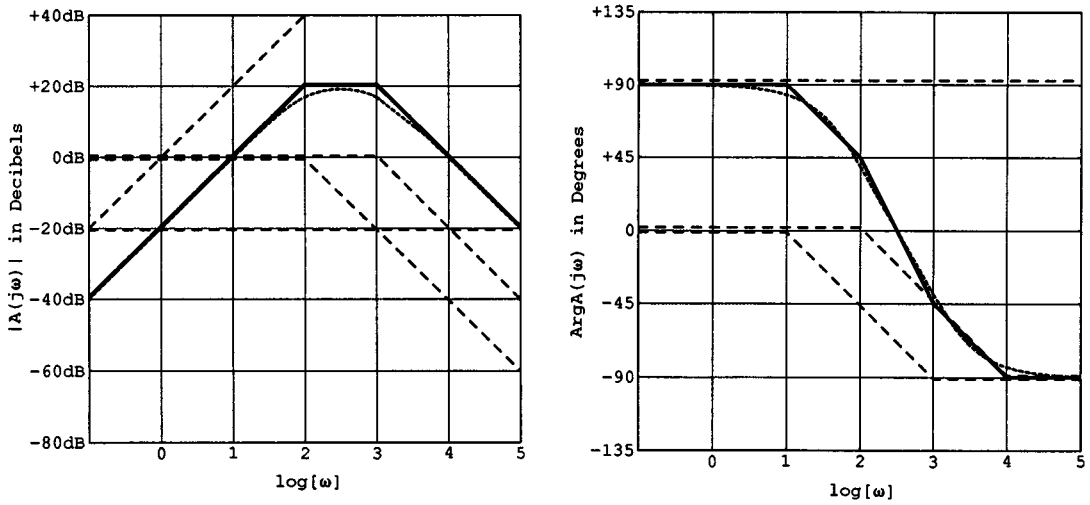


Figure 11.6 Bode diagram of $A(s)$.

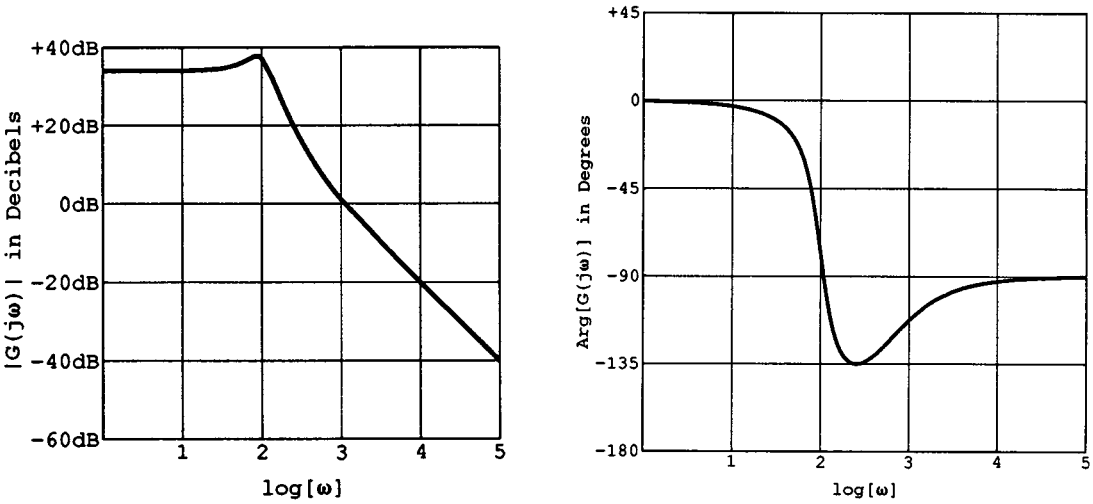


Figure 11.7 Bode diagram of $G(s)$.

11.4 A Comparison of Methods

This chapter concludes with the frequency response of a simple system function plotted in three different ways.

Example 3

$$T(s) = \frac{10^7}{(s + 100)(s + 200)(s + 300)}$$

Figure 11.8 shows the direct, linear frequency response curves for $T(s)$. Corresponding Bode and Nyquist diagrams are shown, respectively, in Figs. 11.9 and 11.10.

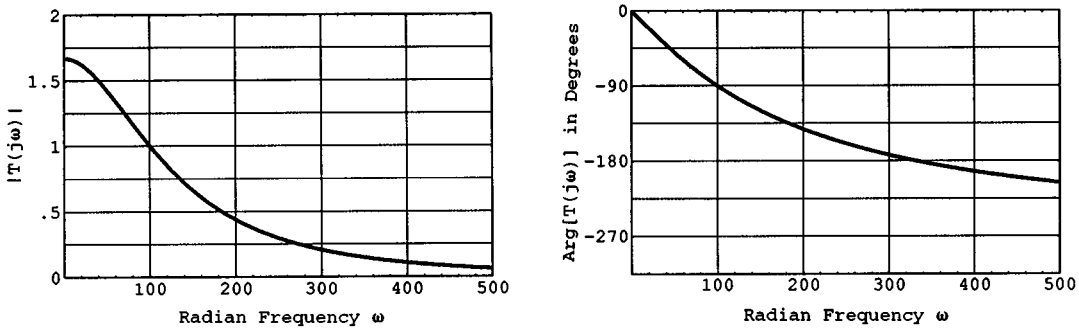


Figure 11.8 Linear frequency response plot of $T(s)$.

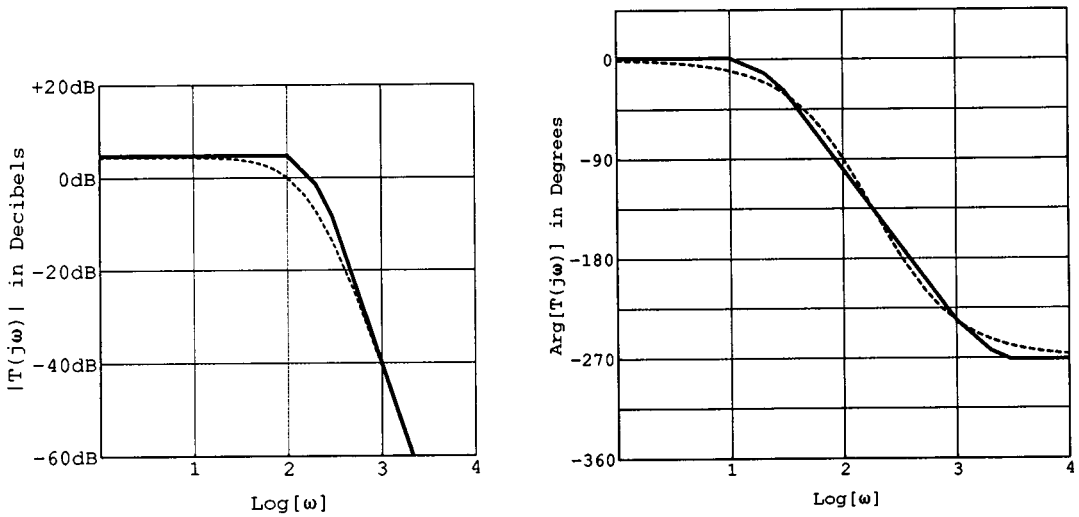


Figure 11.9 Bode diagram of $T(s)$.

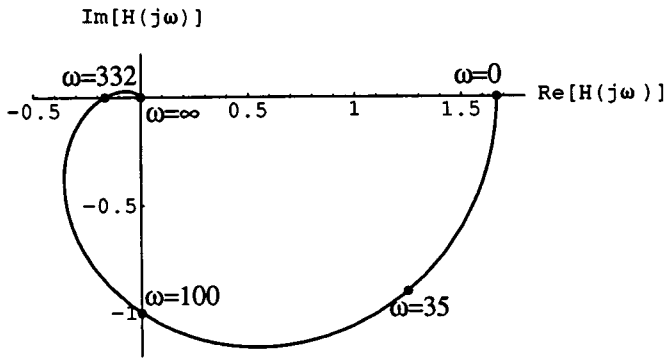


Figure 11.10 Nyquist plot of $T(s)$.

Defining Terms

Bode diagram: A frequency response plot of 20 log gain and phase angle on a log-frequency base.

Breakpoint: A point of abrupt change in slope in the straight line approximation of a Bode magnitude curve.

Damping ratio: The ratio between a system's damping factor (measure of rate of decay of response) and the damping factor when the system is critically damped.

Decade: Synonymous with power of ten. In context, a tenfold change in frequency.

Decibel: A measure of relative size. The decibel gain between voltages V_1 and V_2 is $20 \log_{10}(V_1/V_2)$. The decibel ratio of two powers is $10 \log_{10}(P_1/P_2)$.

Frequency response: The frequency-dependent relation in both gain and phase difference between steady-state sinusoidal inputs and the resultant steady-state sinusoidal outputs.

Nichols chart: Control systems — a plot showing magnitude contours and phase contours of the return transfer function referred to as ordinates of logarithmic loop gain and abscissae of loop phase angle.

Nyquist plot: A parametric frequency response plot with the real part of the transfer function on the abscissa and the imaginary part of the transfer function on the ordinate.

Resonance: The enhancement of the response of a physical system to a steady-state sinusoidal input when the excitation frequency is near a natural frequency of the system.

Related Topics

2.1 Step, Impulse, Ramp, Sinusoidal, Exponential, and DC Signals • 100.3 Frequency Response Methods: Bode Diagram Approach

References

R.C. Dorf, *Modern Control Systems,* 4th ed., Reading, Mass.: Addison-Wesley, 1986.

IEEE Standard Dictionary of Electrical and Electronics Terms, 4th ed., The Institute of Electrical and Electronics Engineers, 1988.

D.E. Johnson, J.R. Johnson, and J.L. Hilburn, *Electric Circuit Analysis,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1992.

B.C. Kuo, *Automatic Control Systems,* 4th ed., Englewood Cliffs, N.J.: Prentice-Hall, 1982.

K. Ogata, *System Dynamics*, Englewood Cliffs, N.J.: Prentice-Hall, 1992.

W.D. Stanley, *Network Analysis with Applications,* Reston, Va.: Reston, 1985.

M.E. Van Valkenburg, *Network Analysis,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Further Information

Good coverage of frequency response techniques can be found in many undergraduate-level electrical engineering textbooks. Refer especially to classical automatic controls or circuit analysis books. Useful information can also be found in books on active network design.

Examples of the application of frequency response methods abound in journal articles ranging over such diverse topics as controls, acoustics, electronics, and communications.

Szidarovszky, F., Bahill, A.T. "Stability Analysis"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

12

Stability Analysis

Ferenc Szidarovszky
University of Arizona

A. Terry Bahill
University of Arizona

12.1 Introduction
12.2 Using the State of the System to Determine Stability
12.3 Lyapunov Stability Theory
12.4 Stability of Time-Invariant Linear Systems
Stability Analysis with State-Space Notation • The Transfer Function Approach
12.5 BIBO Stability
12.6 Physical Examples

12.1 Introduction

In this chapter, which is based on Szidarovszky and Bahill [1992], we first discuss stability in general and then present four techniques for assessing the stability of a system: (1) Lyapunov functions, (2) finding the eigenvalues for state-space notation, (3) finding the location in the complex frequency plane of the poles of the closed-loop transfer function, and (4) proving bounded outputs for all bounded inputs. Proving stability with Lyapunov functions is very general: it works for nonlinear and time-varying systems. It is also good for doing proofs. Proving the stability of a system with Lyapunov functions is difficult, however, and failure to find a Lyapunov function that proves a system is stable does not prove that the system is unstable. The next techniques we present, finding the eigenvalues or the poles of the transfer function, are sometimes difficult, because they require factoring high-degree polynomials. Many commercial software packages are now available for this task, however. We think most engineers would benefit by having one of these computer programs. Jamshidi et al. [1992] and advertisements in technical publications such as the *IEEE Control Systems Magazine* and *IEEE Spectrum* describe many appropriate software packages. The last technique we present, bounded-input, bounded-output stability, is also quite general.

Let us begin our discussion of **stability** and **instability** of systems informally. In an *unstable system* the state can have large variations, and small inputs or small changes in the initial state may produce large variations in the output. A common example of an unstable system is illustrated by someone pointing the microphone of a public address (PA) system at a speaker; a loud high-pitched tone results. Often instabilities are caused by too much gain, so to quiet the PA system, decrease the gain by pointing the microphone away from the speaker. Discrete systems can also be unstable. A friend of ours once provided an example. She was sitting in a chair reading and she got cold. So she went over and turned up the thermostat on the heater. The house warmed up. She got hot, so she got up and turned down the thermostat. The house cooled off. She got cold and turned up the thermostat. This process continued until someone finally suggested that she put on a sweater (reducing the gain of her heat loss system). She did, and was much more comfortable. We modeled this as a discrete system, because she seemed to sample the environment and produce outputs at discrete intervals about 15 minutes apart.

12.2 Using the State of the System to Determine Stability

The stability of a system is defined with respect to a given equilibrium point in state space. If the initial state $\mathbf{x}_0$ is selected at an equilibrium state $\bar{\mathbf{x}}$ of the system, then the state will remain at $\bar{\mathbf{x}}$ for all future time. When the initial state is selected close to an equilibrium state, the system might remain close to the equilibrium state or it might move away. In this section we introduce conditions that guarantee that whenever the system starts near an equilibrium state, it remains near it, perhaps even converging to the equilibrium state as time increases. For simplicity, only time-invariant systems are considered in this section. Time-variant systems are discussed in Section 12.5.

Continuous, time-invariant systems have the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) \tag{12.1}$$

and discrete, time-invariant systems are modeled by the difference equation

$$\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t)) \tag{12.2}$$

Here we assume that $\mathbf{f}$: $X \to \mathbf{R}^n$, where $X \subseteq \mathbf{R}^n$ is the state space. We also assume that function $\mathbf{f}$ is continuous; furthermore, for arbitrary initial state $\mathbf{x}_0 \in X$, there is a unique solution of the corresponding initial value problem $\mathbf{x}(t_0) = \mathbf{x}_0$, and the entire trajectory $\mathbf{x}(t)$ is in X. Assume furthermore that t_0 denotes the initial time period of the system.

It is also known that a vector $\bar{\mathbf{x}} \in X$ is an equilibrium state of the continuous system, Eq. (12.1), if and only if $\mathbf{f}(\bar{\mathbf{x}}) = \mathbf{0}$, and it is an equilibrium state of the discrete system, Eq. (12.2), if and only if $\bar{\mathbf{x}} = \mathbf{f}(\bar{\mathbf{x}})$. In this chapter the equilibrium of a system will always mean the equilibrium *state*, if it is not specified otherwise. In analyzing the dependence of the state trajectory $\mathbf{x}(t)$ on the selection of the initial state $\mathbf{x}_0$ nearby the equilibrium, the following stability types are considered.

Definition 12.1

1. An equilibrium state $\bar{\mathbf{x}}$ is stable if there is an $\varepsilon_0 > 0$ with the following property: For all ε_1, $0 < \varepsilon_1 < \varepsilon_0$, there is an $\varepsilon > 0$ such that if $\| \bar{\mathbf{x}} - \mathbf{x}_0 \| < \varepsilon$, then $\| \bar{\mathbf{x}} - \mathbf{x}(t) \| < \varepsilon_1$, for all $t > t_0$.
2. An equilibrium state $\bar{\mathbf{x}}$ is **asymptotically stable** if it is stable and there is an $\varepsilon > 0$ such that whenever $\| \bar{\mathbf{x}} - \mathbf{x}_0 \| < \varepsilon$, then $\mathbf{x}(t) \to \bar{\mathbf{x}}$ as $t \to \infty$.
3. An equilibrium state $\bar{\mathbf{x}}$ is **globally asymptotically stable** if it is stable and with arbitrary initial state $\mathbf{x}_0 \in X$, $\mathbf{x}(t) \to \bar{\mathbf{x}}$ as $t \to \infty$.

The first definition says an equilibrium state $\bar{\mathbf{x}}$ is stable if the entire trajectory $\mathbf{x}(t)$ is closer to the equilibrium state than any small ε_1, if the initial state $\mathbf{x}_0$ is selected close enough to the equilibrium state. For asymptotic stability, in addition, $\mathbf{x}(t)$ has to converge to the equilibrium state as $t \to \infty$. If an equilibrium state is globally asymptotically stable, then $\mathbf{x}(t)$ converges to the equilibrium state regardless of how the initial state $\mathbf{x}_0$ is selected.

These stability concepts are called **internal**, because they represent properties of the state of the system. They are illustrated in Fig. 12.1.

In the electrical engineering literature, sometimes our stability definition is called marginal stability, and our asymptotic stability is called stability.

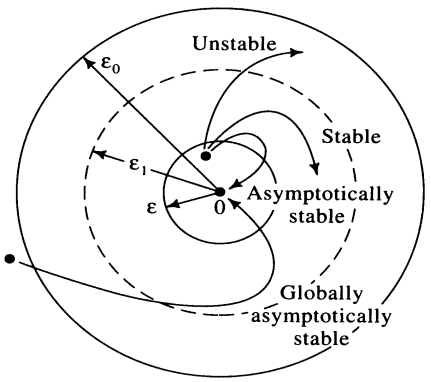


FIGURE 12.1 Stability concepts. (*Source:* F. Szidarovszky and A.T. Bahill, *Linear Systems Theory*, Boca Raton, Fla.: CRC Press, 1992, p. 168. With permission.)

12.3 Lyapunov Stability Theory

Assume that $\bar{\mathbf{x}}$ is an equilibrium state of a continuous or discrete system, and let Ω denote a subset of the state space X such that $\bar{\mathbf{x}} \in \Omega$.

Definition 12.2

A real-valued function V defined on Ω is called a Lyapunov function, if

1. V is continuous;
2. V has a unique global minimum at $\bar{\mathbf{x}}$ with respect to all other points in Ω;
3. for any state trajectory $\mathbf{x}(t)$ contained in Ω, $V(\mathbf{x}(t))$ is nonincreasing in t.

The Lyapunov function can be interpreted as the generalization of the energy function in electrical systems. The first requirement simply means that the graph of V has no discontinuities. The second requirement means that the graph of V has its lowest point at the equilibrium, and the third requirement generalizes the well-known fact of electrical systems, that the energy in a free electrical system with resistance always decreases, unless the system is at rest.

Theorem 12.1

Assume that there exists a Lyapunov function V on the spherical region

$$\Omega = \{\mathbf{x} \mid \| \mathbf{x} - \bar{\mathbf{x}} \| < \varepsilon_0\} \tag{12.3}$$

where $\varepsilon_0 > 0$ is given; furthermore $\Omega \subseteq X$. Then the equilibrium state is stable.

Theorem 12.2

Assume that in addition to the conditions of Theorem 12.1, the Lyapunov function $V(\mathbf{x}(t))$ is strictly decreasing in t, unless $\mathbf{x}(t) = \bar{\mathbf{x}}$. Then the equilibrium state is asymptotically stable.

Theorem 12.3

Assume that the Lyapunov function is defined on the entire state space X, $V(\mathbf{x}(t))$ is strictly decreasing in t unless $\mathbf{x}(t) = \bar{\mathbf{x}}$; furthermore, $V(\mathbf{x})$ tends to infinity as any component of $\mathbf{x}$ gets arbitrarily large in magnitude. Then the equilibrium state is globally asymptotically stable.

Example 12.1

Consider the differential equation

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

The stability of the equilibrium state $(1/\omega, 0)^T$ can be verified directly by using Theorem 12.1 without computing the solution. Select the Lyapunov function

$$V(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{x} - \bar{\mathbf{x}}) = \| \mathbf{x} - \bar{\mathbf{x}} \|_2^2$$

where the Euclidian norm is used.

This is continuous in $\mathbf{x}$; furthermore, it has its minimal (zero) value at $\mathbf{x} = \bar{\mathbf{x}}$. Therefore, to establish the stability of the equilibrium state we have to show only that $V(\mathbf{x}(t))$ is decreasing. Simple differentiation shows that

$$\frac{d}{dt} V(\mathbf{x}(t)) = 2(\mathbf{x} - \bar{\mathbf{x}})^T \cdot \dot{\mathbf{x}} = 2(\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{A}\mathbf{x} + \mathbf{b})$$

with

$$\mathbf{A} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

That is, with $\mathbf{x} = (x_1, x_2)^T$,

$$\begin{aligned} \frac{d}{dt} V(\mathbf{x}(t)) &= 2\left(x_1 - \frac{1}{\omega}, \quad x_2 \right)\begin{pmatrix} \omega x_2 \\ -\omega x_1 + 1 \end{pmatrix} \\ &= 2(\omega x_1 x_2 - x_2 - \omega x_1 x_2 + x_2) = 0 \end{aligned}$$

Therefore, function $V(\mathbf{x}(t))$ is a constant, which is a (not strictly) decreasing function. That is, all conditions of Theorem 12.1 are satisfied, which implies the stability of the equilibrium state.

Theorems 12.1, 12.2, and 12.3 guarantee, respectively, the stability, asymptotic stability, and global asymptotic stability of the equilibrium state, if a Lyapunov function is found. Failure to find such a Lyapunov function does not mean that the system is unstable or that the stability is not asymptotic or globally asymptotic. It only means that you were not clever enough to find a Lyapunov function that proved stability.

12.4 Stability of Time-Invariant Linear Systems

This section is divided into two subsections. In the first subsection the stability of linear time-invariant systems given in state-space notation is analyzed. In the second subsection, methods based on transfer functions are discussed.

Stability Analysis with State-Space Notation

Consider the time-invariant continuous linear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{12.4}$$

and the time-invariant discrete linear system

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{b} \tag{12.5}$$

Assume that $\bar{\mathbf{x}}$ is an equilibrium state, and let $\phi(t,t_0)$ denote the fundamental matrix.

Theorem 12.4

1. The equilibrium state $\bar{\mathbf{x}}$ is stable if and only if $\phi(t,t_0)$ is bounded for $t \geq t_0$.
2. The equilibrium state $\bar{\mathbf{x}}$ is asymptotically stable if and only if $\phi(t,t_0)$ is bounded and tends to zero as $t \to \infty$.

We use the symbol s to denote complex frequency, i.e., $s = \sigma + j\omega$. For specific values of s, such as eigenvalues and poles, we use the symbol λ.

Theorem 12.5

1. If for at least one eigenvalue of A, Re $\lambda_i > 0$ (or $|\lambda_i| > 1$ for discrete systems), then the system is unstable.
2. Assume that for all eigenvalues λ_i of $\mathbf{A}$, Re $\lambda_i \leq 0$ in the continuous case (or $|\lambda_i| \leq 1$ in the discrete case), and all eigenvalues with the property Re $\lambda_i = 0$ (or $|\lambda_i| = 1$) have single multiplicity; then the equilibrium state is stable.
3. The stability is asymptotic if and only if for all i, Re $\lambda_i < 0$ (or $|\lambda_i| < 1$).

Remark 1. Note that Part 2 gives only sufficient conditions for the stability of the equilibrium state. As the following example shows, these conditions are not necessary.

Example 12.2

Consider first the continuous system $\dot{\mathbf{x}} = \mathbf{O}\mathbf{x}$, where $\mathbf{O}$ is the zero matrix. Note that all constant functions $\mathbf{x}(t) \equiv \bar{\mathbf{x}}$ are solutions and also equilibrium states. Since

$$\phi(t, t_0) = e^{\mathbf{O}(t-t_0)} = \mathbf{I}$$

is bounded (being independent of t), all equilibrium states are stable, but $\mathbf{O}$ has only one eigenvalue $\lambda_1 = 0$ with zero real part and multiplicity n, where n is the order of the system.

Consider next the discrete systems $\mathbf{x}(t+1) = \mathbf{I}\mathbf{x}(t)$, when all constant functions $\mathbf{x}(t) \equiv \bar{\mathbf{x}}$ are also solutions and equilibrium states. Furthermore,

$$\phi(t, t_0) = \mathbf{A}^{t-t_0} = \mathbf{I}^{t-t_0} = \mathbf{I}$$

which is obviously bounded. Therefore, all equilibrium states are stable, but the condition of Part 2 of the theorem is violated again, since $\lambda_1 = 1$ with unit absolute value having a multiplicity n.

Remark 2. The following extension of Theorem 12.5 can be proven. The equilibrium state is stable if and only if for all eigenvalues of $\mathbf{A}$, $\text{Re}\,\lambda_i \le 0$ (or $|\lambda_i| \le 1$), and if λ_i is a repeated eigenvalue of A such that $\text{Re}\,\lambda_i = 0$ (or $|\lambda_i| = 1$), then the size of each block containing λ_i in the Jordan canonical form of A is 1×1.

Remark 3. The equilibrium states of inhomogeneous equations are stable or asymptotically stable if and only if the same holds for the equilibrium states of the corresponding homogeneous equations.

Example 12.3

Consider again the continuous system

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

the stability of which was analyzed earlier in Example 12.1 by using the Lyapunov function method. The characteristic polynomial of the coefficient matrix is

$$\varphi(s) = \det \begin{pmatrix} -s & \omega \\ -\omega & -s \end{pmatrix} = s^2 + \omega^2$$

therefore, the eigenvalues are $\lambda_1 = j\omega$ and $\lambda_2 = -j\omega$. Both eigenvalues have single multiplicities, and $\text{Re}\,\lambda_1 = \text{Re}\,\lambda_2 = 0$. Hence, the conditions of Part 2 are satisfied, and therefore the equilibrium state is stable. The conditions of Part 3 do not hold. Consequently, the system is not asymptotically stable.

If a time-invariant system is nonlinear, then the Lyapunov method is the most popular choice for stability analysis. If the system is linear, then the direct application of Theorem 12.5 is more attractive, since the eigenvalues of the coefficient matrix $\mathbf{A}$ can be obtained by standard methods. In addition, several conditions are known from the literature that guarantee the asymptotic stability of time-invariant discrete and continuous systems even without computing the eigenvalues. For examining asymptotic stability, linearization is an alternative approach to the Lyapunov method as is shown here. Consider the time-invariant continuous and discrete systems

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$$

and

$$\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$$

Let $\mathbf{J}(\mathbf{x})$ denote the Jacobian of $\mathbf{f}(\mathbf{x})$, and let $\bar{\mathbf{x}}$ be an equilibrium state of the system. It is known that the method of linearization around the equilibrium state results in the time-invariant linear systems

$$\dot{\mathbf{x}}_\delta(t) = \mathbf{J}(\bar{\mathbf{x}})\mathbf{x}_\delta(t)$$

and

$$\mathbf{x}_\delta(t+1) = \mathbf{J}(\bar{\mathbf{x}})\mathbf{x}_\delta(t)$$

where $\mathbf{x}_\delta(t) = \mathbf{x}(t) - \bar{\mathbf{x}}$. It is also known from the theory of ordinary differential equations that the asymptotic stability of the zero vector in the linearized system implies the asymptotic stability of the equilibrium state $\bar{\mathbf{x}}$ in the original nonlinear system.

For continuous systems the following result has a special importance.

Theorem 12.6

The equilibrium state of a continuous system [Eq. (12.4)] is asymptotically stable if and only if equation

$$\mathbf{A}^T\mathbf{Q} + \mathbf{QA} = -\mathbf{M} \tag{12.6}$$

has positive definite solution $\mathbf{Q}$ with some positive definite matrix $\mathbf{M}$.

We note that in practical applications the identity matrix is almost always selected for $\mathbf{M}$. An initial stability check is provided by the following result.

Theorem 12.7

Let $\varphi(\lambda)=\lambda^n + p_{n-1}\lambda^{n-1} + \ldots + p_1\lambda + p_0$ be the characteristic polynomial of matrix $\mathbf{A}$. Assume that all eigenvalues of matrix $\mathbf{A}$ have negative real parts. Then $p_i > 0$ $(i = 0, 1,..., n-1)$.

Corollary. If any of the coefficients p_i is negative or zero, the equilibrium state of the system with coefficient matrix $\mathbf{A}$ cannot be asymptotically stable. However, the conditions of the theorem do not imply that the eigenvalues of $\mathbf{A}$ have negative real parts.

Example 12.4

For matrix

$$\mathbf{A} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}$$

the characteristic polynominal is $\varphi(s) = s^2 + \omega^2$. Since the coefficient of s^1 is zero, the system of Example 12.3 is not asymptotically stable.

The Transfer Function Approach

The transfer function of the time invariant linear continuous system

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx} \end{aligned} \tag{12.7}$$

and that of the time invariant linear discrete system

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) \end{aligned} \tag{12.8}$$

have the common form

$$\mathbf{TF}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$$

If both the input and output are single, then

$$\mathbf{TF}(s) = \frac{\mathbf{Y}(s)}{\mathbf{U}(s)}$$

or in the familiar electrical engineering notation

$$\mathbf{TF}(s) = \frac{K\mathbf{G}(s)}{1 + K\mathbf{G}(s)\mathbf{H}(s)} \tag{12.9}$$

where K is the gain term in the forward loop, $\mathbf{G}(s)$ represents the dynamics of the forward loop, or the plant, and $\mathbf{H}(s)$ models the dynamics in the feedback loop. We note that in the case of continuous systems s is the variable of the transfer function, and for discrete systems the variable is denoted by z.

After the Second World War systems and control theory flourished. The transfer function representation was the most popular representation for systems. To determine the stability of a system we merely had to factor the denominator of the transfer function (12.9) and see if all of the poles were in the left half of the complex frequency plane. However, with manual techniques, factoring polynomials of large order is difficult. So engineers, being naturally lazy people, developed several ways to determine the stability of a system without factoring the polynomials [Dorf, 1992]. First, we have the methods of Routh and Hurwitz, developed a century ago, that looked at the coefficients of the characteristic polynomial. These methods showed whether the system was stable or not, but they did not show how close the system was to being stable.

What we want to know is for what value of gain, K, and at what frequency, ω, will the denominator of the transfer function (12.9) become zero. Or, when will $K\mathbf{GH} = -1$, meaning, when will the magnitude of $K\mathbf{GH}$ equal 1 with a phase angle of –180 degrees? These parameters can be determined easily with a Bode diagram. Construct a Bode diagram for $K\mathbf{GH}$ of the system, look at the frequency where the phase angle equals –180 degrees, and look up at the magnitude plot. If it is smaller than 1.0, then the system is stable. If it is larger than 1.0, then the system is unstable. Bode diagram techniques are discussed in Chapter 11.

The quantity $K\mathbf{G}(s)\mathbf{H}(s)$ is called the open-loop transfer function of the system, because it is the effect that would be encountered by a signal in one loop around the system if the feedback loop were artificially opened [Bahill, 1981].

To gain some intuition, think of a closed-loop negative feedback system. Apply a small sinusoid at frequency ω to the input. Assume that the gain around the loop, $K\mathbf{GH}$, is 1 or more, and that the phase lag is 180 degrees. The summing junction will flip over the fed back signal and add it to the original signal. The result is a signal that is bigger than what came in. This signal will circulate around this loop, getting bigger and bigger until the real system no longer matches the model. This is what we call instability.

The question of stability can also be answered with Nyquist diagrams. They are related to Bode diagrams, but they give more information. A simple way to construct a Nyquist diagram is to make a polar plot on the complex frequency plane of the Bode diagram. Simply stated, if this contour encircles the –1 point in the complex frequency plane, then the system is unstable. The two advantages of the Nyquist technique are (1) in

addition to the information on Bode diagrams, there are about a dozen rules that can be used to help construct Nyquist diagrams, and (2) Nyquist diagrams handle bizarre systems better, as is shown in the following rigorous statement of the Nyquist stability criterion. The number of clockwise encirclements minus the number of counterclockwise encirclements of the point $s = -1 + j0$ by the Nyquist plot of $K\mathbf{G}(s)\mathbf{H}(s)$ is equal to the number of poles of $\mathbf{Y}(s)/\mathbf{U}(s)$ minus the number of poles of $K\mathbf{G}(s)\mathbf{H}(s)$ in the right half of the s-plane.

The root-locus technique was another popular technique for assessing stability. It furthermore allowed the engineer to see the effects of small changes in the gain, K, on the stability of the system. The root-locus diagram shows the location in the s-plane of the poles of the closed-loop transfer function, $\mathbf{Y}(s)/\mathbf{U}(s)$. All branches of the root-locus diagram start on poles of the open-loop transfer function, $K\mathbf{GH}$, and end either on zeros of the open-loop transfer function, $K\mathbf{GH}$, or at infinity. There are about a dozen rules to help draw these trajectories. The root-locus technique is discussed in Chapter 93.4.

We consider all these techniques to be old fashioned. They were developed to help answer the question of stability without factoring the characteristic polynomial. However, many computer programs are currently available that factor polynomials. We recommend that engineers merely buy one of these computer packages and find the roots of the closed-loop transfer function to assess the stability of a system.

The poles of a system are defined as all values of s such that $s\mathbf{I} - \mathbf{A}$ is singular. The poles of a closed-loop transfer function are exactly the same as the eigenvalues of the system: engineers prefer the term *poles* and the symbol s, and mathematicians prefer the term *eigenvalues* and the symbol λ. We will use s for complex frequency and λ for specific values of s.

Sometimes, some poles could be canceled in the rational function form of $\mathbf{TF}(s)$ so that they would not be explicitly shown. However, even if some poles could be canceled by zeros, we still have to consider all poles in the following criteria which is the statement of Theorem 12.5. The equilibrium state of the continuous system [Eq. (12.7)] with constant input is unstable if at least one pole has a positive real part, and is stable if all poles of $\mathbf{TF}(s)$ have nonpositive real parts and all poles with zero real parts are single. The equilibrium state is asymptotically stable if and only if all poles of $\mathbf{TF}(s)$ have negative real parts; that is, all poles are in the left half of the s-plane. Similarly, the equilibrium state of the discrete system [Eq. (12.8)] with constant input is unstable if the absolute value of at least one pole is greater than one, and is stable if all poles of $\mathbf{TF}(z)$ have absolute values less than or equal to one and all poles with unit absolute values are single. The equilibrium state is asymptotically stable if and only if all poles of $\mathbf{TF}(z)$ have absolute values less than one; that is, the poles are all inside the unit circle of the z-plane.

Example 12.5

Consider again the system

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

which was discussed earlier. Assume that the output equation has the form

$$y = (1, 1)\mathbf{x}$$

Then

$$\mathbf{TF}(s) = \frac{s + \omega}{s^2 + \omega^2}$$

The poles are $j\omega$ and $-j\omega$, which have zero real parts; that is, they are on the imaginary axis of the s-plane. Since they are single poles, the equilibrium state is stable but not asymptotically stable. A system such as this would produce constant amplitude sinusoids at frequency ω. So it seems natural to assume that such systems would be used to build sinusoidal signal generators and to model oscillating systems. However, this is not the case, because (1) zero resistance circuits are hard to make; therefore, most function generators use other

techniques to produce sinusoids; and (2) such systems are not good models for oscillating systems, because most real-world oscillating systems (i.e., biological systems) have energy dissipation elements in them.

More generally, real-world function generators are seldom made from closed-loop feedback control systems with 180 degrees of phase shift, because (1) it would be difficult to get a broad range of frequencies and several waveforms from such systems, (2) precise frequency selection would require expensive high-precision components, and (3) it would be difficult to maintain a constant frequency in such circuits in the face of changing temperatures and power supply variations. Likewise, closed-loop feedback control systems with 180 degrees of phase shift are not good models for oscillating biological systems, because most biological systems oscillate because of nonlinear network properties.

A special stability criterion for single-input, single-output time-invariant continuous systems will be introduced next. Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \quad \text{and} \quad y = \mathbf{c}^{\mathrm{T}}\mathbf{x} \tag{12.10}$$

where $\mathbf{A}$ is an $n \times n$ constant matrix, and $\mathbf{b}$ and $\mathbf{c}$ are constant n-dimensional vectors. The transfer function of this system is

$$TF_1(s) = \mathbf{c}^{\mathrm{T}}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$$

which is obviously a rational function of s. Now let us add negative feedback around this system so that $u = ky$, where k is a constant. The resulting system can be described by the differential equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + k\mathbf{b}\mathbf{c}^{\mathrm{T}}\mathbf{x} = (\mathbf{A} + k\mathbf{b}\mathbf{c}^{\mathrm{T}})\mathbf{x} \tag{12.11}$$

The transfer function of this feedback system is

$$\mathrm{TF}(s) = \frac{\mathrm{TF}_1(s)}{1 - k\mathrm{TF}_1(s)} \tag{12.12}$$

To help show the connection between the asymptotic stability of systems (12.10) and (12.11), we introduce the following definition.

Definition 12.3

Let $r(s)$ be a rational function of s. Then the locus of points

$$L(r) = \{a + jb \,|\, a = Re(r(jv)), \quad b = Im(r(jv)), \quad v \in \mathrm{R}\}$$

is called the *response diagram of r*. Note that $L(r)$ is the image of the imaginary line $\mathrm{Re}(s) = 0$ under the mapping r. We shall assume that $L(r)$ is bounded, which is the case if and only if the degree of the denominator is not less than that of the numerator and r has no poles on the line $\mathrm{Re}(s) = 0$.

Theorem 12.8

The Nyquist stability criterion. Assume that TF_1 has a bounded response diagram $L(TF_1)$. If TF_1 has ν poles in the right half of the s-plane, where $\mathrm{Re}(s) > 0$, then H has $\rho + \nu$ poles in the right half of the s-plane where $\mathrm{Re}(s) > 0$ if the point $1/k + j \cdot 0$ is not on $L(TF_1)$, and $L(TF_1)$ encircles $1/k + j \cdot 0$ ρ times in the clockwise sense.

Corollary. Assume that system (12.10) is asymptotically stable with constant input and that $L(TF_1)$ is bounded and traversed in the direction of increasing ν and has the point $1/k + j \cdot 0$ on its left. Then the feedback system (12.11) is also asymptotically stable.

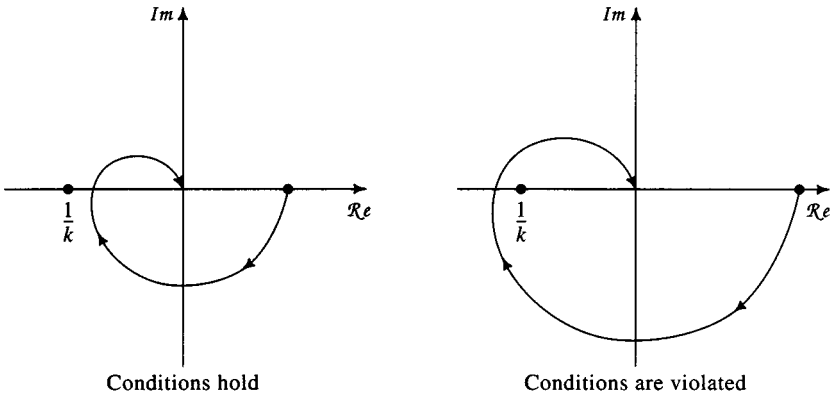


FIGURE 12.2 Illustration of Nyquist stability criteria. (*Source:* F. Szidarovszky and A. T. Bahill, *Linear Systems Theory*, Boca Raton, Fla.: CRC Press, 1992, p.184. With permission.)

This result has many applications, since feedback systems have a crucial role in constructing stabilizers, observers, and filters for given systems. Fig. 12.2 illustrates the conditions of the corollary. The application of this result is especially convenient, if system (12.10) is given and only appropriate values k of the feedback are to be determined. In such cases the locus $L(TF_1)$ has to be computed first, and then the region of all appropriate k values can be determined easily from the graph of $L(TF_1)$.

This analysis has dealt with the closed-loop transfer function, whereas the techniques of Bode, root-locus, etc. use the open-loop transfer function. This should cause little confusion as long as the distinction is kept in mind.

12.5 BIBO Stability

In the previous sections, internal stability of time-invariant systems was examined, i.e., the stability of the state was investigated. In this section the **external stability** of systems is discussed; this is usually called the **BIBO** (*b*ounded-*i*nput, *b*ounded-*o*utput) **stability**. Here we drop the simplifying assumption of the previous section that the system is time-invariant: we will examine time-variant systems.

Definition 12.4

A system is called BIBO stable if for zero initial conditions, a bounded input always evokes a bounded output.

For continuous systems a necessary and sufficient condition for BIBO stability can be formulated as follows.

Theorem 12.9

Let $\mathbf{T}(t,\tau) = (t_{ij}(t,\tau))$ be the weighting pattern, $\mathbf{C}(t)\phi(t,\tau)\mathbf{B}(\tau)$, of the system. Then the continuous time-variant linear system is BIBO stable if and only if the integral

$$\int_{t_0}^{t} \left| t_{ij}(t,\tau) \right| d\tau \tag{12.13}$$

is bounded for all $t > t_0$, i and j.

Corollary. Integrals (12.13) are all bounded if and only if

$$I(t) = \int_{t_0}^{t} \sum_i \sum_j \left| t_{ij}(t,\tau) \right| d\tau \tag{12.14}$$

is bounded for $t \geq t_0$. Therefore, it is sufficient to show the boundedness of only one integral in order to establish BIBO stability.
The discrete counterpart of this theorem can be given in the following way.

Theorem 12.10

Let $\mathbf{T}(t,\tau) = (t_{ij}(t,\tau))$ be the weighting pattern of the discrete linear system. Then it is BIBO stable if and only if the sum

$$I(t) = \sum_{\tau=t_0}^{t-1} \left| t_{ij}(t,\tau) \right| \tag{12.15}$$

is bounded for all $t > t_0$, i and j.
Corollary. The sums (12.15) are all bounded if and only if

$$\sum_{\tau=t_0}^{t-1} \sum_{i} \sum_{j} \left| t_{ij}(t,\tau) \right| \tag{12.16}$$

is bounded. Therefore it is sufficient to verify the boundedness of only one sum in order to establish BIBO stability.

Consider next the time-invariant case, when $\mathbf{A}(t) \equiv \mathbf{A}$, $\mathbf{B}(t) \equiv \mathbf{B}$ and $\mathbf{C}(t) \equiv \mathbf{C}$. From the foregoing theorems and the definition of $\mathbf{T}(t,\tau)$ we have immediately the following sufficient condition.

Theorem 12.11

Assume that for all eigenvalues λ_i of $\mathbf{A}$, $\operatorname{Re}\lambda_i < 0$ (or $|\lambda_i| < 1$). Then the time-invariant linear continuous (or discrete) system is BIBO stable.

Finally, we note that BIBO stability is not implied by an observation that a certain bounded input generates bounded output. All bounded inputs must generate bounded outputs in order to guarantee BIBO stability.

Adaptive-control systems are time-varying systems. Therefore, it is usually difficult to prove that they are stable. Szidarovszky et al. [1990], however, show a technique for doing this. This new result gives a necessary and sufficient condition for the existence of an asymptotically stable model-following adaptive-control system based on the solvability of a system of nonlinear algebraic equations, and in the case of the existence of such systems they present an algorithm for finding the appropriate feedback parameters.

12.6 Physical Examples

In this section we show some examples of stability analysis of physical systems.

1. Consider a simple *harmonic oscillator* constructed of a mass and an ideal spring. Its dynamic response is summarized with

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u$$

In Example 12.3 we showed that this system is stable but not asymptotically stable. This means that if we leave it alone in its equilibrium state, it will remain stationary, but if we jerk on the mass it will oscillate forever. There is no damping term to remove the energy, so the energy will be transferred back and forth between potential energy in the spring and kinetic energy in the moving mass. A good approximation of such a harmonic oscillator is a pendulum clock. The more expensive it is (i.e., the smaller the damping), the less often we have to wind it (i.e., add energy).

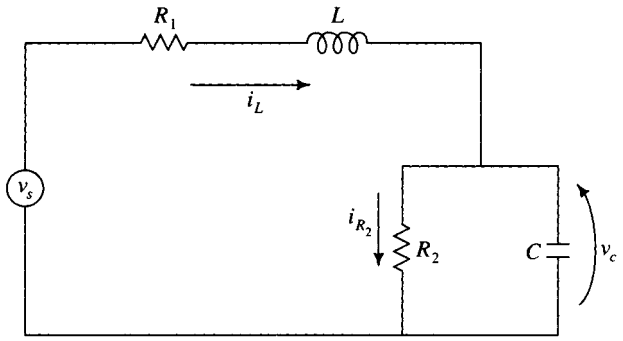


FIGURE 12.3 A simple electrical system. (*Source:* F. Szidarovszky and A. T. Bahill, *Linear Systems Theory*, Boca Raton, Fla.: CRC Press, 1992, p. 125. With permission.)

2. A *linear second-order electrical system* composed of a series connection of an input voltage source, an inductor, a resistor, and a capacitor, with the output defined as the voltage across the capacitor, can be characterized by the second-order equation

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{LCs^2 + RCs + 1}$$

For convenience, let us define

$$\omega_n = \sqrt{\frac{1}{LC}} \quad \text{and} \quad \zeta = \frac{R}{2}\sqrt{\frac{C}{L}}$$

and assume that $\zeta < 1$. With these parameters the transfer function becomes

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n + \omega_n^2}$$

Is this system stable? The roots of the characteristic equation are

$$\lambda_{1,2} = -\zeta\omega_n \pm j\omega_n\sqrt{1-\zeta^2}$$

If $\zeta > 0$, the poles are in the left half of the s-plane, and therefore the system is asymptotically stable. If $\zeta = 0$, as in the previous example, the poles are on the imaginary axis; therefore, the system is stable but not asymptotically stable. If $\zeta < 0$, the poles are in the right half of the s-plane and the system is unstable.

3. An *electrical system* is shown in Fig. 12.3. Simple calculation shows that by defining the state variables

$$x_1 = i_L, \quad x_2 = v_c, \quad \text{and} \quad u = v_s$$

the system can be described by the differential equations

$$\dot{x}_1 = -\frac{R_1}{L}x_1 - \frac{1}{L}x_2 + \frac{1}{L}u$$

$$\dot{x}_2 = \frac{1}{C}x_1 - \frac{1}{CR_2}x_2$$

The characteristic equation has the form

$$\left(-s - \frac{R_1}{L}\right)\left(-s - \frac{1}{CR_2}\right) + \frac{1}{LC} = 0$$

which simplifies as

$$s^2 + s\left(\frac{R_1}{L} + \frac{1}{CR_2}\right) + \left(\frac{R_1}{LCR_2} + \frac{1}{LC}\right) = 0$$

Since R_1, R_2, L, and C are positive numbers, the coefficients of this equation are all positive. The constant term equals $\lambda_1\lambda_2$, and the coefficient of s^1 is $-(\lambda_1 + \lambda_2)$. Therefore

$$\lambda_1 + \lambda_2 < 0 \quad \text{and} \quad \lambda_1\lambda_2 > 0$$

If the eigenvalues are real, then these relations hold if and only if both eigenvalues are negative. If they were positive, then $\lambda_1 + \lambda_2 > 0$. If they had different signs, then $\lambda_1\lambda_2 < 0$. Furthermore, if at least one eigenvalue is zero, then $\lambda_1\lambda_2 = 0$. Assume next that the eigenvalues are complex:

$$\lambda_{1,2} = Re\, s \pm j\, Im\, s$$

Then

$$\lambda_1 + \lambda_2 = 2Re\, s$$

and

$$\lambda_1\lambda_2 = (Re\, s)^2 + (Im\, s)^2$$

Hence $\lambda_1 + \lambda_2 < 0$ if and only if $\text{Re}\, s < 0$.

In summary, the system is asymptotically stable, since in both the real and complex cases the eigenvalues have negative values and negative real parts, respectively.

4. The classical *stick balancing* problem is shown in Fig. 12.4. Simple analysis shows that $y(t)$ satisfies the second-order equation

$$\ddot{y} = \frac{g}{L}(y - u)$$

If one selects $L = 1$, then the characteristic equation has the form

$$s^2 - g = 0$$

So, the eigenvalues are

$$\lambda_{1,2} = \pm\sqrt{g}$$

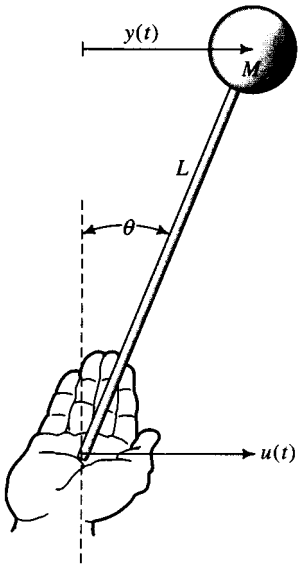


FIGURE 12.4 Stick balancing. (*Source:* F. Szidarovszky and A. T. Bahill, *Linear Systems Theory*, Boca Raton, Fla.: CRC Press, 1992, p. 127. With permission.)

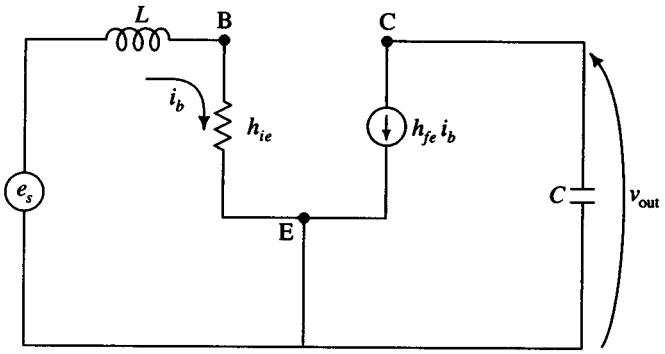


FIGURE 12.5 A model for a simple transistor circuit. (*Source:* F. Szidarovszky and A. T. Bahill, *Linear Systems Theory*, Boca Raton, Fla.: CRC Press 1992, p. 127. With permission.)

One is in the right half of the *s*-plane and the other is in the left half of the *s*-plane, so the system is unstable. This instability is understandable, since without an intelligent input to control the system, if the stick is not upright with zero velocity, it will fall over.

5. A simple *transistor circuit* can be modeled as shown in Fig. 12.5. The state variables are related to the input and output of the circuit: the base current, i_b, is x_1 and the output voltage, v_{out}, is x_2. Therefore,

$$\dot{\mathbf{x}} = \begin{pmatrix} -\dfrac{h_{ie}}{L} & 0 \\ \dfrac{h_{fe}}{C} & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \dfrac{1}{L} \\ 0 \end{pmatrix} e_s \text{ and } \mathbf{c}^T = (0, 1)$$

The **A** matrix looks strange with a column of all zeros, and indeed the circuit does exhibit odd behavior. For example, as we will show, there is no equilibrium state for a unit step input of e_s. This is reasonable, however,

because the model is for mid-frequencies, and a unit step does not qualify. In response to a unit step the output voltage will increase linearly until the model is no longer valid. If e_s is considered to be the input, then the system is

$$\dot{\mathbf{x}} = \begin{pmatrix} -\dfrac{h_{ie}}{L} & 0 \\ \dfrac{h_{fe}}{C} & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \dfrac{1}{L} \\ 0 \end{pmatrix} u$$

If $u(t) \equiv 1$, then at the equilibrium state:

$$\begin{pmatrix} -\dfrac{h_{ie}}{L} & 0 \\ \dfrac{h_{fe}}{C} & 0 \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} + \begin{pmatrix} \dfrac{1}{L} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

That is,

$$-\frac{h_{ie}}{L}\,\bar{x}_1 + \frac{1}{L} = 0$$

$$\frac{h_{fe}}{C}\,\bar{x}_1 = 0$$

Since $h_{fe}/C \neq 0$, the second equation implies that $\bar{x}_1 = 0$, and by substituting this value into the first equation we get the obvious contradiction $1/L = 0$. Hence, with nonzero constant input *no* equilibrium state exists.

Let us now investigate the stability of this system. First let $\tilde{\mathbf{x}}(t)$ denote a fixed trajectory of this system, and let $\mathbf{x}(t)$ be an arbitrary solution. Then the difference $\mathrm{x}_\delta(t) = \mathbf{x}(t) - \tilde{\mathbf{x}}(t)$ satisfies the homogeneous equation

$$\dot{\mathrm{x}}_\delta = \begin{pmatrix} -\dfrac{h_{ie}}{L} & 0 \\ \dfrac{h_{fe}}{C} & 0 \end{pmatrix} \mathrm{x}_\delta$$

This system has an equilibrium $\mathbf{x}_\delta(t) = 0$. Next, the stability of this equilibrium is examined by solving for the poles of the closed-loop transfer function. The characteristic equation is

$$\det \begin{pmatrix} -\dfrac{h_{ie}}{L} - s & 0 \\ \dfrac{h_{fe}}{C} & -s \end{pmatrix} = 0$$

which can be simplified as

$$s^2 + s\,\frac{h_{ie}}{L} + 0 = 0$$

The roots are

$$\lambda_1 = 0 \quad \text{and} \quad \lambda_2 = -\frac{h_{ie}}{L}$$

Therefore, the system is stable but not asymptotically stable. This stability means that for small changes in the initial state the entire trajectory $\mathbf{x}(t)$ remains close to $\tilde{\mathbf{x}}(t)$.

Defining Terms

Asymptotic stability: An equilibrium state $\bar{\mathbf{x}}$ of a system is asymptotically stable if, in addition to being stable, there is an $\varepsilon > 0$ such that whenever $\|\bar{\mathbf{x}} - \mathbf{x}_0\| < \varepsilon$, then $\mathbf{x}(t) \to \bar{\mathbf{x}}$ as $t \to \infty$. A system is asymptotically stable if all the poles of the closed-loop transfer function are in the left half of the *s*-plane (inside the unit circle of the *z*-plane for discrete systems). This is sometimes called *stability.*

BIBO stability: A system is BIBO stable if for zero initial conditions a bounded input always evokes a bounded output.

External stability: Stability concepts related to the input-output behavior of the system.

Global asymptotic stability: An equilibrium state $\bar{\mathbf{x}}$ of a system is globally asymptotically stable if it is stable and with arbitrary initial state $\mathbf{x}_0 \in X$, $\mathbf{x}(t) \to \bar{\mathbf{x}}$ as $t \to \infty$.

Internal stability: Stability concepts related to the state of the system.

Instability: An equilibrium state of a system is unstable if it is not stable. A system is unstable if at least one pole of the closed-loop transfer function is in the right half of the *s*-plane (outside the unit circle of the *z*-plane for discrete systems).

Stability: An equilibrium state $\bar{\mathbf{x}}$ of a system is stable if there is an $\varepsilon_0 > 0$ with the following property: for all ε_1, $0 < \varepsilon_1 < \varepsilon_0$, there is an $\varepsilon > 0$ such that if $\|\bar{\mathbf{x}} - \mathbf{x}_0\| < \varepsilon$, then $\|\bar{\mathbf{x}} - \mathbf{x}(t)\| < \varepsilon_1$ for all $t > t_0$. A system is stable if the poles of its closed-loop transfer function are (1) in the left half of the complex frequency plane, called the *s*-plane (inside the unit circle of the *z*-plane for discrete systems), or (2) on the imaginary axis, and all of the poles on the imaginary axis are single (on the unit circle and all such poles are single for discrete systems). Stability for a system with repeated poles on the $j\omega$ axis (the unit circle) is complicated and is examined in the discussion after Theorem 12.5. In the electrical engineering literature, this definition of stability is sometimes called *marginal stability* and sometimes *stability in the sense of Lyapunov.*

Related Topics

6.2 Applications • 7.2 State Equations in Normal Form • 100.2 Dynamic Response • 100.7 Nonlinear Control Systems

References

A. T. Bahill, *Bioengineering: Biomedical, Medical and Clinical Engineering*, Englewood Cliffs, N.J.:Prentice-Hall, 1981, pp. 214–215, 250–252.

R. C. Dorf, *Modern Control Systems*, 7th ed., Reading, Mass.: Addison-Wesley, 1996.

M. Jamshidi, M. Tarokh, and B. Shafai, *Computer-Aided Analysis and Design of Linear Control Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1992.

F. Szidarovszky and A. T. Bahill, *Linear Systems Theory*, Boca Raton, Fla.: CRC Press, 1992.

F. Szidarovszky, A. T. Bahill, and S. Molnar, "On stable adaptive control systems," *Pure Math. and Appl.*, vol. 1, ser. B, no. 2–3, pp. 115–121, 1990.

Further Information

For further information consult the textbooks *Modern Control Systems* by Dorf [1996] or *Linear Systems Theory* by Szidarovszky and Bahill [1992].

Rollins, J.G., Bendix, P. "Computer Software for Circuit Analysis and Design"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

13[1]

Computer Software for Circuit Analysis and Design

J. Gregory Rollins
Technology Modeling Associates, Inc.

Peter Bendix
LSI Logic Corp.

13.1 Analog Circuit Simulation
Introduction • DC (Steady-State) Analysis • AC Analysis • Transient Analysis • Process and Device Simulation • Process Simulation • Device Simulation • Appendix
13.2 Parameter Extraction for Analog Circuit Simulation
Introduction • MOS DC Models • BSIM Extraction Strategy in Detail

13.1 Analog Circuit Simulation

J. Gregory Rollins

Introduction

Computer-aided simulation is a powerful aid during the design or analysis of electronic circuits and semiconductor devices. The first part of this chapter focuses on analog circuit simulation. The second part covers simulations of semiconductor processing and devices. While the main emphasis is on analog circuits, the same simulation techniques may, of course, be applied to digital circuits (which are, after all, composed of analog circuits). The main limitation will be the size of these circuits because the techniques presented here provide a very detailed analysis of the circuit in question and, therefore, would be too costly in terms of computer resources to analyze a large digital system.

The most widely known and used circuit simulation program is SPICE (simulation program with integrated circuit emphasis). This program was first written at the University of California at Berkeley by Laurence Nagel in 1975. Research in the area of circuit simulation is ongoing at many universities and industrial sites. Commercial versions of SPICE or related programs are available on a wide variety of computing platforms, from small personal computers to large mainframes. A list of some commercial simulator vendors can be found in the Appendix.

It is possible to simulate virtually any type of circuit using a program like SPICE. The programs have built-in elements for resistors, capacitors, inductors, dependent and independent voltage and current sources, diodes, MOSFETs, JFETs, BJTs, transmission lines, transformers, and even transformers with saturating cores in some versions. Found in commercial versions are libraries of standard components which have all necessary

[1]The material in this chapter was previously published by CRC Press in *The Circuits and Filters Handbook,* Wai-Kai Chen, Ed., 1995.

parameters prefitted to typical specifications. These libraries include items such as discrete transistors, op amps, phase-locked loops, voltage regulators, logic integrated circuits (ICs) and saturating transformer cores.

Computer-aided circuit simulation is now considered an essential step in the design of integrated circuits, because without simulation the number of "trial runs" necessary to produce a working IC would greatly increase the cost of the IC. Simulation provides other advantages, however:

- The ability to measure "inaccessible" voltages and currents. Because a mathematical model is used all voltages and currents are available. No loading problems are associated with placing a voltmeter or oscilloscope in the middle of the circuit, with measuring difficult one-shot wave forms, or probing a microscopic die.
- Mathematically ideal elements are available. Creating an ideal voltage or current source is trivial with a simulator, but impossible in the laboratory. In addition, all component values are exact and no parasitic elements exist.
- It is easy to change the values of components or the configuration of the circuit. Unsoldering leads or redesigning IC masks are unnecessary.

Unfortunately, computer-aided simulation has its own problems:

- Real circuits are distributed systems, not the "lumped element models" which are assumed by simulators. Real circuits, therefore, have resistive, capacitive, and inductive parasitic elements present besides the intended components. In high-speed circuits these parasitic elements are often the dominant performance-limiting elements in the circuit, and must be painstakingly modeled.
- Suitable predefined numerical models have not yet been developed for certain types of devices or electrical phenomena. The software user may be required, therefore, to create his or her own models out of other models which are available in the simulator. (An example is the solid-state thyristor which may be created from a NPN and PNP bipolar transistor.)
- The numerical methods used may place constraints on the form of the model equations used.

The following sections consider the three primary simulation modes: DC, AC, and transient analysis. In each section an overview is given of the numerical techniques used. Some examples are then given, followed by a brief discussion of common pitfalls.

DC (Steady-State) Analysis

DC analysis calculates the state of a circuit with fixed (non-time varying) inputs after an infinite period of time. DC analysis is useful to determine the operating point (Q-point) of a circuit, power consumption, regulation and output voltage of power supplies, transfer functions, noise margin and fanout in logic gates, and many other types of analysis. In addition DC analysis is used to find the starting point for AC and transient analysis. To perform the analysis the simulator performs the following steps:

1. All capacitors are removed from the circuit (replaced with opens).
2. All inductors are replaced with shorts.
3. Modified nodal analysis is used to construct the nonlinear circuit equations. This results in one equation for each circuit node plus one equation for each voltage source. Modified nodal analysis is used rather than standard nodal analysis because an ideal voltage source or inductance cannot be represented using normal nodal analysis. To represent the voltage sources, loop equations (one for each voltage source or inductor), are included as well as the standard node equations. The node voltages and voltage source currents, then, represent the quantities which are solved for. These form a vector $\mathbf{x}$. The circuit equations can also be represented as a vector $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.
4. Because the equations are nonlinear, Newton's method (or a variant thereof) is then used to solve the equations.

Example 13.1. **Simulation Voltage Regulator:** We shall now consider simulation of the type 723 voltage regulator IC, shown in Fig. 13.1. We wish to simulate the IC and calculate the sensitivity of the output IV

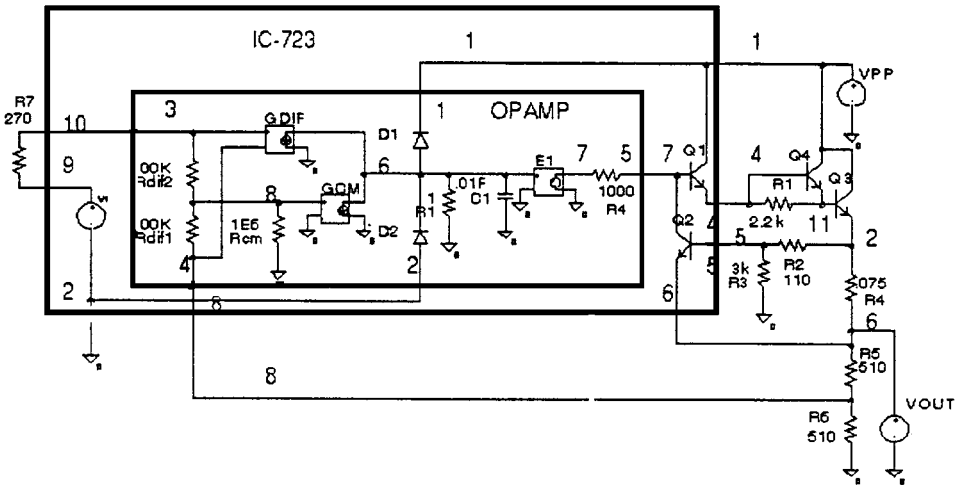


FIGURE 13.1 Regulator circuit to be used for DC analysis, created using PSPICE.

characteristic and verify that the output current follows a "fold-back" type characteristic under overload conditions.

The IC itself contains a voltage reference source and operational amplifier. Simple models for these elements are used here rather than representing them in their full form, using transistors, to illustrate model development. The use of simplified models can also greatly reduce the simulation effort. (For example, the simple op amp used here requires only eight nodes and ten components, yet realizes many advanced features.)

Note in Fig. 13.1 that the numbers next to the wires represent the circuit nodes. These numbers are used to describe the circuit to the simulator. In most SPICE-type simulators the nodes are represented by numbers, with the ground node being node zero. Referring to Fig. 13.2, the 723 regulator and its internal op amp are represented by subcircuits. Each subcircuit has its own set of nodes and components. Subcircuits are useful for encapsulating sections of a circuit or when a certain section needs to be used repeatedly (see next section).

The following properties are modeled in the op amp:

1. Common mode gain
2. Differential mode gain
3. Input impedance
4. Output impedance
5. Dominant pole
6. Output voltage clipping

The input terminals of the op amp connect to a "T" resistance network, which sets the common and differential mode input resistance. Therefore, the common mode resistance is RCM + RDIF = 1.1E6 and the differential mode resistance is RDIF1 + RDIF2 = 2.0E5.

Dependent current sources are used to create the main gain elements. Because these sources force current into a 1-Ω resistor, the voltage gain is Gm*R at low frequency. In the differential mode this gives (GDIF*R1 = 100). In the common mode this gives (GCM*R1*(RCM/(RDIF1 + RCM = 0.0909). The two diodes D1 and D2 implement clipping by preventing the voltage at node 6 from exceeding VCC or going below VEE. The diodes are made "ideal" by reducing the ideality factor n. Note that the diode current is $I_d = I_s[\exp(V_d/(nV_t)) - 1]$, where V_t is the thermal voltage (0.026 V). Thus, reducing n makes the diode turn on at a lower voltage.

A single pole is created by placing a capacitor (C1) in parallel with resistor R1. The pole frequency is therefore given by 1.0/(2*π*R1*C1). Finally, the output is driven by the voltage-controlled voltage source E1 (which has a voltage gain of unity), through the output resistor R4. The output resistance of the op amp is therefore equal to R4.

To observe the output voltage as a function of resistance, the regulator is loaded with a voltage source (VOUT) and the voltage source is swept from 0.05 to 6.0 V. A plot of output voltage vs. resistance can then be obtained

```
Regulator circuit.
* Complete circuit *
* Load source*
vout 6 0
* Power input *
vpp 1 0 11
x1 1 0 4 5 6 7 8 9 10 ic723
* Series Pass transistors *
q3 1 4 11 mq3
q4 1 11 2 mq4
r1 4 11 2.2k
r2 5 2 110
r3 5 0 3k
r4 2 6 0.075
r5 6 8 510
r6 8 0 510
r7 9 10 270
* Control cards *
.op
.model mq3 npn(is=1e-9 bf=30
+ br=5 ikf=50m)
.model mq4 npn(is=1e-6 bf=30
+ br=5 ikf=10)
.dc vout 1 5.5 .01
.plot dc i(vout)
.probe

.subckt ic723 1 2  4 5 6 7 8 9 10
* Type 723 voltage regulator *
x1 1 2 10 8 7 opamp
* Internal voltage reference *
vr 9 2 2.5
q1 3 7 4 mm
q2 7 5 6 mm
.model mm npn (is=1e-12 bf=100
+ br=5)
.ends ic723
* Ideal opamp with limiting
.subckt opamp 1   2   3   4   5
*              vcc vee +in -in out
rdif1 3 8 1e5
rdif2 4 8 1e5
rcm  8 0 1e6
* Common mode gain *
gcm 6 0 8 0 1e-1
* Differential mode gain *
gdif 6 0 4 3 100
r1 6 0 1
* Single pole response *
c1 6 0 .01
d1 6 1 ideal
d2 2 6 ideal
e1 7 0 6 0 1
rout 5 7 1e3
.model ideal d (is=1e-6 n=.01)
.ends opamp
```

FIGURE 13.2 SPICE input listing of regulator circuit shown in Fig. 13.1.

by plotting VOUT vs. VOUT/I(VOUT) (using PROBE in this case; see Fig. 13.3). Note that for this circuit, eventhough a current source would seem a more natural choice, a voltage source must be used as a load rather than a current source because the output characteristic curve is multivalued in current. If a current source were used it would not be possible to easily simulate the entire curve. Of course, many other interesting quantities can be plotted; for example, the power dissipated in the pass transistor can be approximated by plotting IC(Q3)*VC(Q3).

For these simulations PSPICE was used running on an IBM PC. The simulation took < 1 min of CPU time.

Pitfalls. Convergence problems are sometimes experienced if "difficult" bias conditions are created. An example of such a condition is if a diode is placed in the circuit backwards, resulting in a large forward bias voltage, SPICE will have trouble resolving the current. Another difficult case is if a current source were used instead of

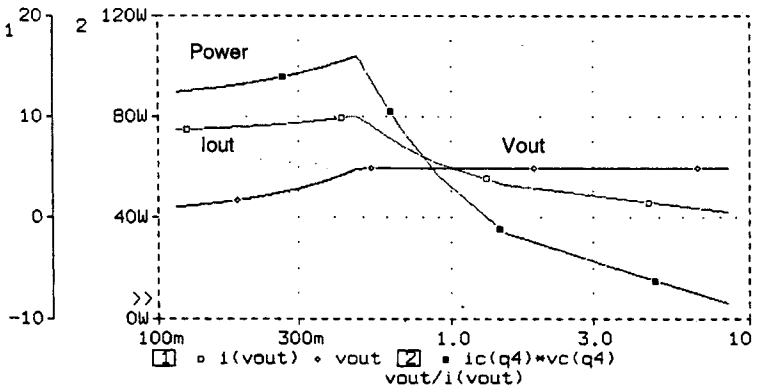


FIGURE 13.3 Output characteristics of regulator circuit using PSPICE.

a voltage to bias the output in the previous example. If the user then tried to increase the output current above 10 A, SPICE would not be able to converge because the regulator will not allow such a large current.

AC Analysis

Ac analysis uses phasor analysis to calculate the frequency response of a circuit. The analysis is useful for calculating the gain. 3 dB frequency input and output impedance, and noise of a circuit as a function of frequency, bias conditions, temperature, etc.

Numerical Method

1. A DC solution is performed to calculate the Q-point for the circuit.
2. A linearized circuit is constructed at the Q point. To do this, all nonlinear elements are replaced by their linearized equivalents. For example, a nonlinear current source $I = aV_1^2 + bV_2^3$ would be replaced by a linear voltage controlled current source $I = V_1(2aV_{1q}) + V_2(3bV_{2q}^2)$.
3. All inductors and capacitors are replaced by complex impedances, and conductances evaluated at the frequency of interest.
4. Nodal analysis is now used to reduce the circuit to a linear algebraic complex matrix. The AC node voltages may now be found by applying an excitation vector (which represents the independent voltage and current sources) and using Gaussian elimination (with complex arithmetic) to calculate the node voltages.

AC analysis does have limitations and the following types of nonlinear or large signal problems cannot be modeled:

1. Distortion due to nonlinearities such as clipping, etc.
2. Slew rate-limiting effects
3. Analog mixers
4. Oscillators

Noise analysis is performed by including noise sources in the models. Typical noise sources include thermal noise in resistors $I_n^2 = 4kT\,\Delta f/R$, and shot $I_n^2 = 2qI_d\,\Delta f$, and flicker noise in semiconductor devices. Here, T is temperature in Kelvins, k is Boltzmann's constant, and Δf is the bandwidth of the circuit. These noise sources are inserted as independent current sources, $\mathrm{In}_j(f)$ into the AC model. The resulting current due to the noise source is then calculated at a user-specified summation node(s) by multiplying by the gain function between the noise source and the summation node $A_{js}(f)$. This procedure is repeated for each noise source and then the contributions at the reference node are root mean squared (RMS) summed to give the total noise at the reference node. The equivalent input noise is then easily calculated from the transfer function between the circuit input and the reference node $A_{is}(f)$. The equation describing the input noise is therefore:

$$I_i = \frac{1}{A_{is}(f)} \sqrt{\sum_j \left[A_{js}(f)\,In_j(f)\right]^2}$$

Example 13.2. **Cascode Amplifier with Macro Models:** Here, we find the gain, bandwidth, input impedance, and output noise of a cascode amplifier. The circuit for the amplifier is shown in Fig. 13.5. The circuit is assumed to be fabricated in a monolithic IC process, so it will be necessary to consider some of the parasitics of the IC process. A cross-section of a typical IC bipolar transistor is shown in Fig. 13.4 along with some of the parasitic elements. These parasitic elements are easily included in the amplifier by creating a "macro model" for each transistor. The macro model is then implemented in SPICE form using subcircuits.

The input to the circuit is a voltage source (VIN), applied differentially to the amplifier. The output will be taken differentially across the collectors of the two upper transistors at nodes 2 and 3. The input impedance of the amplifier can be calculated as VIN/I(VIN) or because VIN = 1.0 just as 1/I(VIN). These quantities are shown plotted using PROBE in Fig. 13.6. It can be seen that the gain of the amplifier falls off at high frequency

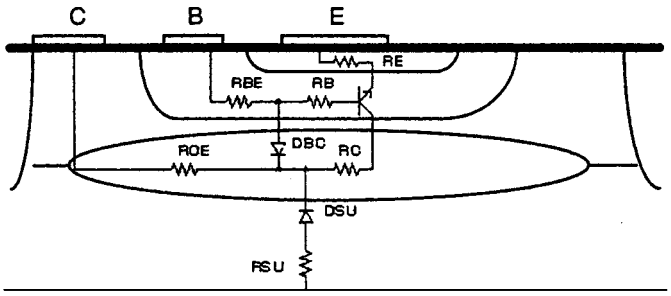


FIGURE 13.4 BJT cross-section with macro model elements.

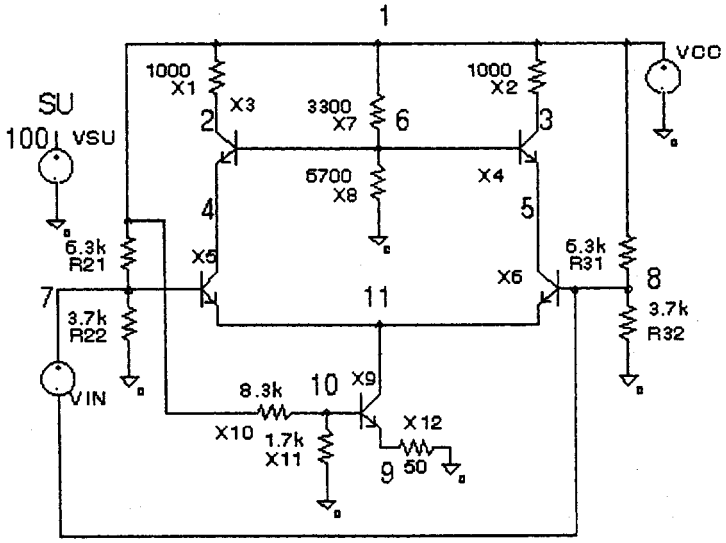


FIGURE 13.5 Cascode amplifier for AC analysis, created using PSPICE.

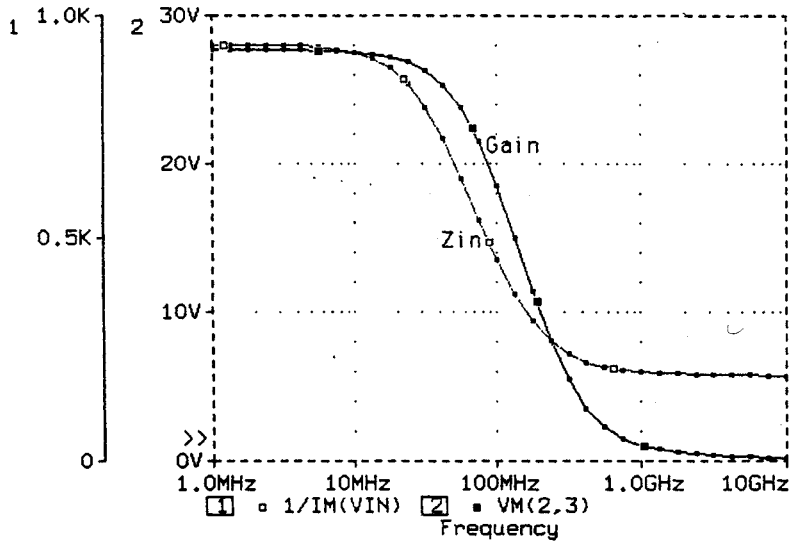


FIGURE 13.6 Gain and input impedance of cascode amplifier.

as expected. The input impedance also drops because parasitic capacitances shunt the input. This example took <1 min on an IBM PC.

Pitfalls. Many novice users will forget that AC analysis is a linear analysis. They will, for example, apply a 1-V signal to an amplifier with 5-V power supplies and a gain of 1000 and be surprised when SPICE tells them that the output voltage is 1000 V. Of course, the voltage generated in a simple amplifier must be less than the power supply voltage, but to examine such clipping effects, transient analysis must be used. Likewise, selection of a proper Q point is important. If the amplifier is biased in a saturated portion of its response and AC analysis is performed, the gain reported will be much smaller than the actual large signal gain.

Transient Analysis

Transient analysis is the most powerful analysis capability of a simulator because the transient response is so hard to calculate analytically. Transient analysis can be used for many types of analysis, such as switching speed, distortion, basic operation of certain circuits like switching power supplies. Transient analysis is also the most CPU intensive and can require 100 or 1000 times the CPU time as a DC or AC analysis.

Numerical Method

In a transient analysis time is discretized into intervals called time steps. Typically the time steps are of unequal length, with the smallest steps being taken during portions of the analysis when the circuit voltages and currents are changing most rapidly. The capacitors and inductors in the circuit are then replaced by voltage and current sources based on the following procedure.

The current in a capacitor is given by $I_c = CdV_c/dt$. The time derivative can be approximated by a difference equation:

$$I_c^k + I_c^{k-1} = 2C\frac{V_c^k - V_c^{k-1}}{t^k - t^{k-1}}$$

In this equation the superscript k represents the number of the time step. Here, k is the time step we are presently solving for and $(k-1)$ is the previous time step. This equation can be solved to give the capacitor current at the present time step.

$$I_c^k = V_c^k\left(2C/\Delta t\right) - V_c^{k-1}\left(2C/\Delta t\right) - I_c^{k-1}.$$

Here, $\Delta t = t^k - t^{k-1}$, or the length of the time step. As time steps are advanced, $V_c^{k-1} \rightarrow V_c^k$; $I_c^{k-1} \rightarrow I_c^k$. Note that the second two terms on the right hand side of the above equation are dependent only on the capacitor voltage and current from the previous time step, and are therefore fixed constants as far as the present step is concerned. The first term is effectively a conductance ($g = 2C/\Delta t$) multiplied by the capacitor voltage, and the second two terms could be represented by an independent current source. The entire transient model for the capacitor therefore consists of a conductance in parallel with two current sources (the numerical values of these are, of course, different at each time step). Once the capacitors and inductors have been replaced as indicated, the normal method of DC analysis is used. One complete DC analysis must be performed for each time point. This is the reason that transient analysis is so CPU intensive. The method outlined here is the trapezoidal time integration method and is used as the default in SPICE.

Example 13.3. **Phase-Locked Loop Circuit:** Figure 13.7 shows the phase-locked loop circuit. The phase detector and voltage-controlled oscillator are modeled in separate subcircuits. Examine the VCO subcircuit and note the PULSE-type current source ISTART connected across the capacitor. The source gives a current pulse 03.E-6 s wide at the start of the simulation to start the VCO running. To start a transient simulation SPICE first computes a DC operating point (to find the initial voltages V_c^{k-1} on the capacitors). As this DC point is a valid, although not necessarily stable, solution, an oscillator will remain at this point indefinitely unless some perturbation is applied to start the oscillations. Remember, this is an ideal mathematical model and no noise

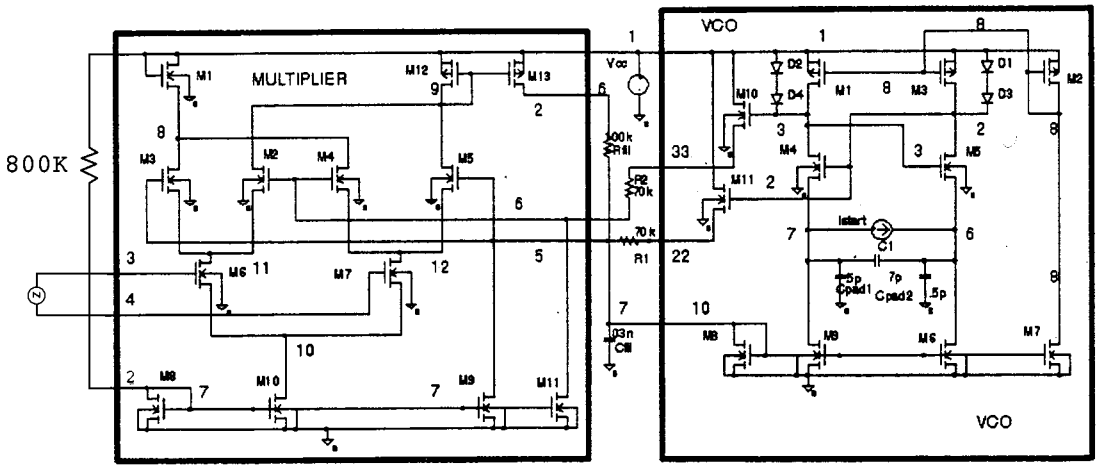


FIGURE 13.7 Phase-locked loop circuit for transient analysis, created with PSPICE.

sources or asymmetries exist that would start a real oscillator—it must be done manually. The capacitor C1 would have to be placed off-chip, and bond pad capacitance (CPAD1 and CPAD2) have been included at the capacitor nodes. Including the pad capacitances is very important if a small capacitor C1 is used for high-frequency operation.

In this example, the PLL is to be used as a FM detector circuit and the FM signal is applied to the input using a single frequency FM voltage source. The carrier frequency is 600 kHz and the modulation frequency is 60 kHz. Figure 13.8 shows the input voltage and the output voltage of the PLL at the VCO output and at the phase detector output. It can be seen that after a brief starting transient, the PLL locks onto the input signal

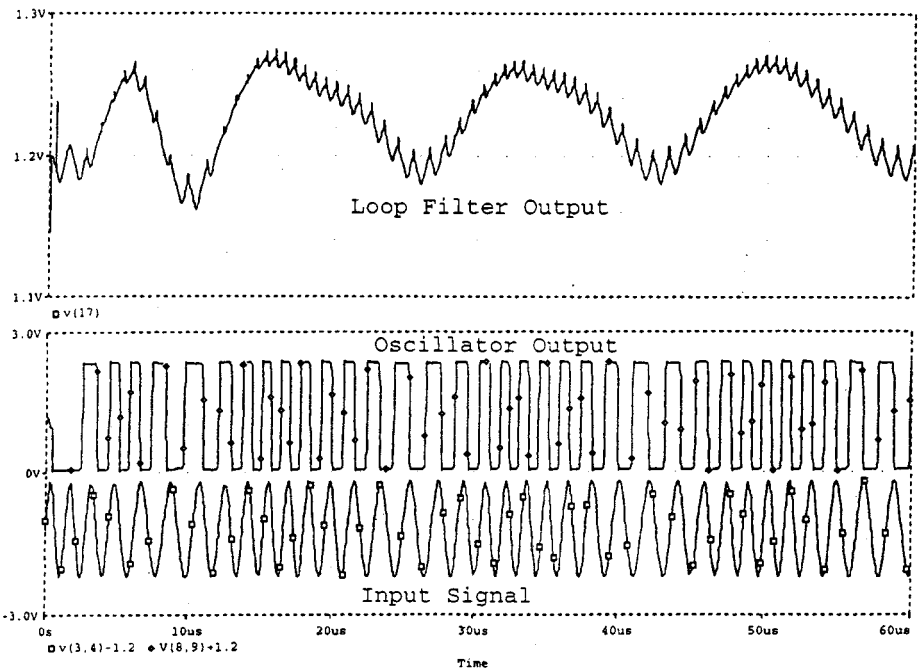


FIGURE 13.8 Transient analysis results of PLL circuit, created using PSPICE.

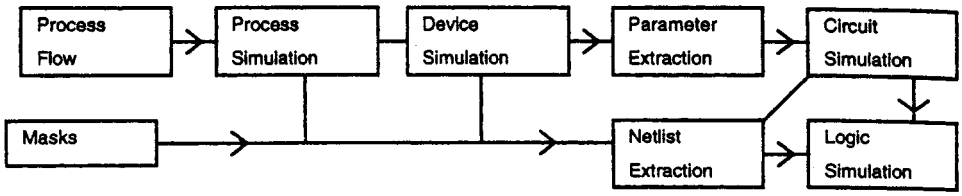


FIGURE 13.9 Data flow for complete process-device-circuit modeling.

and that the phase detector output has a strong 60-kHz component. This example took 251 s on a Sun SPARC-2 workstation (3046 time steps, with an average of 5 Newton iterations per time step).

Pitfalls. Occasionally SPICE will fail and give the message "Timestep too small in transient analysis", which means that the process of Newton iterations at certain time steps could not be made to converge. One of the most common causes of this is the specification of a capacitor with a value that is much too large, for example, specifying a 1-F capacitor instead of a 1 pF capacitor (an easy mistake to make by not adding the "p" in the value specification). Unfortunately, we usually have no way to tell which capacitor is at fault from the type of failure generated other than to manually search the input deck.

Other transient failures are caused by MOSFET models. Some models contain discontinuous capacitances (with respect to voltage) and others do not conserve charge. These models can vary from version to version so it is best to check the user's guide.

Process and Device Simulation

Process and devices simulation are the steps that precede analog circuit simulation in the overall simulation flow (see Fig. 13.9). The simulators are also different in that they are not measurement driven as are analog circuit simulators. The input to a process simulator is the sequence of process steps performed (times, temperatures, gas concentrations) as well as the mask dimensions. The output from the process simulator is a detailed description of the solid-state device (doping profiles, oxide thickness, junction depths, etc.). The input to the device simulator is the detailed description generated by the process simulator (or via measurement). The output of the device simulator is the electrical characteristics of the device (IV curves, capacitances, switching transient curves).

Process and device simulation are becoming increasingly important and widely used during the integrated circuit design process. A number of reasons exist for this:

- As device dimensions shrink, second-order effects can become dominant. Modeling of these effects is difficult using analytical models.
- Computers have greatly improved, allowing time-consuming calculations to be performed in a reasonable amount of time.
- Simulation allows access to impossible to measure physical characteristics.
- Analytic models are not available for certain devices, for example, thyristors, heterojunction devices and IGBTS.
- Analytic models have not been developed for certain physical phenomena, for example, single event upset, hot electron aging effects, latchup, and snap-back.
- Simulation runs can be used to replace split lot runs. As the cost to fabricate test devices increases, this advantage becomes more important.
- Simulation can be used to help device, process, and circuit designers understand how their devices and processes work.

Clearly, process and device simulation is a topic which can be and has been the topic of entire texts. The following sections attempt to provide an introduction to this type of simulation, give several examples showing what the simulations can accomplish, and provide references to additional sources of information.

Process Simulation

Integrated circuit processing involves a number of steps which are designed to deposit (deposition, ion implantation), remove (etching), redistribute (diffusion), or transform (oxidation) the material of which the IC is made. Most process simulation work has been in the areas of diffusion, oxidation, and ion implantation; however, programs are available that can simulate the exposure and development of photo-resist, the associated optical systems, as well as gas and liquid phase deposition and etch.

In the following section a very brief discussion of the governing equations used in SUPREM (from Stanford University, California) will be given along with the results of an example simulation showing the power of the simulator.

Diffusion

The main equation governing the movement of electrically charged impurities (acceptors in this case) in the crystal is the diffusion equation:

$$\frac{\partial C}{\partial t} = \nabla \cdot \left(D\nabla C - \frac{DqC_a}{kT} \mathbf{E} \right)$$

Here, C is the concentration (#/cm^3) of impurities, C_a is the number of electrically active impurities (#/cm^3), q is the electron charge, k is Boltzmann's constant, T is temperature in degrees Kelvin, D is the diffusion constant, and E is the built-in electric field. The built-in electric field E in (V/cm) can be found from:

$$\mathbf{E} = -\frac{kT}{q}\frac{1}{n}\nabla n$$

In this equation n is the electron concentration (#/cm^3), which in turn can be calculated from the number of electrically active impurities (C_a). The diffusion constant (D) is dependent on many factors. In silicon the following expression is commonly used:

$$D = F_{IV}\left(D_x + D_+ \frac{n_i}{n} + D_- \frac{n}{n_i} + D_= \left(\frac{n}{n_i} \right)^2 \right)$$

The four D components represent the different possible charge states for the impurity: (x) neutral, (+) positive, (–) negative, (=) doubly negatively charged. n_i is the intrinsic carrier concentration, which depends only on temperature. Each D component is in turn given by an expression of the type

$$D = A \exp\left(-\frac{B}{kT} \right)$$

Here, A and B are experimentally determined constants, different for each type of impurity (x, +, –, =). B is the activation energy for the process. This expression derives from the Maxwellian distribution of particle energies and will be seen many times in process simulation. It is easily seen that the diffusion process is strongly influenced by temperature. The term F_{IV} is an enhancement factor which is dependent on the concentration of interstitials and vacancies within the crystal lattice (an interstitial is an extra silicon atom which is not located on a regular lattice site; a vacancy is a missing silicon atom which results in an empty lattice site) $F_{IV} \propto C_I + C_v$. The concentration of vacancies, C_v, and interstitials, C_I, are in turn determined by their own diffusion equation:

$$\frac{\partial C_v}{\partial t} = +\nabla \cdot D_V \cdot \nabla C_V - R + G$$

In this equation D_V is another diffusion constant of the form $A \exp(-B/kT)$. R and G represent the recombination and generation of vacancies and interstitials. Note that an interstitial and a vacancy may recombine and in the process destroy each other, or an interstitial and a vacancy pair may be simultaneously generated by knocking a silicon atom off its lattice site. Recombination can occur anywhere in the device via a bulk recombination process $R = A(C_V C_1)\exp(-B/kT)$. Generation occurs where there is damage to the crystal structure, in particular at interfaces where oxide is being grown or in regions where ion implantation has occurred, as the high-energy ions can knock silicon atoms off their lattice sites.

Oxidation

Oxidation is a process whereby silicon reacts with oxygen (or with water) to form new silicon dioxide. Conservation of the oxidant requires the following equation:

$$\frac{dy}{dt} = \frac{\mathbf{F}}{N}$$

Here, F is the flux of oxidant (#/cm^2/s), N is the number of oxidant atoms required to make up a cubic centimeter of oxide, and dy/dt is the velocity with which the Si-SiO_2 interface moves into the silicon. In general the greater the concentration of oxidant (C_0), the faster the growth of the oxide and the greater the flux of oxidant needed at the Si-SiO_2 interface. Thus, $F = k_s C_0$. The flux of oxidant into the oxide from the gaseous environment is given by:

$$F = h(HP_{ox} - C_0)$$

Here H is a constant, P is the partial pressure of oxygen in the gas, and C_0 is the concentration of oxidant in the oxide at the surface and h is of the form $A \exp(-B/kT)$. Finally, the movement of the oxidant within the already existing oxide is governed by diffusion: $\mathbf{F} = D_0 \nabla C$. When all these equations are combined, it is found that (in the one-dimensional case) oxides grow linearly $dy/dt \propto t$ when the oxide is thin and the oxidant can move easily through the existing oxide. As the oxide grows thicker $dy/dt \propto \sqrt{t}$ because the movement of the oxidant through the existing oxide becomes the rate-limiting step.

Modeling two-dimensional oxidation is a challenging task. The newly created oxide must "flow" away from the interface where it is begin generated. This flow of oxide is similar to the flow of a very thick or viscous liquid and can be modeled by a creeping flow equation:

$$\nabla_2 V \propto \nabla P$$

$$\nabla \cdot V = 0$$

V is the velocity at which the oxide is moving and P is the hydrostatic pressure. The second equation results form the incompressibility of the oxide. The varying pressure P within the oxide leads to mechanical stress, and the oxidant diffusion constant D_0 and the oxide growth rate constant k_s are both dependent on this stress. The oxidant flow and the oxide flow are therefore coupled because the oxide flow depends on the rate at which oxide is generated at the interface and the rate at which the new oxide is generated depends on the availability of oxidant, which is controlled by the mechanical stress.

Ion Implantation

Ion implantation is normally modeled in one of two ways. The first involves tables of moments of the final distribution of the ions which are typically generated by experiment. These tables are dependent on the energy and the type of ion being implanted. The second method involves Monte-Carlo simulation of the implantation process. In Monte-Carlo simulation the trajectories of individual ions are followed as they interact with (bounce off) the silicon atoms in the lattice. The trajectories of the ions, and the recoiling Si atoms (which can strike more Si atoms) are followed until all come to rest within the lattice. Typically several thousand trajectories are

simulated (each will be different due to the random probabilities used in the Monte-Carlo method) to build up the final distribution of implanted ions.

Process simulation is always done in the transient mode using time steps as was done with transient circuit simulation. Because partial differential equations are involved, rather than ordinary differential equations, spatial discretization is needed as well. To numerically solve the problem, the differential equations are discretized on a grid. Either rectangular or triangular grids in one, two, or three dimensions are commonly used. This discretization process results in the conversion of the partial differential equations into a set of nonlinear algebraic equations. The nonlinear equations are then solved using a Newton method in a way very similar to the method used for the circuit equations in SPICE.

Example 13.4. **NMOS Transistor:** In this example the process steps used to fabricate a typical NMOS transistor will be simulated using SUPREM-4. These steps are

1. Grow initial oxide (30 min at 1000 K)
2. Deposit nitride layer (a nitride layer will prevent oxidation of the underlying silicon)
3. Etch holes in nitride layer
4. Implant P+ channel stop (boron dose = 5e12, energy = 50 keV)
5. Grow the field oxide (180 min at 1000 K wet O_2)
6. Remove all nitride
7. Perform P channel implant (boron dose = 1e11, energy = 40 keV)
8. Deposit and etch polysilicon for gate
9. Oxidize the polysilicon (30 min at 1000 K, dry O_2)
10. Implant the light doped drain (arsenic dose = 5e13 energy = 50 keV)
11. Deposit sidewall space oxide
12. Implant source and drain (arsenic, dose = 1e15, energy = 200 keV)
13. Deposit oxide layer and etch contact holes
14. Deposit and etch metal

The top 4 μm of the completed structure, as generated by SUPREM-4, is shown in Fig. 13.10. The actual simulation structure used is 200 μm deep to allow correct modeling of the diffusion of the vacancies and interstitials. The gate is at the center of the device. Notice how the edges of the gate have lifted up due to the diffusion of oxidant under the edges of the polysilicon (the polysilicon, as deposited in step 8, is flat). The dashed contours show the concentration of dopants in both the oxide and silicon layers. The short dashes

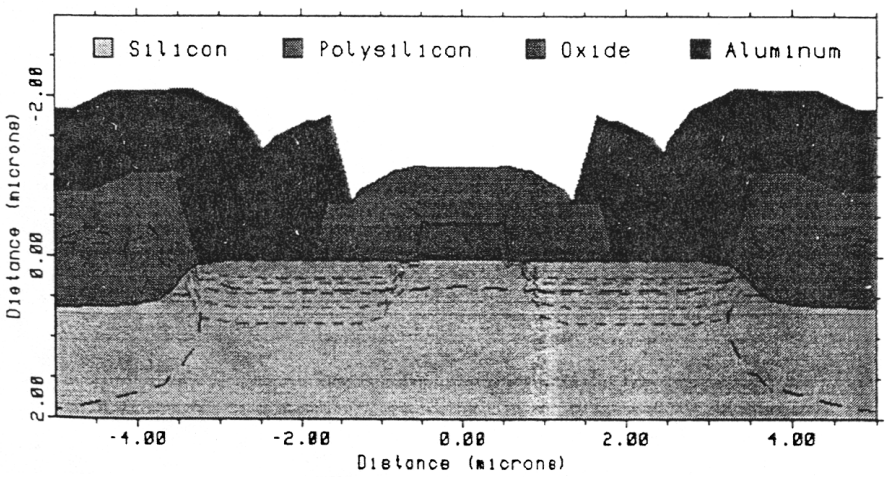


FIGURE 13.10 Complete NMOS transistor cross section generated by process simulation, created with TMA SUPREM-4.

indicate N-type material, while the longer dashes indicate P-type material. This entire simulation requires about 30 min on a Sun SPARC-2 workstation.

Device Simulation

Device simulation uses a different approach from that of conventional lumped circuit models to determine the electrical device characteristics. Whereas with analytic or empirical models all characteristics are determined by fitting a set of adjustable parameters to measured data, device simulators determine the electrical behavior by numerically solving the underlying set of differential equations. The first is the Poisson equation, which describes the electrostatic potential within the device

$$\nabla \cdot \varepsilon \cdot \nabla \Psi = q\left(N_a^- - N_d^+ - p + n - Q_f\right)$$

N_d and N_a are the concentration of donors and acceptors, i.e., the N- and P-type dopants. Q_f is the concentration of fixed charge due, for example, to traps or interface charge. The electron and hole concentrations are given by n and p, respectively, and Ψ is the electrostatic potential.

A set of continuity equations describes the conservation of electrons and holes:

$$\frac{\partial n}{\partial t} = \left(\frac{1}{q}\nabla \cdot \mathbf{J}_n - R + G\right)$$

$$\frac{\partial p}{\partial t} = \left(-\frac{1}{q}\nabla \cdot \mathbf{J}_p - R + G\right)$$

In these equations R and G describe the recombination and generation rates for the electrons and holes. The recombination process is influenced by factors such as the number of electrons and holes present as well as the doping and temperature. The generation rate is also dependent upon the carrier concentrations, but is most strongly influenced by the electric field, with increasing electric fields giving larger generation rates. Because this generation process is included, device simulators are capable of modeling the breakdown of devices at high voltage. $\mathbf{J}_n$ and $\mathbf{J}_p$ are the electron and hole current densities (in amperes per square centimeter). These current densities are given by another set of equations

$$\mathbf{J}_n = q\mu\left(-n\nabla\Psi + \frac{kT_n}{q}\nabla n\right)$$

$$\mathbf{J}_p = q\mu\left(-p\nabla\Psi - \frac{kT_p}{q}\nabla p\right)$$

In this equation k is Boltzmann's constant, μ is the carrier mobility, which is actually a complex function of the doping, n, p, electric field, temperature, and other factors. In silicon the electron mobility will range between 50 and 1000 and the hole mobility will normally be a factor of 2 smaller. In other semiconductors such as gallium arsenide the electron mobility can be as high as 5000. T_n and T_p are the electron and hole mean temperatures, which describe the average carrier energy. In many models these default to the device temperature (300 K). In the first term the current is proportional to the electric field ($\nabla\Psi$), and this term represents the drift of carriers with the electric field. In the second term the current is proportional to the gradient of the carrier concentration (∇n), so this term represents the diffusion of carriers from regions of high concentration to those of low concentration. The model is therefore called the drift-diffusion model.

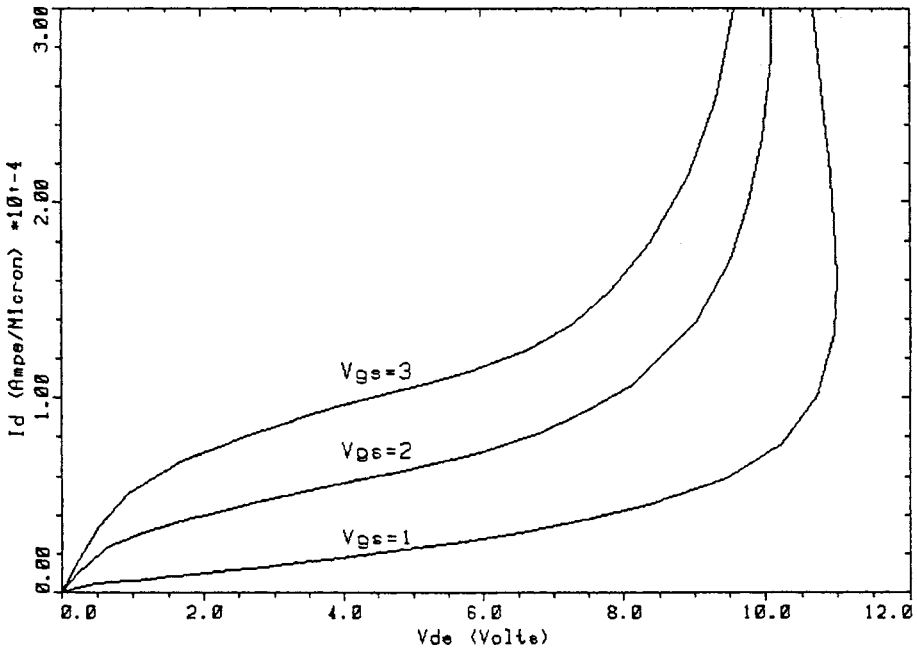


FIGURE 13.11 I_d vs. V_{ds} curves generated by device simulation, created with TMA MEDICI.

In devices in which self-heating effects are important, a lattice heat equation can also be solved to give the internal device temperature:

$$\sigma(T)\frac{\partial T}{\partial t} = H + \nabla \cdot \lambda(T) \cdot \nabla T$$

$$H = -(\mathbf{J}_n + \mathbf{J}_p) \cdot \nabla \Psi + H_R$$

where H is the heat generation term, which includes resistive (Joule) heating as well as recombination heating, H_u. The terms $\sigma(T)$, $\lambda(T)$ represent the specific heat and the thermal conductivity of the material (both temperature dependent). Inclusion of the heat equation is essential in many power device problems.

As with process simulation partial differential equations are involved, therefore, a spatial discretization is required. As with circuit simulation problems, various types of analysis are available:

- Steady state (DC), used to calculate characteristic curves of MOSFETs, BJTs diodes, etc.
- AC analysis, used to calculate capacitances, *Y*-parameters, small signal gains, and *S*-parameters.
- Transient analysis used for calculation of switching and large signal behavior, and special types of analysis such as radiation effects.

Example 13.5. **NMOS IV Curves:** The structure generated in the previous SUPREM-IV simulation is now passed into the device simulator and bias voltages are applied to the gate and drain. Models were included with account for Auger and Shockley Reed Hall recombination, doping and electric field-dependent mobility, and impact ionization. The set of drain characteristics obtained is shown in Fig. 13.11. Observe how the curves bend upward at high V_{ds} as the device breaks down. The V_g, = 1 curve has a negative slope at I_d = 1.5e-4A as the device enters snap-back. It is possible to model this type of behavior because impact ionization is included in the model.

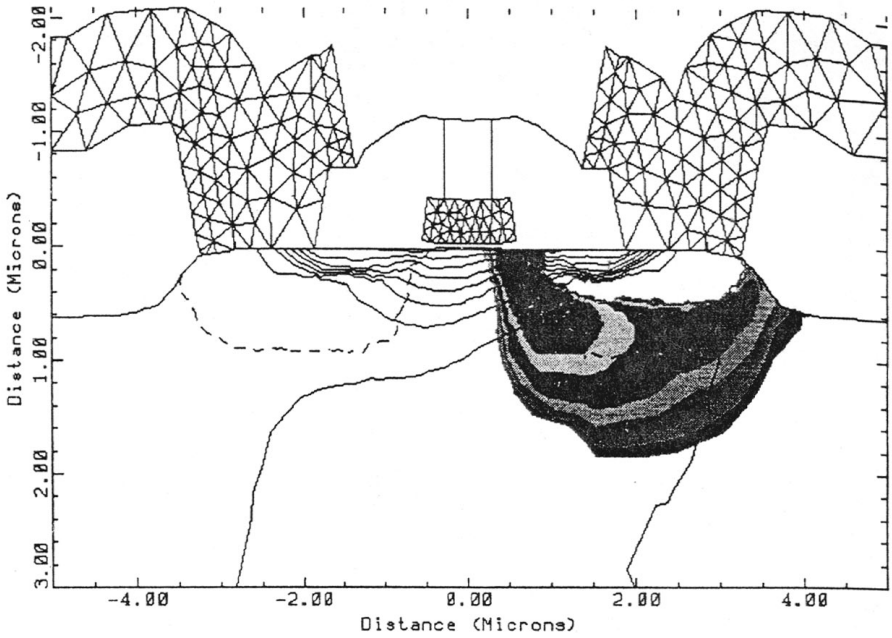


FIGURE 13.12 Internal behavior of MOSFET under bias, created with TMA MEDICI.

Figure 13.12 shows the internal behavior of the device with V_{gs} = 3 V and I_d = 3e-4A. The filled contours indicate impact ionization, with the highest rate being near the edge of the drain right beneath the gate. This is to be expected because this is the region in which the electric field is largest due to the drain depletion region. The dark lines indicate current flow from the source to the drain. Some current also flows from the drain to the substrate. This substrate current consists of holes generated by the impact ionization. The triangular grid used in the simulation can be seen in the source, drain, and gate electrodes. A similar grid was used in the oxide and silicon regions.

Appendix

Circuit Analysis Software

SPICE2, SPICE3: University of California, Berkeley
PSPICE: MicroSim Corporation, Irvine, CA (used in this chapter)
HSPICE: Meta Software, Campbell, CA
IsSPICE: Intusoft, San Pedro, CA
SPECTRE: Cadence Design Systems, San Jose, CA
SABRE: Analogy, Beaverton, OR

Process and Device Simulators

SUPREM-4, PISCES: Stanford University, Palo Alto, CA
MINIMOS: Technical University, Vienna, Austria
SUPREM-4, MEDICI, DAVINCI: Technology Modeling Associates, Palo Alto, CA (used in this chapter)
SEMICAD: Dawn Technologies, Sunnyvale, CA

Related Topics

3.2 Node and Mesh Analysis • 23.1 Processes • 27.1 Ideal and Practical Models

References

P. Antognetti and G. Massobrio, *Semiconductor Device Modeling with SPICE,* New York: McGraw-Hill, 1988.
P. W. Tuinenga, *SPICE, A Guide to Circuit Simulation and Analysis Using PSPICE,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.
J. A. Connelly and P. Choi, *Macromodeling with SPICE,* Englewood Cliffs, N.J.: Prentice-Hall, 1992.
S. Selberherr, *Analysis and Simulation of Semiconductor Devices,* Berlin: Springer-Verlag, 1984.
R. Dutton and Z. Yu, *Technology CAD, Computer Simulation of IC Process and Devices,* Boston: Kluwer Academic, 1993.

13.2 Parameter Extraction for Analog Circuit Simulation

Peter Bendix

Introduction

Definition of Device Modeling

We use various terms such as device characterization, parameter extraction, optimization, and model fitting to address an important engineering task. In all of these, we start with a mathematical model that describes the transistor behavior. The model has a number of parameters which are varied or adjusted to match the IV (current-voltage) characteristics of a particular transistor or set of transistors. The act of determining the appropriate set of model parameters is what we call device modeling. We then use the model with the particular set of parameters that represent our transistors in a circuit simulator such as SPICE[1] to simulate how circuits with our kinds of transistors will behave. Usually the models are supplied by the circuit simulator we chose. Occasionally we may want to modify these models or construct our own models. In this case we need access to the circuit simulator model subroutines as well as the program that performs the device characterization.

Steps Involved in Device Characterization

Device characterization begins with a test chip. Without the proper test chip structures, proper device modeling cannot be done from measured data. A good test chip for MOS technology would include transistors of varying geometries, gate oxide capacitance structures, junction diode capacitance structures, and overlap capacitance structures. This would be a minimal test chip. Additional structures might include ring oscillators and other circuits for checking the AC performance of the models obtained. It is very important that the transistors be well designed and their geometries be chosen appropriate for the technology as well as the desired device model. Although a complete test chip description is beyond the scope of this book, be aware that even perfect device models cannot correct for a poor test chip.

Next we need data that represent the behavior of a transistor or set of transistors of different sizes. these data can come from direct measurement or they can be produced by a device simulator such as PISCES.[2]

It is also possible to use a combination of a process simulator like SUPREM-IV[3] coupled to a device simulator, to provide the simulated results. The benefits of using simulation over measurement are that no expensive measurement equipment or fabricated wafers are necessary. This can be very helpful when trying to predict the device characteristics of a new fabrication process before any wafers have been produced.

Once the measured (or simulated) data are available, parameter extraction software is used to find the best set of model parameter values to fit the data.

[1]SPICE is a circuit simulation program from the Department of Electrical Engineering and Computer Science at the University of California at Berkeley.

[2]PISCES is a process simulation program from the Department of Electrical Engineering at Stanford University, Stanford, CA.

[3]SUPREM-IV is a process simulation program from the Department of Electrical Engineering at Stanford University, Stanford, CA.

Least-Squares Curve Fitting (Analytical)

We begin this section by showing how to do least-squares curve fitting by analytical solutions, using a simple example to illustrate the method. We then mention least-squares curve fitting using numerical solutions in the next section. We can only find analytical solutions to simple problems. The more complex ones must rely on numerical techniques.

Assume a collection of measured data, $m_1, \ldots, m_n$. For simplicity, let these measured data values be functions of a single variable, v, which was varied from v_1 through v_n, measuring each m_i data point at each variable value v_i, i running from 1 to n. For example, the m_i data points might be drain current of an MOS transistor, and the v_i might be the corresponding values of gate voltage. Assume that we have a model for calculating simulated values of the measured data points, and let these simulated values be denoted by $s_1, \ldots, s_n$. We define the least-squares, root mean square (RMS) error as

$$\text{Error}_{\text{rms}} = \left[\frac{\sum_{i=1}^{n} \left\{ \text{weight}_i \left(s_i - m_i \right) \right\}^2}{\sum_{i=1}^{n} \left\{ \text{weight}_i m_i \right\}^2} \right]^{1/2} \tag{13.1}$$

where a weighting term is included for each data point. The goal is to have the simulated data match the measured data as closely as possible, which means we want to minimize the RMS error. Actually, what we have called the RMS error is really the relative RMS error, but the two terms are used synonymously. There is another way of expressing the error, called the absolute RMS error, defined as follows:

$$\text{Error}_{\text{rms}} = \left[\frac{\sum_{i=1}^{n} \left\{ \text{weight}_i \left(s_i - m_i \right) \right\}^2}{\sum_{i=1}^{n} \left\{ \text{weight}_i m_{\min} \right\}^2} \right]^{1/2} \tag{13.2}$$

where we have used the term $m_{\min}$ in the denominator to represent some minimum value of the measured data. The absolute RMS error is usually used when the measured values approach zero to avoid problems with small or zero denominators in (13.1). For everything that follows, we consider only the relative RMS error. The best result is obtained by combining the relative RMS formula with the absolute RMS formula by taking the maximum of the denominator from (13.1) or (13.2).

We have a simple expression for calculating the simulated data points, s_i, in terms of the input variable, v, and a number of model parameters, $p_1, \ldots, p_m$. That is,

$$s_i = f\left(v_i, p_1, \ldots, p_m \right) \tag{13.3}$$

where f is some function. Minimizing the RMS error function is equivalent to minimizing its square. Also, we can ignore the term in the denominator of (13.1) as concerns minimizing, because it is a normalization term. In this spirit, we can define a new error term,

$$\text{Error} = \left(\text{Error}_{\text{rms}} \right)^2 \left[\sum_{i=1}^{n} \left\{ \text{weight}_i m_i \right\}^2 \right] \tag{13.4}$$

and claim that minimizing Error is equivalent to minimizing Error_{rms}. To minimize Error, we set all partial derivatives of it with respect to each model parameter equal to zero; that is, write

$$\frac{\partial(\text{Error})}{\partial p_j} = 0, \qquad \text{for } j = 1, \ldots, m \tag{13.5}$$

Then solve the above equations for the value of p_j.

Least-Square Curve Fitting (Numerical)

For almost all practical applications we are forced to do least-squares curve fitting numerically, because the analytic solutions as previously discussed are not obtainable in closed form. What we are calling least-squares curve fitting is more generally known as nonlinear optimization. Many fine references on this topic are available. We refer the reader to [Gill et al., 1981] for details.

Extraction (as Opposed to Optimization)

The terms "extraction" and "optimization" are, unfortunately, used interchangeably in the semiconductor industry; however, strictly speaking, they are not the same. By optimization, we mean using generalized least-squares curve fitting methods such as the Levenberg-Marquardt algorithm [Gill et al., 1981] to find a set of model parameters. By extraction, we mean any technique that does not use general least-squares fitting methods. This is a somewhat loose interpretation of the term extraction. The main point is that we write the equations we want and then solve them by whatever approximations we choose, as long as these approximations allow us to get the extracted results in closed form. This is parameter extraction.

Extraction vs. Optimization

Extraction has the advantage of being much faster than optimization, but it is not always as accurate. It is also much harder to supply extraction routines for models that are being developed. Each time you make a change in the model, you must make suitable changes in the corresponding extraction routine. For optimization, however, no changes are necessary other than the change in the model itself, because least-squares curve fitting routines are completely general. Also, if anything goes wrong in the extraction algorithm (and no access to the source code is available), almost nothing can be done to correct the problem. With optimization, one can always change the range of data, weighting, upper and lower bounds, etc. A least-squares curve fitting program can be steered toward a correct solution.

Novices at device characterization find least-squares curve fitting somewhat frustrating because a certain amount of user intervention and intuition is necessary to obtain the correct results. These beginners prefer extraction methods because they do not have to do anything. However, after being burned by extraction routines that do not work, a more experienced user will usually prefer the flexibility, control, and accuracy that optimization provides.

Commercial software is available that provides both extraction and optimization together. The idea here is to first use extraction techniques to make reasonable initial guesses and then use these results as a starting point for optimization, because optimization can give very poor results if poor initial guesses for the parameters are used. Nothing is wrong with using extraction techniques to provide initial guesses for optimization, but for an experienced user this is rarely necessary, assuming that the least-squares curve fitting routine is robust (converges well) and the experienced user has some knowledge of the process under characterization. Software that relies heavily on extraction may do so because of the nonrobustness of its optimizer.

These comments apply when an experienced user is doing optimization locally, not globally. For global optimization (a technique we do not recommend), the above comparisons between extraction and optimization are not valid. The following section contains more detail about local vs. global optimization.

Strategies: General Discussion

The most naive way of using an optimization program would be to take all the measured data for all devices, put them into one big file, and fit to all these data with all model parameters simultaneously. Even for a very high quality, robust optimization program the chances of this method converging are slight. Even if the program does converge, it is almost certain that the values of the parameters will be very unphysical. This kind of approach is an extreme case of global optimization. We call any optimization technique that tries to fit with parameters to data outside their region of applicability a global approach. That is, if we try to fit to saturation

region data with linear region parameters such as threshold voltage, mobility, etc., we are using a global approach. In general, we advise avoiding global approaches, although in the strategies described later, sometimes the rules are bent a little.

Our recommended approach is to fit subsets of relevant parameters to corresponding subsets of relevant data in a way that makes physical sense. For example, in the MOS level 3 model, VT0 is defined as the threshold voltage of a long, wide transistor at zero back-bias. It does not make sense to use this parameter to fit to a short channel transistor, or to fit at nonzero back-bias values, or to fit to anywhere outside the linear region. In addition, subsets of parameters should be obtained in the proper order so that those obtained at a later step do not affect those obtained at earlier steps. That is, we would not obtain saturation region parameters before we have obtained linear region parameters because the values of the linear region parameters would influence the saturation region fits; we would have to go back and reoptimize on the saturation region parameters after obtaining the linear region parameters. Finally, never use optimization to obtain a parameter value when the parameter can be measured directly. For example, the MOS oxide thickness, TOX, is a model parameter, but we would never use optimization to find it. Always measure its value directly on a large oxide capacitor provided on the test chip. The recommended procedure for proper device characterization follows:

1. Have all the appropriate structures necessary on your test chip. Without this, the job cannot be performed properly.
2. Always measure whatever parameters are directly measurable. Never use otpimization for these.
3. Fit the subset of parameters to corresponding subsets of data, and do so in physically meaningful ways.
4. Fit parameters in the proper order so that those obtained later do not affect those obtained previously. If this is not possible, iteration may be necessary.

Naturally, a good strategy cannot be mounted if one is not intimately familiar with the model used. There is no substitute for learning as much about the model as possible. Without this knowledge, one must rely on strategies provided by software vendors, and these vary widely in quality.

Finally, no one can provide a completely general strategy applicable to all models and all process technologies. At some point the strategy must be tailored to suit the available technology and circuit performance requirements. This not only requires familiarity with the available device models, but also information from the circuit designers and process architects.

MOS DC Models

Available MOS Models

A number of MOS models have been provided over time with the original circuit simulation program, SPICE. In addition, some commercially available circuit simulation programs have introduced their own proprietary models, most notably HSPICE.[1] This section is concentrated on the standard MOS models provided by UC Berkeley's SPICE, not only because they have become the standard models used by all circuit simulation programs, but also because the proprietary models provided by commercial vendors are not well documented and no source code is available for these models to investigate them thoroughly.

MOS Levels 1, 2, and 3. Originally, SPICE came with three MOS models known as level 1, level 2, and level 3. The level 1 MOS model is a very crude first-order model that is rarely used. The level 2 and level 3 MOS models are extensions of the level 1 model and have been used extensively in the past and present [Vladimirescu and Liu, 1980]. These two models contain about 15 DC parameters each and are usually considered useful for digital circuit simulation down to 1 μm channel length technologies. They can fit the drain current for wide transistors of varying length with reasonable accuracy (about 5% RMS error), but have very little advanced fitting capability for analog application. They have only one parameter for fitting the subthreshold region, and no parameters for fitting the derivative of drain current with respect to drain voltage, G_{ds} (usually considered critical for analog applications). They also have no ability to vary the mobility degradation with back-bias, so the fits to I_{ds} in the saturation region at high back-bias are not very good. Finally, these models do not interpolate well over device

[1]HSPICE is a commercially available, SPICE-like circuit simulation program from Meta Software, Campbell, CA.

geometry; e.g., if a fit it made to a wide-long device and a wide-short device, and then one observes how the models track for lengths between these two extremes, they usually do not perform well. For narrow devices they can be quite poor as well. Level 3 has very little practical advantage over level 2, although the level 2 model is proclaimed to be more physically based, whereas the level 3 model is called semiempirical. If only one can be used, perhaps level 3 is slightly better because it runs somewhat faster and does not have quite such an annoying kink in the transition region from linear to saturation as does level 2.

Berkeley Short-Channel Igfet Model (BSIM). To overcome the many short-comings of level 2 and level 3, the BSIM and BSIM2 models were introduced. The most fundamental difference between these and the level 2 and 3 models is that BSIM and BSIM2 use a different approach to incorporating the geometry dependence [Ouster et al., 1988; Jeng et al., 1987]. In level 2 and 3 the geometry dependence is built directly into the model equations. In BSIM and BSIM2 each parameter (except for a very few) is written as a sum of three terms

$$\text{Parameter} \;-\; \text{Par}_0 + \frac{\text{Par}_L}{L_{\text{eff}}} + \frac{\text{Par}_W}{W_{\text{eff}}}, \tag{13.6}$$

where Par_0 is the zero-order term, Par_L accounts for the length dependence of the parameter, Par_W accounts for the width dependence, and L_{eff} and W_{eff} are the effective channel width and length, respectively. This approach has a large influence on the device characterization strategy, as discussed later. Because of this tripling of the number of parameters and for other reasons as well, the BSIM model has about 54 DC parameters and the BSIM2 model has over 100.

The original goal of the BSIM model was to fit better than the level 2 and 3 models for submicron channel lengths, over a wider range of geometries, in the subthreshold region, and for nonzero back-bias. Without question, BSIM can fit individual devices better than level 2 and level 3. It also fits the subthreshold region better and it fits better for nonzero back-biases. However, its greatest shortcoming is its inability to fit over a large geometry variation. This occurs because (13.6) is a truncated Taylor series in $1/L_{\text{eff}}$ and $1/W_{\text{eff}}$ terms, and in order to fit better over varying geometries, higher power terms in $1/L_{\text{eff}}$ and $1/W_{\text{eff}}$ are needed. In addition, no provision was put into the BSIM model for fitting G_{ds}, so its usefulness for analog applications is questionable. Many of the BSIM model parameters are unphysical, so it is very hard to understand the significance of these model parameters. This has profound implications for generating skew models (fast and slow models to represent the process corners) and for incorporating temperature dependence. Another flaw of the BSIM model is its wild behavior for certain values of the model parameters. If model parameters are not specified for level 2 or 3, they will default to values that will at least force the model to behave well. For BSIM, not specifying certain model parameters, setting them to zero, or various combinations of values can cause the model to become very ill-behaved.

BSIM2. The BSIM2 model was developed to address the shortcomings of the BSIM model. This was basically an extension of the BSIM model, removing certain parameters that had very little effect, fixing fundamental problems such as currents varying the wrong way as a function of certain parameters, adding more unphysical fitting parameters, and adding parameters to allow fitting G_{ds}. BSIM2 does fit better than BSIM, but with more than twice as many parameters as BSIM, it should. However, it does not address the crucial problem of fitting large geometry variations. Its major strengths over BSIM are fitting the subthreshold region better, and fitting G_{ds} better. Most of the other shortcomings of BSIM are also present in BSIM2, and the large number of parameters in BSIM2 makes it a real chore to use in device characterization.

BSIM3. Realizing the shortcomings of BSIM2, UC Berkeley recently introduced the BSIM3 model. This is an unfortunate choice of name because it implies BSIM3 is related to BSIM and BSIM2. In reality, BSIM3 is an entirely new model that in some sense is related more to level 2 and 3 than BSIM or BSIM2. The BSIM3 model abandons the length and width dependence approach of BSIM and BSIM2, preferring to go back to incorporating the geometry dependence directly into the model equations, as do level 2 and 3. In addition, BSIM3 is a more physically based model, with about 30 fitting parameters (the model has many more parameters, but

the majority of these can be left untouched for fitting), making it more manageable, and it has abundant parameters for fitting G_{ds}, making it a strong candidate for analog applications.

It is an evolving model, so perhaps it is unfair to criticize it at this early stage. Its greatest shortcoming is, again, the inability to fit well over a wide range of geometries. It is hoped that future modifications will address this problem. In all fairness, however, it is a large order to ask a model to be physically based, have not too many parameters, be well behaved for all default values of the parameters, fit well over temperature, fit G_{ds}, fit over a wide range of geometries, and still fit individual geometries as well as a model with over 100 parameters, such as BSIM2. Some of these features were compromised in developing BSIM3.

Proprietary Models. A number of other models are available from commercial circuit simulator vendors, the literature, etc. Some circuit simulators also offer the ability to add a researcher's own models. In general, we caution against using proprietary models, especially those which are supplied without source code and complete documentation. Without an intimate knowledge of the model equations, it is very difficult to develop a good device characterization strategy. Also, incorporating such models into device characterization software is almost impossible. To circumvent this problem, many characterization programs have the ability to call the entire circuit simulator as a subroutine in order to exercise the proprietary model subroutines. This can slow program execution by a factor of 20 or more, seriously impacting the time required to characterize a technology. Also, if proprietary models are used without source code, the circuit simulator results can never be checked against other circuit simulators. Therefore, we want to stress the importance of using standard models. If these do not meet the individual requirements, the next best approach is to incorporate a proprietary model whose source code one has access to. This requires being able to add the individual model not only to circuit simulators, but also to device characterization programs; it can become a very large task.

MOS Level 3 Extraction Strategy in Detail

The strategy discussed here is one that we consider to be a good one, in the spirit of our earlier comments. Note, however, that this is not the only possible strategy for the level 3 model. The idea here is to illustrate basic concepts so that this strategy can be refined to meet particular individual requirements.

In order to do a DC characterization, the minimum requirement is one each of the wide-long, wide-short, and narrow-long devices. We list the steps of the procedure and then discuss them in more detail.

STEP 1. Fit the wide-long device in the linear region at zero back-bias at V_{gs} values above the subthreshold region, with parameters VT0 (threshold voltage), U0 (mobility), and THETA (mobility degradation with V_{gs}).

STEP 2. Fit the wide-short device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VT0, LD (length encroachment), and THETA. When finished with this step, replace VT0 and THETA with the values from step 1, but keep the value of LD.

STEP 3. Fit the narrow-long device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VT0, DW (width encroachment), and THETA. When finished with this step, replace VT0 and THETA with the values from step 1, but keep the value of DW.

STEP 4. Fit the wide-short device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters RS and RD (source and drain series resistance).

STEP 5. Fit the wide-long device in the linear region at all back-biases, at V_{gs} values above the subthreshold region, with parameter NSUB (channel doping affects long channel variation of threshold voltage with back-bias).

STEP 6. Fit the wide-short device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameter XJ (erroneously called the junction depth; affects short-channel variation of threshold voltage with back-bias).

STEP 7. Fit the narrow-long device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameter DELTA (narrow channel correction to threshold voltage).

STEP 8. Fit the wide-short device in the saturation region at zero back-bias (or all back-biases) with parameters VMAX (velocity saturation), KAPPA (saturation region slope fitting parameter), and ETA (V_{ds} dependence of threshold voltage).

STEP 9. Fit the wide-short device in the subthreshold region at whatever back-bias and drain voltage is appropriate (usually zero back-bias and low V_{ds}) with parameter NES (subthreshold slope fitting parameter). One may need to fit with VT0 also and then VT0 is replaced after this step with the value of VT0 obtained from step 1.

This completes the DC characterization steps for the MOS level 3 model. One would then go on to do the junction and overlap capacitance terms (discussed later). Note that this model has no parameters for fitting over temperature, although temperature dependence is built into the model that the user cannot control.

In Step 1 VT0, U0, and THETA are defined in the model for a wide-long device at zero back-bias. They are zero-order fundamental parameters without any short or narrow channel corrections. We therefore fit them to a wide-long device. It is absolutely necessary that such a device be on the test chip. Without it, one cannot obtain these parameters properly. The subthreshold region must be avoided also because these parameters do not control the model behavior in subthreshold.

In Step 2 we use LD to fit the slope of the linear region curve, holding U0 fixed from step 1. We also fit with VT0 and THETA because without them the fitting will not work. However, we want only the value of LD that fits the slope, so we throw away VT0 and THETA, replacing them with the values from step 1.

Step 3 is the same as step 2, except that we are getting the width encroachment instead of the length.

In Step 1 the value of THETA that fits the high V_{gs} portion of the wide-long device linear region curve was found. Because the channel length of a long transistor is very large, the source and drain series resistances have almost no effect here, but for a short-channel device, the series resistance will also affect the high V_{gs} portion of the linear region curve. Therefore, in step 4 we fix THETA from step 1 and use RS and RD to fit the wide-short device in the linear region, high V_{gs} portion of the curve.

In Step 5 we fit with NSUB to get the variation of threshold voltage with back-bias. We will get better results if we restrict ourselves to lower values of V_{gs} (but still above subthreshold) because no mobility degradation adjustment exists with back-bias, and therefore the fit may not be very good at higher V_{gs} values for the nonzero back-bias curves.

Step 6 is just like step 5, except we are fitting the short-channel device. Some people think that the value of XJ should be the true junction depth. This is not true. The parameter XJ is loosely related to the junction depth, but XJ is really the short-channel correction to NSUB. Do not be surprised if XJ is not equal to the true junction depth.

Step 7 uses DELTA to make the narrow channel correction to the threshold voltage. This step is quite straightforward.

Step 8 is the only step that fits in the saturation region. The use of parameters VMAX and KAPPA is obvious, but one may question using ETA to fit in the saturation region. The parameter ETA adjusts the threshold voltage with respect to V_{ds}, and as such one could argue that ETA should be used to fit measurements of I_{ds} sweeping V_{gs} and stepping V_{ds} to high values. In doing so, one will corrupt the fit in the saturation region, and usually we want to fit the saturation region better at the expense of the linear region.

Step 9 uses NFS to fit the slope of the $\log(I_{ds})$ vs. V_{gs} curve. Often the value of VT0 obtained from step 1 will prevent one from obtaining a good fit in the subthreshold region. If this happens, try fitting with VT0 and NFS, but replacing the final value of VT0 with that from step 1 at the end, keeping only NFS from this final step.

The above steps illustrate the concepts of fitting relevant subsets of parameters to relevant subsets of data to obtain physical values of the parameters, as well as fitting parameters in the proper order so that those obtained in the later steps will affect those obtained in earlier steps minimally. Please refer to Figs. 13.13 and 13.14 for how the resulting fits typically appear (all graphs showing model fits are provided by the device modeling software package Aurora, from Technology Modeling Associates, Inc., Palo Alto, CA).

An experienced person may notice that we have neglected some parameters. For example, we did not use parameters KP and GAMMA. This means KP will be calculated from U0, and GAMMA will be calculated from NSUB. In a sense U0 and NSUB are more fundamental parameters than KP and GAMMA. For example, KP

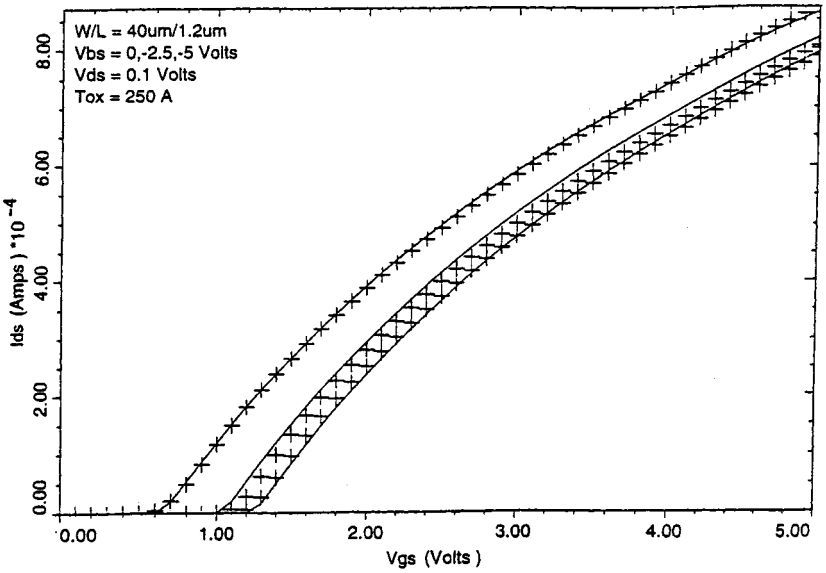


FIGURE 13.13 Typical MOS level 3 linear region measured and simulated plots at various V_{bs} values for a wide-short device.

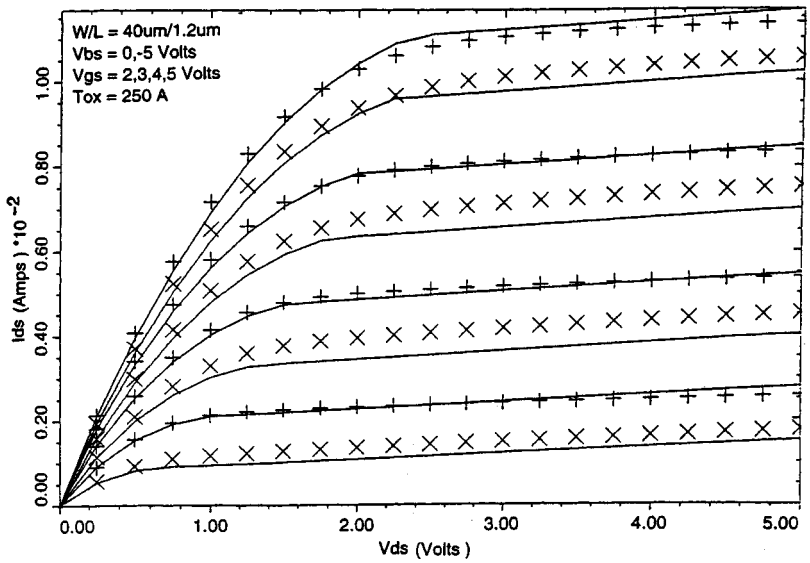


FIGURE 13.14 Typical MOS level 3 saturation region measured and simulated plots at various V_{gs} and V_{bs} values for a wide-short device.

depends on U0 and TOX; GAMMA depends on NSUB and TOX. If one is trying to obtain skew models, it is much more advantageous to analyze statistical distributions of parameters that depend on a single effect than those that depend on multiple effects. KP will depend on mobility and oxide thickness; U0 is therefore a more fundamental parameter. We also did not obtain parameter PHI, so it will be calculated from NSUB. The level 3 model is very insensitive to PHI, so using it for curve fitting is pointless. This illustrates the importance of being very familiar with the model equations. The kind of judgments described here cannot be made without such knowledge.

Test Chip Warnings. The following hints will greatly assist in properly performing device characterization.

1. Include a wide-long device; without this, the results will not be physically correct.
2. All MOS transistors with the same width should be drawn with their sources and drains identical. No difference should be seen in the number of source/drain contacts, contact spacing, source/drain contact overlap, poly gate to contact spacing, etc.
3. Draw devices in pairs. That is, if the wide-long device is W/L = 20/20, make the wide-short device the same width as the wide-long device; e.g., make the short device 20/1, not 19/1. If the narrow-long device is 2/20, make the narrow-short device of the same width; i.e., make is 2/1, not 3/1, and similarly for the lengths. (Make the wide-short and the narrow-short devices have the same length.)

BSIM Extraction Strategy in Detail

All MOS model strategies have basic features in common; namely, fit the linear region at zero back-bias to get the basic zero-order parameters, fit the linear region at nonzero back-bias, fit the saturation region at zero back-bias, fit the saturation region at nonzero back-bias, and then fit the subthreshold region. It is possible to extend the type of strategy we covered for level 3 to the BSIM model, but that is not the way BSIM was intended to be used.

The triplet sets of parameters for incorporating geometry dependence into the BSIM model, (13.6), allow an alternate strategy. We obtain sets of parameters without geometry dependence by fitting to individual devices without using the Par_L and Par_W terms. We do this for each device size individually. This produces sets of parameters relevant to each individual device. So, for device number 1 of width W(1) and length L(1) we would have a value for the parameter VFB which we will call VFB(1); for device number n of width W(n) and length L(n) we will have VFB(n). To get the Par_0, Par_L, and VFB_W we fit to the "data points" VFB(1), . . .,VFB(n) with parameters VFB_0, VFB_L, and VFB_W using (13.6) where L_{eff} and W_{eff} are different for each index, 1 through n.

Note that as L and W become very large, the parameters must approach Par_0. This suggests that we use the parameter values for the wide-long device as the Par_0 terms and only fit the other geometry sizes to get the Par_L and Par_W terms. For example, if we have obtained VFB(1) for our first device which is our wide-long device, we would set VFB_0 = VFB(1), and then fit to VFB(2), . . .,VBF(n) with parameters VFB_L and VFB_W, and similarly for all the other triplets of parameters. In order to use a general least-squares optimization program in this way the software must be capable of specifying parameters as targets, as well as measured data points.

We now list a basic strategy for the BSIM model:

STEP 1. Fit the wide-long device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VFB (flatband voltage), MUZ (mobility), and U0 (mobility degradation), with DL (length encroachment) and DW (width encroachment) set to zero.

STEP 2. Fit the wide-short device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VFB, U0, and DL.

STEP 3. Fit the narrow-long device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VFB, U0, and DW.

STEP 4. Refit the wide-long device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VFB, MUZ, and U0, now that DL and DW are known.

STEP 5. Fit the wide-short device in the linear region at zero back-bias, at V_{gs} values above the subthreshold region, with parameters VFB, RS, and RD. When finished, replace the value of VFB with the value found in step 4.

STEP 6. Fit the wide-long device in the linear region at all back-biases, at V_{gs} values above the subthreshold region, with parameters K1 (first-order body effect), K2 (second-order body effect), U0, and X2U0 (V_{bs} dependence of U0).

STEP 7. Fit the wide-long device in the saturation region at zero back-bias with parameters U0, ETA (V_{ds} dependence of threshold voltage), MUS (mobility in saturation), U1 (V_{ds} dependence of mobility), and X3MS (V_{ds} dependence of MUS).

STEP 8. Fit the wide-long device in the saturation region at all back-biases with parameter X2MS (V_{bs} dependence of MUS).

STEP 9. Fit the wide-long device in the subthreshold region at zero back-bias and low V_{ds} value with parameter N0; then fit the subthreshold region nonzero back-bias low V_{ds} data with parameter NB; and finally fit the subthreshold region data at higher V_{ds} values with parameter ND. Or, fit all the subthreshold data simultaneously with parameters N0, NB, and ND.

Repeat Steps 6 through 10 for all the other geometries, with the result of sets of geometry-independent parameters for each different size device. Then follow the procedure described previously for obtaining the geometry-dependent terms Par_0, Par_L, and Par_W.

In the above strategy we have omitted various parameters either because they have minimal effect or because they have the wrong effect and were modified in the BSIM2 model. Because of the higher complexity of the BSIM model over the level 3 model, many more strategies are possible than the one just listed. One may be able to find variations of the above strategy that suit the individual technology better. Whatever modifications are made, the general spirit of the above strategy probably will remain.

Some prefer to use a more global approach with BSIM, fitting to measured data with Par_L and Par_W terms directly. Although this is certainly possible, it is definitely not a recommended approach. It represents the worst form of blind curve fitting, with no regard for physical correctness or understanding. The BSIM model was originally developed with the idea of obtaining the model parameters via extraction as opposed to optimization. In fact, UC Berkeley provides software for obtaining BSIM parameters using extraction algorithms, with no optimization at all. As stated previously, this has the advantage of being relatively fast and easy. Unfortunately, it does not always work. One of the major drawbacks of the BSIM model is that certain values of the parameters can cause the model to produce negative values of G_{ds} in saturation. This is highly undesirable, not only from a modeling standpoint, but also because of the convergence problems it can cause in circuit simulators. If an extraction strategy is used that does not guarantee non-negative G_{ds}, very little can be done to fix the problem when G_{ds} becomes negative. Of course, the extraction algorithms can be modified, but this is difficult and time consuming. With optimization strategies, one can weight the fitting for G_{ds} more heavily and thus force the model to produce non-negative G_{ds}. We, therefore, do not favor extraction strategies for BSIM, or anything else. As with most things in life, minimal effort provides minimal rewards.

BSIM2 Extraction Strategy

We do not cover the BSIM2 strategy in complete detail because it is very similar to the BSIM strategy, except more parameters are involved. The major difference in the two models is the inclusion of extra terms in BSIM2 for fitting G_{ds} (refer to Fig. 13.15, which shows how badly BSIM typically fits $1/G_{ds}$ vs. V_{ds}). Basically, the BSIM2 strategy follows the BSIM strategy for the extraction of parameters not related to G_{ds}. Once these have been obtained, the last part of the strategy includes steps for fitting to G_{ds} with parameters that account for channel length modulation and hot electron effects. The way this proceeds in BSIM2 is to fit I_{ds} first, and then parameters MU2, MU3, and MU4 are used to fit to $1/G_{ds}$ vs. V_{ds} curves for families of V_{gs} and V_{bs}. This can be a very time consuming and frustrating experience, because fitting to $1/G_{ds}$ is quite difficult. Also, the equations describing how G_{ds} is modeled with MU2, MU3, and MU4 are very unphysical and the interplay between the parameters makes fitting awkward. The reader is referred to Fig. 13.16, which shows how BSIM2 typically fits $1/G_{ds}$ vs. V_{ds}. BSIM2 is certainly better than BSIM but it has its own problems fitting $1/G_{ds}$.

BSIM3 Comments

The BSIM3 model is very new and will undoubtedly change in the future [Huang et al., 1993]. We will not list a BSIM3 strategy here, but focus instead on the features of the model that make it appealing for analog modeling.

BSIM3 has terms for fitting G_{ds} that relate to channel length modulation, drain-induced barrier lowering, and hot electron effects. They are incorporated completely differently from the G_{ds} fitting parameters of BSIM2. In BSIM3 these parameters enter through a generalized Early voltage relation, with the drain current in saturation written as

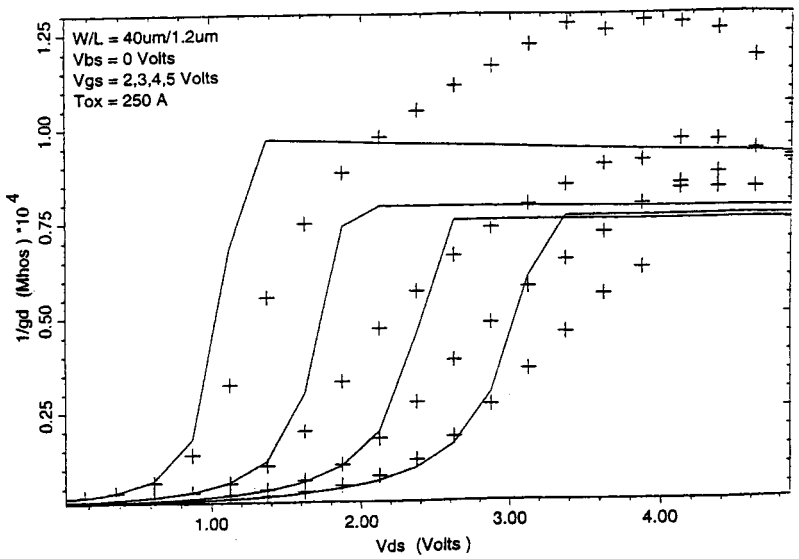


FIGURE 13.15 Typical BSIM $1/G_d$ vs. V_{ds} measured and simulated plots at various V_{gs} values for a wide-short device.

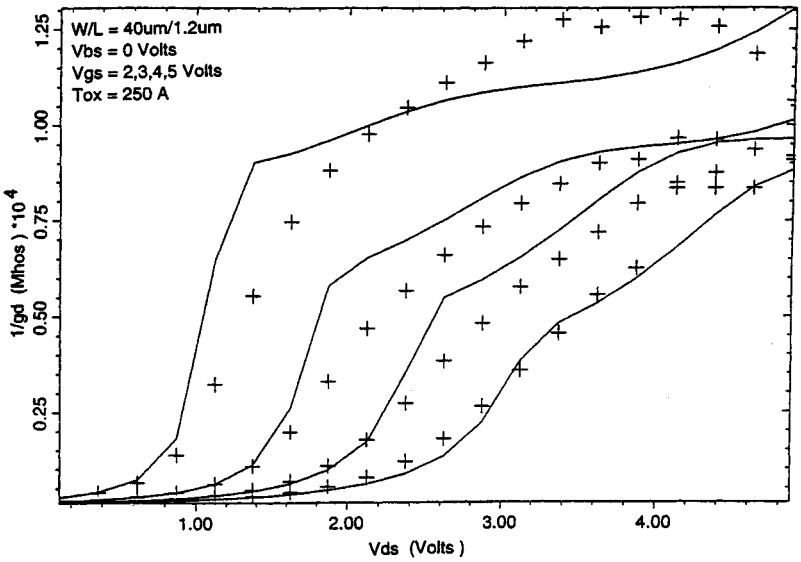


FIGURE 13.16 Typical BSIM2 $1/G_d$ vs. V_{ds} measured and simulated plots at various V_{gs} values for a wide-short device.

$$I_{ds} = I_{d\,\mathrm{sat}}\left[1 + \frac{\left(V_{ds} - V_{d\,\mathrm{sat}}\right)}{V_A}\right] \tag{13.7}$$

where V_A is a generalized Early voltage made up of three terms as

$$\frac{1}{V_A} = \frac{1}{V_{\mathrm{ACLM}}} + \frac{1}{V_{\mathrm{ADIBL}}} + \frac{1}{V_{\mathrm{AHCE}}} \tag{13.8}$$

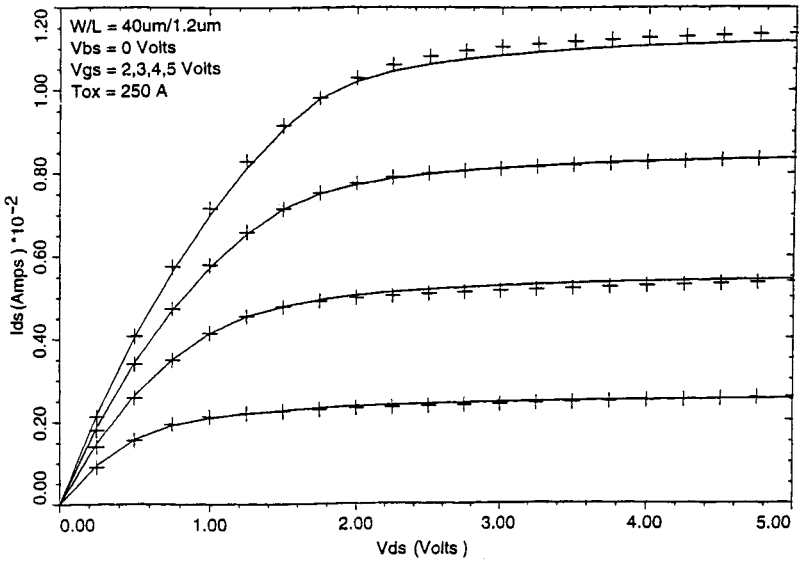


FIGURE 13.17 Typical BSIM3 saturation region measured and simulated plots at various V_{gs} values for a wide-short device.

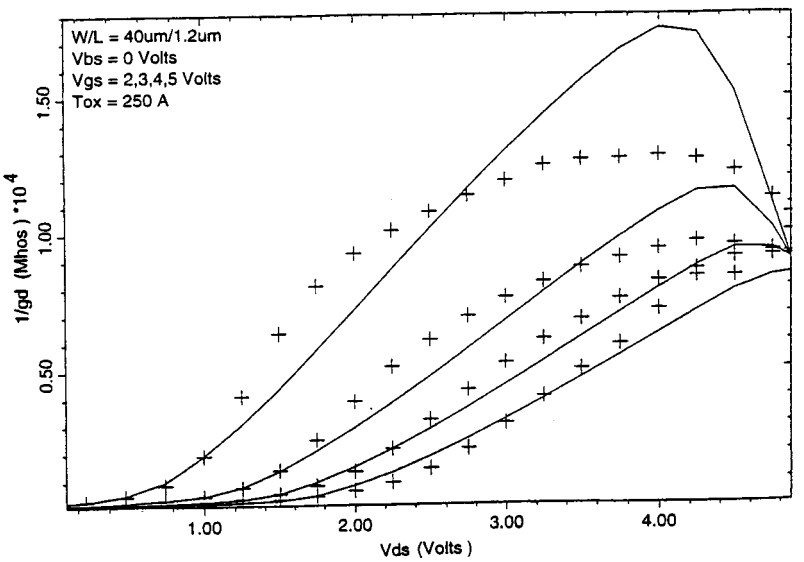


FIGURE 13.18 Typical BSIM3 $1/G_d$ vs. V_{ds} measured and simulated plots at various V_{gs} values for a wide-short device.

with the terms in (13.8) representing generalized Early voltages for channel length modulation (CLM), drain-induced barrier lowering (DIBL), and hot carrier effects (HCE). This formulation is more physically appealing than the one used in BSIM2, making it easier to fit $1/G_{ds}$ vs. V_{ds} curves with BSIM2. Figures 13.17 and 13.18 show how BSIM3 typically fits I_{ds} vs. V_{ds} and $1/G_{ds}$ vs. V_{ds}.

Most of the model parameters for BSIM3 have physical significance so they are obtained in the spirit of the parameters for the level 2 and 3 models. The incorporation of temperature dependence is also easier in BSIM3 because the parameters are more physical. All this, coupled with the fact that about 30 parameters exist for BSIM3 as compared to over 100 for BSIM2, makes BSIM3 a logical choice for analog design. However, BSIM3 is evolving, and shortcomings to the model may still exist that may be corrected in later revisions.

Which MOS Model To Use?

Many MOS models are available in circuit simulators, and the novice is bewildered as to which model is appropriate. No single answer exists, but some questions must be asked before making a choice:

1. What kind of technology am I characterizing?
2. How accurate a model do I need?
3. Do I want to understand the technology?
4. How important are the skew model files (fast and slow parameter files)?
5. How experienced am I? Do I have the expertise to handle a more complicated model?
6. How much time can I spend doing device characterization?
7. Do I need to use this model in more than one circuit simulator?
8. Is the subthreshold region important?
9. Is fitting G_{ds} important?

Let us approach each question with regard to the models available. If the technology is not submicron, perhaps a simpler model such as level 3 is capable of doing everything needed. If the technology is deep submicron, then use a more complicated model such as BSIM, BSIM2, or BSIM3. If high accuracy is required, then the best choice is BSIM3, mainly because it is more physical than all the other models and is capable of fitting better.

For a good physical understanding of the process being characterized. BSIM and BSIM2 are not good choices. These are the least physically based of all the models. The level 2 and 3 models have good physical interpretation for most of the parameters, although they are relatively simple models. BSIM3 is also more physically based, with many more parameters than level 2 or 3, so it is probably the best choice.

If meaningful skew models need to be generated, then BSIM and BSIM2 are very difficult to use, again, because of their unphysical parameter sets. Usually, the simplest physically based model is the best for skew model generation. A more complicated physically based model such as BSIM3 may also be difficult to use for skew model generation.

If the user is inexperienced, none of the BSIM models should be used until the user's expertise improves. Our advice is to practice using simpler models before tackling the harder ones.

If time is critical, the simpler models will definitely be much faster for use in characterization. The more complicated models require more measurements over wider ranges of voltages as well as wider ranges of geometries. This, coupled with the larger number of parameters, means they will take some time with which to work. The BSIM2 model will take longer than all the rest, especially if the G_{ds} fitting parameters are to be used.

The characterization results may need to be used in more than one circuit simulator. For example, if a foundry must supply models to various customers, they may be using different circuit simulators. In this case proprietary models applicable to a single circuit simulator should not be used. Also, circuit designers may want to check the circuit simulation results on more than one circuit simulator. It is better to use standard Berkeley models (level 2, level 3, BSIM, BSIM2, and BSIM3) in such cases.

If the subthreshold region is important, then level 2 or level 3 cannot be used, and probably not even BSIM; BSIM2 or BSIM3 must be used instead. These two models have enough parameters for fitting the subthreshold region.

If fitting G_{ds} is important, BSIM2 and BSIM3 are, again, the only choices. None of the other models have enough parameters for fitting G_{ds}.

Finally, if a very unusual technology is to be characterized, none of the standard models may be appropriate. In this case commercially available specialized models or the user's own models must be used. This will be a large task, so the goals must justify the effort.

Skew Parameter Files

This chapter discussed obtaining model parameters for a single wafer, usually one that has been chosen to represent a typical wafer for the technology being characterized. The parameter values obtained from this wafer correspond to a typical case. Circuit designers also want to simulate circuits with parameter values representing the extremes of process variation, the so-called fast and slow corners, or skew parameter files. These represent the best and worst case of the process variation over time.

Skew parameter values are obtained usually by tracking a few key parameters, measuring many wafers over a long period of time. The standard deviation of these key parameters is found and added to or subtracted from the typical parameter values to obtain the skew models. This method is extremely crude and will not normally produce a realistic skew model. It will almost always overestimate the process spread, because the various model parameters are not independent—they are correlated.

Obtaining realistic skew parameter values, taking into account all the subtle correlations between parameters, is more difficult. In fact, skew model generation is often more an art than a science. Many attempts have been made to utilize techniques from a branch of statistics called multivariate analysis [Dillon and Goldstein, 1984]. In this approach principal component or factor analysis is used to find parameters that are linear combinations of the original parameters. Only the first few of these new parameters will be kept; the others will be discarded because they have less significance. This new set will have fewer parameters than the original set and therefore will be more manageable in terms of finding their skews. The user sometimes must make many choices in the way the common factors are utilized, resulting in different users obtaining different results.

Unfortunately, a great deal of physical intuition is often required to use this approach effectively. To date, we have only seen it applied to the simpler MOS models such as level 3. It is not known if this is a viable approach for a much more complicated model such as BSIM2 [Power et al., 1993].

Related Topic

24.3 The Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET)

References

W. R. Dillon and M. Goldstein, *Multivariate Analysis Methods and Applications,* New York: John Wiley & Sons, 1984.

P. E. Gill, W. Murray, and M. Wright, *Practical Optimization,* Orlando, Fla.: Academic Press, 1981.

J. S. Duster, J.-C. Jeng, P. K. Ko, and C. Hu, "User's Guide for BSIM2 Parameter Extraction Program and The SPICE3 with BSIM Implementation," Electronic Research Laboratory, Berkeley: University of California, 1988.

J.-H. Huang, Z. H. Liu, M.-C. Jeng, P. K. Ko, and C. Hu, "BSIM3 Manual," Berkeley: University of California, 1993.

M.-C. Jeng, P. M. Lee, M. M. Kuo, P. K. Ko, and C. Hu, "Theory, Algorithms, and User's Guide for BSIM and SCALP," Version 2.0, Electronic Research Laboratory, Berkeley: University of California, 1987.

J. A. Power, A. Mathewson, and W. A. Lane, "An Approach for Relating Model Parameter Variabilities to Process Fluctuations," *Proc. IEEE Int. Conf. Microelectronic Test Struct.,* vol. 6, Mar. 1993.

W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C,* Cambridge, U.K.: Cambridge University Press, 1988.

B. J. Sheu, D. L. Scharfetter, P. K. Ko, and M.-C. Jeng, "BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors," *IEEE J. Solid-State Circuits,* vol. SC-22, no. 4, Aug. 1987.

A. Vladimirescu and S. Liu, "The Simulation of MOS Integrated Circuits Using SPICE2," memorandum no. UCB/ERL M80/7, Berkeley: University of California, 1980.

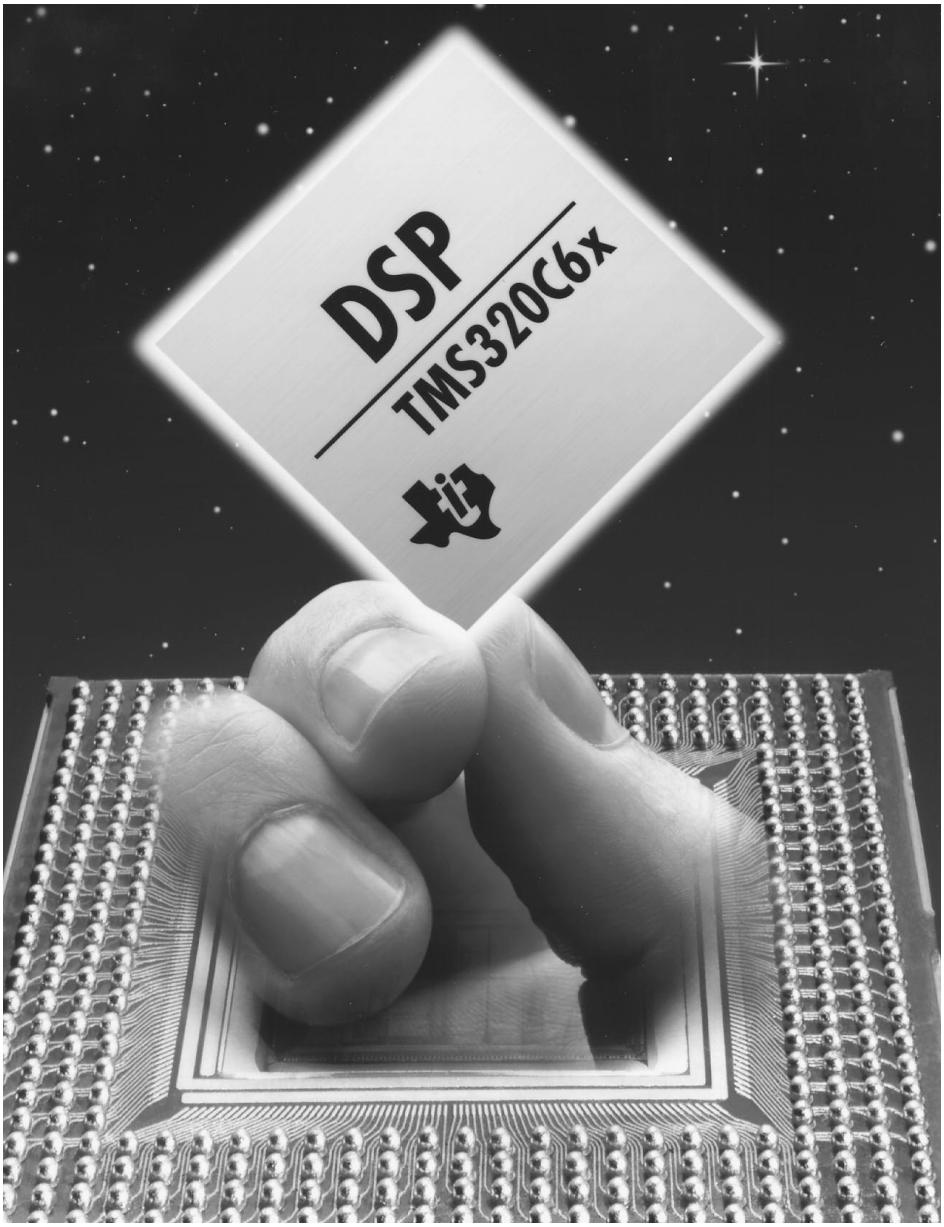
Further Information

Other recommended publications which are useful in device characterization are

L. W. Nagel, "SPICE2: A Computer Program to Simulate Semiconductor Circuits," memorandum no. ERL-M520, Berkeley: University of California, 1975.

G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE,* New York: McGraw-Hill, 1993.

Etter, D. "Section II – Signal Processing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



The world's most powerful digital signal processor, the TMS320C6x, performs at an unprecedented 1600 million-instructions-per-second (MIPS). MIPS are the key measure of a chip's capacity for executing signal processing tasks. This DSP delivers ten times the MIPS performance of any other DSP in history. The TMS320C6x is the first programmable DSP to adopt an advanced Very Long Instruction Word (VLIW) architecture which increases the parallel execution of instructions by packing up to eight 32-bit instructions into a single cycle. The VLIW architecture, combined with the most efficient C compiler ever developed, dramatically improves performance and helps reduce the code development time.

The computing power of the TMS320C6x DSP will change the way new products are designed. The future generations of the TMS320C6x DSP will include devices fabricated with a new 0.18-micron, five-level metal process, operating at speeds beyond 250 MHz. (Photo courtesy of Texas Instruments.)

II

Signal Processing

14 Digital Signal Processing *W.K. Jenkins, A.D. Poularikas, B.W. Bomar, L.M. Smith, J.A. Cadzow*
Fourier Transforms • Fourier Transforms and the Fast Fourier Transform • Design and Implementation of Digital Filters • Signal Restoration

15 Speech Signal Processing *S. McClellan, J.D. Gibson, Y. Ephraim, J.W. Fussell, L.D. Wilcox, M.A. Bush, Y. Gao, B. Ramabhadran, M. Picheny*
Coding, Transmission, and Storage • Speech Enhancement and Noise Reduction • Analysis and Synthesis • Speech Recognition • Large Vocabulary Continuous Speech Recognition

16 Spectral Estimation and Modeling *S.U. Pillai, T.I. Shim, S.N. Batalama, D. Kazakos, F. Daum*
Spectral Analysis • Parameter Estimation • Kalman Filtering

17 Multidimensional Signal Processing *E.J. Delp, J. Allebach, C.A. Bouman, S.A. Rajala, N.K. Bose, L.H. Sibul, W. Wolf, Y-Q Zhang*
Digital Image Processing • Video Signal Processing • Sensor Array Processing • Video Processing Architectures • MPEG-4 Based Multimedia Information System

18 VLSI for Signal Processing *K.K. Parhi, R. Chassaing, B. Bitler*
Special Architectures • Signal Processing Chips and Applications

19 Acoustic Signal Processing *J. Schroeter, S.K. Mehta, G.C. Carter*
Digital Signal Processing in Audio and Electroacoustics • Underwater Acoustical Signal Processing

20 Artificial Neural Networks *J.C. Principe*
Definitions and Scope • Multilayer Perceptrons • Radial Basis Function Networks • Time Lagged Networks • Hebbian Learning and Principal Component Analysis Networks • Competitive Learning and Kohonen Networks

21 Computing Environments for Digital Signal Processing *D.M. Etter*
MATLAB Environment • Example 1: Signal Analysis • Example 2: Filter Design and Analysis • Example 3: Multirate Signal Processing

Delores M. Etter
University of Colorado, Boulder

Signal processing was defined at a meeting in 1991 of the National Science Foundation's MIPS (Microelectronics and Information Processing Systems) Advisory Committee as "the extraction of information-bearing attributes from measured data, and any subsequent transformation of those attributes for the purposes of detection, estimation, classification, or waveform synthesis." If we expand this concise definition, we observe that the signals we typically use in signal processing are functions of time, such as temperature measurements, velocity measurements, voltages, blood pressures, earth motion, and speech signals. Most of these signals are initially continuous signals (also called analog signals) which are measured by sensors that convert energy to electricity. Some of the common types of sensors used for collecting data are microphones, which measure acoustic or sound data; seismometers, which measure earth motion; photocells, which measure light intensity; thermistors, which measure temperature; and oscilloscopes, which measure voltage. When we

work with the continuous electrical signals collected by sensors, we often convert the continuous signal to a digital signal (a sequence of values) with a piece of hardware called an analog-to-digital (A/D) converter. Once we have collected the digital signal, we are ready to use the computer to apply digital signal processing (DSP) techniques to it. These DSP techniques can be designed to perform a number of operations such as:

- Removing noise that is distorting the signal, such as static on a communication line.
- Extracting information from the signal, such as the average value and the power in a signal.
- Separating components of the signal, such as the separation of a band of frequencies that represent the television signal for a specific channel.
- Encoding the information in a more efficient way for transmission, such as the encoding of speech signals into digital signals for transmitting across telephone lines.
- Detecting information in a signal, such as the detection of a surface ship in a sonar signal.

These are just a few of the types of operations that can be performed by signal processing techniques. For some applications, an analog or continuous output signal is needed, and thus a digital-to-analog (D/A) converter is used to convert the modified digital signal to a continuous signal. Another device called a transducer can be used to convert the continuous electrical signal to another form; for example, a speaker converts a continuous electrical signal to an acoustical signal.

In this section the variety and diversity of signal processing is presented from a theoretical point of view, from an implementation point of view, and from an applications point of view. The theoretical point of view includes the development of mathematical models and the development of software algorithms and computer simulations to evaluate and analyze the models both with simulated data and with real data. High-level software tools are important in both the development of new theoretical results and in establishing the validity of the results when applied to real data. The applications determine the way in which the theory is implemented; a key element in the implementation of a signal processing technique relates to whether the technique is applied in real-time (or close to real-time) or whether the processing can be handled off-line. Real-time implementation can use VLSI (very large scale integration) techniques, with commercial DSP chips, or it can involve custom design of chips, MCMs (multichip modules), or ASICs (application-specific integrated circuits). The selection of topics in this section covers the three points of view (theoretical, application, implementation) but should not be assumed to include a complete summary of these topics.

Nomenclature

Symbol	Quantity	Unit	Symbol	Quantity	Unit
AG	array gain	dB	$\phi(K)$	sampled degree phase spectrum	
$A(k)$	sampled amplitude spectrum		$G(e^{j\omega})$	spectral gain function	
C	compression rate		H	entropy	
DFT	discrete Fourier transform		$H(e^{j\omega})$	transfer function of discrete time system	
δ_p	passband ripple				
δ_s	stopband attenuation				
$\delta(t)$	dirac or impulse function		$h(n)$	impulse response	
$\Delta\omega$	transition bandwidth	Hz	η	learning rate parameter	
$E(e^{j\omega})$	Fourier transform of error sequence		$I_n(x)$	modified Bessel function of order n	
f	analog frequency	Hz	L	length of continuous function	s
$f(n)$	sequence		$\mu_x(t)$	ensemble average	
$f(t)$	continuous signal		N	number of sample values	
FFT	fast Fourier transform		ω	digital frequency	rad/s
ϕ	azimuthal angle		Ω	angular frequency	rad

Jenkins, W.K., Poularikas, A.D., Bomar, B.W., Smith, L.M., Cadzow, J.A. "Digital Signal Processing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

14

Digital Signal Processing

14.1 Fourier Transforms
Introduction • The Classical Fourier Transform for CT Signals • Fourier Series Representation of CT Periodic Signals • Generalized Complex Fourier Transform • DT Fourier Transform • Relationship between the CT and DT Spectra • Discrete Fourier Transform

14.2 Fourier Transforms and the Fast Fourier Transform
The Discrete Time Fourier Transform (DTFT) • Relationship to the Z-Transform • Properties • Fourier Transforms of Finite Time Sequences • Frequency Response of LTI Discrete Systems • The Discrete Fourier Transform • Properties of the DFT • Relation between DFT and Fourier Transform • Power, Amplitude, and Phase Spectra • Observations • Data Windowing • Fast Fourier Transform • Computation of the Inverse DFT

14.3 Design and Implementation of Digital Filters
Finite Impulse Response Filter Design • Infinite Impulse Response Filter Design • Finite Impulse Response Filter Implementation • Infinite Impulse Response Filter Implementation

14.4 Signal Restoration
Introduction • Attribute Sets: Closed Subspaces • Attribute Sets: Closed Convex Sets • Closed Projection Operators • Algebraic Properties of Matrices • Structural Properties of Matrices • Nonnegative Sequence Approximation • Exponential Signals and the Data Matrix • Recursive Modeling of Data

W. Kenneth Jenkins
University of Illinois

Alexander D. Poularikas
University of Alabama in Huntsville

Bruce W. Bomar
University of Tennessee Space Institute

L. Montgomery Smith
University of Tennessee Space Institute

James A. Cadzow
Vanderbilt University

14.1 Fourier Transforms

W. Kenneth Jenkins

Introduction

The Fourier transform is a mathematical tool that is used to expand signals into a spectrum of sinusoidal components to facilitate signal analysis and system performance. In certain applications the Fourier transform is used for spectral analysis, or for spectrum shaping that adjusts the relative contributions of different frequency components in the filtered result. In other applications the Fourier transform is important for its ability to decompose the input signal into uncorrelated components, so that signal processing can be more effectively implemented on the individual spectral components. Decorrelating properties of the Fourier transform are important in frequency domain adaptive filtering, subband coding, image compression, and transform coding.

Classical Fourier methods such as the Fourier series and the Fourier integral are used for continuous-time (CT) signals and systems, i.e., systems in which the signals are defined at all values of t on the continuum $-\infty < t < \infty$. A more recently developed set of discrete Fourier methods, including the discrete-time (DT) Fourier transform and the discrete Fourier transform (DFT), are extensions of basic Fourier concepts for DT signals and systems. A DT signal is defined only for integer values of n in the range $-\infty < n < \infty$. The class of DT

Fourier methods is particularly useful as a basis for digital signal processing (DSP) because it extends the theory of classical Fourier analysis to DT signals and leads to many effective algorithms that can be directly implemented on general computers or special-purpose DSP devices.

The Classical Fourier Transform for CT Signals

A CT signal $s(t)$ and its Fourier transform $S(j\omega)$ form a transform pair that are related by Eqs. (14.1) for any $s(t)$ for which the integral (14.1a) converges:

$$S(j\omega) = \int_{-\infty}^{\infty} s(t)e^{-j\omega_t}dt \tag{14.1a}$$

$$s(t) = (1/2\pi)\int_{-\infty}^{\infty} S(j\omega)e^{j\omega_t}d\omega \tag{14.1b}$$

In most literature Eq. (14.1a) is simply called the Fourier transform, whereas Eq. (14.1b) is called the *Fourier integral*. The relationship $S(j\omega) = F\{s(t)\}$ denotes the Fourier transformation of $s(t)$, where $F\{\cdot\}$ is a symbolic notation for the integral operator and where ω is the continuous frequency variable expressed in radians per second. A transform pair $s(t) \leftrightarrow S(j\omega)$ represents a one-to-one invertible mapping as long as $s(t)$ satisfies conditions which guarantee that the Fourier integral converges.

In the following discussion the symbol $\delta(t)$ is used to denote a **CT impulse function** that is defined to be zero for all $t \neq 0$, undefined for $t = 0$, and has unit area when integrated over the range $-\infty < t < \infty$. From Eq. (14.1a) it is found that $F\{\delta(t - t_o)\} = e^{-j\omega t_o}$ due to the well-known sifting property of $\delta(t)$. Similarly, from Eq. (14.1b) we find that $F^{-1}\{2\pi\delta(\omega - \omega_o)\} = e^{j\omega_o t}$, so that $\delta(t - t_o) \leftrightarrow e^{-j\omega t_o}$ and $e^{j\omega_o t} \leftrightarrow 2\pi\delta(\omega - \omega_o)$ are Fourier transform pairs. By using these relationships, it is easy to establish the Fourier transforms of $\cos(\omega_o t)$ and $\sin(\omega_o t)$, as well as many other useful waveforms, many of which are listed in Table 14.1.

The CT Fourier transform is useful in the analysis and design of CT systems, i.e., systems that process CT signals. Fourier analysis is particularly applicable to the design of CT filters which are characterized by Fourier magnitude and phase spectra, i.e., by $|H(j\omega)|$ and $\arg H(j\omega)$, where $H(j\omega)$ is commonly called the frequency response of the filter.

Properties of the CT Fourier Transform

The CT Fourier transform has many properties that make it useful for the analysis and design of linear CT systems. Some of the more useful properties are summarized in this section, while a more complete list of the CT Fourier transform properties is given in Table 14.2. Proofs of these properties are found in Oppenheim et al. [1983] and Bracewell [1986]. Note that $F\{\cdot\}$ denotes the Fourier transform operation, $F^{-1}\{\cdot\}$ denotes the inverse Fourier transform operation, and "$*$" denotes the convolution operation defined as

$$f_1(t) * f_2(t) = \int_{-\infty}^{\infty} f_1(t - t)f_2(t)dt$$

1. *Linearity* (*superposition*): (a and b, complex constants) $F\{af_1(t) + bf_2(t)\} = aF\{f_1(t)\} + bF\{f_2(t)\}$
2. *Time Shifting:* $F\{f(t - t_o)\} = e^{-j\omega t_o}F\{f(t)\}$
3. *Frequency Shifting:* $e^{j\omega_o t}f(t) = F^{-1}\{F(j(\omega - \omega_o))\}$
4. *Time-Domain Convolution:* $F\{f_1(t) * f_2(t)\} = F\{f_1(t)\}F\{f_2(t)\}$
5. *Frequency-Domain Convolution:* $F\{f_1(t)f_2(t)\} = (1/2\pi)F\{f_1(t)\} * F\{f_2(t)\}$
6. *Time Differentiation:* $-j\omega F(j\omega) = F\{d(f(t))/dt\}$
7. *Time Integration:* $$F\left\{\int_{-\infty}^{t} f(t)dt\right\} = (1/j\omega)F(j\omega) + \pi F(0)\delta(\omega)$$

TABLE 14.1 CT Fourier Transform Pairs

Signal	Fourier Transform	Fourier Series Coefficients (if periodic)
$\sum_{k=-\infty}^{+\infty} a_k e^{jk\omega_0 t}$	$2\pi \sum_{k=-\infty}^{+\infty} a_k \delta(\omega - k\omega_0)$	a_k
$e^{j\omega_0 t}$	$2\pi\delta(\omega - \omega_0)$	$a_1 = 1$ $a_k = 0$, otherwise
$\cos \omega_0 t$	$\pi[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]$	$a_1 = a_{-1} = 1/2$ $a_k = 0$, otherwise
$\sin \omega_0 t$	$\frac{\pi}{j}[\delta(\omega - \omega_0) - \delta(\omega + \omega_0)]$	$a_1 = -a_{-1} = 1/2j$ $a_k = 0$, otherwise
$x(t) = 1$	$2\pi\delta(\omega)$	$a_0 = 1, \quad a_k = 0, \ k \neq 0$ (has this Forier series representation for any choice of $T_0 > 0$)
Periodic square wave $x(t) = \begin{cases} 1, & \lvert t \rvert < T_1 \\ 0, & T_1 < \lvert t \rvert \leq \frac{T_0}{2} \end{cases}$ and $x(t + T_0) = x(t)$	$\sum_{k=-\infty}^{+\infty} \frac{2 \sin k\omega_0 T_1}{k} \delta(\omega - k\omega_0)$	$\frac{\omega_0 T_1}{\pi} \operatorname{sinc}\left(\frac{k\omega_0 T_1}{\pi}\right) = \frac{\sin k\omega_0 T_1}{k\pi}$
$\sum_{n=-\infty}^{+\infty} \delta(t - nT)$	$\frac{2\pi}{T} \sum_{k=-\infty}^{+\infty} \delta\left(\omega - \frac{2\pi k}{T}\right)$	$a_k = \frac{1}{T}$ for all k
$x(t) = \begin{cases} 1, & \lvert t \rvert < T_1 \\ 0, & \lvert t \rvert > T_1 \end{cases}$	$2T_1 \operatorname{sinc}\left(\frac{\omega T_1}{\pi}\right) = \frac{2 \sin \omega T_1}{\omega}$	—
$\frac{W}{\pi} \operatorname{sinc}\left(\frac{Wt}{\pi}\right) = \frac{\sin Wt}{\pi t}$	$X(\omega) = \begin{cases} 1, & \lvert \omega \rvert < W \\ 0, & \lvert \omega \rvert > W \end{cases}$	—
$\delta(t)$	1	—
$u(t)$	$\frac{1}{j\omega} + \pi\delta(\omega)$	—
$\delta(t - t_0)$	$e^{-j\omega t_0}$	—
$e^{-at}u(t), \mathcal{R}e\{a\} > 0$	$\frac{1}{a + j\omega}$	—
$te^{-at}u(t), \mathcal{R}e\{a\} > 0$	$\frac{1}{(a + j\omega)^2}$	—
$\frac{t^{n-1}}{(n-1)!} e^{-at}u(t), \mathcal{R}e\{a\} > 0$	$\frac{1}{(a + j\omega)^n}$	—

The above properties are particularly useful in CT system analysis and design, especially when the system characteristics are easily specified in the frequency domain, as in linear filtering. Note that Properties 1, 6, and 7 are useful for solving differential or integral equations. Property 4 (time-domain convolution) provides the

TABLE 14.2 Properties of the CT Fourier Transform

Name	If $\mathcal{F} f(t) = F(j\omega)$, then:
Definition	$F(j\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt$
	$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\omega)e^{j\omega t}d\omega$
Superposition	$\mathcal{F}\left[af_1(t) + bf_2(t)\right] = aF_1(j\omega) + bF_2(j\omega)$
Simplification if:	
(a) $f(t)$ is even	$F(j\omega) = 2\int_0^{\infty} f(t)\cos\omega t\,dt$
(b) $f(t)$ is odd	$F(j\omega) = 2j\int_0^{\infty} f(t)\sin\omega t\,dt$
Negative t	$\mathcal{F} f(-t) = F*(j\omega)$
Scaling:	
(a) Time	$\mathcal{F} f(at) = \frac{1}{\lvert a\rvert}F\left(\frac{j\omega}{a}\right)$
(b) Magnitude	$\mathcal{F} af(t) = aF(j\omega)$
Differentiation	$\mathcal{F}\left[\frac{d^n}{dt^n}f(t)\right] = (j\omega)^n F(j\omega)$
Integration	$\mathcal{F}\left[\int_{-\infty}^{t} f(x)dx\right] = \frac{1}{j\omega}F(j\omega) + \pi F(0)\delta(\omega)$
Time shifting	$\mathcal{F} f(t-a) = F(j\omega)e^{-j\omega a}$
Modulation	$\mathcal{F} f(t)e^{j\omega_0 t} = F\left[j(\omega - \omega_0)\right]$
	$\mathcal{F} f(t)\cos\omega_0 t = \frac{1}{2}\left\{F\left[j(\omega-\omega_0)\right] + F\left[j(\omega+\omega_0)\right]\right\}$
	$\mathcal{F} f(t)\sin\omega_0 t = \frac{1}{2}j\left\{F\left[j(\omega-\omega_0)\right] - F\left[j(\omega+\omega_0)\right]\right\}$
Time convolution	$\mathcal{F}^{-1}\left[F_1(j\omega)F_2(j\omega)\right] = \int_{-\infty}^{\infty} f_1(\tau)f_2(t-\tau)\,d\tau$
Frequency convolution	$\mathcal{F}\left[f_1(t)f_2(t)\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_1(j\lambda)F_2\left[j(\omega-\lambda)\right]d\lambda$

basis for many signal-processing algorithms, since many systems can be specified directly by their impulse or frequency response. Property 3 (frequency shifting) is useful for analyzing the performance of communication systems where different modulation formats are commonly used to shift spectral energy among different frequency bands.

Fourier Spectrum of a CT Sampled Signal

The operation of uniformly sampling a CT signal $s(t)$ at every T seconds is characterized by Eq. (14.2), where $\delta(t)$ is the CT impulse function defined earlier:

$$s_a(t) = \sum_{n=-\infty}^{\infty} s_a(t)\delta(t - nT) = \sum_{n=-\infty}^{\infty} s_a(nT)\delta(t - nT) \tag{14.2}$$

Since $s_a(t)$ is a CT signal, it is appropriate to apply the CT Fourier transform to obtain an expression for the spectrum of the sampled signal:

$$F\{s_a(t)\} = F\left\{\sum_{n=-\infty}^{\infty} s_a(nT)\delta(t - nT)\right\} = \sum_{n=-\infty}^{\infty} s_a(nT)\left[e^{j\omega T}\right]^{-n} \tag{14.3}$$

Since the expression on the right-hand side of Eq. (14.3) is a function of $e^{j\omega T}$, it is customary to express the transform as $F(e^{j\omega T}) = F\{s_a(t)\}$. It will be shown later that if ω is replaced with a normalized frequency $\omega' = \omega/T$, so that $-\pi < \omega' < \pi$, then the right side of Eq. (14.3) becomes identical to the DT Fourier transform that is defined directly for the sequence $s[n] = s_a(nT)$.

Fourier Series Representation of CT Periodic Signals

The classical Fourier series representation of a periodic time domain signal $s(t)$ involves an expansion of $s(t)$ into an infinite series of terms that consist of sinusoidal basis functions, each weighted by a complex constant (Fourier coefficient) that provides the proper contribution of that frequency component to the complete waveform. The conditions under which a periodic signal $s(t)$ can be expanded in a Fourier series are known as the **Dirichlet conditions**. They require that in each period $s(t)$ has a finite number of discontinuities, a finite number of maxima and minima, and that $s(t)$ satisfies the absolute convergence criterion of Eq. (14.4) [Van Valkenburg, 1974]:

$$\int_{-T/2}^{T/2} |s(t)|dt < \infty \tag{14.4}$$

It is assumed throughout the following discussion that the Dirichlet conditions are satisfied by all functions that will be represented by a Fourier series.

The Exponential Fourier Series

If $s(t)$ is a CT periodic signal with period T, then the exponential Fourier series expansion of $s(t)$ is given by

$$s(t) = \sum_{n=-\infty}^{\infty} a_n e^{jn\omega_o t} \tag{14.5a}$$

where $\omega_o = 2\pi/T$ and where the a_n terms are the complex Fourier coefficients given by

$$a_n = (1/T)\int_{-T/2}^{T/2} s(t)e^{-jn\omega_o t}dt \qquad -\infty < n < \infty \tag{14.5b}$$

For every value of t where $s(t)$ is continuous the right side of Eq. (14.5a) converges to $s(t)$. At values of t where $s(t)$ has a finite jump discontinuity, the right side of Eq. (14.5a) converges to the average of $s(t^-)$ and $s(t^+)$, where

$$s(t^-) = \lim_{\varepsilon \to 0} s(t - \varepsilon) \qquad \text{and} \qquad s(t^+) = \lim_{\varepsilon \to 0} s(t + \varepsilon)$$

For example, the Fourier series expansion of the sawtooth waveform illustrated in Fig. 14.1 is characterized by $T = 2\pi$, $\omega_o = 1$, $a_0 = 0$, and $a_n = a_{-n} = A\cos(n\pi)/(jn\pi)$ for $n = 1, 2, \ldots$. The coefficients of the exponential Fourier series given by Eq. (14.5b) can be interpreted as a spectral representation of $s(t)$, since the a_nth coefficient represents the contribution of the $(n\omega_o)$th frequency component to the complete waveform. Since the a_n terms are complex valued, the Fourier domain (spectral) representation has both magnitude and phase spectra. For example, the magnitude of the a_n values is plotted in Fig. 14.2 for the sawtooth waveform of Fig. 14.1. The fact that the a_n terms constitute a discrete set is consistent with the fact that a periodic signal has a **line spectrum**;

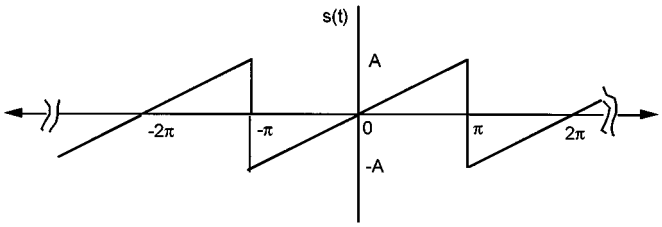


FIGURE 14.1 Periodic CT signal used in Fourier series example.

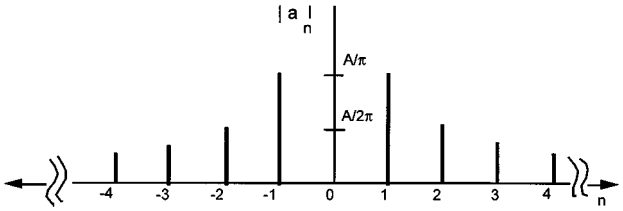


FIGURE 14.2 Magnitude of the Fourier coefficients for the example in Fig. 14.3.

i.e., the spectrum contains only integer multiples of the fundamental frequency ω_o. Therefore, the equation pair given by Eq. (14.5a) and (14.5b) can be interpreted as a transform pair that is similar to the CT Fourier transform for periodic signals. This leads to the observation that the classical Fourier series can be interpreted as a special transform that provides a one-to-one invertible mapping between the discrete-spectral domain and the CT domain.

Trigonometric Fourier Series

Although the complex form of the Fourier series expansion is useful for complex periodic signals, the Fourier series can be more easily expressed in terms of real-valued sine and cosine functions for real-valued periodic signals. In the following discussion it will be assumed that the signal $s(t)$ is real valued for the sake of simplifying the discussion. When $s(t)$ is periodic and real valued it is convenient to replace the complex exponential form of the Fourier series with a **trigonometric expansion** that contains $\sin(\omega_o t)$ and $\cos(\omega_o t)$ terms with corresponding real-valued coefficients [Van Valkenburg, 1974]. The trigonometric form of the Fourier series for a real-valued signal $s(t)$ is given by

$$s(t) = \sum_{n=0}^{\infty} b_n \cos(n\omega_0) + \sum_{n=1}^{\infty} c_n \sin(n\omega_0) \tag{14.6a}$$

where $\omega_o = 2\pi/T$. The b_n and c_n terms are real-valued Fourier coefficients determined by

$$b_0 = (1/T)\int_{-T/2}^{T/2} s(t)\,dt$$

$$b_n = (2/T)\int_{-T/2}^{T/2} s(t)\cos(n\omega_0 t)\,dt, \qquad n = 1, 2, \ldots$$

and

$$c_n = (2/T)\int_{-T/2}^{T/2} s(t)\sin(n\omega_0 t)\,dt, \qquad n = 1, 2, \ldots \tag{14.6b}$$

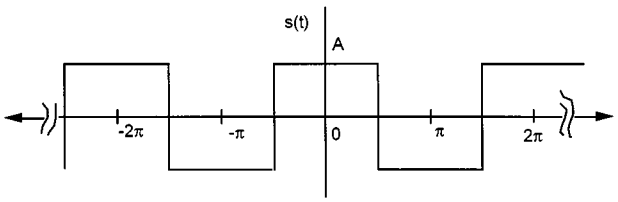


FIGURE 14.3 Periodic CT signal used in Fourier series example 2.

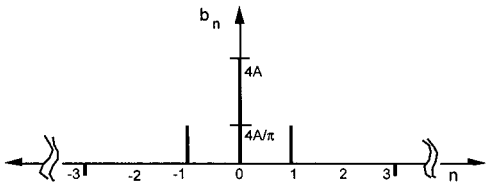


FIGURE 14.4 Fourier coefficients for example of Fig. 14.3.

An arbitrary real-valued signal $s(t)$ can be expressed as a sum of even and odd components, $s(t) = s_{even}(t) + s_{odd}(t)$, where $s_{even}(t) = s_{even}(-t)$ and $s_{odd}(t) = -s_{odd}(-t)$, and where $s_{even}(t) = [s(t) + s(-t)]/2$ and $s_{odd}(t) = [s(t) - s(-t)]/2$. For the trigonometric Fourier series, it can be shown that $s_{even}(t)$ is represented by the (even) cosine terms in the infinite series, $s_{odd}(t)$ is represented by the (odd) sine terms, and b_0 is the dc level of the signal. Therefore, if it can be determined by inspection that a signal has a dc level, or if it is even or odd, then the correct form of the trigonometric series can be chosen to simplify the analysis. For example, it is easily seen that the signal shown in Fig. 14.3 is an even signal with a zero dc level. Therefore, it can be accurately represented by the cosine series with $b_n = 2A \sin(\pi n/2)/(\pi n/2)$, $n = 1, 2, \ldots$, as illustrated in Fig. 14.4. In contrast, note that the sawtooth waveform used in the previous example is an odd signal with zero dc level, so that it can be completely specified by the sine terms of the trigonometric series. This result can be demonstrated by pairing each positive frequency component from the exponential series with its conjugate partner; i.e., $c_n = \sin(n\omega_o t) = a_n e^{jn\omega_0 t} + a_{-n} e^{-jn\omega_0 t}$, whereby it is found that $c_n = 2A \cos(n\pi)/(n\pi)$ for this example. In general, it is found that $a_n = (b_n - jc_n)/2$ for $n = 1, 2, \ldots$, $a_0 = b_0$, and $a_{-n} = a_n^*$. The trigonometric Fourier series is common in the signal processing literature because it replaces complex coefficients with real ones and often results in a simpler and more intuitive interpretation of the results.

Convergence of the Fourier Series

The Fourier series representation of a periodic signal is an approximation that exhibits mean-squared convergence to the true signal. If $s(t)$ is a periodic signal of period T and $s'(t)$ denotes the Fourier series approximation of $s(t)$, then $s(t)$ and $s'(t)$ are equal in the mean-squared sense if

$$\text{mse} = \int_{-T/2}^{T/2} \left| s(t) - s'(t) \right|^2 dt = 0 \tag{14.7}$$

Even when Eq. (14.7) is satisfied, **mean-squared error** (mse) convergence does not guarantee that $s(t) = s'(t)$ at every value of t. In particular, it is known that at values of t where $s(t)$ is discontinuous the Fourier series converges to the average of the limiting values to the left and right of the discontinuity. For example, if t_0 is a point of discontinuity, then $s'(t_0) = [s(t_0^-) + s(t_0^+)]/2$, where $s(t_0^-)$ and $s(t_0^+)$ were defined previously (note that at points of continuity, this condition is also satisfied by the very definition of continuity). Since the Dirichlet conditions require that $s(t)$ have at most a finite number of points of discontinuity in one period, the set S_t such that $s(t) \neq s'(t)$ within one period contains a finite number of points, and S_t is a set of measure zero in the formal mathematical sense. Therefore, $s(t)$ and its Fourier series expansion $s'(t)$ are *equal almost everywhere*, and $s(t)$ can be considered identical to $s'(t)$ for analysis in most practical engineering problems.

The condition described above of convergence almost everywhere is satisfied only in the limit as an infinite number of terms are included in the Fourier series expansion. If the infinite series expansion of the Fourier series is truncated to a finite number of terms, as it must always be in practical applications, then the approximation will exhibit an oscillatory behavior around the discontinuity, known as the **Gibbs phenomenon** [Van Valkenburg, 1974]. Let $s'_N(t)$ denote a truncated Fourier series approximation of $s(t)$, where only the terms in Eq. (14.5a) from $n = -N$ to $n = N$ are included if the complex Fourier series representation is used or where only the terms in Eq. (14.6a) from $n = 0$ to $n = N$ are included if the trigonometric form of the Fourier series is used. It is well known that in the vicinity of a discontinuity at t_0 the Gibbs phenomenon causes $s'_N(t)$ to be a poor approximation to $s(t)$. The peak magnitude of the Gibbs oscillation is 13% of the size of the jump discontinuity $s(t_0^-) - s(t_0^+)$ regardless of the number of terms used in the approximation. As N increases, the region which contains the oscillation becomes more concentrated in the neighborhood of the discontinuity, until, in the limit as N approaches infinity, the Gibbs oscillation is squeezed into a single point of mismatch at t_0. The Gibbs phenomenon is illustrated in Fig. 14.5, where an ideal low-pass frequency response is approximated by an impulse response function that has been limited to having only N nonzero coefficients, and hence the Fourier series expansion contains only a finite number of terms.

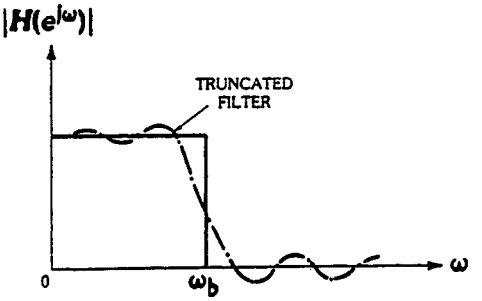


FIGURE 14.5 Gibbs phenomenon in a low-pass digital filter caused by truncating the impulse response to N terms.

If $s'(t)$ in Eq. (14.7) is replaced by $s'_N(t)$ it is important to understand the behavior of the error mse_N as a function of N, where

$$\text{mse}_N = \int_{-T/2}^{T/2} \left| s(t) - s'_N(t) \right|^2 dt \tag{14.8}$$

An important property of the Fourier series is that the exponential basis functions $e^{jn\omega_o t}$ (or $\sin(n\omega_o t)$ and $\cos(n\omega_o t)$ for the trigonometric form) for $n = 0, \pm 1, \pm 2, \ldots$ (or $n = 0, 1, 2, \ldots$ for the trigonometric form) constitute an **orthonormal set**; i.e., $t_{nk} = 1$ for $n = k$, and $t_{nk} = 0$ for $n \neq k$, where

$$t_{nk} = (1/T)\int_{-T/2}^{T/2} \left(e^{-jn\omega_o t}\right)\left(e^{jn\omega_o t}\right) dt \tag{14.9}$$

As terms are added to the Fourier series expansion, the orthogonality of the basis functions guarantees that the error decreases monotonically in the mean-squared sense, i.e., that mse_N monotonically decreases as N is increased. Therefore, when applying Fourier series analysis, including more terms always improves the accuracy of the signal representation.

Fourier Transform of Periodic CT Signals

For a periodic signal $s(t)$ the CT Fourier transform can then be applied to the Fourier series expansion of $s(t)$ to produce a mathematical expression for the "line spectrum" that is characteristic of periodic signals:

$$F\{s(t)\} = F\left\{\sum_{n=\infty}^{\infty} a_n e^{jn\omega_o t}\right\} = 2\pi \sum_{n=-\infty}^{\infty} a_n \delta(\omega - n\omega_o) \tag{14.10}$$

The spectrum is shown in Fig. 14.6. Note the similarity between the spectral representation of Fig. 14.6 and the plot of the Fourier coefficients in Fig. 14.2, which was heuristically interpreted as a line spectrum. Figures 14.2 and

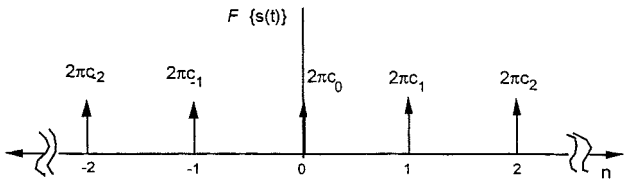


FIGURE 14.6 Spectrum of the Fourier representation of a periodic signal.

14.6 are different, but equivalent, representations of the Fourier line spectrum that is characteristic of periodic signals.

Generalized Complex Fourier Transform

The CT Fourier transform characterized by Eqs. (14.11a) and (14.11b) can be generalized by considering the variable $j\omega$ to be the special case of $u = \sigma + j\omega$ with $\sigma = 0$, writing Eqs. (14.11) in terms of u, and interpreting u as a complex frequency variable. The resulting complex Fourier transform pair is given by Eqs. (14.11a) and (14.11b):

$$s(t) = (1/2\pi j)\int_{\sigma - j\infty}^{\sigma + j\infty} S(u)e^{jut}du \tag{14.11a}$$

$$s(u) = \int_{-\infty}^{\infty} s(t)e^{-jut}dt \tag{14.11b}$$

The set of all values of u for which the integral of Eq. (14.11b) converges is called the region of convergence, denoted ROC. Since the transform $S(u)$ is defined only for values of u within the ROC, the path of integration in Eq. (14.11a) must be defined by s so the entire path lies within the ROC. In some literature this transform pair is called the *bilateral Laplace transform* because it is the same result obtained by including both the negative and positive portions of the time axis in the classical Laplace transform integral. The complex Fourier transform (bilateral Laplace transform) is not often used in solving practical problems, but its significance lies in the fact that it is the most general form that represents the place where Fourier and Laplace transform concepts merge. Identifying this connection reinforces the observation that Fourier and Laplace transform concepts share common properties because they are derived by placing different constraints on the same parent form.

DT Fourier Transform

The DT Fourier transform (DTFT) is obtained directly in terms of the sequence samples $s[n]$ by taking the relationship obtained in Eq. (14.3) to be the definition of the DTFT. By letting $T = 1$ so that the sampling period is removed from the equations and the frequency variable is replaced with a normalized frequency $\omega' = \omega T$, the DTFT pair is defined by Eqs. (14.12). In order to simplify notation it is not customary to distinguish between ω and ω', but rather to rely on the context of the discussion to determine whether ω refers to the normalized ($T = 1$) or to the unnormalized ($T \neq 1$) frequency variable.

$$S(e^{j\omega'}) = \sum_{n=-\infty}^{\infty} s[n]e^{-j\omega' n} \tag{14.12a}$$

$$s[n] = (1/2\pi)\int_{-\pi}^{\pi} S(e^{j\omega'})e^{jn\omega'}d\omega' \tag{14.12b}$$

TABLE 14.3 Some Basic DTFT Pairs

Sequence	Fourier Transform
1. $\delta[n]$	1
2. $\delta[n - n_0]$	$e^{-j\omega n_0}$
3. $1\ (-\infty < n < \infty)$	$\sum_{k=-\infty}^{\infty} 2\pi\delta(\omega + 2\pi k)$
4. $a^n u[n]\ (\lvert a\rvert < 1)$	$\frac{1}{1 - ae^{-j\omega}}$
5. $u[n]$	$\frac{1}{1 - e^{-j\omega}} + \sum_{k=-\infty}^{\infty} \pi\delta(\omega + 2\pi k)$
6. $(n + 1)a^n u[n]\ (\lvert a\rvert < 1)$	$\frac{1}{(1 - ae^{-j\omega})^2}$
7. $\frac{r^n \sin \omega_p(n + 1)}{\sin \omega_p} u[n]\ (\lvert r\rvert < 1)$	$\frac{1}{1 - 2r\cos\omega_p e^{-j\omega} + r^2 e^{-j2\omega}}$
8. $\frac{\sin \omega_c n}{\pi n}$	$X(e^{j\omega}) = \begin{cases} 1, & \lvert\omega\rvert < \omega_c, \\ 0, & \omega_c < \lvert\omega\rvert \le \pi \end{cases}$
9. $x[n] = \begin{cases} 1, & 0 \le n \le M \\ 0, & \text{otherwise} \end{cases}$	$\frac{\sin[\omega(M + 1)/2]}{\sin(\omega/2)} e^{-j\omega M/2}$
10. $e^{j\omega_0 n}$	$\sum_{k=-\infty}^{\infty} 2\pi\delta(\omega - \omega_0 + 2\pi k)$
11. $\cos(\omega_0 n + \phi)$	$\pi\sum_{k=-\infty}^{\infty} [e^{j\phi}\delta(\omega - \omega_0 + 2\pi k) + e^{-j\phi}\delta(\omega + \omega_0 + 2\pi k)]$

The spectrum $S(e^{j\omega'})$ is periodic in ω' with period 2π. The fundamental period in the range $-\pi < \omega' \le \pi$, sometimes referred to as the baseband, is the useful frequency range of the DT system because frequency components in this range can be represented unambiguously in sampled form (without aliasing error). In much of the signal-processing literature the explicit primed notation is omitted from the frequency variable. However, the explicit primed notation will be used throughout this section because there is a potential for confusion when so many related Fourier concepts are discussed within the same framework.

By comparing Eqs. (14.3) and (14.12a), and noting that $\omega' = \omega T$, we see that

$$F\{s_a(t)\} = \text{DTFT}\{s[n]\} \tag{14.13}$$

where $s[n] = s(t)|_{t = nT}$. This demonstrates that the spectrum of $s_a(t)$ as calculated by the CT Fourier transform is identical to the spectrum of $s[n]$ as calculated by the DTFT. Therefore, although $s_a(t)$ and $s[n]$ are quite different sampling models, they are equivalent in the sense that they have the same Fourier domain representation. A list of common DTFT pairs is presented in Table 14.3. Just as the CT Fourier transform is useful in CT signal system analysis and design, the DTFT is equally useful for DT system analysis and design.

In the same way that the CT Fourier transform was found to be a special case of the complex Fourier transform (or bilateral Laplace transform), the DTFT is a special case of the *bilateral z-transform* with $z = e^{j\omega' t}$. The more general bilateral z-transform is given by

$$S(z) = \sum_{n=-\infty}^{\infty} s[n] z^{-n} \tag{14.14a}$$

$$s[n] = (1/2\pi j) \int_C S(z) z^{n-1} dz \tag{14.14b}$$

where C is a counterclockwise contour of integration which is a closed path completely contained within the ROC of $S(z)$. Recall that the DTFT was obtained by taking the CT Fourier transform of the CT sampling model $s_a(t)$. Similarly, the bilateral z-transform results by taking the bilateral Laplace transform of $s_a(t)$. If the lower limit on the summation of Eq. (14.14a) is taken to be $n = 0$, then Eqs. (14.14a) and (14.14b) become the one-sided z-transform, which is the DT equivalent of the one-sided Laplace transform for CT signals.

Properties of the DTFT

Since the DTFT is a close relative of the classical CT Fourier transform, it should come as no surprise that many properties of the DTFT are similar to those of the CT Fourier transform. In fact, for many of the properties presented earlier there is an analogous property for the DTFT. The following list parallels the list that was presented in the previous section for the CT Fourier transform, to the extent that the same property exists. A more complete list of DTFT pairs is given in Table 14.4:

1. *Linearity* (*superposition*): (a and b, complex constants) $\text{DTFT}\{af_1[n] + bf_2[n]\} = a\text{DTFT}\{f_1[n]\} + b\text{DTFT}\{f_2[n]\}$
2. *Index Shifting:* $\text{DTFT}\{f[n - n_o]\} = e^{-j\omega n_o}\text{DTFT}\{f[n]\}$
3. *Frequency Shifting:* $e^{j\omega_o n} f[n] = \text{DTFT}^{-1}\{F(j(\omega - \omega_o))\}$
4. *Time-Domain Convolution:* $\text{DTFT}\{f_1[n] * f_2[n]\} = \text{DTFT}\{f_1[n]\}\ \text{DTFT}\{f_2[n]\}$
5. *Frequency-Domain Convolution:* $\text{DTFT}\{f_1[n]\ f_2[n]\} = (1/2\pi)\text{DTFT}\{f_1[n]\} * \text{DTFT}\{f_2[n]\}$
6. *Frequency Differentiation:* $nf[n] = \text{DTFT}^{-1}\{dF(j\omega)/d\omega\}$

Note that the time-differentiation and time-integration properties of the CT Fourier transform do not have analogous counterparts in the DTFT because time-domain differentiation and integration are not defined for DT signals. When working with DT systems practitioners must often manipulate difference equations in the frequency domain. For this purpose Property 1 (linearity) and Property 2 (index shifting) are important. As with the CT Fourier transform, Property 4 (time-domain convolution) is very important for DT systems because it allows engineers to work with the frequency response of the system in order to achieve proper shaping of the input spectrum, or to achieve frequency selective filtering for noise reduction or signal detection. Also, Property 3 (frequency shifting) is useful for the analysis of modulation and filtering common in both analog and digital communication systems.

Relationship between the CT and DT Spectra

Since DT signals often originate by sampling a CT signal, it is important to develop the relationship between the original spectrum of the CT signal and the spectrum of the DT signal that results. First, the CT Fourier transform is applied to the CT sampling model, and the properties are used to produce the following result:

$$F\{s_a(t)\} = F\left\{s_a(t) \sum_{n=-\infty}^{\infty} \delta(t - nT)\right\} = (1/2\pi) S(j\omega) F\left\{\sum_{n=-\infty}^{\infty} \delta(t - nT)\right\} \tag{14.15}$$

Table 14.4 Properties of the DTFT

Sequence $x[n]$ $y[n]$	Fourier Transform $X(e^{j\omega})$ $Y(e^{j\omega})$
1. $ax[n] + by[n]$	$aX(e^{j\omega}) + bY(e^{j\omega})$
2. $x[n - n_d]$ $(n_d$ an integer$)$	$e^{-j\omega n_d}X(e^{j\omega})$
3. $e^{j\omega_0 n}x[n]$	$X(e^{j(\omega-\omega_0)})$
4. $x[-n]$	$X(e^{-j\omega})$ if $x[n]$ real
	$X*(e^{j\omega})$
5. $nx[n]$	$j\dfrac{dX(e^{j\omega})}{d\omega}$
6. $x[n]*y[n]$	$X(e^{j\omega})Y(e^{j\omega})$
7. $x[n]y[n]$	$\dfrac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\theta})Y(e^{j(\omega-\theta)})d\theta$
Parseval's Theorem	
8. $\sum_{n=-\infty}^{\infty}\lvert x[n]\rvert^2 = \dfrac{1}{2\pi}\int_{-\pi}^{\pi}\lvert X(e^{j\omega})\rvert^2 d\omega$	
9. $\sum_{n=-\infty}^{\infty} x[n]y*[n] = \dfrac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})Y*(e^{j\omega}d\omega)$	

In this section it is important to distinguish between ω and ω', so the explicit primed notation is used in the following discussion where needed for clarification. Since the sampling function (summation of shifted impulses) on the right-hand side of the above equation is periodic with period T it can be replaced with a CT Fourier series expansion as follows:

$$S(e^{j\omega T}) = F\{s_a(t)\} = (1/2\pi)S(j\omega)F\left\{\sum_{n=-\infty}^{\infty}(1/T)e^{j(2\pi/T)nt}\right\}$$

Applying the frequency-domain convolution property of the CT Fourier transform yields

$$S(e^{j\omega T}) = (1/2\pi)\sum_{n=-\infty}^{\infty} S(j\omega) * (2\pi/T)\delta(\omega - (2\pi/T)n) = (1/T)\sum_{n=-\infty}^{\infty} S(j[\omega - n\omega_s]) \quad (14.16a)$$

where $\omega_s = (2\pi/T)$ is the sampling frequency (rad/s). An alternate form for the expression of Eq. (14.16a) is

$$S(e^{j\omega'}) = (1/T)\sum_{n=-\infty}^{\infty} S(j[(\omega' - n2\pi)/T]) \quad (14.16b)$$

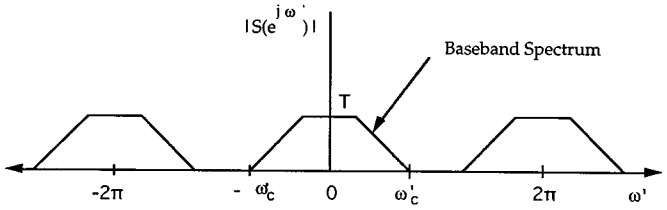


FIGURE 14.7 Relationship between the CT and DT spectra.

where $\omega' = \omega T$ is the normalized DT frequency axis expressed in radians. Note that $S(e^{j\omega T}) = S(e^{j\omega'})$ consists of an infinite number of replicas of the CT spectrum $S(j\omega)$, positioned at intervals of $(2\pi/T)$ on the ω axis (or at intervals of 2π on the ω' axis), as illustrated in Fig. 14.7. If $S(j\omega)$ is band limited with a bandwidth ω_c and if T is chosen sufficiently small so that $\omega_s > 2\omega_c$, then the DT spectrum is a copy of $S(j\omega)$ (scaled by $1/T$) in the baseband. The limiting case of $\omega_s = 2\omega_c$ is called the **Nyquist sampling frequency**. Whenever a CT signal is sampled at or above the Nyquist rate, no aliasing distortion occurs (i.e., the baseband spectrum does not overlap with the higher-order replicas) and the CT signal can be exactly recovered from its samples by extracting the baseband spectrum of $S(e^{j\omega'})$ with an ideal low-pass filter that recovers the original CT spectrum by removing all spectral replicas outside the baseband and scaling the baseband by a factor of T.

Discrete Fourier Transform

To obtain the DFT the continuous-frequency domain of the DTFT is sampled at N points uniformly spaced around the unit circle in the z-plane, i.e., at the points $\omega_k = (2\pi k/N)$, $k = 0, 1, \ldots, N-1$. The result is the DFT transform pair defined by Eqs. (14.17a) and (14.17b). The signal $s[n]$ is either a finite-length sequence of length N or it is a periodic sequence with period N.

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{-j2\pi kn/N} \qquad k = 0, 1, \ldots, N-1 \tag{14.17a}$$

$$s[n] = (1/N) \sum_{n=0}^{N-1} S[k] e^{j2\pi kn/N} \qquad n = 0, 1, \ldots, N-1 \tag{14.17b}$$

Regardless of whether $s[n]$ is a finite-length or a periodic sequence, the DFT treats the N samples of $s[n]$ as though they characterize one period of a periodic sequence. This is an important feature of the DFT, and one that must be handled properly in signal processing to prevent the introduction of artifacts. Important properties of the DFT are summarized in Table 14.5. The notation $(k)_N$ denotes k modulo N, and $R_N[n]$ is a rectangular window such that $R_N[n] = 1$ for $n = 0, \ldots, N-1$, and $R_N[n] = 0$ for $n < 0$ and $n \geq N$. The transform relationship given by Eqs. (14.17a) and (14.17b) is also valid when $s[n]$ and $S[k]$ are periodic sequences, each of period N. In this case, n and k are permitted to range over the complete set of real integers, and $S[k]$ is referred to as the discrete Fourier series (DFS). The DFS is developed by some authors as a distinct transform pair in its own right [Oppenheim and Schafer, 1975]. Whether or not the DFT and the DFS are considered identical or distinct is not very important in this discussion. The important point to be emphasized here is that the DFT treats $s[n]$ as though it were a single period of a periodic sequence, and all signal processing done with the DFT will inherit the consequences of this assumed periodicity.

Properties of the DFT

Most of the properties listed in Table 14.5 for the DFT are similar to those of the z-transform and the DTFT, although there are some important differences. For example, Property 5 (time-shifting property), holds for circular shifts of the finite-length sequence $s[n]$, which is consistent with the notion that the DFT treats $s[n]$ as one period of a periodic sequence. Also, the multiplication of two DFTs results in the **circular convolution**

TABLE 14.5 Properties of the Discrete Fourier Transform (DFT)

Finite-Length Sequence (Length N)	N-Point DFT (Length N)
1. $x[n]$	$X[k]$
2. $x_1[n], x_2[n]$	$X_1[k], X_2[k]$
3. $ax_1[n] + bx_2[n]$	$aX_1[k] + bX_2[k]$
4. $X[n]$	$Nx[(-k)_N]$
5. $x[(n-m)_N]$	$W_N^{km}X[k]$
6. $W_N^{-\ell n}x[n]$	$X[(k-\ell)_N]$
7. $\sum_{m=0}^{N-1} x_1(m)x_2[(n-m)_N]$	$X_1[k]X_2[k]$
8. $x_1[n]x_2[n]$	$\frac{1}{N}\sum_{\ell=0}^{N-1} X_1(\ell)X_2[(k-\ell)_N]$
9. $x^*[n]$	$X^*[(-k)_N]$
10. $x^*[(-n)_N]$	$X^*[k]$
11. $\mathcal{R}e\{x[n]\}$	$X_{ep}[k] = \frac{1}{2}\{X[(k)_N] + X^*[(-k)_N]\}$
12. $j\mathcal{I}m\{x[n]\}$	$X_{op}[k] = \frac{1}{2}\{X[(k)_N] - X^*[(-k)_N]\}$
13. $x_{ep}[n] = \frac{1}{2}\{x[n] + x^*[(-n)_N]\}$	$\mathcal{R}e\{X[k]\}$
14. $x_{op}[n] = \frac{1}{2}\{x[n] - x^*[(-n)_N]\}$	$j\mathcal{I}m\{X[k]\}$
Properties 15–17 apply only when $x[n]$ is real.	
15. Symmetry properties	$\begin{cases} X[k] = X^*[(-k)_N] \\ \mathcal{R}e\{X[k]\} = \mathcal{R}e\{X[(-k)_N]\} \\ \mathcal{I}m\{X[k]\} = \mathcal{I}m\{X[(-k)]_N\} \\ \lvert X[k]\rvert = \lvert X[(-k)_N]\rvert \\ \measuredangle\{X[k]\} = -\measuredangle\{X[(-k)_N]\} \end{cases}$
16. $x_{ep}[n] = \frac{1}{2}\{x[n] + x[(-n)_N]\}$	$\mathcal{R}e\{X[k]\}$
17. $x_{op}[n] = \frac{1}{2}\{x[n] - x[(-n)_N]\}$	$j\mathcal{I}m\{X[k]\}$

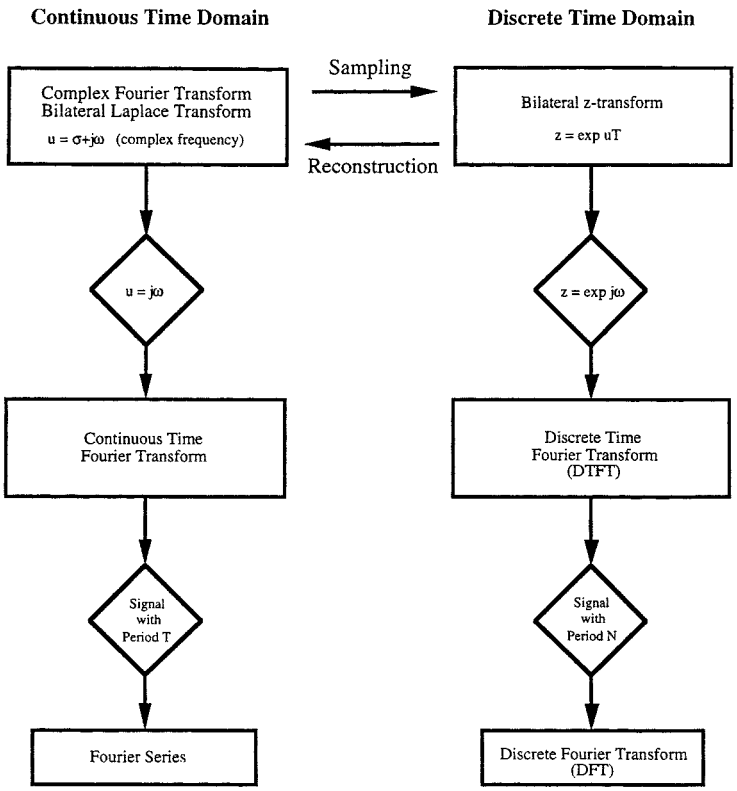


FIGURE 14.8 Functional relationships among various Fourier transforms.

of the corresponding DT sequences, as specified by Property 7. This latter property is quite different from the *linear convolution* property of the DTFT. Circular convolution is simply a linear convolution of the periodic extensions of the finite sequences being convolved, where each of the finite sequences of length N defines the structure of one period of the periodic extensions.

For example, suppose it is desired to implement the following finite impulse response (FIR) digital filter,

$$y[n] = \sum_{k=0}^{N-1} h[k]s[n-k] \tag{14.18}$$

the output of which is obtained by transforming $h[n]$ and $s[n]$ into $H[k]$ and $S[k]$ via the DFT (FFT), multiplying the transforms point-wise to obtain $Y[k] = H[k]S[k]$, and then using the inverse DFT (FFT) to obtain $y[n] = \text{IDFT}\{Y[k]\}$. If $s[n]$ is a finite sequence of length M, then the result of the circular convolution implemented by the DFT will correspond to the desired linear convolution if and only if the block length of the DFT is chosen so that $N_{\text{DFT}} \geq N + M - 1$ and both $h[n]$ and $s[n]$ are padded with zeros to form blocks of length N_{DFT}.

Relationships among Fourier Transforms

Figure 14.8 illustrates the functional relationships among the various forms of CT and DT Fourier transforms that have been discussed in the previous sections. The family of CT is shown on the left side of Fig. 14.8, whereas the right side of the figure shows the hierarchy of DTFTs. Fourier transforms. The complex Fourier transform is identical to the bilateral Laplace transform, and it is at this level that the classical Laplace transform techniques and the Fourier transform techniques become identical.

Defining Terms

Continuous-time (CT) impulse function: A generalized function $\delta(t)$ defined to be zero for all $t \neq 0$, undefined at $t = 0$, and having the special property that $\int_{-\infty}^{\infty} \delta(t)dt = 1$.

Circular convolution: A convolution of finite-length sequences in which the shifting operation is performed circularly within the finite support interval. Alternatively called *periodic convolution.*

Dirichlet conditions: Conditions that must be satisfied in order to expand a periodic signal $s(t)$ in a Fourier series: each period of $s(t)$ must have a finite number of discontinuities, a finite number of maxima and minima, and $\int_{-T/2}^{T/2} |s(t)|dt < \infty$ must be satisfied, where T is the period.

Gibbs phenomenon: Oscillatory behavior of Fourier series approximations in the vicinity of finite jump discontinuities.

Line spectrum: A common term for Fourier transforms of periodic *signals* for which the spectrum has nonzero components only at integer multiples of the fundamental frequency.

Mean-squared error (mse): A measure of "closeness" between two functions given by

$$\text{mse} = \frac{1}{T}\int_{-T/2}^{T/2} |f_1(t) - f_2(t)|^2 dt$$

where T is the period.

Nyquist sampling frequency: Minimum sampling frequency for which a CT signal $s(t)$ can be perfectly reconstructed from a set of uniformly spaced samples $s(nT)$.

Orthonormal set: A countable set of functions for which every pair in the set is mathematically orthogonal according to a valid norm, and for which each element of the set has unit length according to the same norm. The Fourier basis functions form an orthonormal set according to the mse norm.

Trigonometric expansion: A Fourier series expansion for a real-valued signal in which the basis functions are chosen to be $\sin(n\omega_o t)$ and $\cos(n\omega_o t)$

Related Topic

16.1 Spectral Analysis

References

R. N. Bracewell, *The Fourier Transform,* 2nd ed., New York: McGraw-Hill, 1986.

W. K. Jenkins, "Fourier series, Fourier transforms, and the discrete Fourier transform," in *The Circuits and Filters Handbook,* Chen, (ed.), Boca Raton, Fla.: CRC Press, 1995.

A. V. Oppenheim, A. S. Willsky, and I. T. Young, *Signals and Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1983.

A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1975.

M. E., VanValkenburg, *Network Analysis*, Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Further Information

A more thorough treatment of the complete family of CT and DT Fourier transform concepts is given in Jenkins [1995]. This article emphasizes the parallels between CT and DT Fourier concepts.

An excellent treatment of Fourier waveform analysis is given by D. C. Munson, Jr., in chapter 7 of *Reference Data for Engineers: Radio, Electronics, Computers, and Communications*, 8th ed., M. E. Van Valkenburg (ed.,) Carmel, Ind.: SAMS Publishing Co., 1993.

A classic reference on the CT Fourier transform is Bracewell [1986].

14.2 Fourier Transforms and the Fast Fourier Transform

Alexander D. Poularikas

The Discrete Time Fourier Transform (DTFT)

The discrete time Fourier transform of a signal $\{f(n)\}$ is defined by

$$\mathcal{F}_{dt}\left\{f(n)\right\} \equiv F(\omega) \equiv F\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} f(n)e^{-j\omega n} \tag{14.19}$$

and its inverse discrete time Fourier transform (IDTFT) is give by

$$f(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(\omega)e^{j\omega n}d\omega \tag{14.20}$$

The amplitude and phase spectra are periodic with a period of 2π and thus the frequency range of any discrete signal is limited to the range $(-\pi,\pi]$ or $(0,2\pi]$.

Example 1
Find the DTFT of the sequence $f(n) = 0.8^n$ for $n = 0,1,2,3\ldots$

Solution
From (14.19), we write

$$F(\omega) = \sum_{n=0}^{\infty} 0.8^n e^{-j\omega n} = \sum_{n=0}^{\infty}\left(0.8e^{-j\omega}\right)^n = \frac{1}{1 - 0.8e^{-j\omega}} \tag{14.21}$$

$$\left|F(\omega)\right| = \frac{1}{\sqrt{1.64 - 1.6\cos\omega}};\ Arg\,F(\omega) = \tan^{-1}\left(\frac{0.8\sin\omega}{1 - 0.8\cos\omega}\right) \tag{14.22}$$

If we set $\omega = -\omega$ in the last two equations we find that the amplitude is an even function and the argument is an odd function.

Relationship to the Z-Transform

$$F(z)\Big|_{z=e^{j\omega}} = \sum_{n=-\infty}^{\infty} f(n)z^{-n}\Big|_{z=e^{j\omega}}$$

Properties

Table 14.6 tabulates the DTFT properties of discrete time sequences.

Fourier Transforms of Finite Time Sequences

The trancated Fourier transform of a sequence is given by

$$F_N(\omega) = \sum_{n=0}^{N-1} f(n)e^{-j\omega n} = \sum_{n=-\infty}^{\infty} f(n)w(n)e^{-j\omega n} = \frac{1}{2\pi}F(\omega) * W(\omega) \tag{14.23}$$

TABLE 14.6

Property	Time Domain	Frequency Domain
Linearity	$af_1(n) + bf_2(n)$	$aF_1(\omega)+bF_2(\omega)$
Time Shifting	$f(n - n_0)$	$e^{-j\omega n_0}F(\omega)$
Time Reversal	$f(-n)$	$F(-\omega)$
Convolution	$f_1(n)*f_2(n)$	$F_1(\omega)F_2(\omega)$
Frequency Shifting	$e^{j\omega_0 n}f(n)$	$F(\omega - \omega_0)$
Time Multiplication	$nf(n)$	$-z\left.\dfrac{dF(z)}{dz}\right\|_{z=e^{j\omega}}$
Modulation	$f(n)\cos\omega_0 n$	$\frac{1}{2}F(\omega - \omega_0) + \frac{1}{2}F(\omega + \omega_0)$
Correlation	$f_1(n)\bullet f_2(n)$	$F_1(\omega)F_2(-\omega)$
Parseval's Formula	$\sum_{n=-\infty}^{\infty}\|f(n)\|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\|F(\omega)\|^2 d\omega$	

where $w(n)$ is a window function that extends from $n = 0$ to $n = N - 1$. If the value of the sequence is unity for all n's, the window is known as the rectangular one. From (14.23) we observe that the trancation of a sequence results in a smoothered version of the exact spectrum.

Frequency Response of LTI Discrete Systems

A first order LTI discrete system is described by the difference equation

$$y(n) + a_1 y(n - 1) = b_0 x(n) + b_1 x(n - 1)$$

The DTFT of the above equation is given by

$$Y(\omega) + a_1 e^{-j\omega} Y(\omega) = b_0 X(\omega) + b_1 e^{-j\omega} X(\omega)$$

from which we write the system function

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} = \frac{b_0 + b_1 e^{-j\omega}}{1 + a_1 e^{-j\omega}}$$

To approximate the continuous time Fourier transform using the DTFT we follow the following steps:

1. Select the time interval T such that $F(\omega_c) \approx 0$ for all $|\omega_c| > \pi/T$. ω_c designates the frequency of a continuous time function.
2. Sample $f(t)$ at times nT to obtain $f(nT)$.
3. Compute the DFT using the sequence $\{Tf(nT)\}$.
4. The resulting approximation is then $F(\omega_c) \approx F(\omega)$ for $-\pi/T < \omega_c < \pi/T$.

The Disrete Fourier Transform

One of the methods, and one that is used extensively, calls for replacing continuous Fourier transforms by an equivalent *discrete Fourier transform* (DFT) and then evaluating the DFT using the discrete data. However, evaluating a DFT with 512 samples (a small number in most cases) requires more than 1.5×10^6 mathematical operations. It was the development of the *fast Fourier transform* (**FFT**), a computational technique that reduces

the number of mathematical operations in the evaluation of the DFT to $N \log_2 (N)$ (approximately 2.5×10^4 operations for the 512-point case mentioned above), that makes DFT an extremely useful tool in most all fields of science and engineering.

A data sequence is available only with a finite time window from $n = 0$ to $n = N - 1$. The transform is discretized for N values by taking samples at the frequencies $2\pi/NT$, where T is the time interval between sample points. Hence, we define the DFT of a sequence of N samples for $0 \le k \le N-1$ by the relation

$$\begin{aligned} F(k\Omega) \doteq \mathcal{F}_d\{f(nT)\} &= T\sum_{n=0}^{N-1} f(nT)e^{-j2\pi nkT/NT} \\ &= T\sum_{n=0}^{N-1} f(nT)e^{-j\Omega Tnk} \quad n = 0, 1, \ldots, N-1 \end{aligned} \tag{14.24}$$

where N = number of sample values, T = sampling time interval, $(N-1)T$ = signal length, $f(nT)$ = sampled form of $f(t)$ at points nT, $\Omega = (2\pi/T)1/N = \omega_s/N$ = frequency sampling interval, $e^{-i\Omega T} = N$th principal root of unity, and $j = \sqrt{-1}$. The inverse DFT is given by

$$\begin{aligned} f(nT) \doteq \mathcal{F}_d^{-1}\{F(k\Omega)\} &= \frac{1}{NT}\sum_{k=0}^{N-1} F(k\Omega)e^{j2\pi nkT/NT} \\ &= \frac{1}{NT}\sum_{k=0}^{N-1} F(k\Omega)e^{i\Omega Tnk} \end{aligned} \tag{14.25}$$

The sequence $f(nT)$ can be viewed as representing N consecutive samples $f(n)$ of the continuous signal, while the sequence $F(k\Omega)$ can be considered as representing N consecutive samples $F(k)$ in the frequency domain. Therefore, Eqs. (14.24) and (14.25) take the compact form

$$\begin{aligned} F(k) \doteq \mathcal{F}_d\{f(n)\} &= \sum_{n=0}^{N-1} f(n)e^{-j2\pi nk/N} \\ &= \sum_{n=0}^{N-1} f(n)W_N^{nk} \quad k = 0, \ldots, N-1 \end{aligned} \tag{14.26}$$

$$\begin{aligned} f(n) \doteq \mathcal{F}_d^{-1}\{F(k)\} &= \frac{1}{N}\sum_{k=0}^{N-1} F(k)e^{j2\pi nk/N} \\ &= \sum_{k=0}^{N-1} F(k)W_N^{-nk} \quad k = 0, \ldots, N-1 \end{aligned} \tag{14.27}$$

where

$$W_N = e^{-j2\pi/N} \quad j = \sqrt{-1}$$

An important property of the DFT is that $f(n)$ and $F(k)$ are uniquely related by the transform pair (14.26) and (14.27).

We observe that the functions W^{kn} are N-periodic; that is,

$$W_N^{kn} = W_N^{k(n+N)} \qquad k, n = 1, \pm 1, \pm 2, \ldots \tag{14.28}$$

As a consequence, the sequences $f(n)$ and $F(k)$ as defined by (14.26) and (14.27) are also N-periodic.

It is generally convenient to adopt the convention

$$\{f(n)\} \leftrightarrow \{F(k)\} \tag{14.29}$$

to represent the transform pair (14.26) and (14.27).

Properties of the DFT

A detailed discussion of the properties of DFT can be found in the cited references at the end of this section. In what follows we consider a few of these properties that are of value for the development of the FFT.

1. *Linearity:*

$$\{af(n) + by(n)\} \leftrightarrow \{aF(k)\} + \{bY(k)\} \tag{14.30}$$

2. *Complex conjugate:* If $f(n)$ is real, $N/2$ is an integer and $\{f(n)\} \leftrightarrow \{F(k)\}$, then

$$F\left(\frac{N}{2} + l\right) = F^*\left(\frac{N}{2} - l\right) \quad l = 0, 1, \ldots, \frac{N}{2} \tag{14.31}$$

   where $F^*(k)$ denotes the complex conjugate of $F(k)$. The preceding identity shows the folding property of the DFT.

3. *Reversal:*

$$\{f(-n)\} \leftrightarrow \{F(-k)\} \tag{14.32}$$

4. *Time shifting:*

$$\{f(n + l)\} \leftrightarrow \{W^{-lk} F(k)\} \tag{14.33}$$

5. *Convolution of real sequences:* If

$$y(n) = \frac{1}{N} \sum_{l=0}^{N-1} f(l)h(n - l) \quad n = 0, 1, \ldots, N - 1 \tag{14.34}$$

   then

$$\{y(n)\} \leftrightarrow \{F(k)\ H(k)\} \tag{14.35}$$

6. *Correlation of real sequences:* If

$$y(n) = \frac{1}{N} \sum_{l=0}^{N-1} f(l)h(n + l) \quad n = 0, 1, \ldots, N - 1 \tag{14.36}$$

   then

$$\{y(n)\} \leftrightarrow \{F(r)\ H^*(k)\} \tag{14.37}$$

7. *Symmetry:*

$$\left\{\frac{1}{N}F(n)\right\} \leftrightarrow \{f(-k)\} \tag{14.38}$$

8. Parseval's theorem:

$$\sum_{n=0}^{N-1}\left|f(n)\right|^2 = \frac{1}{N}\sum_{k=0}^{N-1}\left|F(k)\right|^2 \tag{14.39}$$

where $|F(k)| = F(k)\ F^*(k)$.

Example 1

Verify Parseval's theorem for the sequence $\{f(n)\} = \{1, 2, -1, 3\}$.

Solution. With the help of (14.26) we obtain

$$\begin{aligned} F(k)\Big|_{k=0} &= F(0) = \sum_{n=0}^{3} f(n)e^{-j(2\pi/4)kn}\Big|_{k=0} \\ &= (1e^{-j(\pi/2)0\cdot 0} + 2e^{-j(\pi/2)0\cdot 1} - e^{-j(\pi/2)0\cdot 2} + 3e^{-j(\pi/2)0\cdot 3}) \\ &= 5 \end{aligned}$$

Similarly, we find

$$F(1) = 2 + j \quad F(2) = -5 \quad F(3) = 2 - j$$

Introducing these values in (14.39) we obtain

$$1^2 + 2^2 + (-1)^2 + 3^2 = 1/4[5^2 + (2 + j)(2 - j) + 5^2 + (2 - j)(2 + j)] \quad \text{or} \quad 15 = 60/4$$

which is an identity, as it should have been.

Relation between DFT and Fourier Transform

The sampled form of a continuous function $f(t)$ can be represented by N equally spaced sampled values $f(n)$ such that

$$f(n) = f(nT) \quad n = 0,1,\ldots, N-1 \tag{14.40}$$

where T is the sampling interval. The length of the continuous function is $L = NT$, where $f(N) = f(0)$.

We denote the sampled version of $f(t)$ by $f_s(t)$, which may be represented by a sequence of impulses. Mathematically it is represented by the expression

$$f_s(t) = \sum_{n=0}^{N-1}[Tf(n)]\delta(t - nT) \tag{14.41}$$

where $\delta(t)$ is the Dirac or impulse function.

Taking the Fourier transform of $f_s(t)$ in (14.41) we obtain

$$\begin{aligned} F_s(\omega) &= T\int_{\infty}^{\infty}\sum_{n=0}^{N-1} f(n)\delta(t - nT)e^{-j\omega t}\, dt \\ &= T\sum_{n=0}^{N-1} f(n)\int_{\infty}^{\infty}\delta(t - nT)e^{-j\omega t}dt \\ &= T\sum_{n=0}^{N-1} f(n)e^{-j\omega nT} \end{aligned} \tag{14.42}$$

Equation (14.42) yields $F_s(\omega)$ for all values of ω. However, if we are only interested in the values of $F_s(\omega)$ at a set of discrete equidistant points, then (14.42) is expressed in the form [see also (14.24)]

$$F_s(k\Omega) = T\sum_{n=0}^{N-1} f(n)e^{-jkn\Omega T} \qquad k = 0, \pm 1, \pm 2, \ldots, \pm N/2 \tag{14.43}$$

where $\Omega = 2\pi/L = 2\pi/NT$. Therefore, comparing (14.26) and (14.43) we observe that we can find $F(\omega)$from $F_s(\omega)$ using the relation

$$F(k) = F_s(\omega)\Big|_{\omega = k\Omega} \tag{14.44}$$

Power, Amplitude, and Phase Spectra

If $f(t)$ represents voltage or current waveform supplying a load of 1 Ω, the left-hand side of Parseval's theorem (14.39) represents the power dissipated in the 1-Ω resistor. Therefore, the right-hand side represents the power contributed by each harmonic of the spectrum. Thus the DFT **power spectrum** is defined as

$$P(k) = F(k)F^*(k) = |F(k)|^2 \qquad k = 0, 1, \ldots, N - 1 \tag{14.45}$$

For real $f(n)$ there are only $(N/2 + 1)$ independent DFT spectral points as the complex conjugate property shows (14.31). Hence we write

$$P(k) = |F(k)|^2 \qquad k = 0, 1, \ldots, N/2 \tag{14.46}$$

The *amplitude spectrum* is readily found from that of a power spectrum, and it is defined as

$$A(k) = |F(k)| \qquad k = 0, 1, \ldots, N - 1 \tag{14.47}$$

The power and amplitude spectra are invariant with respect to shifts of the data sequence $\{f(n)\}$.
The **phase spectrum** of a sequence $\{f(n)\}$ is defined as

$$\phi_f(k) = \tan^{-1}\frac{\mathrm{Im}\{F(k)\}}{\mathrm{Re}\{F(k)\}} \qquad k = 0, 1, \ldots, N - 1 \tag{14.48}$$

As in the case of the power spectrum, only $(N/2 + 1)$ of the DFT phase spectral points are independent for real $\{f(n)\}$. For a real sequence $\{f(n)\}$ the power spectrum is an *even function* about the point $k = N/2$ and the phase spectrum is an *odd function* about the point $k = N/2$.

Observations

1. The frequency spacing $\Delta\omega$ between coefficients is

$$\Delta\omega = \Omega = \frac{2\pi}{NT} = \frac{\omega_s}{N} \qquad \text{or} \qquad \Delta f = \frac{1}{NT} = \frac{f_s}{N} = \frac{1}{T_0} \tag{14.49}$$

2. The reciprocal of the record length defines the frequency resolution.
3. If the number of samples N is fixed and the sampling time is increased, the record length and the precision of frequency resolution is increased. When the sampling time is decreased, the opposite is true.
4. If the record length is fixed and the sampling time is decreased (N increases), the resolution stays the same and the computed accuracy of $F(n\Omega)$ increases.
5. If the record length is fixed and the sampling time is increased (N decreases), the resolution stays the same and the computed accuracy of $F(n\Omega)$ decreases.

Data Windowing

To produce more accurate frequency spectra it is recommended that the data are weighted by a **window** function. Hence, the new data set will be of the form$\{f(n)\ w(n)\}$. The following are the most commonly used windows:

1. Triangle (Fejer, Bartlet) window:

$$w(n) = \begin{cases} \dfrac{n}{N/2} & n = 0, 1, \ldots, \dfrac{N}{2} \\ w(N-n) & n = \dfrac{N}{2}, \ldots, N-1 \end{cases} \tag{14.50}$$

2. $\text{Cos}^{\alpha}(x)$ windows:

$$\begin{aligned} w(n) &= \sin^2\left(\frac{n}{N}\pi\right) \\ &= 0.5\left[1 - \cos\left(\frac{2n}{N}\pi\right)\right] \qquad n = 0, 1 \ldots, N-1 \qquad \alpha = 2 \end{aligned} \tag{14.51}$$

   This window is also called the raised cosine or Hamming window.
3. Hamming window:

$$w(n) = 0.54 - 0.46\ \cos\left(\frac{2\pi}{N}n\right) \qquad n = 0, 1, \ldots, N-1 \tag{14.52}$$

4. Blackman window:

$$w(n) = \sum_{m=0}^{K} (-1)^m a_m \cos\left(2\pi m \frac{n}{N}\right) \qquad n = 0, 1, \ldots, N-1 \qquad K \le \frac{N}{2} \tag{14.53}$$

   for $K = 2$, $a_0 = 0.42$, $a_1 = 0.50$, and $a_2 = 0.08$.

TABLE 14.7

No. of Terms in (13.30)	Maximum Sidelobe, dB	Parameter Values			
		a_0	a_1	a_2	a_3
3	–70.83	0.42323	0.49755	0.07922	
3	–62.05	0.44959	0.49364	0.05677	
4	–92	0.35875	0.48829	0.14128	0.01168
4	–74.39	0.40217	0.49703	0.09892	0.00188

5. Blackman-Harris window. Harris used a gradient search technique to find three- and four-term expansion of (14.53) that either minimized the maximum sidelobe level for fixed mainlobe width, or traded mainlobe width versus minimum sidelobe level (see Table 14.7)
6. Centered Gaussian window:

$$w(n) = \exp\left[-\frac{1}{2}\alpha\left(\frac{n}{N/2}\right)^2\right] \qquad 0 \le |n| \le \frac{N}{2} \qquad \alpha = 2, 3, \ldots \tag{14.54}$$

   As α increases, the mainlobe of the frequency spectrum becomes broader and the sidelobe peaks become lower.
7. Centered Kaiser-Bessel window:

$$w(n) = \frac{I_0\left[\pi a\sqrt{1.0 - \left(\frac{n}{N/2}\right)^2}\right]}{I_0(\pi\alpha)} \qquad 0 \le |n| \le \frac{N}{2} \tag{14.55}$$

   where

$$\begin{aligned} I_0(x) &= \text{zero-order modified Bessel function} \\ &= \sum_{k=0}^{\infty}\left(\frac{(x/2)^k}{k!}\right)^2 \\ k_! &= 1 \times 2 \times 3 \times \cdots \times k \\ \alpha &= 2,\ 2.5,\ 3 \qquad \text{(typical values)} \end{aligned} \tag{14.56}$$

Fast Fourier Transform

One of the approaches to speed the computation of the DFT of a sequence is the *decimation-in-time* method. This approach is one of breaking the N-point transform into two $(N/2)$-point transforms, breaking each $(N/2)$-point transform into two $(N/4)$-point transforms, and continuing the above process until we obtain the two-point transform. We start with the DFT expression and factor it into two DFTs of length $N/2$:

$$\begin{aligned} F(k) = &\sum_{n=0}^{N-2} f(n)W_N^{kn} \qquad n \text{ even} \\ &+ \sum_{n=1}^{N-1} f(n)W_N^{kn} \qquad n \text{ odd} \end{aligned} \tag{14.57}$$

Letting $n = 2m$ in the first sum and $n = 2m + 1$ in the second, (14.57) becomes

$$F(k) = \sum_{m=0}^{(N/2)-1} f(2m)W_N^{2mk} + \sum_{m=0}^{(N/2)-1} f(2m+1)W_N^{(2m+1)k} \tag{14.58}$$

However, because of the identities

$$W_N^{2mk} = (W_N^2)^{mk} = e^{-j(2\pi/N)2mk} = e^{-j(4\pi mk/N)} = W_{N/2}^{mk} \tag{14.59}$$

and the substitution $f(2m) = f_1(m)$ and $f(2m + 1) = f_2(m)$, $m = 0, 1, \ldots, N/2 - 1$, takes the form

$$\begin{aligned} F(k) = {} & \sum_{m=0}^{(N/2)-1} f_1(m)W_{N/2}^{mk} && \frac{N}{2}-\text{point DFT of even-indexed sequence} \\ & + W_N^k \sum_{m=0}^{(N/2)-1} f_2(m)W_{N/2}^{mk} && \frac{N}{2}-\text{point DFT of odd-indexed sequence} \\ & && k = 0, \ldots, N/2 - 1 \end{aligned} \tag{14.60}$$

We can also write (14.60) in the form

$$\begin{aligned} F(k) &= F_1(k) + W_N^k F_2(k) && k = 0, 1, \ldots, N/2 - 1 \\ F\left(k + \frac{N}{2}\right) &= F_1(k) + W_N^{k+N/2} F_2(k) \\ &= F_1(k) - W_N^k F_2(k) && k = 0, 1, \ldots, N/2 - 1 \end{aligned} \tag{14.61}$$

where $W_N^{k+N/2} = -W_N^k$ and $W_{N/2}^{m(k+N/2)} = W_{N/2}^{mk}$. Since the DFT is periodic, $F_1(k) = F_1(k + N/2)$ and $F_2(k) = F_2(k + N/2)$.

We next apply the same procedure to each $N/2$ samples, where $f_{11}(m) = f_1(2m)$ and $f_{21}(m) = f_2(2m + 1)$, $m = 0,1, \ldots, (N/4) - 1$. Hence,

$$F_1(k) = \sum_{m=0}^{(N/4)-1} f_{11}(m)W_{N/4}^{mk} + W_N^{2k} \sum_{m=0}^{(N/4)-1} f_{21}(m)W_{N/4}^{mk} \qquad k = 0, 1, \ldots, N/4 - 1 \tag{14.62}$$

or

$$\begin{aligned} F_1(k) &= F_{11}(k) + W_N^{2k} F_{21}(k) \\ F_1\left(k + \frac{N}{4}\right) &= F_{11}(k) - W_N^{2k} F_{21}(k) \quad k = 0, 1, \ldots, N/4 - 1 \end{aligned} \tag{14.63}$$

Therefore, each one of the sequences f_1 and f_2 has been split into two DFTs of length $N/4$.

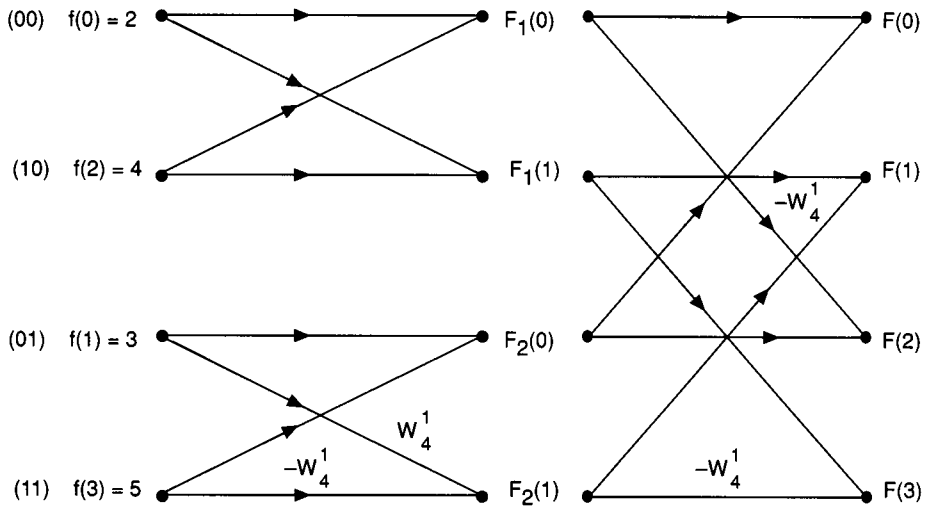


FIGURE 14.9 Illustration of Example 2.

Example 2

To find the FFT of the sequence {2, 3, 4, 5} we first bit reverse the position of the elements from their priority {00, 01, 10, 11} to {00, 10, 01, 11} position. The new sequence is {2, 4, 3, 5} (see also Fig. 14.9). Using (14.60) and (14.61) we obtain

$$F_1(0) = \sum_{m=0}^{1} f_1(m)W_2^{m0} = f_1(0)W_2^0 + f_1(1)W_2^0 = f(0)\cdot 1 + f(2)\cdot 1$$

$$F_1(1) = \sum_{m=0}^{1} f_1(m)W_2^{m\cdot 1} = f_1(0)W_2^{0\cdot 1} + f_1(1)W_2^1 = f(0) + f(2)(-j)$$

$$F_2(0) = W_4^0\sum_{m=0}^{1} f_2(m)W_2^{m\cdot 0} = f_2(0)W_2^0 + f_2(1)W_2^0 = f(1) + f(3)$$

$$F_2(1) = W_4^1\sum_{m=0}^{1} f_2(m)W_2^{m\cdot 1} = W_4^1\left[f(1)W_2^0 + f(3)W_2^1\right] = W_4^1 f(1) - W_4^1 f(3)$$

From (14.61) the output is

$$\begin{aligned} F(0) &= F_1(0) + W_4^0 F_2(0)\\ F(1) &= F_1(1) + W_4^1 F_2(1)\\ F(2) &= F_1(0) - W_4^0 F_2(0)\\ F(3) &= F_1(1) - W_4^1 F_2(1) \end{aligned}$$

Computation of the Inverse DFT

To find the inverse FFT using an FFT algorithm, we use the relation

$$f(n) = \frac{[FFT(F^*(k))]^*}{N} \tag{14.64}$$

TABLE 14.8 FFT Subroutine

```
      SUBROUTINE FOUR1 (DATA, NN, ISIGN)
            Replaces DATA by its discrete Fourier transform, if SIGN is input as 1; or replaces DATA by NN times its inverse
            discrete Fourier transform, if ISIGN is input as –1. DATA is a complex array of length NN or, equivalently, a real
            array of length 2*NN. NN must be an integer power of 2.
      REAL*8 WR, WI, WPR, WPI, WTEMP, THETA                    Double precision for the trigonometric recurrences.
      DIMENSION DATA (2*NN)
      N=2*NN
      J=1
      DO 11 I=1, N, 2                                          This is the bit-reversal section of the routine.
         IF (J.GT.I) THEN                                      Exchange the two complex numbers.
            TEMPR=DATA(J)
            TEMPI=DATA(J+1)
            DATA(J)=DATA(I)
            DATA(J+1)=DATA(I+1)
            DATA(I)=TEMPR
            DATA(I+1)=TEMPI
         ENDIF
         M=N/2
1        IF ((M.GE.2).AND. (J.GT.M)) THEN
                  J=J-M
                  M=M/2
            GO TO 1
            ENDIF
            J=J+M
11       CONTINUE
         MMAX=2                                               Here begins the Danielson-Lanczos section of the routine.
2           IF (N.GT.MMAX) THEN                               Outer loop executed log2 NN times.
            ISTEP=2*MMAX
            THETA=6.28318530717959D0/(ISIGN*MMAX)             Initialize for the trigonometric recurrence.
            WPR=-2.D0*DSIN(0.5D0*THETA)**2
            WPI=DSIN(THETA)
            WR=1.D0
            WI=0.D0
            DO 13 M=1,MMAX,2                                  Here are the two nested inner loops.
               DO 12 I=M,N,ISTEP
               J=I+MMAX                                       This is the Danielson-Lanczos formula:
               TEMPR=SNGL(WR)*DATA(J)-SNGL(WI)*DATA(J+1)
               TEMPI=SNGL(WR)*DATA(J+1)+SNGL(WI)*DATA(J)
               DATA(J)=DATA(I)-TEMPR
               DATA(J+1)=DATA(I+1)-TEMPI
               DATA(I)=DATA(I)+TEMPR
               DATA(I+1)=DATA(I+1)+TEMPI
12          CONTINUE
            WTEMP=WR                                          Trigonometric recurrence.
            WR=WR*WPR-WI*WPI+WR
            WI=WI*WPR+WTEMP*WPI+WI
13       CONTINUE
         MMAX=STEP
      GO TO 2
      ENDIF
      RETURN
      END
```

Source: ©1986 Numerical Recipes Software. From *Numerical Recipes: The Art of Scientific Computing*, published by Cambridge University Press. Used by permission.

For other transforms and their fast algorithms the reader should consult the references given at the end of this section.

Table 14.8 gives the FFT subroutine for fast implementation of the DFT of a finite sequence.

Defining Terms

FFT: A computational technique that reduces the number of mathematical operations in the evaluation of the discrete Fourier transform (DFT) to $N \log_2 N$.

Phase spectrum: All phases associated with the spectrum harmonics.

Power spectrum: A power contributed by each harmonic of the spectrum.

Window: Any appropriate function that multiplies the data with the intent to minimize the distortions of the Fourier spectra.

Related Topic

14.1 Fourier Transforms

References

A. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing*, New York: Springer-Verlag, 1975.

E. R. Blahut, *Fast Algorithms for Digital Signal Processing*, Reading, Mass.: Addison-Wesley, 1987.

E. O. Bringham, *The Fast Fourier Transform*, Englewood Cliffs, N.J.: Prentice-Hall, 1974.

F. D. Elliot, *Fast Transforms, Algorithms, Analysis, Applications*, New York: Academic Press, 1982.

H. J. Nussbaumer, *Fast Fourier Transform and Convolution Algorithms*, New York: Springer-Verlag, 1982.

A. D. Poularikas and S. Seely, *Signals and System*. 2nd ed., Melbourne, FL: Krieger Publishing, 1995.

Further Information

A historical overview of the fast Fourier transform can be found in J.W. Cooley, P.A.W. Lewis, and P.D. Welch, "Historical notes on the fast Fourier transform," *IEEE Trans. Audio Electroacoust.*, vol. AV-15, pp. 76–79, June 1967.

Fast algorithms appear frequently in the monthly magazine *Signal Processing*, published by The Institute of Electrical and Electronics Engineers.

14.3 Design and Implementation of Digital Filters

Bruce W. Bomar and L. Montgomery Smith

A *digital filter* is a linear, shift-invariant system for computing a **discrete output sequence** from a **discrete input sequence**. The input/output relationship is defined by the *convolution sum*

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)x(n-m)$$

where $x(n)$ is the input sequence, $y(n)$ is the output sequence, and $h(n)$ is the *impulse response* of the filter. The filter is often conveniently described in terms of its frequency characteristics that are given by the *transfer function* $H(e^{j\omega})$. The impulse response and transfer function are a Fourier transform pair:

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n} \qquad -\pi \le \omega \le \pi$$

$$h(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\omega})e^{j\omega n}\, d\omega \qquad -\infty \le n \le \infty$$

Closely related to the Fourier transform of $h(n)$ is the z-transform defined by

$$H(z) = \sum_{n=-\infty}^{\infty} h(n)z^{-n}$$

The Fourier transform is then the z-transform evaluated on the unit circle in the z-plane ($z = e^{j\omega}$). An important property of the z-transform is that $z^{-1}H(z)$ corresponds to $h(n-1)$, so z^{-1} represents a one-sample delay, termed a *unit delay*.

In this section, attention will be restricted to *frequency-selective* filters. These filters are intended to pass frequency components of the input sequence in a given band of the spectrum while blocking the rest. Typical frequency-selective filter types are *low-pass, high-pass, bandpass,* and *band-reject*. Other special-purpose filters exist, but their design is an advanced topic that will not be addressed here. In addition, special attention is given to *causal* filters, that is, those for which the impulse response is identically zero for negative n and thus can be realized in real time. Digital filters are further separated into two classes depending on whether the impulse response contains a finite or infinite number of nonzero terms.

Finite Impulse Response Filter Design

The objective of **finite impulse response (FIR) filter** design is to determine $N + 1$ coefficients

$$h(0),\ h(1),\ \ldots,\ h(N)$$

so that the transfer function $H(e^{j\omega})$ approximates a desired frequency characteristic $H_d(e^{j\omega})$. All other impulse response coefficients are zero. An important property of FIR filters for practical applications is that they can be designed to be *linear phase*, that is, the transfer function has the form

$$H(e^{j\omega}) = A(e^{j\omega})e^{-j\omega N/2}$$

where the amplitude $A(e^{j\omega})$ is a real function of frequency. The desired transfer function can be similarly written

$$H_d(e^{j\omega}) = A_d(e^{j\omega})e^{-j\omega N/2}$$

where $A_d(e^{j\omega})$ describes the amplitude of the desired frequency-selective characteristics. For example, the amplitude frequency characteristics of an ideal low-pass filter are given by

$$A_d(e^{j\omega}) = \begin{cases} 1 & \text{for } |\omega| \le \omega_c \\ 0 & \text{otherwise} \end{cases}$$

where ω_c is the *cutoff frequency* of the filter.

A linear phase characteristic ensures that a filter has a constant group delay independent of frequency. Thus, all frequency components in the signal are delayed by the same amount, and the only signal distortion introduced is that imposed by the filter's frequency-selective characteristics. Since a FIR filter can only approximate a desired frequency-selective characteristic, some measures of the accuracy of approximation are needed to describe the quality of the design. These are the *passband ripple* δ_p, the *stopband attenuation* δ_s, and the *transition bandwidth* $\Delta\omega$. These quantities are illustrated in Fig. 14.10 for a prototype low-pass filter. The passband ripple gives the maximum deviation from the desired amplitude (typically unity) in the region where the input signal spectral components are desired to be passed unattenuated. The stopband attenuation gives the maximum deviation from zero in the region where the input signal spectral components are desired to be blocked. The transition bandwidth gives the width of the spectral region in which the frequency characteristics

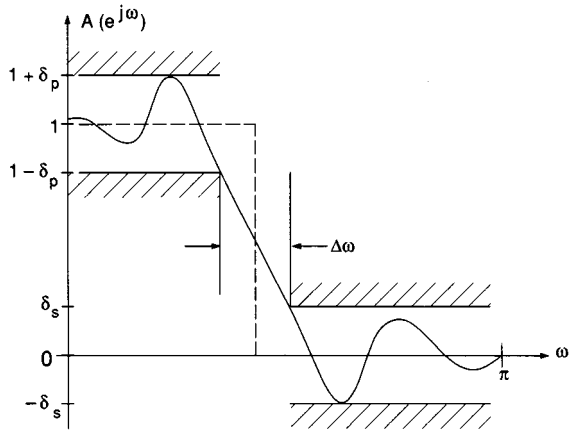


FIGURE 14.10 Amplitude frequency characteristics of a FIR low-pass filter showing definitions of passband ripple δ_p, stopband attenuation δ_s, and transition bandwidth $\Delta\omega$.

of the transfer function change from the passband to the stopband values. Often, the passband ripple and stopband attenuation are specified in decibels, in which case their values are related to the quantities δ_p and δ_s by

$$\text{Passband ripple (dB)} = P = -20 \log_{10} (1 - \delta_p)$$

$$\text{Stopband attenuation (dB)} = S = -20 \log_{10} \delta_s$$

FIR Filter Design by Windowing

The windowing design method is a computationally efficient technique for producing nonoptimal filters. Filters designed in this manner have equal passband ripple and stopband attenuation:

$$\delta_p = \delta_s = \delta$$

The method begins by finding the impulse response of the desired filter from

$$h_d(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} A_d(e^{j\omega}) e^{j\omega (n - N/2)} d\omega$$

For ideal low-pass, high-pass, bandpass, and band-reject frequency-selective filters, the integral can be solved in closed form. The impulse response of the filter is then found by multiplying this ideal impulse response with a window $w(n)$ that is identically zero for $n < 0$ and for $n > N$:

$$h(n) = h_d(n) w(n) \quad n = 0, 1, \ldots, N$$

Some commonly used windows are defined as follows:

1. Rectangular (truncation)

$$w(n) = \begin{cases} 1 & \text{for } 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

2. Hamming

$$w(n) = \begin{cases} 0.54 - 0.46 \cos \dfrac{2\pi n}{N} & \text{for } 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

3. Kaiser

$$w(n) = \begin{cases} \dfrac{I_0\left(\beta\sqrt{1 - [(2n - N)/N]^2}\right)}{I_0(\beta)} & \text{for } 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

In general, windows that slowly taper the impulse response to zero result in lower passband ripple and a wider transition bandwidth. Other windows (e.g., Hamming, Blackman) are also sometimes used but not as often as those shown above.

Of particular note is the Kaiser window where $I_0(.)$ is the 0th-order modified Bessel function of the first kind and β is a shape parameter. The proper choice of N and β allows the designer to meet given passband ripple/stopband attenuation and transition bandwidth specifications. Specifically, using S, the stopband attenuation in dB, the filter order must satisfy

$$N = \frac{S - 8}{2.285\Delta\omega}$$

Then, the required value of the shape parameter is given by

$$\beta = \begin{cases} 0 & \text{for } S < 21 \\ 0.5842(S - 21)^{0.4} + 0.07886(S - 21) & \text{for } 21 \le S \le 50 \\ 0.1102(S - 8.7) & \text{for } S > 50 \end{cases}$$

As an example of this design technique, consider a low-pass filter with a cutoff frequency of $\omega_c = 0.4\pi$. The ideal impulse response for this filter is given by

$$h_d(n) = \frac{\sin[0.4\pi(n - N/2)]}{\pi(n - N/2)}$$

Choosing $N = 8$ and a Kaiser window with a shape parameter of $\beta = 0.5$ yields the following impulse response coefficients:

$$h(0) = h(8) = -0.07568267$$

$$h(1) = h(7) = -0.06236596$$

$$h(2) = h(6) = 0.09354892$$

$$h(3) = h(5) = 0.30273070$$

$$h(4) = 0.40000000$$

Design of Optimal FIR Filters

The accepted standard criterion for the design of optimal FIR filters is to minimize the maximum value of the error function

$$E(e^{j\omega}) = W_d(e^{j\omega})\left|A_d(e^{j\omega}) - A(e^{j\omega})\right|$$

over the full range of $-\pi \le \omega \le \pi$. $W_d(e^{j\omega})$ is a desired weighting function used to emphasize specifications in a given frequency band. The ratio of the deviation in any two bands is inversely proportional to the ratio of their respective weighting.

A consequence of this optimization criterion is that the frequency characteristics of optimal filters are *equiripple:* although the maximum deviation from the desired characteristic is minimized, it is reached several times in each band. Thus, the passband and stopband deviations oscillate about the desired values with equal amplitude in each band. Such approximations are frequently referred to as *minimax* or *Chebyshev* approximations. In contrast, the maximum deviations occur near the band edges for filters designed by windowing.

Equiripple FIR filters are usually designed using the *Parks-McClellan* computer program [Parks and Burrus, 1987], which uses the *Remez exchange algorithm* to determine iteratively the *extremal frequencies* at which the maximum deviations in the error function occur. A listing of this program along with a detailed description of its use is available in several references including Parks and Burrus [1987] and DSP Committee [1979]. The program is executed by specifying as inputs the desired band edges, gain for each band (usually 0 or 1), band weighting, and FIR length. If the resulting filter has too much ripple in some bands, those bands can be weighted more heavily and the filter redesigned. Details on this design procedure are discussed in Rabiner [1973], along with approximate design relationships which aid in selecting the filter length needed to meet a given set of specifications.

Although we have focused attention on the design of frequency-selective filters, other types of FIR filters exist. For example, the Parks-McClellan program will also design linear-phase FIR filters for differentiating broadband signals and for approximating the Hilbert transform of such signals. A simple modification to this program permits arbitrary magnitude responses to be approximated with linear-phase filters. Other design techniques are available that permit the design of FIR filters which approximate an arbitrary complex response [Parks and Burrus, 1987; Chen and Parks, 1987], and, in cases where a nonlinear phase response is acceptable, design techniques are available that give a shorter impulse response length than would be required by a linear-phase design [Goldberg et al., 1981].

As an example of an equiripple filter design, an 8th-order low-pass filter with a passband $0 \le \omega \le 0.3\pi$, a stopband $0.5\pi \le \omega \le \pi$, and equal weighting for each band was designed. The impulse response coefficients generated by the Parks-McClellan program were as follows:

$$h(0) = h(8) = -0.06367859$$

$$h(1) = h(7) = -0.06912276$$

$$h(2) = h(6) = 0.10104360$$

$$h(3) = h(5) = 0.28574990$$

$$h(4) = 0.41073000$$

These values can be compared to those for the similarly specified filter designed in the previous subsection using the windowing method.

Infinite Impulse Response Filter Design

An **infinite impulse response (IIR) digital filter** requires less computation to implement than a FIR digital filter with a corresponding frequency response. However, IIR filters cannot generally achieve a perfect linear-phase response and are more susceptible to **finite wordlength effects.**

Techniques for the design of IIR analog filters are well established. For this reason, the most important class of IIR digital filter design techniques is based on forcing a digital filter to behave like a reference analog filter. This can be done in several different ways. For example, if the analog filter impulse response is $h_a(t)$ and the digital filter impulse response is $h(n)$, then it is possible to make $h(n) = h_a(nT)$, where T is the sample spacing of the digital filter. Such designs are referred to as *impulse-invariant* [Parks and Burrus, 1987]. Likewise, if $g_a(t)$ is the unit step response of the analog filter and $g(n)$ is the unit step response of the digital filter, it is possible to make $g(n) = g_a(nT)$, which gives a *step-invariant* design [Parks and Burrus, 1987].

The step-invariant and impulse-invariant techniques perform a time domain matching of the analog and digital filters but can produce aliasing in the frequency domain. For frequency-selective filters it is better to attempt matching frequency responses. This task is complicated by the fact that the analog filter response is defined for an infinite range of frequencies ($\Omega = 0$ to ∞), while the digital filter response is defined for a finite range of frequencies ($\omega = 0$ to π). Therefore, a method for mapping the infinite range of analog frequencies Ω into the finite range from $\omega = 0$ to π, termed the *bilinear transform,* is employed.

Bilinear Transform Design of IIR Filters

Let $H_a(s)$ be the Laplace transform transfer function of an analog filter with frequency response $H_a(j\,\Omega)$. The bilinear transform method obtains the digital filter transfer function $H(z)$ from $H_a(s)$ using the substitution

$$s = \frac{2(1 - z^{-1})}{T(1 + z^{-1})}$$

That is,

$$H(z) = H_a(s)\big|_{s = \frac{2}{T}\,\frac{1 - z^{-1}}{1 + z^{-1}}}$$

This maps analog frequency Ω to digital frequency ω according to

$$\omega = 2\tan^{-1}\frac{\Omega T}{2}$$

thereby warping the frequency response $H_a(j\,\Omega)$ and forcing it to lie between 0 and π for $H(e^{j\omega})$. Therefore, to obtain a digital filter with a cutoff frequency of ω_c it is necessary to design an analog filter with cutoff frequency

$$\Omega_c = \frac{2}{T}\tan\frac{\omega_c}{2}$$

This process is referred to as *prewarping* the analog filter frequency response to compensate for the warping of the bilinear transform. Applying the bilinear transform substitution to this analog filter will then give a digital filter that has the desired cutoff frequency.

Analog filters and hence IIR digital filters are typically specified in a slightly different fashion than FIR filters. Figure 14.11 illustrates how analog and IIR digital filters are usually specified. Notice by comparison to Fig. 14.10 that the passband ripple in this case never goes above unity, whereas in the FIR case the passband ripple is specified about unity.

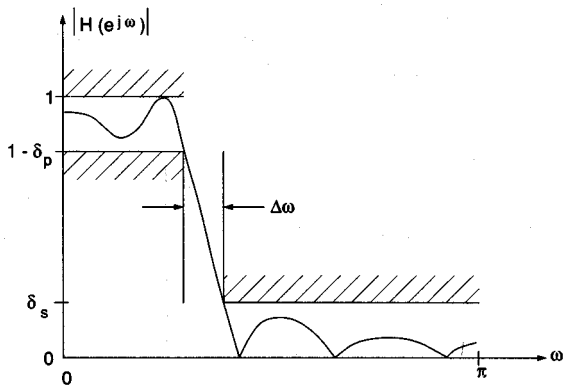


FIGURE 14.11 Frequency characteristics of an IIR digital low-pass filter showing definitions of passband ripple δ_p, stopband attenuation δ_s, and transition bandwith $\Delta\omega$.

Four basic types of analog filters are generally used to design digital filters: (1) Butterworth filters that are maximally flat in the passband and decrease monotonically outside the passband, (2) Chebyshev filters that are equiripple in the passband and decrease monotonically outside the passband, (3) inverse Chebyshev filters that are flat in the passband and equiripple in the stopband, and (4) elliptic filters that are equiripple in both the passband and stopband. Techniques for designing these analog filters are covered elsewhere [see, for example, Van Valkenberg, 1982] and will not be considered here.

To illustrate the design of an IIR digital filter using the bilinear transform, consider the design of a second-order Chebyshev low-pass filter with 0.5 dB of passband ripple and a cutoff frequency of $\omega_c = 0.4\,\pi$. The sample rate of the digital filter is to be 5 Hz, giving $T = 0.2$ s. To design this filter we first design an analog Chebyshev low-pass filter with a cutoff frequency of

$$\Omega_c = \frac{2}{0.2}\tan 0.2\pi = 7.2654 \text{ rad/s}$$

This filter has a transfer function

$$H(s) = \frac{0.9441}{1 + 0.1249s + 0.01249s^2}$$

Substituting

$$s = \frac{2}{0.2}\,\frac{z-1}{z+1}$$

gives

$$H(z) = \frac{0.2665(z+1)^2}{z^2 - 0.1406z + 0.2695}$$

Computer programs are available that accept specifications on a digital filter and carry out all steps required to design the filter, including prewarping frequencies, designing the analog filter, and performing the bilinear transform. Two such programs are given in the references [Parks and Burrus, 1987; Antoniou, 1979].

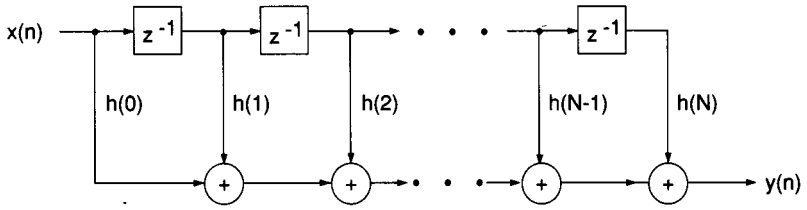


FIGURE 14.12 A direct-form implementation of a FIR filter.

Design of Other IIR Filters

For frequency-selective filters, the bilinear transformation of an elliptic analog filter provides an optimal equiripple design. However, if a design other than standard low-pass, high-pass, band-pass, or bandstop is needed or if it is desired to approximate an arbitrary magnitude or group delay characteristic, some other design technique is needed. Unlike the FIR case, there is no standard IIR design program for obtaining optimal approximations to an arbitrary response.

Four techniques that have been used for designing optimal equiripple IIR digital filters are [Parks and Burrus, 1987] (1) minimizing the L_p norm of the weighted difference between the desired and actual responses, (2) linear programming, (3) iteratively using the Remez exchange algorithm on the numerator and denominator of the transfer function, and (4) the differential correction algorithm. A computer program for implementing the first method is available in DSP Committee [1979].

Finite Impulse Response Filter Implementation

For FIR filters, the convolution sum represents a computable process, and so filters can be implemented by directly programming the arithmetic operations. Nevertheless, some options are available that may be preferable for a given processor architecture, and means for reducing computational loads exist. This section outlines some of these methods and presents schemes for FIR filter realization.

Direct Convolution Methods

The most obvious method for the implementation of FIR filters is to directly evaluate the sum of products in the convolution sum:

$$y(n) = h(0)x(n) + h(1)x(n-1) + \ldots + h(N)x(n-N)$$

The block diagram for this is shown in Fig. 14.12. This method involves storing the present and previous N values of the input, multiplying each sample by the corresponding impulse response coefficient, and summing the products to compute the output. This method is referred to as a *tapped delay line* structure.

A modification to this approach is suggested by writing the convolution as

$$y(n) = h(0)x(n) + \sum_{m=1}^{N} h(m)x(n-m)$$

In this approach, the output is computed by adding the product of $h(0)$ with the present input sample to a previously computed sum of products and updating a set of N sums of products with the present input sample value. The signal flow graph for this method is shown in Fig. 14.13.

FIR filters designed to have linear phase are usually obtained by enforcing the symmetry constraint

$$h(n) = h(N-n)$$

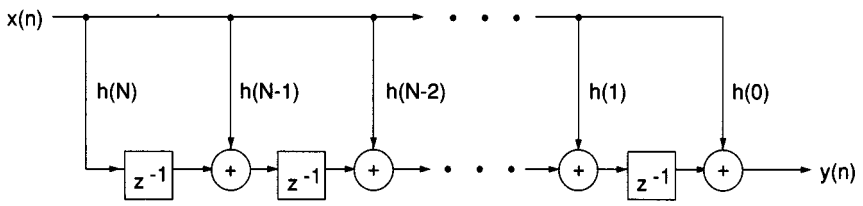


FIGURE 14.13 Another direct-form implementation of a FIR filter.

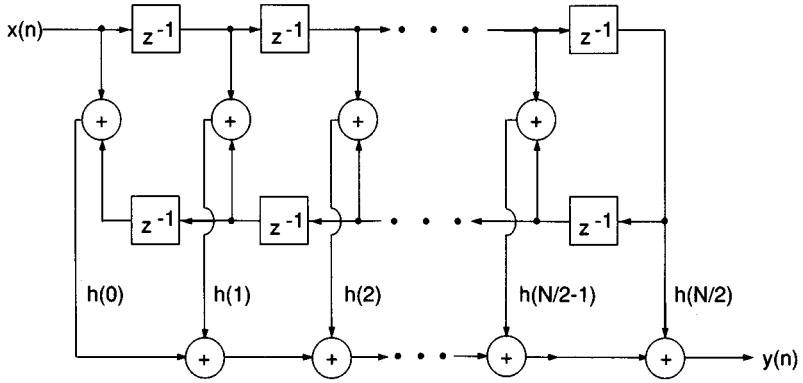


FIGURE 14.14 Implementation of a linear-phase FIR filter for even *N*.

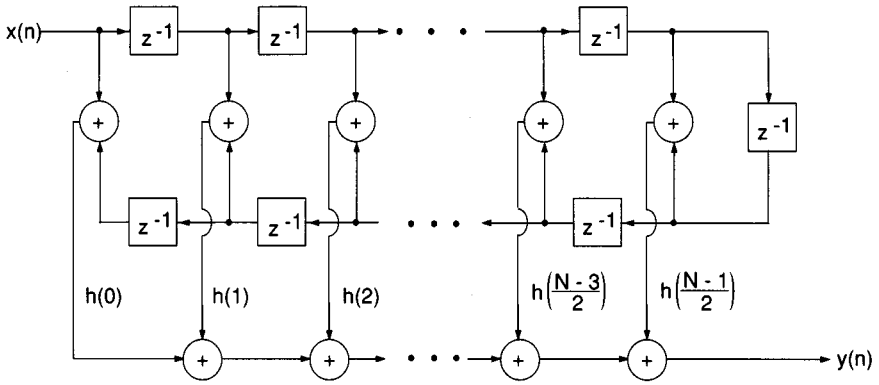


FIGURE 14.15 Implementation of a linear-phase FIR filter for odd *N*.

For these filters, the convolution sum can be written

$$y(n) = \begin{cases} \displaystyle\sum_{m=0}^{N/2-1} h(m)[x(n-m) + x(n+m-N)] + h\left(\frac{N}{2}\right)x\left(n-\frac{N}{2}\right) & N \text{ even} \\ \displaystyle\sum_{m=0}^{(N-1)/2} h(m)[x(n-m) + x(n+m-N)] & N \text{ odd} \end{cases}$$

Implementation of the filter according to these formulas reduces the number of multiplications by approximately a factor of 2 over direct-form methods. The block diagrams for these filter structures are shown in Figs. 14.14 and 14.15.

Implementation of FIR Filters Using the Discrete Fourier Transform

A method for implementing FIR filters that can have computational advantages over direct-form convolution involves processing the input data in blocks using the discrete Fourier transform (DFT) via the *overlap-save* method. The computational advantage arises primarily from use of the fast Fourier transform (FFT) algorithm (discussed in Section 14.2) to compute the DFTs of the individual data blocks. In this method, the input data sequence $\{x(n); -\infty < n < \infty\}$ is divided into L-point blocks

$$x_i(n) \quad 0 \le n \le L-1 \quad -\infty < i < \infty$$

where $L > N+1$, the length of the FIR filter. The L-point DFT of the impulse response is precomputed from

$$H[k] = \sum_{n=0}^{L-1} h(n)e^{-j2\pi kn/L} \qquad k = 0, 1, \ldots, L-1$$

where square brackets are used to distinguish the DFT from the continuous-frequency transfer function of the filter $H(e^{j\omega})$. Then, the DFT of each data block is computed according to

$$X_i[k] = \sum_{n=0}^{L-1} x_i(n)e^{-j2\pi kn/L} \qquad k = 0, 1, \ldots, L-1$$

These two complex sequences are multiplied together term by term to form the DFT of the output data block:

$$Y_i[k] = H[k]X_i[k] \quad k = 0, 1, \ldots, L-1$$

and the output data block is computed by the inverse DFT:

$$y_i(n) = \frac{1}{L}\sum_{k=0}^{L-1} Y_i[k]e^{j2\pi kn/L} \qquad n = 0, 1, \ldots, L-1$$

However, the output data block computed in this manner is the *circular convolution* of the impulse response of the filter and the input data block given by

$$y_i(n) = \sum_{m=0}^{N} h(m)x_i((n-m) \text{ modulo } L)$$

Thus, only the output samples from $n = N$ to $n = L-1$ are the same as those that would result from the convolution of the impulse response with the infinite-length data sequence $x(n)$. The first N data points are corrupted and must therefore be discarded. So that the output data sequence does not have N-point "gaps" in it, it is therefore necessary to *overlap* the data in adjacent input data blocks. In carrying out the processing, samples from block to block are *saved* so that the last N points of the ith data block $x_i(n)$ are the same as the first N points of the following data block $x_{i+1}(n)$. Each processed L-point data block thus produces $L-N$ output samples.

Another technique of block processing of data using DFTs is the *overlap-add* method in which $L-N$-point blocks of input data are zero-padded to L points, the resulting output blocks are overlapped by N points, and corresponding samples added together. This method requires more computation than the overlap-save method and is somewhat more difficult to program. Therefore, its usage is not as widespread as the overlap-save method.

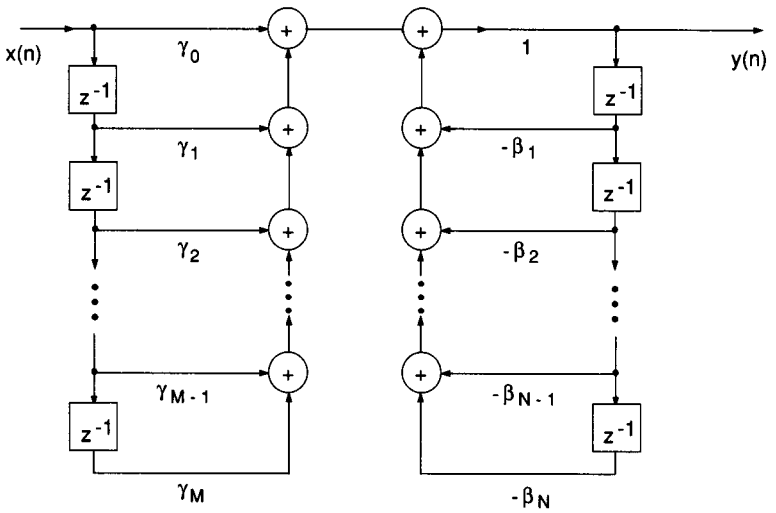


FIGURE 14.16 Direct-form I realization.

Infinite Impulse Response Filter Implementation

Direct-Form Realizations

For an IIR filter the convolution sum does not represent a computable process. Therefore, it is necessary to examine the general transfer function, which is given by

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\gamma_0 + \gamma_1 z^{-1} + \gamma_2 z^{-2} + \cdots + \gamma_M z^{-M}}{1 + \beta_1 z^{-1} + \beta_2 z^{-2} + \cdots + \beta_N z^{-N}}$$

where $Y(z)$ is the z-transform of the filter output $y(n)$ and $X(z)$ is the z-transform of the filter input $x(n)$. The unit-delay characteristic of z^{-1} then gives the following *difference equation* for implementing the filter:

$$y(n) = \gamma_0 x(n) + \gamma_1 x(n-1) + \ldots + \gamma_M x(n-M) - \beta_1 y(n-1) - \ldots - \beta_N y(n-N)$$

When calculating $y(0)$, the values of $y(-1)$, $y(-2)$, . . ., $y(-N)$ represent initial conditions on the filter. If the filter is started in an initially relaxed state, then these initial conditions are zero.

Figure 14.16 gives a block diagram realizing the filter's difference equation. This structure is referred to as the *direct-form I* realization. Notice that this block diagram can be separated into two parts, giving two cascaded networks, one of which realizes the filter zeros and the other the filter poles. The order of these networks can be reversed without changing the transfer function. This results in a structure where the two strings of delays are storing the same values, so a single string of delays of length $max(M, N)$ is sufficient, as shown in Fig. 14.17. The realization of Fig. 14.17 requires the minimum number of z^{-1} delay operations and is referred to as the *direct-form II* realization.

Cascade and Parallel Realizations

The transfer function of an IIR filter can always be factored into the product of second-order transfer functions as

$$H(z) = C\prod_{k=1}^{K} \frac{1 + a_{1k} z^{-1} + a_{2k} z^{-2}}{1 + b_{1k} z^{-1} + b_{2k} z^{-2}} = C\prod_{k=1}^{K} H_k(z)$$

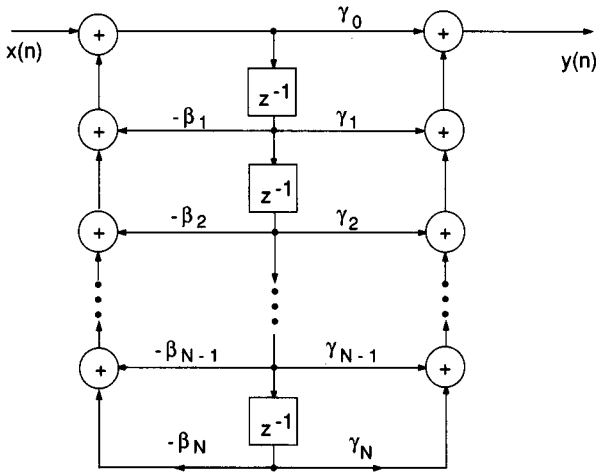


FIGURE 14.17 Direct-form II realization.

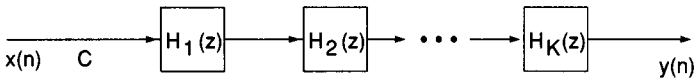


FIGURE 14.18 Cascade realization of an IIR filter.

where we have assumed $M = N$ in the original transfer function and where K is the largest integer contained in $(N + 1)/2$. If N is odd, the values of a_{2k} and b_{2k} in one term are zero. The realization corresponding to this transfer function factorization is shown in Fig. 14.18. Each second-order $H_k(z)$ term in this realization is referred to as a *biquad*. The digital filter design programs in Parks and Burrus [1987] and Antoniou [1979] give the filter transfer function in factored form.

If the transfer function of an IIR filter is written as a partial-fraction expansion and first-order sections with complex-conjugate poles are combined, $H(z)$ can be expressed in the form

$$H(z) = D + \sum_{k=1}^{K} \frac{\alpha_{0k} + \alpha_{1k} z^{-1}}{1 + b_{1k} z^{-1} + b_{2k} z^{-2}} = D + \sum_{k=1}^{K} G_k(z)$$

This results in the parallel realization of Fig. 14.19.

Finite Wordlength Effects in IIR Filters

Since practical digital filters must be implemented with limited-precision arithmetic, four types of finite wordlength effects result: (1) roundoff noise, (2) coefficient quantization error, (3) overflow oscillations, and (4) limit cycles. *Round-off noise* is that error in the filter output which results from rounding (or truncating) calculations within the filter. This error appears as low-level noise at the filter output. *Coefficient quantization error* refers to the deviation of a practical filter's frequency response from the ideal due to the filter's coefficients being represented with finite precision. The term *overflow oscillation*, sometimes also referred to as *adder overflow limit cycle*, refers to a high-level oscillation that can exist in an otherwise stable filter due to the nonlinearity associated with the overflow of internal filter calculations. A *limit cycle*, sometimes referred to as a *multiplier round-off limit cycle*, is a low-level oscillation that can exist in an otherwise stable filter as a result of the nonlinearity associated with rounding (or truncating) internal filter calculations. Overflow oscillations and limit cycles require recursion to exist and do not occur in nonrecursive FIR filters.

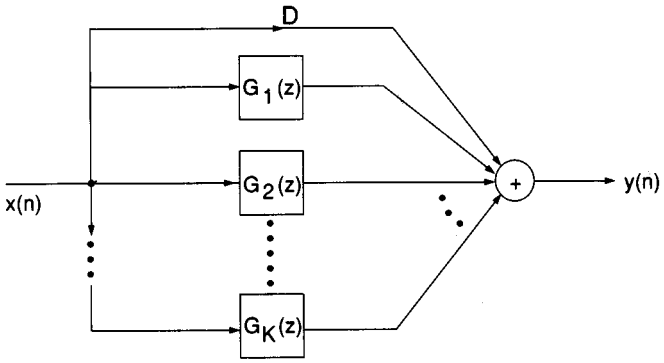


FIGURE 14.19 Parallel realization of an IIR filter.

The direct-form I and direct-form II IIR filter realizations generally have very poor performance in terms of all finite wordlength effects. Therefore, alternative realizations are usually employed. The most common alternatives are the cascade and parallel realizations where the direct-form II realization is used for each second-order section. By simply factoring or expanding the original transfer function, round-off noise and coefficient quantization error are significantly reduced. A further improvement is possible by implementing the cascade or parallel sections using *state-space* realizations [Roberts and Mullis, 1987]. The price paid for using state-space realizations is an increase in the computation required to implement each section. Another realization that has been used to reduce round-off noise and coefficient quantization error is the *lattice realization* [Parks and Burrus, 1987] which is usually formed directly from the unfactored and unexpanded transfer function.

Overflow oscillations can be prevented in several different ways. One technique is to employ floating-point arithmetic that renders overflow virtually impossible due to the large dynamic range which can be represented. In fixed-point arithmetic implementations it is possible to *scale* the calculations so that overflow is impossible [Roberts and Mullis, 1987], to use saturation arithmetic [Ritzerfeld, 1989], or to choose a realization for which overflow transients are guaranteed to decay to zero [Roberts and Mullis, 1987].

Limit cycles can exist in both fixed-point and floating-point digital filter implementations. Many techniques have been proposed for testing a realization for limit cycles and for bounding their amplitude when they do exist. In fixed-point realizations it is possible to prevent limit cycles by choosing a state-space realization for which any internal transient is guaranteed to decay to zero and then using magnitude truncation of internal calculations in place of rounding [Diniz and Antoniou, 1986].

Defining Terms

Discrete sequence: A set of values constituting a signal whose values are known only at distinct sampled points. Also called a digital signal.

Filter design: The process of determining the coefficients of a difference equation to meet a given frequency or time response characteristic.

Filter implementation: The numerical method or algorithm by which the output sequence is computed from the input sequence.

Finite impulse response (FIR) filter: A filter whose output in response to a unit impulse function is identically zero after a given bounded number of samples. A FIR filter is defined by a linear constant-coefficient difference equation in which the output depends only on the present and previous sample values of the input.

Finite wordlength effects: Perturbations of a digital filter output due to the use of finite precision arithmetic in implementing the filter calculations. Also called quantization effects.

Infinite impulse response (IIR) filter: A filter whose output in response to a unit impulse function remains nonzero for indefinitely many samples. An IIR filter is defined by a linear constant-coefficient difference equation in which the output depends on the present and previous samples of the input and on previously computed samples of the output.

Related Topics

8.1 Introduction • 17.3 Sensor Array Processing • 21.2 Example 1: Signal Analysis

References

A. Antoniou, *Digital Filters: Analysis, Design, and Applications,* 2nd ed., New York: McGraw-Hill, 1993.

X. Chen and T. W. Parks, "Design of FIR filters in the complex domain," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-35, pp. 144–153, 1987.

P. S. R. Diniz and A. Antoniou, "More economical state-space digital filter structures which are free of constant-input limit cycles," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 807–815, 1986.

DSP Committee, IEEE ASSP (eds.), *Programs for Digital Signal Processing*, New York: IEEE Press, 1979.

E. Goldberg, R. Kurshan, and D. Malah, "Design of finite impulse response digital filters with nonlinear phase response," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-29, pp. 1003–1010, 1981.

T. W. Parks and C. S. Burrus, *Digital Filter Design*, New York: Wiley, 1987.

L. R. Rabiner, "Approximate design relationships for low-pass FIR digital filters," *IEEE Trans. Audio Electroacoust.* vol. AU-21, pp. 456–460, 1973.

J. H. F. Ritzerfeld, "A condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures using saturation," *IEEE Trans. Circuits Syst.*, vol. CAS-36, pp. 1049–1057, 1989.

R. A. Roberts and C. T. Mullis, *Digital Signal Processing,* Reading, Mass.: Addison-Wesley, 1987.

M. E. Van Valkenberg, *Analog Filter Design*, New York: Holt, Rinehart, Winston, 1982.

Further Information

The monthly magazine *IEEE Transactions on Circuits and Systems II* routinely publishes articles on the design and implementation of digital filters. Finite wordlength effects are discussed in articles published in the April 1988 issue (pp. 365–374) and in the February 1992 issue (pp. 90–98).

Another journal containing articles on digital filters is the *IEEE Transactions on Signal Processing.* Overflow oscillations and limit cycles are discussed in the August 1978 issue (pp. 334–338).

The bimonthly journal *IEEE Transactions on Instrumentation and Measurement* also contains related information. The use of digital filters for integration and differentiation is discussed in the December 1990 issue (pp. 923–927).

14.4 Signal Restoration

James A. Cadzow

The concept of **signal restoration** has been applied with success to a number of fundamental problems found in interdisciplinary applications. In the typical signal restoration problem, there is given empirical data as symbolically designated by $\mathbf{x}$ that corresponds to corrupted measurements made on an underlying signal being monitored. By using attributes (properties) known to be or hypothesized as being possessed by the monitored signal, it is possible to devise signal restoration algorithms that effectively remove the corruption to give useful approximations of the underlying signal being monitored. Many of the more widely used restoration algorithms take the form of a sequence of *successive projections* as specified by

$$\mathbf{x}_n = P_1 P_2 \ldots P_m \mathbf{x}_{n-1}$$

In this algorithm, P_k designates the projection operator corresponding to the set of signals possessing attribute k while $\mathbf{x}_n$ denotes the enhanced signal at the nth iteration in which the initial point is $\mathbf{x}_o = \mathbf{x}$. In this section, we will give a historical perspective on the evolution of the **method of successive projections** as well as provide implementations of some of the most important projection operators.

Members of the class of signal restoration algorithms that are describable by a successive projection operator are primarily distinguished by the restrictions placed on the **attribute sets**. The earliest versions of these algorithms were highly restrictive and required that the attribute sets be closed subspaces. Unfortunately, this requirement severely limited the types of restoration problems that could be addressed. In recognition of this limitation, subsequent methods of successive projections algorithms eased the requirements on the attribute sets to that of being **closed convex sets** and eventually to the projection operators being closed mappings. This progression of less restrictive requirements significantly expands the types of signal processing problems that are amenable to signal restoration by successive projections.

Introduction

In a typical signal processing application, one is given empirically gathered data that arises from measurements made on a signal(s) characterizing a phenomenon under study. These measurements are invariably corrupted due to imperfections arising from the measurement instrumentation and environmental influences. To recover a reasonable facsimile of the original signal being monitored, it is generally necessary to process the measurement data in a manner that takes into account all information known about the monitored signal, the instrument dynamics, and the nature of the corruption. Although it is generally impossible to obtain a perfect recovery, remarkable approximations can be made in several important applications by employing the concept of *signal restoration* (or *signal recovery*). In signal restoration, *a priori* knowledge concerning the underlying signal's intrinsic nature may be effectively used to strip away the corruption in the measurement data. The philosophy behind this approach is to modify the measurement data to the smallest extent possible so that the modified data possesses the prescribed properties known (or hypothesized) as being possessed by the underlying signal(s). This modification then serves as a cleansing process whereby the measurement corruption is effectively removed.

Metric Space Formulation

To avail ourselves of the extensive analysis tools found in algebra, it is useful to formulate the basic signal restoration problem in a general metric space setting. This has the desirable effect of enabling us to treat a wide variety of applications in a single setting. With this in mind, the measurement signals are taken to lie in a *metric space* which is composed of a set of elements designated by X and a metric $d(\mathbf{x},\mathbf{y})$ that measures the distance between any two elements $\mathbf{x},\mathbf{y} \in X$. The elements of the set X are taken to possess a common form such as being composed of all $n \times 1$ real-valued vectors or all complex-valued continuous functions defined on a given interval. Moreover, the distance metric identifying this space must satisfy the axioms associated with a distance measure.[1] We will interchangeably refer to the elements of the metric space as vectors or signals. Depending on the nature of a particular application, the signals can take on such disparate forms as being composed of all real- (or complex-) valued $n \times 1$ vectors, $m \times n$ matrices, infinite-length sequences, continuous-time functions, and so forth.

Example 1. In digital signal processing, the two most commonly employed metric spaces correspond to the set of all real-valued $n \times 1$ vectors as designated by R^n and the set of all real-valued $m \times n$ matrices as designated by $R^{m \times n}$. The elements of a vector contained in R^n typically correspond to samples made of a one-dimensional signal. On the other hand, the elements of a matrix contained in $R^{m \times n}$ might correspond to the brightness levels of the pixels in a rectangular image or the entries of a data matrix formed from samples of a one-dimensional signal. It often happens in engineering applications that the data under analysis is complex valued. To treat such cases, we will have need to consider the set of all complex-valued $n \times 1$ tuples as designated by C^n and the set of all complex-valued $m \times n$ matrices as denoted by $C^{m \times n}$.

To complete the metric space description of spaces R^n and C^n, it is necessary to introduce a distance metric. The most commonly employed distance measure for either space is the Euclidean-induced metric

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n} |y(k) - x(k)|^2} \tag{14.65}$$

[1]The mapping $X \times X \rightarrow R$ designated by $d(\mathbf{x},\mathbf{y})$ is said to be a distance metric if it satisfies the four axioms: (1) $d(\mathbf{x},\mathbf{y})$ is nonnegative real-valued, (2) $d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$, (3) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$, and (4) $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$.

In a similar fashion, the *Frobenius norm* distance metric on the spaces $R^{m\times n}$ and $C^{m\times n}$ is commonly used where

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |Y(i,j) - X(i,j)|^2} \tag{14.66}$$

It is a simple matter to show that either of these measures satisfies the axioms of a distance metric.

Attribute Sets

As indicated previously, the concept of signal restoration is directed towards applications in which a measurement process introduces inevitable distortion in a signal that is being monitored. The fundamental philosophy underlying signal restoration is based on the hypothesis that the signal under observation is known or presumed to lie in a restricted *attribute set C*, *X*. This attribute set is composed of all signals in *X* that possess a prescribed set of attributes (or properties). The measurement process, however, results in a measurement signal that is perturbed outside this set. The set *C* employed in a signal restoration problem is often decomposable as the intersection of a finite number of basic attribute sets that are each describable by elementary attributes (or properties). We will designate these basic attribute sets as

$$C_k = \{\mathbf{x} \in X : \mathbf{x} \text{ has attribute } \mathcal{A}_k\} \quad \text{for } 1 \le k \le m \tag{14.67}$$

We will hereafter refer to the intersection of these basic attribute subsets as the composite attribute set since its constituent elements satisfy all the required attributes believed to be possessed by the signal being monitored.

$$C = C_1 \cap C_2 \cap \ldots \cap C_m \tag{14.68}$$

The usefulness of signal restoration is critically dependent on one's ability to identify all essential basic sets C_k describing the underlying information signal.

Once the composite attribute set *C* has been identified, the fundamental signal restoration problem entails finding a signal contained in this set that lies closest to the measurement signal **x** in the underlying distance metric sense. This gives rise to the following optimization problem:

$$\min_{\mathbf{y}\in C} d(\mathbf{x},\mathbf{y}) \tag{14.69}$$

In finding that signal contained in *C* which lies closest to **x**, we are in effect seeking the smallest perturbation of the measured signal **x** which results in a modified signal that possesses the hypothesized properties of the signal being monitored. Implicit in this approach is the assumption that the measurement process introduces the smallest distortion possible compatible with the measured data. Although this assumption is generally violated in most applications, it is reasonably accurate so that the solution to signal restoration problem (14.69) typically provides for a useful reconstruction of the signal being monitored.

Example 2. To illustrate the concept of signal properties, let us consider the autocorrelation sequence associated with the wide-sense stationary time series $\{x(n)\}$. This two-sided sequence is formally defined by

$$r_{xx}(n) = E\{x(n+m)\,\bar{x}(m)\} \quad \text{for } n = 0, \pm 1, \pm 2 \quad \ldots \tag{14.70}$$

where *E* designates the *expected value operator*. It is well known that this autocorrelation sequence satisfies the two attributes of being conjugate symmetric and having a nonnegative Fourier transform. The set of signals associated with these two attributes are then formally given by

$$C_{\text{cs}} = \{\text{set of sequences } \{x(n)\} \text{ such that } x(n) = \bar{x}(-n)\}$$

$$C_{\text{nnd}} = \{\text{set of sequences } \{x(n)\} \text{ such that } X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \ge 0\}$$

It can be shown that the set of conjugate-symmetric sequences C_{cs} is a closed vector space. Moreover, the set of nonnegative definite sequences C_{nnd} is a closed convex cone subset of C_{cs}.

Normed Vector Space Formulation

The primary objective of this paper is to present several algorithms which have been proposed for solving optimization problem (14.69) using an iterative approach. These algorithms take on a similar form and are distinguished by the algebraic restrictions placed on the underlying attribute subsets. To provide for a satisfactory mathematical characterization of the restoration problem and its solution, it is essential that we provide an algebraic structure to the underlying metric space. In particular, the ability to add vectors and multiply vectors by scalars provides a useful means for interpreting the measurement signal and more importantly considerably increases the arsenal of analysis tools at our disposal. Fortunately, little loss in generality is incurred by introducing these two algebraic operations since in most signal processing applications of interest there exists an intuitively obvious means for their implementation. For example, if the metric space is taken to be R^n (or C^n), then the sum of any two vectors $\mathbf{x},\mathbf{y} \in R^n$ (or C^n) has as its kth component $x(k) + y(k)$, while the scalar product $a\mathbf{x}$ has as its kth component $a\mathbf{x}(k)$ for $1 \le k \le n$.

One of the most important benefits of posing the signal restoration problem in a vector space setting is that of providing a widely invoked model for the measured signal (vector) as specified by

$$\mathbf{x} = \mathbf{s} + \mathbf{w} \tag{14.71}$$

where $\mathbf{s}$ designates the signal being monitored and $\mathbf{w}$ represents measurement error. From our previous discussion, it has been hypothesized that the monitored signal lies in each of the attribute sets so that $\mathbf{s} \in C_k$ for $1 \le k \le m$. To effectively recover $\mathbf{s}$ from $\mathbf{x}$ using signal restoration, it is tacitly assumed that the measurement error vector $\mathbf{w}$ has features which are distinctly different from those specified by these m attribute subsets.

The metric needed to measure the distance between any two elements in vector space X often follows in a natural fashion from the basic structure of the vector space. In particular, many of the vector spaces encountered in typical applications have an underlying vector norm measure.[2] A natural choice of a distance metric on a normed vector space is specified by

$$d(\mathbf{x},\mathbf{y}) = \| \mathrm{x} - \mathrm{y} \| \tag{14.72}$$

This distance metric is said to be induced by the norm defined on the underlying normed vector space X. A solution to the signal restoration problem in a normed vector space setting therefore requires solving the minimization problem

$$\min_{\mathbf{y} \in C} \| \mathbf{x} - \mathbf{y} \| \tag{14.73}$$

Example 3. As indicated previously, the vector spaces R^n and C^n play prominent roles in signal processing applications. The class of l_p norms as defined by

$$\| \mathbf{x} \|_p = \left[\sum_{k=1}^{n} | x(k) |^p \right]^{1/p} \tag{14.74}$$

are commonly used where the number p is restricted to lie in the interval $1 \le p < \infty$. Three of the most widely employed choices of the norm index are $p = 1$, 2, and ∞. In a similar manner, the l_p-induced norm for any matrix $A \in R^{m \times n}$ (or $C^{m \times n}$) is specified by

$$\| A \|_p = \max_{\mathbf{x}^*\mathbf{x}=1} \| A\mathbf{x} \|_p \tag{14.75}$$

[2]A mapping between X and R as designated by $\|\mathrm{x}\|$ is said to be a *norm* if it satisfies the axioms (1) $\|\mathrm{x}\| \ge 0$ for all $\mathrm{x} \in \mathrm{X}$ and $\|\mathrm{x}\| = 0$ if and only if $\mathrm{x} = 0$; (2) $\|\mathrm{x} + \mathrm{y}\| \le \|\mathrm{x}\| + \|\mathrm{y}\|$ for all $\mathrm{x}, \mathrm{y} \in \mathrm{X}$; (3) $\|\alpha\mathrm{x}\| = |\alpha| \, \|\mathrm{x}\|$ for all scalars α and all $\mathrm{x} \in \mathrm{X}$.

where the number p is again restricted to lie in the interval $1 \le p < \infty$.

Hilbert Space Setting

Much of the original research in signal restoration assumed that the underlying set of signals X is a complete inner product (i.e., a Hilbert) space. An inner product space is a vector space upon which there is defined an inner product $\langle \mathbf{x},\mathbf{y} \rangle$ which maps any two vectors $\mathbf{x},\mathbf{y} \in X$ into a scalar such that the axioms of an inner product axiom are satisfied.[3] This inner product induces a natural norm distance metric as specified by

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{14.76}$$

An inner product space is said to be complete if every Cauchy sequence contained in that space converges to an element of that space.[4] Our interest in Hilbert spaces arises in recognition of the fact that many important signal processing problems can be naturally formulated in such a setting.

A variety of algorithms have been proposed for iteratively solving minimization problem (14.73) when the signals are taken to lie in a Hilbert space. These algorithms differentiate themselves by the assumptions made on the characteristics of the attribute sets $C_1, C_2, \ldots, C_m$ in which the unknown signal being monitored is assumed to be contained. Of particular interest is the situation in which these sets are constrained to be made up exclusively of (1) closed subspaces or translated closed subspaces (i.e., linear varieties) and (2) closed convex sets. Some of the more important theoretical results characterizing these cases are examined in the next two subsections.

Attribute Sets: Closed Subspaces

Much of the original research in signal restoration was concerned with the highly restrictive case in which each of the individual attribute sets C_k is a closed subspace. The composite attribute set C as formed from their set intersection (14.68) must therefore also be a closed subspace. For this special case there exists a useful analytical characterization of the solution to the signal restoration problem (14.73). The notion of vector orthogonality is central to this discussion. In particular, the two vectors $\mathbf{x},\mathbf{y} \in X$ are said to be orthogonal if their inner product is zero, that is,

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0 \tag{14.77}$$

Furthermore, if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, it follows that the squared inner product-induced norm (14.76) of their vector difference satisfies

$$\| x - y \|^2 = \| x \|^2 + \| y \|^2 \tag{14.78}$$

which is commonly known as the general *Pythagorean theorem*. This theorem is readily proven by direct substitution and using the orthogonality of the two vectors. With these preliminaries completed, the celebrated projection theorem is now given [Luenberger, 1969].

Theorem 1. Let C be a closed subspace of Hilbert space X. Corresponding to any vector $\mathbf{x} \in X$, there is a unique vector $\mathbf{x}^o \in C$ such that $\|\mathbf{x} - \mathbf{x}^o\| \le \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{y} \in C$. Furthermore, a necessary and sufficient condition that $\mathbf{x}^o \in C$ be the unique minimization vector is that $\mathbf{x} - \mathbf{x}^o$ is orthogonal to every vector in C.

When the attribute sets C_k are each closed spaces, it follows that their intersection gives rise to a composite attribute set C which is also a closed subspace. The above theorem indicates that the solution to the signal restoration problem (14.73) is unique. It will be useful to interpret this solution from an algebraic viewpoint.

---

[3]The axioms of an inner product are (1) $\langle \mathbf{x},\mathbf{y} \rangle = \langle \overline{\mathbf{y},\mathbf{x}} \rangle$, (2) $\langle \mathbf{x}+\mathbf{y},\mathbf{z} \rangle = \langle \mathbf{x},\mathbf{z} \rangle + \langle \mathbf{y},\mathbf{z} \rangle$, (3) $\langle \alpha\mathbf{x},\mathbf{y} \rangle = \alpha\langle \mathbf{x},\mathbf{y} \rangle$, and (4) $\langle \mathbf{x},\mathbf{x} \rangle \ge 0$ and $\langle \mathbf{x},\mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

[4]The sequence $\{\mathbf{x}_n\}$ contained in X is said to be Cauchy if for every $\epsilon > 0$ there is an integer $N(\epsilon)$ such that $d(x_n, x_m) < \epsilon$ for all m and $n > N(\epsilon)$.

Specifically, the concept of orthogonal projection operator plays a central role in characterizing the basic nature of the vector $\mathbf{x}^o$ described in Theorem 1. The association of the given vector $\mathbf{x}$ with its unique approximating vector $\mathbf{x}^o \in C$ is notationally specified by

$$\mathbf{x}^o = P_C \mathbf{x} \tag{14.79}$$

It is straightforwardly shown that this one-to-one association P_C possesses the three properties of being:

1. Linear so that $P_C(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha P_C\mathbf{x} + \beta P_C\mathbf{y}$
2. Idempotent so that $P_C^2 = P_C$
3. Self-adjacent so that $P_C^\star = P_C$

A mapping that possesses these three properties is commonly referred to as an *orthogonal projection operator.* The mapping P_C designates the *orthogonal projection* of vector space X onto the closed subspace C. The term orthogonal arises from the observation that every vector in subspace C is orthogonal to the associated error vector $\mathbf{x} - P_C\mathbf{x}$. The concept of orthogonal projection operators is of fundamental importance in optimization theory and has many important practical and theoretical implications.

There exists a convenient means for obtaining a solution to the signal restoration problem (14.73) when the composite attribute set C is a finite-dimensional subspace. The following theorem summarizes the main points of this solution procedure.

Theorem 2. Let the nonempty closed subspace C of Hilbert space X be composed of all vectors which are linear combinations of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$. Corresponding to any vector $\mathbf{x} \in X$, the unique vector that minimizes $\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{y} \in C$ is of the form

$$\mathbf{x}^o = \sum_{k=1}^{q} a_k \mathbf{x}_k \tag{14.80}$$

The a_k coefficients in this linear combination are the components of any solution to the consistent system of linear equations

$$G\mathbf{a} = \mathbf{b} \tag{14.81}$$

where G is the $q \times q$ Gram matrix whose (i,j)th component is specified by $\langle \mathbf{x}_j, \mathbf{x}_i \rangle$ and $\mathbf{b}$ is the $q \times 1$ vector whose ith component is given by $\langle \mathbf{x}, \mathbf{x}_i \rangle$.

If the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$, that span subspace C are linearly independent, then the Gram matrix G is invertible and the linear system of equations (14.81) has a unique solution for the coefficient vector $\mathbf{a}$. On the other hand, this linear system of equations will have an infinity of coefficient vector solutions when the $\{\mathbf{x}_k\}$ vectors are linearly dependent. Nevertheless, each of these coefficient vector solutions lead to the same unique optimum solution (14.80).

Example 4. One of the most commonly studied signal restoration and signal recovery problems is concerned with the task of estimating the signal component in a noise-contaminated measurement as described by the relationship

$$\mathbf{x} = H\mathbf{a} + \mathbf{w} \tag{14.82}$$

In this expression the measurement signal $\mathbf{x}$ is assumed to lie in ϵC^m, H is a known matrix contained in $C^{m \times n}$ of full rank n, and $\mathrm{w} \in C^m$ is an unobserved noise vector. It is now desired to find a vector of form $\mathrm{y} = H\mathrm{a}$ that best approximates the measurement x in the sense of minimizing the quadratic criterion

$$f(\mathbf{a}) = [\mathbf{x} - H\mathbf{a}]^\star W[\mathbf{x} - H\mathbf{a}] \tag{14.83}$$

In this criterion W is a positive-definite Hermitian matrix, and the asterisk symbol designates the operation of complex transposition. It is to be noted that if $\mathbf{w}$ is a Gaussian random vector with covariance matrix W^{-1}, then the minimization of functional (14.83) corresponds to the maximum likelihood estimation of the signal component $H\mathbf{a}$.

Examination of this problem formulation reveals that it is of the same form treated in Theorem 1 in which the Hilbert space is C^n. In view of criterion (14.83) that is to be minimized, the inner product identifying this Hilbert space is taken to be $<\mathbf{x}, \mathbf{y}> \equiv \mathbf{x}^* W\mathbf{y}$. It is a simple matter to show that this measure satisfies the three axioms of an inner product. The closed subspace C corresponds to the set of all vectors that are expressible as $H\mathbf{a}$ (i.e., the range space of matrix H). Since matrix H is taken to have full column rank n, it follows that C has dimension n. If in a given application the elements of the coefficient vector $\mathbf{a}$ can take on complex values, it is readily shown that the orthogonal projection matrix P_C is specified by

$$P_C = H[H^* WH]^{-1} H^* W \tag{14.84}$$

Furthermore, the unique vector $\mathbf{a}^o \in C^n$ which minimizes functional (14.83) is obtained by using the projection relationship $H\mathbf{a}^o = P_C \mathbf{x}$ to yield

$$\mathbf{a}^o = [H^* WH]^{-1} H^* W\mathbf{x} \tag{14.85}$$

On the other hand, if the coefficient vector $\mathbf{a}$ is restricted to lie in R^n, the required orthogonal projection operation takes the form

$$P_C \mathbf{x} = H[\mathrm{Re}\{H^* WH\}]^{-1}\mathrm{Re}\{H^* W\mathbf{x}\} \tag{14.86}$$

where Re designates the "real part of" operator. Moreover, the unique real-valued coefficient vector minimizing functional (14.83) is specified by

$$\mathbf{a}^o = [\mathrm{Re}\{H^* WH\}]^{-1}\mathrm{Re}\{H^* W\mathbf{x}\} \tag{14.87}$$

A solution to the general signal restoration problem for the case in which the individual attribute sets are closed subspaces formally entails determining the orthogonal projection operator P_C on the closed composite attribute subspace (14.81). Unfortunately, an analytical expression for P_C is typically intractable even when the orthogonal projection operators P_k defined on each of the individual attribute subspaces C_k are readily constructed. The natural question arises as to whether it is possible to use these individual orthogonal projection operators to generate P_C. In recognition of this need, J. von Neumann developed a method for iteratively constructing the required projection operator [von Neumann, 1950] for the case of two ($m = 2$) closed subspaces (p. 55, theorem 14.7). This result was later extended by Halperin [1962] for the multiple attribute closed subspace case as described in the following theorem.

Theorem 3. Let P_k denote the orthogonal projection operators onto the closed subspaces C_k for $k = 1,2, ..., m$ of Hilbert space X. Moreover, let C designate the nonempty closed subspace formed by the intersection of these closed subspaces so that $C = C_1 \cap C_2 \cap \ldots \cap C_m$. If P_C designates the orthogonal projection matrix onto closed subspace C and $T = P_1 P_2 \ldots P_m$, it then follows that T^n converges strongly to P_C, that is,

$$\lim_{n \to \infty} \left\| T^n - P_C \right\| = 0 \tag{14.88}$$

This theorem indicates that repeated applications of operator T converge to the required orthogonal projection operator P_C. The practicality of this result arises from the observation that it is often possible to synthesize the individual orthogonal projection operators P_k but not the composite orthogonal projection operator P_C. This capability was illustrated in Theorem 2 for the case in which the subspaces were finite dimensional. To solve

the signal restoration problem (14.73) when the attribute sets are each closed subspaces, we then simply use the following iterative scheme

$$\mathbf{x}_n = T\mathbf{x}_{n-1} \quad \text{for } n = 1, 2, 3, \ldots \tag{14.89}$$

in which the initial point is taken to be the measurement signal so that $\mathbf{x}_o = \mathbf{x}$. The sequence of signals thereby generated is guaranteed to converge to the unique solution of the signal restoration problem.

Linear Variety Property Sets

Theorem 3 is readily generalized to the case in which the individual attribute sets are closed linear varieties. A set contained in vector space X is said to be a linear variety if it is a translation of a subspace contained in X. More specifically, if C is a subspace of X and $\mathbf{u}$ is a fixed vector contained in X, then the associated linear variety is specified by

$$\begin{aligned} V &= \mathbf{u} + C \\ &= \{\mathbf{x} \in X : \mathbf{x} = \mathbf{u} + \mathbf{y} \text{ for all } \mathbf{y} \in C\} \end{aligned} \tag{14.90}$$

It is to be noted that the vector $\mathbf{u}$ used in this linear variety formulation is not unique. In fact, any vector contained in V could have been used in place of $\mathbf{u}$. When subspace C is closed, there exists a unique vector contained in V of minimum norm that is of particular interest in many applications. It is formally specified by

$$\mathbf{u}^o = P_C\mathbf{u} \tag{14.91}$$

where P_C designates the orthogonal projection operator onto the closed subspace C. Vector $\mathbf{u}^o$ represents that unique vector contained in the closed linear variety V which lies closest to the origin in the inner product-induced norm sense. With these thoughts in mind, the following lemma is readily proven.

Lemma 1. Let $V_k = u_k + C_k$ be closed linear varieties associated with the closed subspaces C_k and vectors u_k contained in Hilbert space X for $k = 1, 2, \ldots, m$. Moreover, let V designate the nonempty closed linear variety formed by the intersection of these closed linear varieties so that $V = V_1 \cap V_2 \cap \ldots \cap V_m$. Corresponding to any vector $\mathbf{x} \in X$, the vector contained in V that lies closest to $\mathbf{x}$ in the inner product-induced sense is the limit of the sequence generated according to

$$\mathbf{x}_n = T_1 T_2 \ldots T_m \mathbf{x}_{n-1} \quad \text{for } n = 1, 2, 3, \ldots \tag{14.92}$$

where $\mathbf{x}_o = \mathbf{x}$. The operators appearing in this expression are formally defined by

$$T_k\mathrm{y} = P_{C_k}\mathrm{y} + \{I - P_{C_k}\}\mathrm{u}_k \qquad \text{for } k = 1, 2, \ldots, m \tag{14.93}$$

Attribute Sets: Closed Convex Sets

Although the case in which the signal attribute sets are closed subspaces is of theoretical interest, it is typically found to be too restrictive for most practical applications. As we will now see, however, it is possible to extend these concepts to the more general case of closed convex attribute sets. The set C is said to be convex if for any two vectors $\mathbf{x},\mathbf{y} \in C$ their convex sum as defined by $\lambda\,\mathbf{x} + (1 - \lambda)\mathbf{y}$ is also contained in C for all $0 \le \lambda \le 1$. The ability to broaden the class of attribute sets to include closed convex sets significantly expands the class of problems that can be treated by signal restoration. The following well-known functional analysis theorem provides the framework for this development.

Theorem 4. Let C be a nonempty closed convex set of Hilbert space X. Corresponding to any vector $\mathbf{x} \in X$, there is a unique vector $\mathbf{x}^o \in C$ such that $\|\mathbf{x} - \mathbf{x}^o\| \le \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{y} \in C$. Furthermore, a necessary and sufficient condition that $\mathbf{x}^o$ be the unique minimizing vector is that $\langle \mathbf{x} - \mathbf{x}^o, \mathbf{y} - \mathbf{x}^o \rangle \le 0$ for all $\mathbf{y} \in C$.

From this theorem it is seen that there exists a one-to-one correspondence between a general vector $\mathbf{x} \in X$ and its closest approximation in the closed convex set C. This mapping is here operationally designated by

$$\mathbf{x}^o = P_C \mathbf{x} \tag{14.94}$$

and we will refer to P_C as the projection operator onto the closed convex set C. It is important to note that if the closed convex set C is not a subspace or a linear variety, then this associated projection operator is no longer linear. Moreover, if $\mathbf{x} \in C$, then $P_C \mathbf{x} = \mathbf{x}$, so that every vector of C is a *fixed point* of operator P_C.[5] Thus, the fixed points of P_C and the closed convex set C are equivalent. These concepts were used by Bregman [1965] and others to prove a type of convergence of the successive projection algorithm. The following theorem summarizes their results.

Theorem 5. Let P_k designate the projection operators onto the closed convex sets C_k for $k = 1, 2, \ldots, m$ of Hilbert space X and $C = C_1 \cap C_2 \cap \ldots \cap C_m$ be their nonempty closed convex set intersection. Furthermore, for every $\mathbf{x} \in X$ consider the sequence of successive projections as generated according to

$$\mathbf{x}_n = P_1 P_2 \quad \ldots P_m \mathbf{x}_{n-1} \quad \text{for } n = 1, 2, 3, \ldots \tag{14.95}$$

in which $\mathbf{x}^o = \mathbf{x}$. It then follows that this sequence converges:

1. Weakly to a point in the set intersection C[6]
2. Strongly to a point in the set intersection C if at least one of the sets C_k is bounded and compact[7]

The weak-point convergence theorem (1) was first proven by Bregman [1965], while the strong-point convergence theorem (2) was developed by Stiles [1965]. It is important to appreciate what this theorem does and does not say. Specifically, although it ensures that sequence (14.95) converges to a vector contained in the set intersection C, this convergent point need not minimize the original signal restoration criterion (14.73). This is the price paid when considering the more general case of closed convex attribute sets. Nonetheless, it is found that the vector to which sequence (14.95) converges often provides a satisfactory approximation.

To improve the convergence of algorithm (14.95), Gubin et al. introduced an overrelaxation modification that extends the projections beyond the boundary of the attribute sets [1967]. This overrelaxation approach was also adopted by Youla [1978] who proposed the algorithm

$$\mathbf{x}_n = T_1 T_2 \ldots T_m \mathbf{x}_{n-1} \quad \text{for } n = 1, 2, 3, \ldots \tag{14.96}$$

in which operators $T_k = [I + \lambda_k (P_k - I)]$ are employed. They proved that the sequence so generated converges weakly to a point of C for any choice of the relaxation constants λ_k in the open interval $0 < \lambda_k < 2$. Moreover, if at least one of the closed convex sets C_k is contained in a finite-dimensional subspace, then the convergence is strong.

In summary, the successive projection algorithm provides a useful signal-processing tool for the case in which the individual attribute sets are closed convex sets. Its primary deficiency is that although the signal sequence (14.95) so generated converges to an element of C, it need not converge to the closest approximation of the data signal $\mathbf{x}$ contained in C. Thus, the successive projection algorithm generally fails to provide a solution to the signal restoration problem (14.73). In recognition of this shortcoming, Dykstra [1983] developed an algorithm which does provide an algorithmic solution. This algorithm was further studied by Dykstra and Boyle [1987], Han [1988], and Gaffke and Mathar [1989]. The formulation as given by Gaffke and Mathar is now summarized.

[5]The vector $\mathbf{x}$ is said to be a fixed point of operator T if $T\mathbf{x} = \mathbf{x}$.

[6]The sequence $\{\mathbf{y}_n\}$ is said to converge weakly to $\mathbf{y}$ if $< y_n - \mathbf{y}, \mathbf{z} >$ converges to zero for every $\mathbf{z} \in X$ as n becomes unbounded.

[7]The sequence $\{\mathbf{y}_n\}$ is said to converge strongly to $\mathbf{y}$ if $\| y_n - y \|$ approaches zero as n becomes unbounded.

Theorem 6. Let P_k designate the projection operators onto the closed convex sets C_k for $k = 1, 2, \ldots, m$ of Hilbert space X in which the closed convex set intersection $C = C_1 \cap C_2 \cap \ldots \cap C_m$ is nonempty. Let the two sets of m vector sequences $\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \ldots, \mathbf{x}_m^{(k)}$ and $\mathbf{y}_1^{(k)}, \mathbf{y}_2^{(k)}, \ldots, \mathbf{y}_m^{(k)}$ for $k = 1,2,3, \ldots$ be generated according to

$$\begin{aligned} \mathbf{x}_i^{(k)} &= P_i\left(\mathbf{x}_{i-1}^{(k)} - \mathbf{y}_i^{(k-1)}\right) && \text{for } i = 1, 2, \ldots, m \\ \mathbf{x}_0^{(k+1)} &= \mathbf{x}_m^{(k)} \\ \mathbf{y}_i^{(k)} &= \mathbf{x}_i^{(k)} - \left(\mathbf{x}_{i-1}^{(k)} - d_i^{(k-1)}\right) && \text{for } i = 1, 2, \ldots, m \end{aligned} \tag{14.97}$$

in which the initial conditions are taken to be $\mathbf{y}_1^{(o)} = \mathbf{y}_2^{(o)} = \ldots = \mathbf{y}_m^{(o)} = \mathbf{0}$ and $\mathbf{x}_m^{(o)} = \mathbf{x}$. It then follows that the sequence $\{\mathbf{x}_m^k\}$ converges to the unique point in C that lies closest to $\mathbf{x}$ in the sense of minimizing functional (14.73).

This algorithm provides a useful means for iteratively finding a solution to the signal restoration problem for the case of closed convex sets.

Closed Projection Operators

A solution to the signal restoration problem (14.73) is generally intractable unless restrictive assumptions are made on the constituent attribute sets. In the previous two subsections, the method of successive projections and its variations were presented for iteratively finding a solution when the underlying attribute sets are closed subspaces, closed linear varieties, or closed convex sets. Unfortunately, some of the more important attribute sets encountered in signal processing do not fall into any of these categories. This is illustrated by the case in which the Hilbert space is taken to be $C^{m \times n}$ and one of the attribute sets corresponds to all $m \times n$ matrices which have rank q where $q < \min(m,n)$. It is readily shown that this set is neither a subspace nor a linear variety, nor is it convex. Thus, use of the extremely important rank q attribute set cannot be justified for any of the algorithms considered up to this point. This is a serious shortcoming when it is realized that this attribute set is used so extensively in many contemporary signal processing applications.

To provide a viable method for approximating a solution to the signal restoration problem for nonconvex attribute sets, we shall now broaden the approach taken. The signals are again assumed to lie in a metric space X with distance metric $d(\mathbf{x},\mathbf{y})$. Furthermore, it is assumed that it is possible to solve each of the individual *projection operator* problems as defined by

$$P_k(\mathbf{x}) = \{\mathbf{y} : d(x, y) = \min_{\mathbf{z} \in C_k} d(\mathbf{x}, \mathbf{z})\} \qquad \text{for } 1 \le k \le m \tag{14.98}$$

for an arbitrary signal $\mathbf{x} \in X$. These sets consist of all elements in C_k that lie closest to $\mathbf{x}$ in the distance-metric sense for $1 \le k \le m$. It is to be noted that the projection operators $P_k(\mathbf{x})$ are generally nonlinear and not one to one as was the case in the previous two subsections. When determining the individual projection operators (14.98), the fundamental issues of the existence and uniqueness of solution need to be addressed. It is tacitly assumed that the signal attributes and metrics under consideration are such that at least one solution exists and the solution(s) may be obtained in a reasonably simple fashion. Fortunately, the generation of solutions imposes no serious restrictions for many commonly invoked attribute sets. Moreover, many relevant signal attributes are characterized by the fact that more than one solution to optimization problems (14.98) exists.

The projection operators (14.98) are unusual in that they need not be of the traditional point-to-point variety as was the case when the attribute set is a closed subspace, closed linear variety or closed convex set. For general C_k sets, P_k is a point-to-set mapping. The concept of a *closed mapping* is of importance when extending the notion of signal restoration to nonconvex sets. A closed mapping is a generalization of the notion of continuity as applied to standard point-to-point mappings [Zangwill, 1969], that is:

Definition 1. The point-to-set projection mapping P is said to be closed at x ∈ X if the assumptions

1. $\mathbf{x}_k \rightarrow \mathbf{x}$ with $\mathbf{x}_k \in X$
2. $\mathbf{y}_k \rightarrow \mathbf{y}$ with $\mathbf{y}_k \in P(\mathbf{x}_k)$

imply that $\mathbf{y} \in P(\mathbf{x})$. The point-to-set **projection mapping** P is said to be closed on the set X_1 if it is closed at each point in X_1.

Signal Restoration Algorithm

We are now in a position to describe the signal restoration algorithm for the case of general attribute sets. This shall be formally done in the format of the following theorem [Cadzow, 1988].

Theorem 7. Let P_k be the projection operators associated with the attribute sets C_k contained in metric space X for $k = 1, 2, \ldots, m$. For any signal $\mathbf{x} \in X$, let the sequence $\{\mathbf{x}_k\}$ be generated according to

$$\mathbf{x}_k \in P_1 P_2 \ldots P_m(\mathbf{x}_{k-1}) \quad \text{for } k \geq 1 \tag{14.99}$$

in which the initial signal is specified by $\mathbf{x}_o = \mathbf{x}$. A subsequence of this sequence always exists which converges to an element of the set intersection $C = C_1 \cap C_2 \cap \ldots \cap C_m$ provided that: (1) the $d(\mathbf{x}_k, \mathbf{x}_r) < d(\mathbf{x}_{k-1}, \mathbf{x}_r)$, where $\mathbf{x}_r \in X$ designates a reference signal which is often the origin of X, and (2) the set of signals $\mathbf{y} \in X$ that satisfy the inequality $d(\mathbf{y}, \mathbf{x}_r) \leq d(\mathbf{x}, \mathbf{x}_r)$ defines a closed and bounded set.

A casual examination of signal restoration algorithm (14.99) indicates that it is of the same form as the sequence of projections algorithms described in the previous two subsections. It distinguishes itself from those algorithms in that the attribute sets C_k need not be closed subspaces, closed linear varieties, or closed convex sets. The proposed algorithm also distinguishes itself from several other signal restoration algorithms in that the metric d need not be inner product induced. These can be important considerations in specific applications. As an example, it has been conjectured by several authors that the l_1 norm provides for a more effective error measure when the data set has outliers (i.e., unrepresentative data). Signal restoration algorithm (14.99) can be directly applied to such problems since we have not restricted the metric. It must be observed, however, that the nature of the individual projection operators P_k is typically most easily characterized when the metric employed is inner product induced.

It is useful to represent the multiple mapping signal restoration algorithm by the *composite mapping* as defined by

$$P = P_1 P_2 \ldots P_m \tag{14.100}$$

The process of generating the solution set $P(\mathbf{x}_{k-1})$ from $\mathbf{x}_{k-1}$ is to be interpreted in the following sequential manner. First, the set $P_m(\mathbf{x}_{k-1})$ is found. This set consists of all signals possessing the mth attribute that lie closest to $\mathbf{x}_{k-1}$ in the given signal metric. Next, the set $P_{m-1}(P_m(\mathbf{x}_{k-1}))$ is formed and consists of all signals possessing the $(m-1)$th attribute that lie closest to each of the signals in set $P_m(\mathbf{x}_{k-1})$. It is to be noted that although each of the signals in $P_{m-1}(P_m(\mathbf{x}_{k-1}))$ possess the $(m-1)$th attribute, they need not possess the mth attribute. This process is continued in this fashion until the set $P(\mathbf{x}_{k-1})$ is generated. Finally, we arbitrarily select one signal from $P(\mathbf{x}_{k-1})$ to be equal to $\mathbf{x}_k$. When the individual projection mappings P_k are each point-to-point mappings, the signal $\mathbf{x}_k$ generated in this fashion will be unique.

Signal restoration algorithm (14.99) has been applied to many fundamental signal processing problems. It has produced effective results that often exceed those achieved by more traditional methods. The ultimate utility of signal restoration is dependent on the user's innovativeness in generating signal attributes that distinguish the underlying signal from the corruption in the data. In many applications, matrix descriptions of the data under analysis arise in a natural manner. With this in mind, we will now explore some salient matrix properties and how they can be used in typical signal restoration applications.

Algebraic Properties of Matrices

Many of the more important and interesting applications of signal restoration are related to the vector space $C^{m\times n}$. Matrices contained in $C^{m\times n}$ may occur in a natural manner as exemplified by digital images where the nonnegative elements of the matrix correspond to the brightness levels of associated pixels. The underlying signal restoration problem in such cases is commonly referred to as *image reconstruction* or *image restoration.* In other examples, however, the matrix under consideration may be a by-product of an associated data analysis solution routine. For example, in approximating a finite-length time series as a sum of weighted exponentials, one often forms an associated *data matrix* from the time series elements. Whatever the case, the matrix under analysis is typically corrupted in some manner, and it is desired to remove this corruption in order to recover the underlying information-bearing matrix. In using signal restoration for this purpose, it is necessary to employ attributes associated with the information-bearing signal. These attributes are normally of an algebraic or a structural description. In this subsection we will examine two of the more widely invoked algebraic attributes, and some commonly employed structural attributes are examined in the next subsection.

Singular Value Decomposition

The **singular value decomposition** (SVD) of a real- or complex-valued matrix plays an increasingly important role in contemporary signal processing applications. For a generally complex-valued matrix $A \in C^{m\times n}$, its associated SVD representation takes the form of the following sum of r weighted outer products,

$$A = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\star} \tag{14.101}$$

where r designates the rank of matrix A. In this SVD representation, the σ_k are positive *singular values* that are ordered in the monotonic fashion $\sigma_{k+1} \le \sigma_k$, while the $\mathbf{u}_k$ are the $m \times 1$ pairwise orthogonal *left singular vectors* and the $\mathbf{v}_k$ are the $n \times 1$ pairwise orthogonal *right singular vectors.* Upon examination of SVD representation (14.101) it is seen that the mn components of matrix A are equivalently replaced by the $r(m + n + 1)$ elements corresponding to the SVD singular values and singular vectors. For low-rank matrices [i.e., $r(1 + m + n) < mn$], the SVD provides for a more efficient representation of a matrix. This observation has been effectively used for the data compression of digital images. Furthermore, the concept of a low rank data matrix plays a vital role in the modeling of empirical data as a sum of weighted exponential signals. With these thoughts in mind, the important concept of low rank approximation of matrices is now examined.

Reduced-Rank Approximation

In various applications, it is frequently desired to find a matrix of rank q that best approximates a given matrix $A \in C^{m\times n}$, where $q < \text{rank}\{A\}$. If the Frobenius metric (14.66) is used to measure the distance between two matrices, it is well known that the best rank-q matrix approximation of A is obtained by simply dropping all but the q largest singular-valued weighted outer products in SVD representation (14.101), that is,

$$A^{(q)} = \sum_{k=1}^{q} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\star} \tag{14.102}$$

where it is tacitly assumed that $q \le r$. From this expression, it follows that the mapping from A into $A^{(q)}$ is one-to-one when $\sigma_q > \sigma_{q+1}$ but is one-to-many (infinity) if $\sigma_q = \sigma_{q+1}$. The special case in which $\sigma_q = \sigma_{q+1}$ therefore results in a point-to-set mapping and gives rise to subtle issues which are addressed in the following theorem [Mittelmann and Cadzow, 1987].

Theorem 8. Let the $m \times n$ matrix $A \in C^{m\times n}$ have the SVD decomposition (14.101) in which $r = \text{rank}(A)$. The best rank-q Frobenius norm approximation of A as given by expression (14.102) is unique if and only if $\sigma_q > \sigma_{q+1}$. The projection operator $P^{(q)}$ from A into $A^{(q)}$ as designated by

$$A^{(q)} = P^{(q)} A \tag{14.103}$$

is nonlinear and closed. Furthermore, this mapping is continuous for $\sigma_q \neq \sigma_{q+1}$ and is not continuous when $\sigma_q = \sigma_{q+1}$.

When applying the reduced-rank approximation of a matrix as specified by relationship (14.103), it is desirable that the gap between the so-called *signal-level* singular values and *noise-level* singular values be large (i.e., $\sigma_q - \sigma_{q+1} >> 0$). If this is true, then the issues of nonuniqueness and continuity of mapping do not arise. Unfortunately, in many challenging applications, this gap is often small, and one must carefully examine the consequences of this fact on the underlying problem being addressed. For example, in modeling empirical data as a sum of exponentials, this gap is typically small, which in turn leads to potential undesirable algorithmic sensitivities. We will examine the exponential modeling problem in a later subsection.

Positive-Semidefinite Matrices

Positive-semidefinite matrices frequently arise in applications related to random and deterministic time series. For example, if $\mathbf{x} \in C^n$ is a vector whose components are random variables, it is well known that the associated $n \times n$ *correlation matrix* as defined by $R_{xx} = E\{\mathbf{x}\mathbf{x}^*\}$ is positive-semidefinite Hermitian, where E designates the expected value operator. Similarly, *orthogonal projection* matrices are often used to describe vector subspaces that identify signals present in empirical data. An orthogonal projection matrix is a positive-semidefinite matrix which has the additional requirements of being idempotent (i.e., $A^2 = A$) and Hermitian (i.e., $A^* = A$). It is well known that the eigenvalues associated with an orthogonal projection matrix are exclusively equal to zero and one. With these examples serving as motivation, we shall now examine some of the salient algebraic characteristics of positive-semidefinite matrices.

The $n \times n$ matrix $A \in C^{n \times n}$ is said to be positive semidefinite if the associated quadratic form inequality as specified by

$$\mathbf{x}^* A\mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in C^n \tag{14.104}$$

is satisfied for all vectors $x \in C^n$. Furthermore, if the only vector that causes this quadratic form to be zero is the zero vector, then A is said to be *positive definite*. Since this quadratic form is real valued, we can further infer that any positive-semidefinite matrix must be Hermitian so that $A^* = A$. Moreover, using elementary reasoning it directly follows that the set of positive-definite matrices contained in C^n is a closed convex cone.

In many practical applications, there is given empirical time series data to be analyzed. This analysis is often predicated on one's having knowledge of the time series-associated correlation matrices or orthogonal projection matrices. Since such knowledge is generally unknown, these matrices must be estimated from the empirical data under analysis. These estimates, however, are almost always in error. For example, the estimate $\hat{R}_{xx}$ of a correlation matrix R_{xx} is often Hermitian but not positive semidefinite. To mitigate the effects of these errors, an intuitively appealing procedure would be to find a matrix lying close to $\hat{R}_{xx}$ that possesses the two prerequisite properties of being (1) Hermitian and (2) positive semidefinite. The concept of signal restoration can be used for this purpose if it is possible to develop a closed-form expression for the operator that maps a general Hermitian matrix into the closest positive-semidefinite Hermitian matrix in the Frobenius matrix norm sense. As is now shown, a simple expression for this operator is available using the SVD of the Hermitian matrix being approximated.

With these thoughts in mind, we will now consider the generic problem of finding a positive-semidefinite matrix that lies closest to a given Hermitian matrix $A \in C^{n \times n}$. In those applications where matrix A is not Hermitian, then this matrix is first replaced by its Hermitian component as defined by $(A + A^*)/2$ and then the closest positive-semidefinite matrix to this Hermitian component is found.[8] The problem at hand is readily solved by first making an eigenanalysis of the Hermitian matrix A, that is,

$$A\mathbf{x}_k = \lambda_k \mathbf{x}_k \quad \text{for} \quad 1 \leq k \leq n \tag{14.105}$$

[8]Any matrix A can be represented as the sum of a Hermitian matrix and a skew Hermitian matrix using the decomposition $A = (A + A^*)/2 + (A - A^*)/2$.

Since A is Hermitian, its eigenvalues must all be real and there exists a full set of n associated eigenvectors that can always be chosen orthonormal [i.e., $\mathbf{x}_k^* \mathbf{x}_m = \delta(k-m)$]. With this characterization, the following theorem is readily proven.

Theorem 9. Let A be a Hermitian matrix contained in $C^{n\times n}$ whose eigenanalysis is specified by relationship (14.105). Furthermore, let the eigenvalues be ordered in the monotonically nonincreasing fashion $\lambda_k \geq \lambda_{k+1}$ in which the first p eigenvalues are positive and the last $n-p$ are nonpositive. It then follows that the SVD of Hermitian matrix A can be decomposed as

$$\begin{aligned} A &= \sum_{k=1}^{p} \lambda_k \mathbf{x}_k \mathbf{x}_k^* + \sum_{k=p+1}^{n} \lambda_k \mathbf{x}_k \mathbf{x}_k^* \\ &= A^+ + A^- \end{aligned} \tag{14.106}$$

The Hermitian matrix A^+ is positive semidefinite of rank p, while the Hermitian matrix A^- is negative semidefinite. Furthermore, the unique positive-semidefinite matrix that lies closest to A in the Frobenius and Euclidean norm sense is given by the truncated SVD mapping

$$A^+ = P^+(A) = \sum_{k=1}^{p} \lambda_k \mathbf{x}_k \mathbf{x}_k^* \tag{14.107}$$

The projection operator P^+ is closed and continuous. Furthermore, an idempotent Hermitian (i.e., an orthogonal projection) matrix which lies closest to A in the Frobenius and Euclidean norm sense is specified by

$$A^{op} = P^{op}(A) = \sum_{k:\lambda_k \geq 0.5} \mathbf{x}_k \mathbf{x}_k^* \tag{14.108}$$

This closest idempotent Hermitian matrix is unique provided that none of the eigenvalues of A are equal to 0.5. Moreover, projection operator P^{op} is closed for any distribution of eigenvalues.

Examination of this theorem indicates that the left and right singular vectors of a Hermitian matrix corresponding to its positive eigenvalues are each equal to the associated eigenvector while those corresponding to its negative eigenvalues are equal to the associated eigenvector and its negative image. Furthermore, any Hermitian matrix may be uniquely decomposed into the sum of a positive- and negative-semidefinite Hermitian matrix as specified by (14.106).

This theorem's proof is a direct consequence of the fact that the Frobenius and Euclidean norm of the matrix A and Q^*AQ are equal for any unitary matrix Q. Upon setting Q equal to the $n \times n$ matrix whose columns are equal to the n orthonormal eigenvectors of matrix A, it follows that the Frobenius (Euclidean) norm of the matrices $A-B$ and $Q^*[A-B]Q$ are equal. From this equality the optimality of positive-definite matrix (14.107) immediately follows since Q^*AQ is equal to the diagonal matrix with the eigenvalues of A as its diagonal components. The closest positive-semidefinite matrix (14.107) is obtained by simply truncating the SVD to the positive singular value outer products. Similarly, the closest orthogonal projection matrix is obtained by replacing each singular value by 1 if the singular value is greater than or equal to 0.5 and by 0 otherwise.

Structural Properties of Matrices

In various applications, a matrix under consideration is known to have its elements functionally dependent on a relatively small set of parameters. A brief listing of some of the more commonly used matrix classes so characterized is given in Table 14.9. In each case, there exists a relatively simple relationship for the elements

TABLE 14.9 Structured Matrices

Matrix Class	Matrix Elements
Hermitian	$A(i, j) = \overline{A}(j, i)$
Toeplitz	$A(i + 1, j + 1) = A(i, j)$
Hankel	$A(i + 1, j) = A(i, j + 1)$
Circulant	$A(i + 1, j) = A(i, j - 1)$ with $A(i + 1, 1) = A(i,n)$
Vandermonde	$A(i, j) = A(2, j)^{i-1}$

of the matrix. For example, an $m \times n$ Toeplitz matrix is completely specified by the $m + n - 1$ elements of its first row and first column. We now formalize this concept.

Definition 2. Let $a_{ij}(\theta_1, \theta_2, \ldots, \theta_p)$ for $1 \le i \le m$ and $1 \le j \le n$ designate a set of mn functions that are dependent on a set of real-valued parameters $\theta_1, \theta_2, \ldots, \theta_p$ in which $p < mn$. Furthermore, consider the matrix $A \in C^{m \times n}$ whose elements are given by

$$A(i, j) = a_{ij}(\theta_1, \theta_2, \ldots, \theta_p) \quad \text{for } 1 \le i \le m \text{ and } 1 \le j \le n + 1 \tag{14.109}$$

for a specific choice of the parameters $\theta_1, \theta_2, \ldots, \theta_p$. These p parameters shall be compactly represented by the parameter vector $\boldsymbol{\theta} \in R^p$. The set of all matrices that can be represented in this fashion is designated by $\mathcal{M}$ and is said to have a structure induced by the functions $a_{ij}(\boldsymbol{\theta})$ and to have p degrees of freedom. If the functions $a_{ij}(\boldsymbol{\theta})$ are linearly dependent on the parameters, the matrix class $\mathcal{M}$ is said to have a linear structure.

In what is to follow, we will be concerned with the task of optimally approximating a given matrix $B \in C^{m \times n}$ by a matrix with a specific linear structure. For purposes of description, the specific class of matrices to be considered is denoted by $\mathcal{L}$ and its constituent matrices are functionally dependent on the real-valued parameter vector $\boldsymbol{\theta} \in R^p$. The matrix approximation problem is formally expressed as

$$\min_{\boldsymbol{\theta} \in R^p} \left\| B - A(\boldsymbol{\theta}) \right\|_F \tag{14.110}$$

where $A(\boldsymbol{\theta})$ designates a general matrix contained in $\mathcal{L}$ that is linearly dependent on the parameter vector $\boldsymbol{\theta}$.

It is possible to utilize standard analysis techniques to obtain a closed-form solution to matrix approximation problem (14.110). To begin this analysis, it is useful to represent matrix B by the $mn \times 1$ vector $\mathbf{b}_c$ whose elements are formed by concatenating the column vectors of B. This concatenation mapping is one to one and onto and is therefore invertible. Namely, given B there exists a unique concatenated vector image $\mathbf{b}_c$, and vice versa. It further follows that the Euclidean norm of $\mathbf{b}_c$ and the Frobenius norm of B are equal, that is,

$$\left\| \mathrm{b_c} \right\|_2 = \left\| \mathrm{B} \right\|_\mathrm{F} \tag{14.111}$$

Using this norm equivalency, it follows that the original matrix approximation problem (14.110) can be equivalently expressed as

$$\min_{\boldsymbol{\theta} \in R^p} \left\| B - A(\boldsymbol{\theta}) \right\|_F = \min_{\boldsymbol{\theta} \in R^p} \left\| \mathbf{b}_c - \mathbf{a}_c(\boldsymbol{\theta}) \right\|_2 \tag{14.112}$$

where $\mathbf{b}_c$ and $\mathbf{a}_c(\boldsymbol{\theta})$ designate the concatenated vector representations for matrices B and A($\boldsymbol{\theta}$), respectively. Since each element of matrix A($\boldsymbol{\theta}$) $\in \mathcal{L}$ is linearly dependent on the parameter vector $\boldsymbol{\theta}$, it follows that there exists a unique mn × p matrix L such that

$$a_c(\boldsymbol{\theta}) = \mathrm{L}\boldsymbol{\theta} \tag{14.113}$$

is the concatenated representation for matrix $A(\boldsymbol{\theta}) \in \mathcal{L}$. Thus, the original matrix approximation problem can be equivalently expressed as

$$\min_{\boldsymbol{\theta} \in R^p} \| \mathbf{b}_c - L\boldsymbol{\theta} \|_2 \tag{14.114}$$

This problem, however, is seen to be quadratic in the parameter vector $\boldsymbol{\theta}$, and an optimum parameter vector is obtained by solving the associated consistent system of normal equations

$$L^* L\boldsymbol{\theta}^o = \mathrm{L}^*\mathbf{b}_c \tag{14.115}$$

In many cases, the matrix product L^*L is invertible, thereby rendering a unique solution to these equations. Whatever the case, the associated vector representation for the optimum matrix contained in $\mathcal{L}$ is given by $\mathbf{a}_c(\boldsymbol{\theta}^o) = \mathrm{L}\boldsymbol{\theta}^o$. Finally, the corresponding optimum approximating matrix $\mathrm{A}(\boldsymbol{\theta}^o)$ is simply obtained by reversing the column vector concatenation mapping that generates $\mathbf{a}_c(\boldsymbol{\theta}^o)$.

Example 5. To illustrate the above procedure, let us consider the specific case of the class of real 3×2 Toeplitz matrices. A general parametric representation for a matrix in this class and its associated concatenated vector equivalent is given by

$$A(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 & \theta_2 \\ \theta_3 & \theta_1 \\ \theta_4 & \theta_3 \end{bmatrix} \leftrightarrow \mathbf{a}_c(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 \\ \theta_3 \\ \theta_4 \\ \theta_2 \\ \theta_1 \\ \theta_3 \end{bmatrix} \tag{14.116}$$

It then follows that the matrix mapping the parameter vector $\boldsymbol{\theta}$ into $\mathbf{a}_c(\boldsymbol{\theta})$is given by

$$\mathbf{a}_c(\boldsymbol{\theta}) = L\boldsymbol{\theta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \tag{14.117}$$

It is seen that the structure matrix mapping L has a full column rank of four. Finally, the unique solution to relationship (14.115) for the optimum parameter vector used for representing the closest Euclidean norm approximation is specified by

$$\boldsymbol{\theta}^o = [L^* L]^{-1} L^* \mathbf{b}_c = \begin{bmatrix} \frac{1}{2}[B(1,1) + B(2,2)] \\ B(1,2) \\ \frac{1}{2}[B(2,1) + B(3,3)] \\ B(3,2) \end{bmatrix} \tag{14.118}$$

It is clear that the entities $A(\boldsymbol{\theta})$ and $\mathbf{a}_c(\boldsymbol{\theta})$are equivalent. Moreover, the class of real 3×2 Toeplitz matrices is seen to have four degrees of freedom (i.e., the real parameters θ_1, θ_2, θ_3, and θ_4).

It is readily established using the above arguments that the best Toeplitz approximation to a matrix B is obtained by first determining the mean value of each of its diagonals and then using these mean values as entries of the best approximating Toeplitz diagonals. Let us formalize this important result.

Theorem 10. Let C_T and C_H designate the set of all Toeplitz and Hankel matrices contained in the space $C^{m\times n}$, respectively. It follows that C_T and C_H are each complex $(m + n - 1)$-dimensional subspaces of $C^{n\times n}$. Furthermore, the Toeplitz matrix A_T which best approximates $A \in C^{m\times n}$ in the Frobenius norm sense has the constant element along its kth diagonal specified by

$$\alpha_k = \text{mean}[\mathbf{d}_k] \quad \text{for } -n + 1 \leq k \leq m - 1 \tag{14.119}$$

In this expression, mean $[\mathbf{d}_k]$ designates the arithmetic mean of vector $\mathbf{d}_k$ whose components correspond to the elements of matrix A along its kth diagonal. In particular, vector $\mathbf{d}_o$ has as its components the elements of the main diagonal [i.e, elements $A(1,1)$, $A(2,2)$, etc.], vector $\mathbf{d}_1$ has as its components the elements of the diagonal immediately below the main diagonal, vector $\mathbf{d}_{-1}$ has as its components the elements of the diagonal immediately above the main diagonal, and so forth. The projection operator P_T that maps A into A_T as governed by relationship (14.119) is designated by

$$A_T = P_T A \tag{14.120}$$

and is linear and one to one.

Similarly, the Hankel matrix A_H that lies closest to A in the Frobenius norm sense has the constant element along its kth antidiagonal specified by

$$\beta_k = \text{mean}[\mathbf{a}_k] \quad \text{for } -n + 1 \leq k \leq m - 1 \tag{14.121}$$

The components of $\mathbf{a}_k$ correspond to the elements of matrix A along its kth antidiagonal in which vector $\mathbf{a}_o$ corresponds to the main antidiagonal [i.e., elements $A(1,n)$, $A(2,n-1)$, etc.], vector $\mathbf{a}_1$ to the antidiagonal immediately below the main antidiagonal, and so forth. The projection operator P_H mapping A into A_H as governed by relationship (14.121) is designated by

$$A_H = P_H A \tag{14.122}$$

and is linear and one to one.

It is interesting to note that relationships (14.120) and (14.122) which identify the closest approximating Toeplitz and Hankel matrices are very much dependent on the Frobenius measure of matrix size. If a different metric had been incorporated, then different expressions for the best approximating Toeplitz and Hankel matrix approximations would have arisen. For example, in applications in which data outliers are anticipated, it is often beneficial to use the l_1-induced norm. In this case, the elements α_k and β_k are replaced by the median value of the kth diagonal and antidiagonal of matrix A, respectively. Similarly, if the l_∞-induced norm were used, the elements α_k and β_k would be replaced by the midpoint of the largest and smallest elements of the kth diagonal and the kth antidiagonal of matrix A, respectively.

In pure sinusoidal modeling applications, the concept of forward and backward prediction is often used. The data matrix arising from a forward-backward modeling will then have a block Toeplitz-Hankel structure. This in turn gives rise to the signal restoration task of finding a matrix of block Toeplitz-Hankel structure that most closely approximates a given matrix. The results of Theorem 10 can be trivially extended to treat this case.

Nonnegative Sequence Approximation

Two related fundamental problems which arise in various signal processing applications are that of finding (1) a nonnegative-definite sequence which lies closest to a given sequence, or (2) a nonnegative-definite matrix which lies closest to a given matrix. For example, in many commonly employed spectral estimation algorithms, estimates of a time series autocorrelation sequence are either explicitly or implicitly computed from a finite-length sample of the time series. It is well known that the autocorrelation sequence associated with a wide-sense stationary time series has a nonnegative-definite Fourier transform. However, the process of forming the autocorrelation lag estimates from empirical data often results in lag estimates whose Fourier transform can be negative. With this application (and others) in mind, we will now briefly explore some basic theory related to nonnegative-definite sequences and then employ the signal restoration algorithm to solve the second problem posed above.

To begin our development, the sequence $\{x(n)\}$ is said to be *nonnegative definite* if its Fourier transform is real nonnegative, that is,

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \geq 0 \qquad \text{for } \omega \in [0, 2\pi] \tag{14.123}$$

As one might suspect, nonnegative-definite time series possess a number of salient properties which distinguish themselves from more general time series. The following theorem provides an insight into some of the more important properties.

Theorem 11. Let C_{nnd} designate the set of all finite- and infinite-length nonnegative-definite time series. It follows that C_{nnd} is a closed convex cone whose vertex is located at the origin. Moreover, every time series contained in this cone is conjugate symmetric so that $x(-n) = \bar{x}(n)$ for all integers n. Furthermore, the data matrix of order k formed from any nonnegative-definite time series $\{x(n)\} \in C_{\text{nnd}}$ and having the Hermitian-Toeplitz structure given by

$$X_k = \begin{bmatrix} x(0) & \bar{x}(1) & \cdots & \bar{x}(k) \\ x(1) & x(0) & \cdots & \bar{x}(k-1) \\ \vdots & \vdots & \vdots & \vdots \\ x(k) & x(k-1) & \cdots & x(0) \end{bmatrix} \tag{14.124}$$

is positive semidefinite for all nonnegative integer values of the order parameter k.

If $\{x(n)\}$ is a nonnegative sequence of length $2q+1$, then the zeros of its z-transform always occur in conjugate reciprocal pairs, that is,

$$X(z) = \sum_{n=-q}^{q} x(n)z^{-n} = \alpha \prod_{k=1}^{q} [1 - z_k z][1 - \bar{z}_k z^{-1}] \tag{14.125}$$

where α is a positive scalar. In addition, there will exist scalars $b_0, b_1, \ldots, b_q$ such that this $(2q+1)$-length nonnegative sequence can be represented by

$$x(n) = \sum_{k=0}^{q-n} b_k b_{k+n} \qquad \text{for } 1 \leq n \leq q \tag{14.126}$$

A more detailed treatment of this topic is found in Cadzow and Sun [1986]. This theorem must be carefully interpreted in order not to inappropriately infer properties possessed by nonnegative sequences. For instance, it is not true that the positive semidefiniteness of a finite number of data matrices X_k implies that the underlying time series is nonnegative. As an example, the length-three sequence with elements $x(0) = 3$, $x(1) = x(-1) = 2$ is not positive although the data matrix X_1 is positive definite. In a similar fashion, this theorem does not indicate that the symmetric truncation of a nonnegative sequence is itself nonnegative.

Nonnegative-Definite Toeplitz-Hermitian Matrix Approximation

It is possible to employ the concept of signal restoration in a straightforward fashion to iteratively solve the two problems considered at the beginning of this section. We will demonstrate a solution procedure by treating the problem in which it is desired to find that unique nonnegative-definite Toeplitz-Hermitian matrix which lies closest to a given matrix $X \in C^{n \times n}$. To employ the concept of signal restoration to this problem, we now identify the two attribute sets in which the required approximating matrix must lie, namely,

$$C_+ = \{Y \in C^{n \times n} \quad \text{which are nonnegative}\} \tag{14.127}$$

$$C_{TH} = \{Y \in C^{n \times n} \quad \text{for which } Y \text{ is Toeplitz and Hermitian}\}$$

Relationship (14.107) provides the mapping corresponding to the attribute set C_+. Implementation of the mapping associated with attribute set C_{TH} is a straightforward modification of Toeplitz mapping (14.120). In particular, the mean value of the elements of the two diagonals equispaced k units above and below the main diagonal is employed to determine the constant element used in the Toeplitz-Hermitian approximation where $0 \le k \le n - 1$.

Since the attribute sets C_+ and C_{TH} are both closed convex sets, the sequence of successive projections algorithm as specified by

$$X_k = P_+ P_{TH} X_{k-1} \quad \text{for } k = 1, 2, \ldots \tag{14.128}$$

can be employed. The initial matrix for this algorithm is set equal to the matrix being approximated (i.e., $X_o = X$). This algorithm produces a matrix sequence that is guaranteed to converge to a positive-semidefinite matrix with the required Toeplitz-Hermitian structure. Unfortunately, this solution need not be the closest matrix contained in $C_+ \cap C_{TH}$ that lies closest to X in the Frobenius norm sense. To obtain the optimum approximating matrix contained in $C_+ \cap C_{TH}$, we can alternatively employ the algorithm described by Relationship (14.97).

Example 6. We will now illustrate the point that the sequence of vectors generated by the successive projections algorithm need not always converge to the closest vector in a closed convex set. This will be accomplished by considering the task of finding a positive-semidefinite Toeplitz matrix that lies closest to given matrix. The given matrix X is here taken to be

$$X = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$$

Although this matrix is positive semidefinite, it is not Toeplitz, thereby necessitating the use of a signal restoration algorithm for finding a suitably close positive-semidefinite Toeplitz-Hermitian matrix approximation. The sequence of successive projections algorithm (14.128) has guaranteed convergence to a positive-semidefinite Toeplitz matrix approximation. To be assured of finding the positive-semidefinite Toeplitz matrix that lies closest to X in the Frobenius norm sense, however, it is generally necessary to employ the Dykstra algorithm (14.97). Using the given matrix X as the initial condition, the positive-semidefinite Toeplitz matrices to which these two algorithms converge are found to be

$$X_{\text{ssp}} = \begin{bmatrix} 3.0811 & 2.9230 \\ 2.9230 & 3.0811 \end{bmatrix} \quad \text{and} \quad X_{\text{dyk}} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}$$

Clearly, the positive-semidefinite Toeplitz matrix approximation X_{dyk} lies slightly closer to X in the Frobenius norm sense than does X_{ssp}. Convergence was deemed to have occurred when the normed matrix error $\| X_n - X \| / \| X \|$ became less than 10^{-9}. The successive projection algorithms and Dykstra's algorithm took two and three iterations, respectively, to reach this normed error level.

Exponential Signals and the Data Matrix

In various applications, the basic objective is to approximate a finite sample of a time series by a linear combination of real- and/or complex-valued exponential signals. The set of data to be modeled is taken to be

$$x(1),\ x(2),\ \ldots,\ x(N) \tag{14.129}$$

where N designates the length of the data. It is well known that this data set can be exactly modeled by an exponential signal of order p or less if and only if there exists a nontrivial set of coefficients $a_o, a_1, \ldots, a_p$ such that the following homogeneous relationship is satisfied:

$$a_o x_n + a_1 x_{n-1} + \ldots + a_p x_{n-p} = 0 \quad \text{for } p + 1 \le n \le N \tag{14.130}$$

Upon examination of these relationships, it is clear that nontrivial solutions will always exist when the number of equations is fewer than the number of unknowns (i.e., $N - p < p$). Most data modeling applications, however, are concerned with the distinctly overdetermined case in which $N - p >> p$.

From the above comments, it is apparent that a characterization of the exponential data modeling problem can be obtained by analyzing the linear homogeneous relationships (14.130). It will be convenient to compactly represent these ideal relationships in the vector format

$$X\mathbf{a} = \mathbf{0} \tag{14.131}$$

where $\mathbf{a}$ is the $(p + 1) \times 1$ coefficient vector with elements a_k and X is the corresponding $(N - p) \times (p + 1)$ *data matrix* as specified by

$$\mathbf{X} = \begin{bmatrix} x(p+1) & x(p) & \cdots & x(1) \\ x(p+2) & x(p+1) & \cdots & x(2) \\ \vdots & \vdots & \vdots & \vdots \\ x(N) & x(N-1) & \cdots & x(N-p) \end{bmatrix} \tag{14.132}$$

This data matrix is seen to have a *Toeplitz* structure since the elements along each of its diagonals are equal. Furthermore, if relationship (14.131) is to have a nontrivial solution, it is clear that the rank of data matrix X must be equal to or less than p. These salient attributes play a critical role in various exponential modeling algorithms, and they are now formalized.

Theorem 12. The data set $\{x_1, x_2, \ldots, x_N\}$ is exactly representable as a qth-order exponential signal if and only if the associated $(N - p) \times (p + 1)$ Toeplitz-structured data matrix (14.132) has exactly q nonzero singular values provided that $q \le p$ and $N - p > p$.

The exponential modeling characterization spelled out in this theorem is only applicable to data that is exactly represented by an exponential model. In most practical applications, however, it is found that the data being analyzed can only be approximately represented by an exponential model of reasonably small order. For such situations, it is conceptually possible to employ the concept of signal restoration to slightly perturb the given data set so that the perturbed data set is exactly represented by a qth-order exponential model. To achieve this objective, we need to introduce signal attributes that facilitate this goal. From Theorem 12, it is apparent that the ideal data matrix should be contained in the two attribute sets

$$C^{(q)} = \{Y \in C^{(N-p)\times(p+1)} \quad \text{which have rank } q\}$$

$$C_T = \{Y \in C^{(N-p)\times(p+1)} \quad \text{which have Toeplitz structure}\}$$

The attribute set C_T is a closed subspace and therefore possesses a prerequisite property needed for signal restoration. On the other hand, attribute set $C^{(q)}$ is not convex, which seemingly precludes us from using the sequence of successive projections algorithm for signal restoration. Theorem 8, however, indicates that the associated rank-q operator $P^{(q)}$ is closed. We may therefore employ the sequence of successive projections algorithm to effect the desired signal restoration. This algorithm takes the form

$$X_k = P_T P^{(q)}(X_{k-1}) \quad \text{for } k \geq 1 \tag{14.133}$$

where the projection operators P_T and $P^{(q)}$ are described in Theorems 10 and 8, respectively. The initial data matrix used in this iterative scheme is set equal to the given data matrix (14.132), that is, $X_o = X$.

To implement algorithm (14.133), we first generate the rank-q approximation of the data matrix X. The corresponding matrix $P^{(q)}(X)$ is generally found to be non-Toeplitz in structure. To recover the prerequisite Toeplitz structure, we next apply projection operator P_T to matrix $P^{(q)}(X)$ to complete the first iteration of the signal restoration algorithm. It is generally found that this new Toeplitz-structured data matrix $X_1 = P_T P^{(q)}(X)$ has full rank. It is closer to a rank-q matrix, however, than was the original data matrix X. The first iteration has therefore led to a data matrix whose elements comprise a data sequence that is more compatible with a qth-order exponential model. Often, this first iteration is sufficient in many modeling applications.

To obtain a data sequence that is exactly representable by a qth-order exponential model, we continue this iterative process in an obvious manner. In particular, one sequentially computes the data matrices $X_{k+1} = P_T P^{(q)}(X_k)$ for $k = 0, 1, 2, \ldots$ until the rank of data matrix X_{k+1} is deemed sufficiently close to q. Since the projection operator $P^{(q)}$ and P_T are each closed, we are assured that this iterative process will eventually converge to a Toeplitz-structured data matrix of rank q. It has been empirically determined that the algorithmic process converges in a rapid fashion and typically takes from three to ten iterations for small-dimensioned matrices. Furthermore, the resulting enhanced data matrix has data elements that generally provide a better representation of the underlying signal components than did the original data. The restoration process has therefore effectively stripped away noise that contaminates the original data. We will now examine a special case of data restoration that has important practical applications.

Sinusoidal Signal Identification

In a surprisingly large number of important signal processing applications, the primary objective is that of identifying sinusoidal components in noise-corrupted data. For example, multiple plane waves incident on an equispaced linear array produce complex sinusoidal steering vectors. To identify sinusoidal signals in data, a widely employed procedure is to first form the data matrix whose upper and lower halves correspond to the forward and backward prediction equations associated with the data, respectively. If the data under analysis is specified by $x(1), x(2), \ldots, x(N)$, the associated forward-backward data matrix then takes the form

$$X_{fb} = \begin{bmatrix} X \\ \cdots \\ J_{N-p}\overline{X}J_{p+1} \end{bmatrix} \tag{14.134}$$

In this expression, the forward data matrix X is given by (14.132) while J_n designates the *order reversal matrix* whose elements are all zero except for ones which appear along its main antidiagonal [i.e., $J_n(i,j) = \delta(i, n + 1 - j)$]. The matrix $J_{N-p}\overline{X}J_{p+1}$ appearing in the lower half of the data matrix (14.134) corresponds to the backward prediction equations. The matrices X and $J_{N-p}\overline{X}J_{p+1}$ are seen to have a Toeplitz and Hankel structure, respectively. The combined forward-backward data matrix X_{fb} is therefore said to have a block Toeplitz-Hankel structure.

If the data is noise-free and composed of q complex sinusoids, then the block Toeplitz-Hankel data matrix X_{fb} has rank q. Various procedures for identifying the q (with $q < p$) complex sinusoidal signal components when noise is present have been proposed. Two related SVD-based methods that appeared at the same time have proven effective for this purpose and are now briefly described. In each method, the forward-backward data matrix (14.134) is first decomposed as

$$X_{fb} = [\mathbf{x}_1 \; X_r] \tag{14.135}$$

where $\mathbf{x}_1$ denotes the first column of X_{fb} and X_r its remaining p columns. In the method developed by the author [Cadzow, 1982], the rank-q approximation of the total forward-backward data matrix X is first determined using the truncated SVD (i.e., $X_{fb}^{(q)}$). Finally, the related coefficient vector is then specified by

$$\mathbf{a}_c^o = -[X_r^{(q)}]^\dagger \mathbf{x}_1^{(q)} \tag{14.136}$$

where † designates the pseudo matrix inverse operator while $\mathbf{x}_1^{(q)}$ and $X_r^{(q)}$ are the first and remaining p columns, respectively, of the rank-q approximation matrix $X^{(q)}$. In a very similar fashion, the Tufts-Kumaresan method is obtained by first determining the rank-q approximation of submatrix X_r which is here denoted by $X_{rkt}^{(q)}$ [Tufts and Kumaresan, 1982]. The corresponding coefficient vector is then given by

$$\mathbf{a}_{kt}^o = -[X_{rkt}^{(q)}]^\dagger \mathbf{x}_1 \tag{14.137}$$

It is to be noted that although these two coefficient vectors are similar, the latter approach excludes the first column of X in the rank-q approximation. As such, it does not achieve the full benefits of the SVD decomposition and therefore typically yields marginally poorer performance, as the example to follow illustrates. In both methods, the component sinusoids may be graphically identified by peaks that appear in the *detection functional*

$$d(f) = \frac{1}{\left| \sum_{n=0}^{p} a_n^o e^{j2\pi fn} \right|} \tag{14.138}$$

Rank-Reduced Data Matrix Enhancement

Although these algorithms are effective in identifying sinusoidal components, the application of signal restoration can improve their performance. In particular, one simply applies the signal restoration algorithm (14.133) with mapping P_T replaced by P_{TH}. The restoration algorithm for determining that rank-q data matrix with the block Toeplitz-Hankel structure (14.134) that approximates X is then given by

$$X_k = P_{TH} P^{(q)}(X_{k-1}) \quad \text{for } k \geq 1 \tag{14.139}$$

The mapping $P_{TH}(X)$ determines the block Toeplitz-Hankel matrix that lies closest to matrix X in the Frobenius norm. Implementation of $P_{TH}(X)$ is realized in a fashion similar to $P_T(X)$. The modified matrix achieved through iteration (14.139) is then used in expression (14.136) or (14.137) to provide an enhanced coefficient vector estimate.

Example 7. To illustrate the effectiveness of signal restoration preprocessing let us consider the following data set

$$x(n) = e^{j2\pi(0.25)n} + e^{j2\pi(0.26)n} + e^{j2\pi(0.29)n} + w(n) \quad 1 \leq n \leq 24 \tag{14.140}$$

where $w(n)$ is Gaussian white noise whose real and imaginary components have standard deviation 0.05. When estimation procedures (14.136) and (14.137) are applied to the original data with $p = 17$ (the choice advocated by Tufts and Kumaresan [1982]) and $m = 3$, the spectral estimates shown in Fig. 14.20 arise. Each estimate

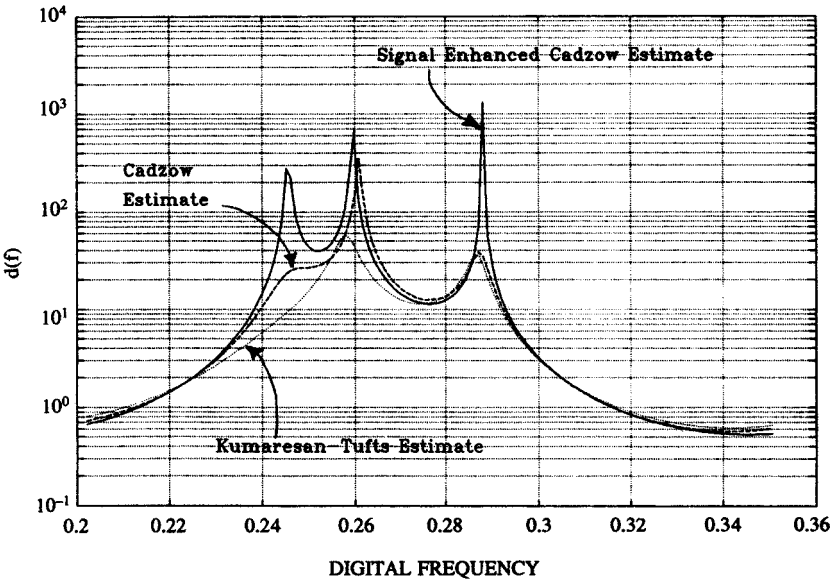


FIGURE 14.20 Sinusoid estimates.

produces two clear peaks with the Cadzow estimate also providing a hint of a third sinusoidal peak. When signal restoration is applied to the original block Toeplitz-Hankel data matrix (14.134), the enhanced Cadzow estimate also shown in Fig. 14.20 arises. This enhanced estimate clearly identifies the three sinusoids and their corresponding frequency estimates in an accurate fashion. The advantages accrued through signal restoration are made evident by this example.

Subsequence Restoration

It is possible to obtain further performance improvements in modeling data as a linear combination of exponentials by employing the concept of *data decimation*. As an example, any data sequence may be decomposed into two subsequences that are composed of its even and odd samples, respectively. If the data under analysis is exactly modeled as a linear combination of exponential signals, it is a simple matter to establish that the associated even and odd decimated subsequences are similarly characterized. The component exponentials in the even and odd decimated subsequences are found to equal the square of those in the original data. This decomposition procedure can be continued in an obvious fashion to generate an additional three subsequences, each of which is composed of every third sample of the original data, and so forth. This data decimation procedure has been combined with the signal restoration technique presented in this subsection to effect improved estimation performance. The interested reader will find this approach described in Cadzow and Wilkes [1991].

Recursive Modeling of Data

The linear recursive modeling of excitation-response data is of interdisciplinary interest in a variety of applications. For purposes of simplicity, we will only deal here with the case in which the data is dependent on a single time variable. The procedure to be described, however, is readily extended to the multidimensional time variable case. In the one-dimensional time case, there is given the pair of data sequences

$$(x(n),y(n)) \quad \text{for} \quad 0 \le n \le N \tag{14.141}$$

We will refer to $x(n)$ and $y(n)$ as being the excitation and response sequences, respectively. Without loss of generality, the measurement time interval has been selected to be $[0,N]$. The pair of data sequences (14.141)

is said to be recursively related if there exist a_k and b_k coefficients such that the following recursive relationship is satisfied:

$$y(n) + \sum_{k=1}^{p} a_k y(n-k) = \sum_{k=0}^{q} b_k x(n-k) \quad \text{for } 0 \le n \le N \tag{14.142}$$

In specifying the time interval over which this recursive relationship holds to be $0 \le n \le N$, it has been tacitly assumed that the sequence pairs are identically zero prior to $n = 0$. If this is not the case, then the time interval over which relationship (14.142) holds must be changed to $\max(p,q) \le n \le N$. In the analysis to follow, it is assumed that the appropriate time interval is $0 \le n \le N$. Modification of this analysis for the time interval $\max(p,q) \le n \le N$ is straightforward and not given.

It will be convenient to represent recursive relationships (14.142) in a matrix format so as to draw upon algebraic attributes that characterize an associated data matrix. This matrix representation takes the form

$$\begin{bmatrix} y(0) & 0 & \cdots & 0 \\ y(1) & y(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ y(N) & y(N-1) & \cdots & y(N-p) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x(1) & x(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ x(N) & x(N-1) & \cdots & x(N-q) \end{bmatrix} \begin{bmatrix} b_o \\ b_1 \\ \vdots \\ b_q \end{bmatrix}$$

or, equivalently,

$$Y_p \mathbf{a}_p = X_q \mathbf{b}_q \tag{14.143}$$

In this latter representation, Y_p and X_q are referred to as the $(N+1) \times (p+1)$ *response matrix* and the $(N+1) \times (q+1)$ *excitation matrix*, respectively. Similarly, $\mathbf{a}_p$ and $\mathbf{b}_q$ are the recursive coefficient vectors identifying the recursive operator. With this preliminary development, the basic attributes characterizing recursively related data are now formally spelled out [see Cadzow and Solomon, 1988].

Theorem 13. Let the excitation-response data $(x(n),y(n))$ for $0 \le n \le N$ be related through a reduced-order recursive relationship $(\tilde{p}, \tilde{q})$ in which $\tilde{p} \le p$ and $\tilde{q} \le q$. It then follows that the extended-order recursive relationship (14.143) will always have a solution. Moreover, if the excitation and response matrices are full rank so that

$$\text{rank}[X_q] = q + 1 \quad \text{and} \quad \text{rank}[Y_p] = p + 1 \tag{14.144}$$

then all solutions are expressible as

$$\begin{bmatrix} \mathbf{a}_p \\ \cdots \\ \mathbf{b}_q \end{bmatrix} = \sum_{k=1}^{s} \alpha_k \mathbf{v}_k \tag{14.145}$$

with the α_k parameters selected to ensure that the first component of $\mathbf{a}_p$ is one as required. The upper limit in this sum is given by $s = 1 + \min(p - \tilde{p}, q - \tilde{q})$ while the vectors $\mathbf{v}_k$ correspond to the s eigenvectors associated with zero eigenvalue of multiplicity s of matrix $D_{p,q}^{*} D_{p,q}$ where $D_{p,q}$ is the $(N+1) \times (p+q+2)$ composite data matrix

$$D_{p,q} = [\, Y_p \,\vdots\, -X_q \,] \tag{14.146}$$

Furthermore, the transfer function associated with any solution to the system of equations (14.145) reduces to (after pole-zero cancelation) the underlying reduced-order transfer function of order $(\tilde{p}, \tilde{q})$.

When using the algebraic characteristics of the composite matrix to form a rational model of empirical data pairs, there is much to be gained by using an over-ordered model. By taking an over-ordered approach, the recursive model parameters are made less sensitive to quirks in the empirical data. A more detailed explanation of this concept is found in Cadzow and Solomon [1988].

Signal-Enhanced Data Modeling

From the above development, it follows that when the observed data $\{(x(n),y(n))\}$ are perfectly represented by a recursive relationship of order $(\tilde{p}, \tilde{q})$, the composite data matrix will satisfy the two attributes

- $D_{p,q}$ is a block Toeplitz matrix
- $D_{p,q}$ has nullity $s = 1 + \min(p - \tilde{p}, q - \tilde{p})$

In most practical applications, the given data observations are not perfectly represented by a low-order recursive relationship. This is typically manifested in the composite data matrix being full rank. To use the concept of signal restoration to achieve a suitably good approximate recursive model, we could suitably modify the given excitation-response data so that the modified data has an associated composite data matrix which satisfies the above two attributes. The signal restoration algorithm associated with this objective is given by

$$D_k = P_T P^{(p+q+2-s)}(D_{k-1}) \quad \text{for } k \geq 1 \tag{14.147}$$

where the initial composite matrix $D_o = [Y_p \vdots -X_q]$ has the given original excitation-response data as entries. We have dropped the subscript p,q in the composite data matrix to simplify the notation.

The signal restoration theorem ensures that the composite data matrix sequence (14.147) will contain a subsequence that converges to a composite data matrix which satisfies the prerequisite block Toeplitz–nullity s attributes. The recursive coefficient vectors as specified by (14.145) when applied to the convergent composite data matrix will typically give a satisfactory model of the data. It should be noted that in some applications it is known that either the excitation or the response data is accurate and should not be perturbed when applying the operator $P^{(p+q+2-s)}$. This is readily accomplished by inserting the original block after the rank reduction projection mapping $P^{(p+q+2-s)}$ has been applied to D_{k-1}. This is illustrated in the following numerical example.

Example 8. Let us apply the above signal restoration procedure to model recursively an unknown system when the input signal $x(n)$ and a noisy observation of the output signal $y(n)$ are available. The previously described signal-enhanced data modeling technique will be used with $p = q$ for simplicity. Clearly, if the excitation-response data were noiseless and the unknown system could be modeled by an autoregressive moving average (ARMA) (p,p) system, $D_{p,p}$ would not have full rank. The presence of noise in the response data will cause $D_{p,p}$ to have full rank, and the signal restoration algorithm will be applied to produce a block Toeplitz matrix having nullity $s = 1$. Since the input data is known exactly, its block will be inserted after each low rank approximation step. From (14.145) above it is clear that the resulting solution for the $\mathbf{a}_p$ and $\mathbf{b}_p$ coefficients will consist of the eigenvector associated with the zero eigenvalue of $D^*_{p,p}D_{p,p}$.

The system to be identified has the following ARMA relationship:

$$y_a(n) - 1.5y_a(n-1) + 0.7y_a(n-2) = x(n-1) + 0.5x(n-2) \tag{14.148}$$

and the observed output is $y(n) = y_a(n) + w(n)$, where $w(n)$ is the measurement noise at the output. In this example, the input signal is zero-mean unit variance white Gaussian noise. The signal-to-noise ratio at the output is 12 dB, and 300 samples of the input and output signals are used. The results for $p = 2$ are shown in Fig. 14.21; the true frequency response is given by the solid line, the dotted line is the solution that would result if no signal restoration were performed, and the dashed line depicts the solution after signal restoration (25 iterations).

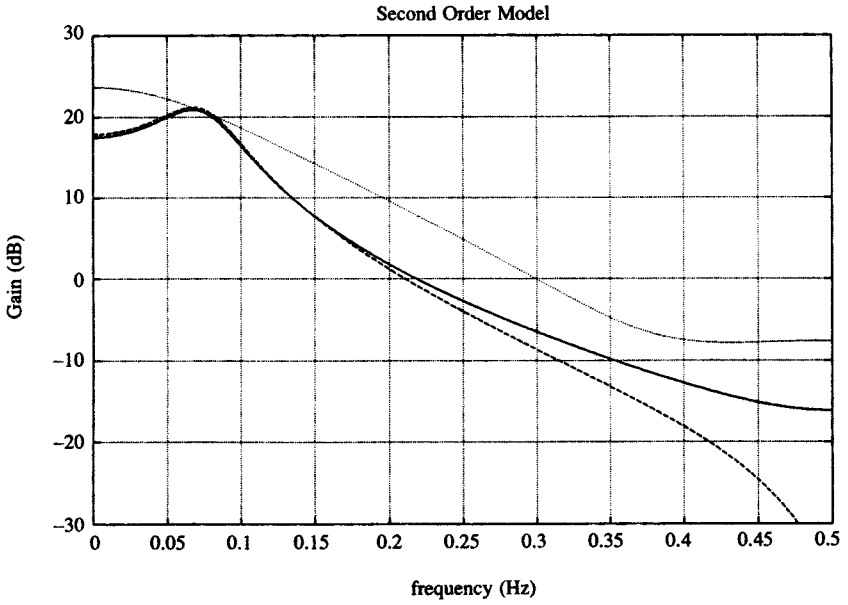


FIGURE 14.21 Second-order model results. —: true; ---: signal enhanced; ····: no enhancement.

Conclusion

The signal restoration algorithm has been shown to provide a useful means for solving a variety of important signal processing problems. In addition to the problems described in this chapter, it has been successfully applied to the missing data problem, deconvolution, and high-dimensional filter synthesis. Very useful results can be achieved by innovatively introducing signal attributes that characterize the underlying information signals.

Defining Terms

Attribute set: A set of vectors (signals) lying in a metric space that possess prescribed properties.

Closed convex sets: A set of vectors C such that if $\mathbf{x},\mathbf{y} \in C$ then $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in C$ for all $0 \le \lambda \le 1$.

Method of successive projections: An iterative procedure for modifying a signal so that the modified signal has properties which match an ideal objective.

Projection mapping: A mathematical procedure for determining a vector (signal) lying in a prescribed set that lies closest to a given vector.

Signal restoration: The restoring of data that has been corrupted by instrumentation dynamics and noise.

Singular value decomposition: A procedure for representing a matrix as a sum of positive weighted orthogonal outer products.

Structured matrix set: A set of common dimensioned matrices that have a prescribed algebraic structure (e.g., Toeplitz, Hankel, Hermitian).

Related Topic

21.1 MATLAB Environment

References

L.M. Bregman, "The method of successive projection for finding the common point of convex sets," *Soviet Mathematics-Doklady,* vol. 6, pp. 688–692, 1965.

J.A. Cadzow, "Signal enhancement: A composite property mapping algorithm," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. ASSP-36, no. 1, pp. 49–62, January 1988.

J.A. Cadzow, "Spectral estimation: An overdetermined rational model equation approach," *Proc. IEEE, Special Issue on Spectral Analysis,* pp. 907–939, September 1982.
J.A. Cadzow and O.M. Solomon, "Algebraic approach to system identification," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. ASSP-34, no. 3, pp. 462–469, June 1988.
J.A. Cadzow and Y. Sun, "Sequences with positive semidefinite Fourier transforms," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. ASSP-34, no. 6, pp. 1502–1510, December 1986.
J.A. Cadzow and D.M. Wilkes, "Enhanced rational signal modeling," *Signal Processing,* vol. 25, no. 2, pp. 171–188, November 1991.
P.L. Combettes and H.J. Trussel, "Method of successive projections for finding a common point of sets in metric spaces," *JOTA,* vol. 67, no. 3, pp. 487–507, December 1990.
R.L. Dykstra, "An algorithm for restricted least squares regression," *Journal Amer. Stat. Assoc.*, vol. 78, pp. 837–842, 1983.
R.L. Dykstra and J.P. Boyle, "An algorithm for least squares projection onto the intersection of translated, convex cones," *Journal Statistical Plann. Inference,* vol. 15, pp. 391–399, 1987.
N. Gaffke and R. Mathar, "A cyclic projection algorithm via duality," *Metrika,* vol. 36, pp. 29–54, 1989.
L.G. Gubin, B.T. Polyak, and E.V. Raik, "The method of projections for finding the common point of sets," *USSR Computational Mathematics and Mathematical Physics,* vol. 7, pp. 1–24, 1967.
I. Halperin, "The product of projection operators," *Acta Scientiarum Mathematicarum,* vol. 23, pp. 96–99, 1962.
S.-P. Han, "A successive projection method," *Mathematical Programming,* vol. 40, pp. 1–14, 1988.
D.G. Luenberger, *Optimization by Vector Space Methods,* New York: John Wiley, 1969.
H.D. Mittelmann and J.A. Cadzow, "Continuity of closest rank-p approximations to matrices," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-35, no. 8, pp. 1211–1212, August 1987.
N. Ottavy, "Strong convergence of projection-like methods in Hilbert spaces," *J. Optimization Theory and Applications,* vol. 56, pp. 433–461, 1988.
W.J. Stiles, "Closest point maps and their product," *Nieuw Archief voor Wiskunde,* vol. 13, pp. 212–225, 1965.
H.J. Trussel and M.R. Civanlar, "The feasible solution in signal restoration," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. ASSP-32, pp. 201–212, 1984.
D.W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proc. IEEE, Special Issue on Spectral Analysis,* pp. 975–989, September 1982.
J. von Neumann, *Functional Operators,* vol. 2 (Annals of Mathematics Studies, no. 22), Princeton, N.J., 1950. Reprinted from mimeographed lecture notes first distributed in 1933.
D.C. Youla, "Generalized image restoration by the method of alternating orthogonal projections," *IEEE Trans. Circuits and Systems,* vol. CAS-25, pp. 694–702, Sept. 1978.
W.I. Zangwill, *Nonlinear Programming: A Unified Approach,* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Further Information

The monthly *IEEE Transactions on Acoustics, Speech, and Signal Processing* frequently publishes articles on the theory and application of signal restoration and recovery. Signal restoration concepts were discussed in articles published in the January 1988 issue (pp. 49–62), in the March 1989 issue (pp. 393–401), and in the May 1990 issue (pp. 778–786).

The *IEEE Transactions on Circuits and Systems* also publishes signal restoration application papers. Examples are to be found in the September 1975 issue (pp. 735–742) and the September 1978 issue (pp. 694–702).

Image restoration articles appear in the *IEEE Transactions on Medical Imaging* as illustrated by the articles that appeared in the October 1992 issue (pp. 81–94) and the January 1984 issue (pp. 91–98).

McClellan, S., Gibson, J.D., Ephraim, Y., Fussell, J.W., Wilcox, L.D., Bush, M.A., Gao, Y., Ramabhadran, B., Picheny, M. "Speech Signal Processing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

Stan McClellan
University of Alabama at Birmingham

Jerry D. Gibson
Texas A&M University

Yariv Ephraim
AT&T Bell Laboratories
George Mason University

Jesse W. Fussell
Department of Defense

Lynn D. Wilcox
FX Palo Alto Lab

Marcia A. Bush
Xerox Palo Alto Research Center

Yuqing Gao
IBM
T.J. Watson Research Center

Bhuvana Ramabhadran
IBM
T.J. Watson Research Center

Michael Picheny
IBM
T.J. Watson Research Center

15

Speech Signal Processing

15.1 Coding, Transmission, and Storage
General Approaches • Model Adaptation • Analysis-by-Synthesis • Particular Implementations • Speech Quality and Intelligibility • Standardization • Variable Rate Coding • Summary and Conclusions

15.2 Speech Enhancement and Noise Reduction
Models and Performance Measures • Signal Estimation • Source Coding • Signal Classification • Comments

15.3 Analysis and Synthesis
Analysis of Excitation • Fourier Analysis • Linear Predictive Analysis • Homomorphic (Cepstral) Analysis • Speech Synthesis

15.4 Speech Recognition
Speech Recognition System Architecture • Signal Pre-Processing • Dynamic Time Warping • Hidden Markov Models • State-of-the-Art Recognition Systems

15.5 Large Vocabulary Continuous Speech Recognition
Overview of a Speech Recognition System • Hidden Markov Models As Acoustic Models for Speech Recognition • Speaker Adaptation • Modeling Context in Continuous Speech • Language Modeling • Hypothesis Search • State-of-the-Art Systems • Challenges in Speech Recognition • Applications

15.1 Coding, Transmission, and Storage

Stan McClellan and Jerry D. Gibson

Interest in speech coding is motivated by a wide range of applications, including commercial telephony, digital cellular mobile radio, military communications, voice mail, speech storage, and future personal communications networks. The goal of speech coding is to represent speech in digital form with as few bits as possible while maintaining the intelligibility and quality required for the particular application. At higher bit rates, such as 64 and 32 kbits/s, achieving good quality and intelligibility is not too difficult, but as the desired bit rate is lowered to 16 kbits/s and below, the problem becomes increasingly challenging. Depending on the application, many difficult constraints must be considered, including the issue of complexity.

For example, for the 32-kbits/s speech coding standard, the ITU-T[1] not only required highly intelligible, high-quality speech, but the coder also had to have low delay, withstand independent bit error rates up to 10^{-2}, have acceptable performance degradation for several synchronous or asynchronous tandem connections, and pass some voiceband modem signals. Other applications may have different criteria. Digital cellular mobile radio in the U.S. has no low delay or voiceband modem signal requirements, but the speech data rates required are under 8 kbits/s and the transmission medium (or channel) can be very noisy and have relatively long fades. These considerations affect the speech coder chosen for a particular application.

As speech coder data rates drop to 16 kbits/s and below, perceptual criteria taking into account human auditory response begin to play a prominent role. For *time domain coders,* the perceptual effects are incorporated using a frequency-weighted error criterion. The *frequency-domain coders* include perceptual effects by allocating

[1]International Telecommunications Union, Telecommunications Standardization Sector, formerly the CCITT.

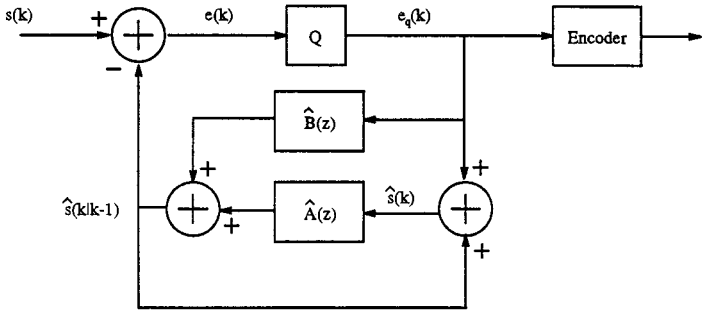


FIGURE 15.1 Differential encoder transmitter with a pole-zero predictor.

The focus of this article is the contrast among the three most important classes of speech coders that have representative implementations in several international standards—**time-domain coders, frequency-domain coders**, and **hybrid coders**. In the following, we define these classifications, look specifically at the important characteristics of representative, general implementations of each class, and briefly discuss the rapidly changing national and international standardization efforts related to speech coding.

General Approaches

Time Domain Coders and Linear Prediction

Linear Predictive Coding (LPC) is a modeling technique that has seen widespread application among time-domain speech coders, largely because it is computationally simple and applicable to the mechanisms involved in speech production. In LPC, general spectral characteristics are described by a parametric model based on estimates of autocorrelations or autocovariances. The model of choice for speech is the all-pole or *autoregressive* (*AR*) *model*. This model is particularly suited for voiced speech because the vocal tract can be well modeled by an all-pole transfer function. In this case, the estimated LPC model parameters correspond to an AR process which can produce waveforms very similar to the original speech segment. Differential Pulse Code Modulation (DPCM) coders (i.e., ITU-T G.721 ADPCM [CCITT, 1984]) and LPC vocoders (i.e., U.S. Federal Standard 1015 [National Communications System, 1984]) are examples of this class of time-domain predictive architecture. Code Excited Coders (i.e., ITU-T G728 [Chen, 1990] and U.S. Federal Standard 1016 [National Communications System, 1991]) also utilize LPC spectral modeling techniques.[1]

Based on the general spectral model, a predictive coder formulates an estimate of a future sample of speech based on a weighted combination of the immediately preceding samples. The error in this estimate (the *prediction residual*) typically comprises a significant portion of the data stream of the encoded speech. The residual contains information that is important in speech perception and cannot be modeled in a straightforward fashion. The most familiar form of predictive coder is the classical Differential Pulse Code Modulation (DPCM) system shown in Fig. 15.1. In DPCM, the predicted value at time instant k, $\hat{s}(k|k-1)$, is subtracted from the input signal at time k, $s(k)$, to produce the prediction error signal $e(k)$. The prediction error is then approximated (*quantized*) and the quantized prediction error, $e_q(k)$, is coded (represented as a binary number) for transmission to the receiver. Simultaneously with the coding, $e_q(k)$ is summed with $\hat{s}(k|k-1)$ to yield a reconstructed version of the input sample, $\hat{s}(k)$. Assuming no channel errors, an identical reconstruction, distorted only by the effects of quantization, is accomplished at the receiver. At both the transmitter and receiver, the predicted value at time instant $k+1$ is derived using reconstructed values up through time k, and the procedure is repeated.

The first DPCM systems had $\hat{B}(z) = 0$ and $\hat{A}(z) = \sum_{i=1}^{N} a_i z^{-i}$, where $\{a_i, i = 1 \ldots N\}$ are the LPC coefficients and z^{-1} represents unit delay, so that the predicted value was a weighted linear combination of previous reconstructed values, or

[1]However, codebook excitation is generally described as a *hybrid* coding technique.

$$\hat{s}\left(k\middle|k-1\right) = \sum_{i=1}^{N} a_i \hat{s}\left(k-i\right). \tag{15.1}$$

Later work showed that letting $\hat{B}(z) = \sum_{j=1}^{M} b_j z^{-j}$ improves the perceived quality of the reconstructed speech[1] by shaping the spectrum of the quantization noise to match the speech spectrum, as well as improving noisy-channel performance [Gibson, 1984]. To produce high-quality, highly intelligible speech, it is necessary that the quantizer and predictor parameters be adaptive to compensate for nonstationarities in the speech waveform.

Frequency Domain Coders

Coders that rely on spectral decomposition often use the usual set of sinusoidal basis functions from signal theory to represent the specific short-time spectral content of a segment of speech. In this case, the approximated signal consists of a linear combination of sinusoids with specified amplitudes and arguments (frequency, phase). For compactness, a countable subset of harmonically related sinusoids may be used. The two most prominent types of frequency domain coders are subband coders and multi-band coders.

Subband coders digitally filter the speech into nonoverlapping (as nearly as possible) frequency bands. After filtering, each band is decimated (effectively sampled at a lower rate) and coded separately using PCM, DPCM, or some other method. At the receiver, the bands are decoded, upsampled, and summed to reconstruct the speech. By allocating a different number of bits per sample to the subbands, the perceptually more important frequency bands can be coded with greater accuracy. The design and implementation of subband coders and the speech quality produced have been greatly improved by the development of digital filters called quadrature mirror filters (QMFs) [Johnston, 1980] and polyphase filters. These filters allow subband overlap at the encoder, which causes aliasing, but the reconstruction filters at the receiver can be chosen to eliminate the aliasing if quantization errors are small.

Multi-band coders perform a similar function by characterizing the contributions of individual sinusoidal components to the short-term speech spectrum. These parameters are then quantized, coded, transmitted, and used to configure a bank of tuned oscillators at the receiver. Outputs of the oscillators are mixed in proportion to the distribution of spectral energy present in the original waveform. An important requirement of multi-band coders is a capability to precisely determine perceptually significant spectral components and track the evolution of their energy and phase. Recent developments related to multi-band coding emphasize the use of harmonically related components with carefully intermixed spectral regions of bandlimited white noise. Sinusoidal Transform Coders (STC) and Multi-Band Excitation coders (MBE) are examples of this type of frequency domain coders.

Model Adaptation

Adaptation algorithms for coder predictor or quantizer parameters can be loosely grouped based on the signals that are used as the basis for adaptation. Generally, *forward adaptive* coder elements analyze the input speech (or a filtered version of it) to characterize predictor coefficients, spectral components, or quantizer parameters in a blockwise fashion. *Backward adaptive* coder elements analyze a reconstructed signal, which contains quantization noise, to adjust coder parameters in a sequential fashion. Forward adaptive coder elements can produce a more efficient model of speech signal characteristics, but introduce delay into the coder's operation due to buffering of the signal. Backward adaptive coder elements do not introduce delay, but produce signal models that have lower fidelity with respect to the original speech due to the dependence on the noisy reconstructed signal. Most low-rate coders rely on some form of forward adaptation. This requires moderate to high delay in processing for accuracy of parameter estimation (autocorrelations/autocovariances for LPC-based coders, sinusoidal resolution for frequency-domain coders). The allowance of significant delay for many coder architectures has enabled a spectrally matched pre- or post-processing step to reduce apparent quantization noise and provide significant perceptual improvements. Perceptual enhancements combined with **analysis-by-synthesis** optimization, and enabled by recent advances in high-power computing architectures such as digital signal processors, have tremendously improved speech coding results at medium and low rates.

[1] In this case, the predicted value is $\hat{s}(k|\,k-1) = \sum_{i=1}^{N} a_i \hat{s}\left(k-i\right) + \sum_{i=1}^{M} b_j e_q\left(k-j\right)$.

Analysis-by-Synthesis

A significant drawback to traditional "instantaneous" coding approaches such as DPCM lies in the perceptual or subjective relevance of the distortion measure and the signals to which it is applied. Thus, the advent of analysis-by-synthesis coding techniques poses an important milestone in the evolution of medium- to low-rate speech coding. An analysis-by-synthesis coder chooses the coder excitation by minimizing distortion between the original signal and the set of synthetic signals produced by every possible codebook excitation sequence.

In contrast, time-domain predictive coders must produce an estimated prediction residual (innovations sequence) to drive the spectral shaping filter(s) of the LPC model, and the classical DPCM approach is to quantize the residual sequence directly using scalar or vector quantizers. The incorporation of frequency-weighted distortion in the optimization of analysis-by-synthesis coders is significant in that it de-emphasizes (increases the tolerance for) quantization noise surrounding spectral peaks. This effect is perceptually transparent since the ear is less sensitive to error around frequencies having higher energy [Atal and Schroeder, 1979].

This approach has resulted in significant improvements in low-rate coder performance, and recent increases in processor speed and power are crucial enabling techniques for these applications. Analysis-by-synthesis coders based on linear prediction are generally described as hybrid coders since they fall between waveform coders and vocoders.

Particular Implementations

Currently, three coder architectures dominate the fields of medium and low-rate speech coding:

- *Code-Excited Linear Prediction* (*CELP*): an LPC-based technique which optimizes a vector of excitation samples (and/or pitch filter and lag parameters) using analysis-by-synthesis.
- *Multi-Band Excitation* (*MBE*): a direct spectral estimation technique which optimizes the spectral reconstruction error over a set of subbands using analysis-by-synthesis.
- *Mixed-Excitation Linear Prediction* (*MELP*): an optimized version of the traditional LPC vocoder which includes an explicit multiband model of the excitation signal.

Several realizations of these approaches have been adopted nationally and internationally as standard speech coding architectures at rates below 16 kbits/s (i.e., G.728, IMBE, U.S. Federal Standard 1016, etc.). The success of these implementations is due to LPC-based analysis-by-synthesis with a perceptual distortion criterion or short-time frequency-domain modeling of a speech waveform or LPC residual. Additionally, the coders that operate at lower rates all benefit from forward adaptation methods which produce efficient, accurate parameter estimates.

CELP

The general CELP architecture is described as a blockwise analysis-by-synthesis selection of an LPC excitation sequence. In low-rate CELP coders, a forward-adaptive linear predictive analysis is performed at 20 to 30 msec intervals. The gross spectral characterization is used to reconstruct, via linear prediction, candidate speech segments derived from a constrained set of plausible filter excitations (the "codebook"). The excitation vector that produces the synthetic speech segment with smallest perceptually weighted distortion (with respect to the original speech) is chosen for transmission. Typically, the excitation vector is optimized more often than the LPC spectral model. The use of *vectors* rather than *scalars* for the excitation is significant in bit-rate reduction. The use of perceptual weighting in the CELP reconstruction stage and analysis-by-synthesis optimization of the dominant low-frequency (pitch) component are key concepts in maintaining good quality encoded speech at lower rates. CELP-based speech coders are the predominant coding methodologies for rates between 4 kbits/s and 16 kbits/s due to their excellent subjective performance. Some of the most notable are detailed below.

- ITU-T Recommendation G.728 (LD-CELP) [Chen, 1990] is a low delay, backward adaptive CELP coder. In G.728, a low algorithmic delay (less than 2.5 msec) is achieved by using 1024 candidate excitation sequences, each only 5 samples long. A 50th-order LPC spectral model is used, and the coefficients are backward-adapted based on the transmitted excitation.
- The speech coder standardized by the CTIA for use in the U.S. (time-division multiple-access) 8 kbits/s digital cellular radio systems is called vector sum excited linear prediction (VSELP) [Gerson and Jasiuk,

1990]. VSELP is a forward-adaptive form of CELP where two excitation codebooks are used to reduce the complexity of encoding.

- Other approaches to complexity reduction in CELP coders are related to "sparse" codebook entries which have few nonzero samples per vector and "algebraic" codebooks which are based on integer lattices [Adoul and Lamblin, 1987]. In this case, excitation code vectors can be constructed on an as-needed basis instead of being stored in a table. ITU-T standardization of a CELP algorithm which uses lattice-based excitations has resulted in the 8 kbps G.729 (ACELP) coder.
- U.S. Federal Standard 1016 [National Communications System, 1991] is a 4.8 kbps CELP coder. It has both long- and short-term linear predictors which are forward adaptive, and so the coder has a relatively large delay (100 msec). This coder produces highly intelligible, good-quality speech in a variety of environments and is robust to independent bit errors.

Below about 4 kbps, the subjective quality of CELP coders is inferior to other architectures. Much research in variable-rate CELP implementations has resulted in alternative coder architectures which adjust their coding rates based on a number of channel conditions or sophisticated, speech-specific cues such as phonetic segmentation [Wang and Gersho, 1989; Paksoy et al., 1993]. Notably, most variable-rate CELP coders are implementations of finite-state CELP wherein a vector of speech cues controls the evolution of a state-machine to prescribe mode-dependent bit allocations for coder parameters. With these architectures, excellent speech quality at average rates below 2 kbps has been reported.

MBE

The MBE coder [Hardwick and Lim, 1991] is an efficient frequency-domain architecture partially based on the concepts of sinusoidal transform coding (STC) [McAulay and Quatieri, 1986]. In MBE, the instantaneous spectral envelope is represented explicitly by harmonic estimates in several subbands. The performance of MBE coders at rates below 4 kbps is generally "better" than that of CELP-based schemes.

An MBE coder decomposes the instantaneous speech spectrum into subbands centered at harmonics of the fundamental glottal excitation (pitch). The spectral envelope of the signal is approximated by samples taken at pitch harmonics, and these harmonic amplitudes are compared to adaptive thresholds (which may be determined via analysis-by-synthesis) to determine subbands of high spectral activity. Subbands that are determined to be "voiced" are labeled, and their energies and phases are encoded for transmission. Subbands having relatively low spectral activity are declared "unvoiced". These segments are approximated by an appropriately filtered segment of white noise, or a locally dense collection of sinusoids with random phase. Careful tracking of the evolution of individual spectral peaks and phases in successive frames is critical in the implementation of MBE-style coders.

An efficient implementation of an MBE coder was adopted for the International Maritime Satellite (INMARSAT) voice processor, and is known as Improved-MBE, or IMBE [Hardwick and Lim, 1991]. This coder optimizes several components of the general MBE architecture, including grouping neighboring harmonics for subband voicing decisions, using non-integer pitch resolution for higher speaker fidelity, and differentially encoding the log-amplitudes of voiced harmonics using a DCT-based scheme. The IMBE coder requires high delay (about 80 msec), but produces very good quality encoded speech.

MELP

The MELP coder [McCree and Barnwell, 1995] is based on the traditional LPC vocoder model where an LPC synthesis filter is excited by an impulse train (voiced speech) or white noise (unvoiced speech). The MELP excitation, however, has characteristics that are more similar to natural human speech. In particular, the MELP excitation can be a mixture of (possibly aperiodic) pulses with bandlimited noise. In MELP, the excitation spectrum is explicitly modeled using Fourier series coefficients and bandpass voicing strengths, and the time-domain excitation sequence is produced from the spectral model via an inverse transform. The synthetic excitation sequence is then used to drive an LPC synthesizer which introduces formant spectral shaping.

Common Threads

In addition to the use of analysis-by-synthesis techniques and/or LPC modeling, a common thread between low-rate, forward adaptive CELP, MBE, and MELP coders is the dependence on an estimate of the fundamental glottal frequency, or pitch period. CELP coders typically employ a pitch or long-term predictor to characterize

the glottal excitation. MBE coders estimate the fundamental frequency and use this estimate to focus subband decompositions on harmonics. MELP coders perform pitch-synchronous excitation modeling. Overall coder performance is enhanced in the CELP and MBE coders with the use of sub-integer lags [Kroon and Atal, 1991]. This is equivalent to performing pitch estimation using a signal sampled at a higher sampling rate to improve the precision of the spectral estimate. Highly precise glottal frequency estimation improves the "naturalness" of coded speech at the expense of increased computational complexity, and in some cases increased bit rate.

Accurate characterization of pitch and LPC parameters can also be used to good advantage in postfiltering to reduce apparent quantization noise. These filters, usually derived from forward-adapted filter coefficients transmitted to the receiver as side-information, perform post-processing on the reconstructed speech which reduces perceptually annoying noise components [Chen and Gersho, 1995].

Speech Quality and Intelligibility

To compare the performance of two speech coders, it is necessary to have some indicator of the intelligibility and quality of the speech produced by each coder. The term *intelligibility* usually refers to whether the output speech is easily understandable, while the term *quality* is an indicator of how natural the speech sounds. It is possible for a coder to produce highly intelligible speech that is low quality in that the speech may sound very machine-like and the speaker is not identifiable. On the other hand, it is unlikely that unintelligible speech would be called high quality, but there are situations in which perceptually pleasing speech does not have high intelligibility. We briefly discuss here the most common measures of intelligibility and quality used in formal tests of speech coders.

DRT

The diagnostic rhyme test (DRT) was devised by Voiers [1977] to test the intelligibility of coders known to produce speech of lower quality. Rhyme tests are so named because the listener must determine which consonant was spoken when presented with a pair of rhyming words; that is, the listener is asked to distinguish between word pairs such as meat-beat, pool-tool, saw-thaw, and caught-taught. Each pair of words differs on only one of six phonemic attributes: voicing, nasality, sustention, sibilation, graveness, and compactness. Specifically, the listener is presented with one spoken word from a pair and asked to decide which word was spoken. The final DRT score is the percent responses computed according to $P = \frac{1}{T}(R - W) \times 100$, where R is the number correctly chosen, W is the number of incorrect choices, and T is the total of word pairs tested. Usually, $75 \leq \text{DRT} \leq 95$, with a *good* being about 90 [Papamichalis, 1987].

MOS

The **Mean Opinion Score (MOS)** is an often-used performance measure [Jayant and Noll, 1984]. To establish a MOS for a coder, listeners are asked to classify the quality of the encoded speech in one of five categories: excellent (5), good (4), fair (3), poor (2), or bad (1). Alternatively, the listeners may be asked to classify the coded speech according to the amount of perceptible distortion present, i.e., imperceptible (5), barely perceptible but not annoying (4), perceptible and annoying (3), annoying but not objectionable (2), or very annoying and objectionable (1). The numbers in parentheses are used to assign a numerical value to the subjective evaluations, and the numerical ratings of all listeners are averaged to produce a MOS for the coder. A MOS between 4.0 and 4.5 usually indicates high quality.

It is important to compute the variance of MOS values. A large variance, which indicates an unreliable test, can occur because participants do not known what categories such as *good* and *bad* mean. It is sometimes useful to present examples of good and bad speech to the listeners before the test to calibrate the 5-point scale [Papamichalis, 1987]. The MOS values for a variety of speech coders and noise conditions are given in [Daumer, 1982].

DAM

The diagnostic acceptability measure (DAM) developed by Dynastat [Voiers, 1977] is an attempt to make the measurement of speech quality more systematic. For the DAM, it is critical that the listener crews be highly trained and repeatedly calibrated in order to get meaningful results. The listeners are each presented with encoded sentences taken from the Harvard 1965 list of phonetically balanced sentences, such as "Cats and dogs

TABLE 15.1 Speech Coder Performance Comparisons

Algorithm (acronym)	Standardization		Rate	Subjective		
	Body	Identifier	kbits/s	MOS	DRT	DAM
μ-law PCM	ITU-T	G.711	64	4.3	95	73
ADPCM	ITU-T	G.721	32	4.1	94	68
LD-CELP	ITU-T	G.728	16	4.0	94[a]	70[a]
RPE-LTP	GSM	GSM	13	3.5	—	—
VSELP	CTIA	IS-54	8	3.5	—	—
CELP	U.S. DoD	FS-1016	4.8	3.13[b]	90.7[b]	65.4[b]
IMBE	Inmarsat	IMBE	4.1	3.4	—	—
LPC-10e	U.S. DoD	FS-1015	2.4	2.24[b]	86.2[b]	50.3[b]

[a] Estimated.
[b] From results of 1996 U.S. DoD 2400 bits/s vocoder competition.

each hate the other" and "The pipe began to rust while new". The listener is asked to assign a number between 0 and 100 to characteristics in three classifications—signal qualities, background qualities, and total effect. The ratings of each characteristic are weighted and used in a multiple nonlinear regression. Finally, adjustments are made to compensate for listener performance. A typical DAM score is 45 to 55%, with 50% corresponding to a *good* system [Papamichalis, 1987].

The perception of "good quality" speech is a highly individual and subjective area. As such, no single performance measure has gained wide acceptance as an indicator of the quality and intelligibility of speech produced by a coder. Further, there is no substitute for subjective listening tests under the actual environmental conditions expected in a particular application. As a rough guide to the performance of some of the coders discussed here, we present the DRT, DAM, and MOS values in Table 15.1, which is adapted from [Spanias, 1994; Jayant, 1990]. From the table, it is evident that at 8 kbits/s and above, performance is quite good and that the 4.8 kbits/s CELP has substantially better performance than LPC-10e.

Standardization

The presence of international, national, and regional speech coding **standards** ensures the interoperability of coders among various implementations. As noted previously, several standard algorithms exist among the classes of speech coders. The ITU-T (formerly CCITT) has historically been a dominant factor in international standardization of speech coders, such as G.711, G.721, G.728, G.729, etc. Additionally, the emergence of digital cellular telephony, personal communications networks, and multimedia communications has driven the formulation of various national or regional standard algorithms such as the GSM full and half-rate standards for European digital cellular, the CTIA full-rate TDMA and CDMA algorithms and their half-rate counterparts for U.S. digital cellular, full and half-rate Pitch-Synchronous CELP [Miki et al., 1993] for Japanese cellular, as well as speech coders for particular applications [ITU-TS, 1991].

The standardization efforts of the U.S. Federal Government for secure voice channels and military applications have a historically significant impact on the evolution of speech coder technology. In particular, the recent re-standardization of the DoD 2400 bits/s vocoder algorithm has produced some competing algorithms worthy of mention here. Of the classes of speech coders represented among the algorithms competing to replace LPC-10, several implementations utilized STC or MBE architectures, some used CELP architectures, and others were novel combinations of multiband-excitation with LPC modeling [McCree and Barnwell, 1995] or pitch-synchronous prototype waveform interpolation techniques [Kleijn, 1991].

The final results of the U.S. DoD standard competition are summarized in Table 15.2 for "quiet" and "office" environments. In the table, the column labeled "FOM" is the overall Figure of Merit used by the DoD Digital Voice Processing Consortium in selecting the coder. The FOM is a unitless combination of *complexity* and *performance* components, and is measured with respect to FS-1016. The complexity of a coder is a weighted combination of memory and processing power required. The performance of a coder is a weighted combination of four factors: quality (Q—measured via MOS), intelligibility (I—measured via DRT), speaker recognition (R), and communicability (C). Recognizability and communicability for each coder were measured by tests

TABLE 15.2 Speech Coder Performance Comparisons Taken from Results of 1996 U.S. DoD 2400 bits/s Vocoder Competition

Algorithm (acronym)	FOM	Rank	Best	Quiet			Office		
				MOS	DRT	DAM	MOS	DRT	DAM
MELP	2.616	1	I	3.30	92.3	64.5	2.96	91.2	52.7
PWI	2.347	2	Q	3.28	90.5	70.0	2.88	88.4	55.5
STC	2.026	3	R	3.08	89.9	63.8	2.82	91.5	54.1
IMBE	2.991	*	C	2.89	91.4	62.3	2.71	91.1	52.4
CELP	0.0	N/A	—	3.13	90.7	65.4	2.89	89.0	56.1
LPC-10e	–9.19	N/A	—	2.24	86.2	50.3	2.09	85.2	48.4

* Ineligible due to failure of the quality (MOS) criteria minimum requirements (better than CELP) in both quiet and office environments.

comparing processed vs. unprocessed data, and effectiveness of communication in application-specific cooperative tasks [Schmidt-Nielsen and Brock, 1996; Kreamer and Tardelli, 1996]. The MOS and DRT scores were measured in a variety of common DoD environments. Each of the four "finalist" coders ranked first in one of the four categories examined (Q,I,R,C), as noted in the table.

The results of the standardization process were announced in April, 1996. As indicated in Table 15.2, the new 2400 bits/s Federal Standard vocoder replacing LPC-10e is a version of the Mixed Excitation Linear Prediction (MELP) coder which uses several specific enhancements to the basic MELP architecture. These enhancements include multi-stage VQ of the formant parameters based on frequency-weighted bark-scale spectral distortion, direct VQ of the first 10 Fourier coefficients of the excitation using bark-weighted distortion, and a gain coding technique which is robust to channel errors [McCree et al., 1996].

Variable Rate Coding

Previous standardization efforts and discussion here have centered on fixed-rate coding of speech where a fixed number of bits are used to represent speech in digital form per unit of time. However, with recent developments in transmission architectures (such as CDMA), the implementation of *variable-rate* speech coding algorithms has become feasible. In **variable-rate coding**, the average data rate for conversational speech can be reduced by a factor of at least 2.

A variable-rate speech coding algorithm has been standardized by the CTIA for wideband (CDMA) digital mobile cellular telephony under IS-95. The algorithm, QCELP [Gardner et al., 1993], is the first practical variable-rate speech coder to be incorporated in a digital cellular system. QCELP is a multi-mode, CELP-type analysis-by-synthesis coder which uses blockwise spectral energy measurements and a finite-state machine to switch between one of four configurations. Each configuration has a fixed rate of 1, 2, 4, or 8 kbits/s with a predetermined allocation of bits among coder parameters (coefficients, gains, excitation, etc.). The subjective performance of QCELP in the presence of low background noise is quite good since the bit allocations per-mode and mode-switching logic are well-suited to high-quality speech. In fact, QCELP at an average rate of 4 kbits/s has been judged to be MOS-equivalent to VSELP, its 8 kbits/s, fixed-rate cellular counterpart. A time-averaged encoding rate of 4 to 5 kbits/s is not uncommon for QCELP, however the average rate tends toward the 8 kbits/s maximum in the presence of moderate ambient noise. The topic of robust fixed-rate and variable-rate speech coding in the presence of significant background noise remains an open problem.

Much recent research in speech coding below 8 kbits/s has focused on multi-mode CELP architectures and efficient approaches to source-controlled mode selection [Das et al., 1995]. Multimode coders are able to quickly invoke a coding scheme and bit allocation specifically tailored to the local characteristics of the speech signal. This capability has proven useful in optimizing perceptual quality at low coding rates. In fact, the majority of algorithms currently proposed for half-rate European and U.S. digital cellular standards, as well as many algorithms considered for rates below 2.4 kbits/s are multimode coders. The direct coupling between variable-rate (multimode) speech coding and the CDMA transmission architecture is an obvious benefit to both technologies.

Summary and Conclusions

The availability of general-purpose and application-specific digital signal processing chips and the ever-widening interest in digital communications have led to an increasing demand for speech coders. The worldwide desire to establish standards in a host of applications is a primary driving force for speech coder research and development. The speech coders that are available today for operation at 16 kbits/s and below are conceptually quite exotic compared with products available less than 10 years ago. The re-standardization of U.S. Federal Standard 1015 (LPC-10) at 2.4 kbits/s with performance constraints similar to those of FS-1016 at 4.8 kbits/s is an indicator of the rapid evolution of speech coding paradigms and VLSI architectures.

Other standards to be established in the near term include the European (GSM) and U.S. (CTIA) half-rate speech coders for digital cellular mobile radio. For the longer term, the specification of standards for forthcoming mobile personal communications networks will be a primary focus in the next 5 to 10 years.

In the preface to their book, Jayant and Noll [1984] state that "our understanding of speech and image coding has now reached a very mature point …" As of 1997, this statement rings truer than ever. The field is a dynamic one, however, and the wide range of commercial applications demands continual progress.

Defining Terms

Analysis-by-synthesis: Constructing several versions of a waveform and choosing the best match.
Predictive coding: Coding of time-domain waveforms based on a (usually) linear prediction model.
Frequency domain coding: Coding of frequency-domain characteristics based on a discrete time-frequency transform.
Hybrid coders: Coders that fall between waveform coders and vocoders in how they select the excitation.
Standard: An encoding technique adopted by an industry to be used in a particular application.
Mean Opinion Score (MOS): A popular method for classifying the quality of encoded speech based on a five-point scale.
Variable-rate coders: Coders that output different amounts of bits based on the time-varying characteristics of the source.

Related Topics

17.1 Digital Image Processing • 21.4 Example 3: Multirate Signal Processing

References

A. S. Spanias, "Speech coding: A tutorial review," *Proc. IEEE,* 82, 1541–1575, October 1994.

A. Gersho, "Advances in speech and audio compression," *Proc. IEEE,* 82, June 1994.

W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis,* Amsterdam, Holland: Elsevier, 1995.

CCITT, "32-kbit/s adaptive differential pulse code modulation (ADPCM)," *Red Book,* III.3, 125–159, 1984.

National Communications System, Office of Technology and Standards, *Federal Standard 1015: Analog to Digital Conversion of Voice by 2400 bit/second Linear Predictive Coding,* 1984.

J.-H. Chen, "High-quality 16 kb/s speech coding with a one-way delay less than 2 ms," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Albuquerque, NM, pp. 453–456, April 1990.

National Communications System, Office of Technology and Standards, *Federal Standard 1016: Telecommunications: Analog to Digital Conversion of Radio Voice by 4800 bit/second Code Excited Linear Prediction* (*CELP*), 1991.

J. Gibson, "Adaptive prediction for speech encoding," *IEEE ASSP Magazine,* 1, 12–26, July 1984.

J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Denver, CO, pp. 291–294, April 1980.

B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing,* ASSP-27, 247–254, June 1979.

I. Gerson and M. Jasiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kb/s," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Albuquerque, NM, pp. 461–464, April 1990.

J.-P. Adoul and C. Lamblin, "A comparison of some algebraic structures for CELP coding of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Dallas, TX, pp. 1953–1956, April 1987.

S. Wang and A. Gersho, "Phonetically-based vector excitation coding of speech at 3.6 kbps," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Glasgow, Scotland, pp. 49–52, May 1989.

E. Paksoy, K. Srinivasan, and A. Gersho, "Variable rate speech coding with phonetic segmentation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Minneapolis, MN, pp. II.155–II.158, April 1993.

J. Hardwick and J. Lim, "The application of the IMBE speech coder to mobile communications," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 249–252, May 1991.

R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing,* 34, 744–754, August 1986.

A. McCree and T. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Processing,* 3, 242–250, July 1995.

P. Kroon and B. S. Atal, "On improving the performance of pitch predictors in speech coding systems," in *Advances in Speech Coding,* B. S. Atal, V. Cuperman, and A. Gersho, Eds., Boston, Mass: Kluwer, 1991, pp. 321–327.

J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech and Audio Processing,* 3, 59–71, January 1995.

W. Voiers, "Diagnostic evaluation of speech intelligibility," in *Speech Intelligibility and Recognition,* M. Hawley, Ed., Stroudsburg, Pa.: Dowden, Hutchinson, and Ross, 1977.

P. Papamichalis, *Practical Approaches to Speech Coding,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

N. S. Jayant and P. Noll, *Digital Coding of Waveforms,* Englewood Cliffs, N.J.: Prentice-Hall, 1984.

W. Daumer, "Subjective evaluation of several different speech coders," *IEEE Trans. Commun.,* COM-30, 655–662, April 1982.

W. Voiers, "Diagnostic acceptability measure for speech communications systems," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* 204–207, 1977.

N. Jayant, "High-quality coding of telephone speech and wideband audio," *IEEE Communications Magazine,* 28, 10–20, January 1990.

S. Miki, K. Mano, H. Ohmuro, and T. Moriya, "Pitch synchronous innovation CELP (PSI-CELP)," *Proc. European Conf. Speech Comm. Technol.,* Berlin, Germany, pp. 261–264, September 1993.

ITU-TS Study Group XV, *Draft recommendation AV.25Y—Dual Rate Speech Coder for Multimedia Telecommunication Transmitting at 5.3 & 6.4 kbit/s,* December 1991.

W. Kleijn, "Continuous representations in linear predictive coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 201–204, 1991.

A. Schmidt-Nielsen and D. Brock, "Speaker recognizability testing for voice coders," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing,* pp. 1149–1152, April 1996.

E. Kreamer and J. Tardelli, "Communicability testing for voice coders," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 1153–1156, April 1996.

A. McCree, K. Truong, E. George, T. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 200–203, April 1996.

W. Gardner, P. Jacobs, and C. Lee, "QCELP: A variable rate speech coder for CDMA digital cellular," in *Speech and Audio Coding for Wireless Networks,* B. S. Atal, V. Cuperman, and A. Gersho, Eds., Boston, Mass.: Kluwer, 1993, pp. 85–92.

A. Das, E. Paksoy, and A. Gersho, "Multimode and variable-rate coding of speech," in *Speech Coding and Synthesis,* W. B. Kleijn and K. K. Paliwal, Eds., Amsterdam: Elsevier, 1995, pp. 257–288.

Further Information

For further information on the state of the art in speech coding, see the articles by Spanias [1994] and Gersho [1994], and the book *Speech Coding and Synthesis* by Kleijn and Paliwal [1995].

15.2 Speech Enhancement and Noise Reduction

Yariv Ephraim

Voice communication systems are susceptible to interfering **signals** normally referred to as **noise.** The interfering signals may have harmful effects on the performance of any speech communication system. These effects depend on the specific system being used, on the nature of the noise and the way it interacts with the clean signal, and on the relative intensity of the noise compared to that of the signal. The latter is usually measured by the **signal-to-noise ratio** (SNR), which is the ratio of the power of the signal to the power of the noise.

The speech communication system may simply be a recording which was performed in a noisy environment, a standard digital or analog communication system, or a speech recognition system for human-machine communication. The noise may be present at the input of the communication system, in the channel, or at the receiving end. The noise may be correlated or uncorrelated with the signal. It may accompany the clean signal in an additive, multiplicative, or any other more general manner. Examples of noise sources include competitive speech; background sounds like music, a fan, machines, door slamming, wind, and traffic; room reverberation; and white Gaussian channel noise.

The ultimate goal of **speech enhancement** is to minimize the effects of the noise on the performance of speech communication systems. The performance measure is system dependent. For systems which comprise recordings of noisy speech, or standard analog communication systems, the goal of speech enhancement is to improve perceptual aspects of the noisy signal. For example, improving the **quality** and **intelligibility** of the noisy signal are common goals. Quality is a subjective measure which reflects on the pleasantness of the speech or on the amount of effort needed to understand the speech material. Intelligibility, on the other hand, is an objective measure which signifies the amount of speech material correctly understood. For standard digital communication systems, the goal of speech enhancement is to improve perceptual aspects of the encoded speech signal. For human-machine speech communication systems, the goal of speech enhancement is to reduce the error rate in recognizing the noisy speech signals.

To demonstrate the above ideas, consider a "hands-free" cellular radio telephone communication system. In this system, the transmitted signal is composed of the original speech and the background noise in the car. The background noise is generated by an engine, fan, traffic, wind, etc. The transmitted signal is also affected by the radio channel noise. As a result, noisy speech with low quality and intelligibility is delivered by such systems. The background noise may have additional devastating effects on the performance of this system. Specifically, if the system encodes the signal prior to its transmission, then the performance of the speech coder may significantly deteriorate in the presence of the noise. The reason is that speech coders rely on some statistical model for the clean signal, and this model becomes invalid when the signal is noisy. For a similar reason, if the cellular radio system is equipped with a speech recognizer for automatic dialing, then the error rate of such recognizer will be elevated in the presence of the background noise. The goals of speech enhancement in this example are to improve perceptual aspects of the transmitted noisy speech signals as well as to reduce the speech recognizer error rate.

Other important applications of speech enhancement include improving the performance of:

1. Pay phones located in noisy environments (e.g., airports)
2. Air-ground communication systems in which the cockpit noise corrupts the pilot's speech
3. Teleconferencing systems where noise sources in one location may be broadcasted to all other locations
4. Long distance communication over noisy radio channels

The problem of speech enhancement has been a challenge for many researchers for almost three decades. Different solutions with various degrees of success have been proposed over the years. An excellent introduction to the problem, and review of the systems developed up until 1979, can be found in the landmark paper by Lim and Oppenheim [1979]. A panel of the National Academy of Sciences discussed in 1988 the problem and various ways to evaluate speech enhancement systems. The panel's findings were summarized in Makhoul et al. [1989]. Modern statistical approaches for speech enhancement were recently reviewed in Boll [1992] and Ephraim [1992].

In this section the principles and performance of the major speech enhancement approaches are reviewed, and the advantages and disadvantages of each approach are discussed. The signal is assumed to be corrupted by additive statistically independent noise. Only a single noisy version of the clean signal is assumed available for enhancement. Furthermore, it is assumed that the clean signal cannot be preprocessed to increase its robustness prior to being affected by the noise. Speech enhancement systems which can either preprocess the clean speech signal or which have access to multiple versions of the noisy signal obtained from a number of microphones are discussed in Lim [1983].

This presentation is organized as follows. In the second section the speech enhancement problem is formulated and commonly used models and performance measures are presented. In the next section signal estimation for improving perceptual aspects of the noisy signal is discussed. In the fourth section source coding techniques for noisy signals are summarized, and the last section deals with recognition of noisy speech signals. Due to the limited number of references (10) allowed in this publication, tutorial papers are mainly referenced. Appropriate credit will be given by pointing to the tutorial papers which reference the original papers.

Models and Performance Measures

The goals of speech enhancement as stated in the first section are to improve perceptual aspects of the noisy signal whether the signal is transmitted through analog or digital channels and to reduce the error rate in recognizing noisy speech signals. Improving perceptual aspects of the noisy signal can be accomplished by estimating the clean signal from the noisy signal using perceptually meaningful estimation performance measures. If the signal has to be encoded for transmission over digital channels, then source coding techniques can be applied to the given noisy signal. In this case, a perceptually meaningful fidelity measure between the clean signal and the encoded noisy signal must be used. Reducing error rate in speech communication systems can be accomplished by applying optimal signal classification approaches to the given noisy signals. Thus the speech enhancement problem is essentially a set of signal estimation, source coding, and signal classification problems.

The probabilistic approach for solving these problems requires explicit knowledge of the performance measure as well as the probability laws of the clean signal and noise process. Such knowledge, however, is not explicitly available. Hence, mathematically tractable performance measures and statistical models which are believed to be meaningful are used. In this section we briefly review the most commonly used statistical models and performance measures.

The most fundamental model for speech signals is the Gaussian **autoregressive (AR) model.** This model assumes that each 20- to 40-msec segment of the signal is generated from an excitation signal which is applied to a linear time-invariant all-pole filter. The excitation signal comprises a mixture of white Gaussian noise and a periodic sequence of impulses. The period of that sequence is determined by the pitch period of the speech signal. This model is described in Fig. 15.2. Generally, the excitation signal represents the flow of air through the vocal cords and the all-pole filter represents the vocal tract. The model for a given sample function of speech

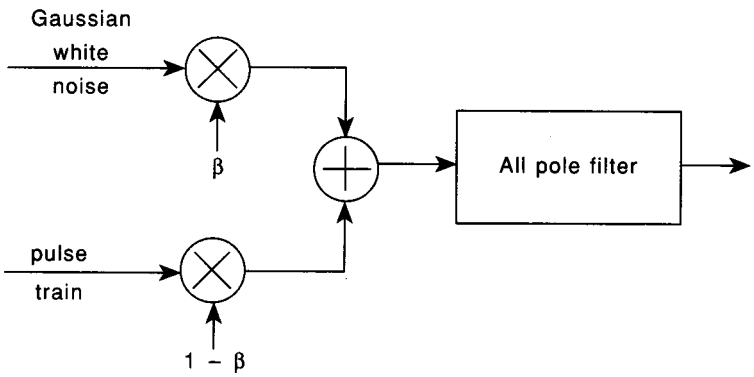


FIGURE 15.2 Gaussian autoregressive speech model.

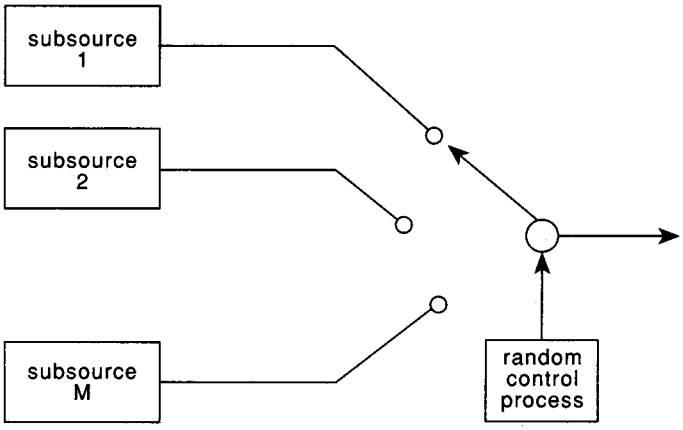


FIGURE 15.3 Composite source model.

signals, which is composed of several consecutive 20- to 40-msec segments of that signal, is obtained from the sequence of AR models for the individual segments. Thus, a linear time-varying AR model is assumed for each sample function of the speech signal. This model, however, is slowly varying in accordance with the slow temporal variation of the articulatory system. It was found that a set of approximately 2048 prototype AR models can reliably represent all segments of speech signals. The AR models are useful in representing the short time spectrum of the signal, since the spectrum of the excitation signal is white. Thus, the set of AR models represents a set of 2048 spectral prototypes for the speech signal.

The time-varying AR model for speech signals lacks the "memory" which assigns preference to one AR model to follow another AR model. This memory could be incorporated, for example, by assuming that the individual AR models are chosen in a Markovian manner. That is, given an AR model for the current segment of speech, certain AR models for the following segment of speech will be more likely than others. This results in the so-called *composite source model* (CSM) for the speech signal.

A block diagram of a CSM is shown in Fig. 15.3. In general, this model is composed of a set of M vector subsources which are controlled by a switch. The position of the switch at each time instant is chosen randomly, and the output of one subsource is provided. The position of the switch defines the state of the source at each time instant. CSMs for speech signals assume that the subsources are Gaussian AR sources, and the switch is controlled by a Markov chain. Furthermore, the subsources are usually assumed statistically independent and the vectors generated from each subsource are also assumed statistically independent. The resulting model is known as a **hidden Markov model** (HMM) [Rabiner, 1989] since the output of the model does not contain the states of the Markovian switch.

The performance measure for speech enhancement is task dependent. For signal estimation and coding, this measure is given in terms of a distortion measure between the clean signal and the estimated or the encoded signals, respectively. For signal classification applications the performance measure is normally the probability of misclassification. Commonly used distortion measures are the mean-squared error (MSE) and the Itakura-Saito distortion measures. The Itakura-Saito distortion measure is a measure between two power spectral densities, of which one is usually that of the clean signal and the other of a model for that signal [Gersho and Gray, 1991]. This distortion measure is normally used in designing speech coding systems and it is believed to be perceptually meaningful. Both measures are mathematically tractable and lead to intuitive estimation and coding schemes. Systems designed using these two measures need not be optimal only in the MSE and the Itakura-Saito sense, but they may as well be optimal in other more meaningful senses (see a discussion in Ephraim [1992]).

Signal Estimation

In this section we review the major approaches for speech signal estimation given noisy signals.

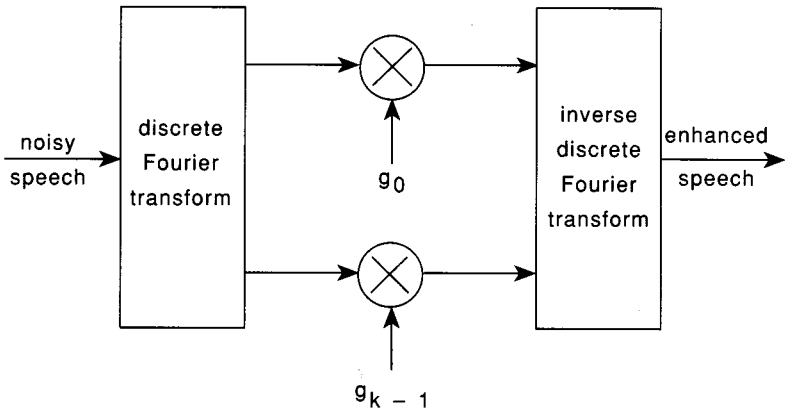


FIGURE 15.4 Spectral subtraction signal estimator.

Spectral Subtraction

The spectral subtraction approach [Weiss, 1974] is the simplest and most intuitive and popular speech enhancement approach. This approach provides estimates of the clean signal as well as of the short time spectrum of that signal. Estimation is performed on a frame-by-frame basis, where each frame consists of 20–40 msec of speech samples. In the spectral subtraction approach the signal is Fourier transformed, and spectral components whose variance is smaller than that of the noise are nulled. The surviving spectral components are modified by an appropriately chosen gain function. The resulting set of nulled and modified spectral components constitute the spectral components of the enhanced signal. The signal estimate is obtained from inverse Fourier transform of the enhanced spectral components. The short time spectrum estimate of the signal is obtained from squaring the enhanced spectral components. A block diagram of the spectral subtraction approach is shown in Fig. 15.4.

Gain functions motivated by different perceptual aspects have been used. One of the most popular functions results from linear minimum MSE (MMSE) estimation of each spectral component of the clean signal given the corresponding spectral component of the noisy signal. In this case, the value of the gain function for a given spectral component constitutes the ratio of the variances of the clean and noisy spectral components. The variance of the clean spectral component is obtained by subtracting an assumed known variance of the noise spectral component from the variance of the noisy spectral component. The resulting variance is guaranteed to be positive by the nulling process mentioned above. The variances of the spectral components of the noise process are normally estimated from silence portions of the noisy signal.

A family of spectral gain functions proposed in Lim and Oppenheim [1979] is given by

$$g_n = \left(\frac{|Z_n|^a - bE\left\{|V_n|^a\right\}}{|Z_n|^a} \right)^c \qquad n = 1, \ldots, N \tag{15.2}$$

where Z_n and V_n denote the nth spectral components of the noisy signal and the noise process, respectively, and $a > 0$, $b \geq 0$, $c > 0$. The MMSE gain function is obtained when $a = 2$, $b = 1$, and $c = 1$. Another commonly used gain function in the spectral subtraction approach is obtained from using $a = 2$, $b = 1$, and $c = 1/2$. This gain function results from estimating the spectral magnitude of the signal and combining the resulting estimate with the phase of the noisy signal. This choice of gain function is motivated by the relative importance of the spectral magnitude of the signal compared to its phase. Since both cannot be simultaneously optimally estimated [Ephraim, 1992], only the spectral magnitude is optimally estimated, and combined with an estimate of the complex exponential of the phase which does not affect the spectral magnitude estimate. The resulting estimate

of the phase can be shown to be the phase of the noisy signal within the HMM statistical framework. Normally, the spectral subtraction approach is used with $b = 2$, which corresponds to an artificially elevated noise level.

The spectral subtraction approach has been very popular since it is relatively easy to implement; it makes minimal assumptions about the signal and noise; and when carefully implemented, it results in reasonably clear enhanced signals. A major drawback of the spectral subtraction enhancement approach, however, is that the residual noise has annoying tonal characteristics referred to as "musical noise." This noise consists of narrowband signals with time-varying frequencies and amplitudes. Another major drawback of the spectral subtraction approach is that its optimality in any given sense has never been proven. Thus, no systematic methodology for improving the performance of this approach has been developed, and all attempts to achieve this goal have been based on purely heuristic arguments. As a result, a family of spectral subtraction speech enhancement approaches have been developed and experimentally optimized.

In a recent work [Ephraim et al., 1995] a version of the spectral subtraction was shown to be a signal subspace estimation approach which is asymptomatically optimal (as the frame length approaches infinity) in the linear MMSE sense.

Empirical Averages

This approach attempts to estimate the clean signal from the noisy signal in the MMSE sense. The conditional mean estimator is implemented using the conditional sample average of the clean signal given the noisy signal. The sample average is obtained from appropriate training sequences of the clean and noisy signals. This is equivalent to using the sample distribution or the histogram estimate of the probability density function (pdf) of the clean signal given the noisy signal. The sample average approach is applicable for estimating the signal as well as functionals of that signal, e.g., the spectrum, the logarithm of the spectrum, and the spectral magnitude.

Let $\{Y_t, t = 0, \ldots, T\}$ be a training data from the clean signal, where Y_t is a K-dimensional vector in the Euclidean space R^K. Let $\{Z_t, t = 0, \ldots, T\}$ be a training data from the noisy signal, where $Z_t \in R^K$. The sequence $\{Z_t\}$ can be obtained by adding a noise training sequence $\{V_t, t = 0, \ldots, T\}$ to the sequence of clean signals $\{Y_t\}$. Let $z \in R^K$ be a vector of the noisy signal from which the vector y of the clean signal is estimated. Let $\mathbf{Y}(z) = \{Y_t: Z_t = z, t = 0, \ldots, T\}$ be the set of all clean vectors from the training data of the clean signal which could have resulted in the given noisy observation z The cardinality of this set is denoted by $|\mathbf{Y}(z)|$. Then, the sample average estimate of the conditional mean of the clean signal y given the noisy signal z is given by

$$\begin{aligned} \hat{y} &= E\{y|z\} \\ &= \frac{\int yp(y,z)dy}{\int p(y,z)dy} \\ &= \frac{1}{|\mathbf{Y}(z)|} \sum_{\{Y_t \in \mathbf{Y}(z)\}} Y_t \end{aligned} \tag{15.3}$$

Obviously, this approach is only applicable for signals with finite alphabet since otherwise the set $Y(z)$ is empty with probability one. For signals with continuous pdf's, the approach can be applied only if those signals are appropriately quantized.

The sample average approach was first applied for enhancing speech signals by Porter and Boll in 1984 [Boll, 1992]. They, however, considered a simpler situation in which the noise true pdf was assumed known. In this case, enhanced signals with residual noise characterized as being a blend of wideband noise and musical noise were obtained. The balance between the two types of residual noise depended on the functional of the clean signal which was estimated.

The advantages of the sample average approach are that it is conceptually simple and it does not require *a priori* assumptions about the form of the pdf's of the signal and noise. Hence, it is a nonparametric estimation approach. This approach, however, has three major disadvantages. First, the estimator does not utilize any speech specific information such as the periodicity of the signal and the signal's AR model. Second, the training

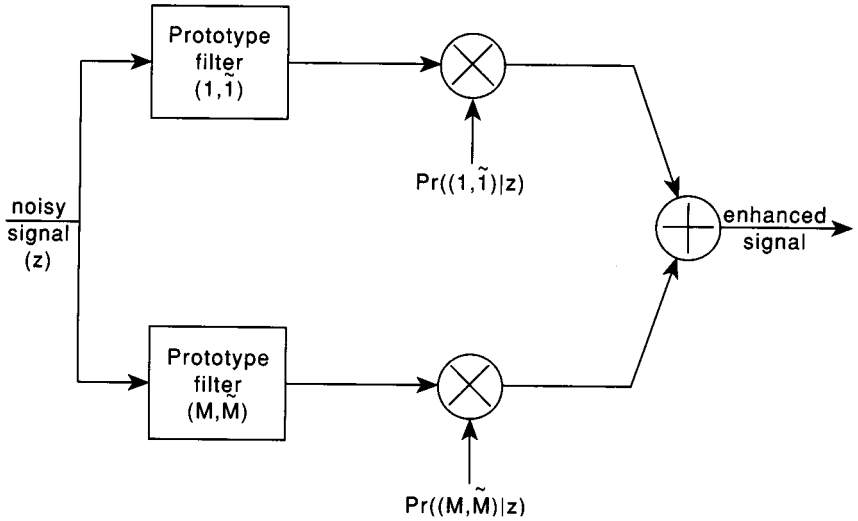


FIGURE 15.5 HMM-based MMSE signal estimator.

sequences from the signal and noise must be available at the speech enhancement unit. Furthermore, these training sequences must be applied for each newly observed vector of the noisy signal. Since the training sequences are normally very long, the speech enhancement unit must have extensive memory and computational resources. These problems are addressed in the model-based approach described next.

Model-Based Approach

The model-based approach [Ephraim, 1992] is a Bayesian approach for estimating the clean signal or any functional of that signal from the observed noisy signal. This approach assumes CSMs for the clean signal and noise process. The models are estimated from training sequences of those processes using the maximum likelihood (ML) estimation approach. Under ideal conditions the ML model estimate is consistent and asymptotically efficient. The ML model estimation is performed using the expectation-maximization (EM) or the Baum iterative algorithm [Rabiner, 1989; Ephraim, 1992]. Given the CSMs for the signal and noise, the clean signal is estimated by minimizing the expected value of the chosen distortion measure. The model-based approach uses significantly more statistical knowledge about the signal and noise compared to either the spectral subtraction or the sample average approaches.

The MMSE signal estimator is obtained from the conditional mean of the clean signal given the noisy signal. If $y_t \in R^K$ denotes the vector of the speech signal at time t, and z_0^t denotes the sequence of K-dimensional vectors of noisy signals $\{z_0, \ldots, z_t\}$ from time $\tau = 0$ to $\tau = t$, then the MMSE estimator of y_t is given by

$$\begin{aligned} \hat{y}_t &= E\{y_t \,|\, z_0^t\} \\ &= \sum_{\bar{x}_t} P(\bar{x}_t \,|\, z_0^t) E\{y_t \,|\, z_t, \bar{x}_t\} \end{aligned} \tag{15.4}$$

where $\bar{x}_t$ denotes the composite state of the noisy signal at time t. This state is given by $\bar{x}_t \triangleq (x_t, \tilde{x}_t)$, where x_t is the Markov state of the clean signal at time t and $\tilde{x}_t$ denotes the Markov state of the noise process at the same time instant t. The MMSE estimator, Eq. (15.4), comprises a weighted sum of conditional mean estimators for the composite states of the noisy signal, where the weights are the probabilities of those states given the noisy observed signal. A block diagram of this estimator is shown in Fig. 15.5.

The probability $P(\bar{x}_t \,|\, z_0^t)$ can be efficiently calculated using the forward recursion associated with HMMs. For CSMs with Gaussian subsources, the conditional mean $E\{y_t \,|\, z_t, \bar{x}_t\}$ is a linear function of the noisy vector z_t, given by

$$E(y_t|z_t, \bar{x}_t) = S_{xt}(S_{xt} + S_{\bar{x}t})^{-1} z_t \triangleq H_{\bar{x}t} z_t \tag{15.5}$$

where S_{xt} and $S_{\bar{x}t}$ denote the covariance matrices of the Gaussian subsources associated with the Markov states x_t and $\bar{x}_t$, respectively. Since, however, $P(\bar{x}_t | z_0^t)$ is a nonlinear function of the noisy signal z_0^t, the MMSE signal estimator $\hat{y}_t$ is a nonlinear function of the noisy signal z_0^t.

The MMSE estimator, Eq. (15.4), is intuitively appealing. It uses a predesigned set of filters $\{H_{\bar{x}_t}\}$ obtained from training data of speech and noise. Each filter is optimal for a pair of subsources of the CSMs for the clean signal and the noise process. Since each subsource represents a subset of signals from the corresponding source, each filter is optimal for a pair of signal subsets from the speech and noise. The set of predesigned filters covers all possible pairs of speech and noise signal subsets. Hence, for each noisy vector of speech there must exist an optimal filter in the set of predesigned filters. Since, however, a vector of the noisy signal could possibly be generated from any pair of subsources of the clean signal and noise, the most appropriate filter for a given noisy vector is not known. Consequently, in estimating the signal vector at each time instant, all filters are tried and their outputs are weighted by the probabilities of the filters to be correct for the given noisy signal. Other strategies for utilizing the predesigned set of filters are possible. For example, at each time instant only the most likely filter can be applied to the noisy signal. This approach is more intuitive than that of the MMSE estimation. It was first proposed in Drucker [1968] for a five-state model which comprises subsources for fricatives, stops, vowels, glides, and nasals. This approach was shown by Ephraim and Merhav [Ephraim, 1992] to be optimal only in an asymptotic MMSE sense.

The model-based MMSE approach provides reasonably good enhanced speech quality with significantly less structured residual noise than the spectral subtraction approach. This performance was achieved for white Gaussian input noise at 10 dB input SNR using 512-2048 filters. An improvement of 5–6 dB in SNR was achieved by this approach. The model-based approach, however, is more elaborate than the spectral subtraction approach, since it involves two steps of training and estimation, and training must be performed on sufficiently long data. The MMSE estimation approach is usually superior to the asymptotic MMSE enhancement approach. The reason is that the MMSE approach applies a "soft decision" rather than a "hard decision" in choosing the most appropriate filter for a given vector of the noisy signal.

A two-state version of the MMSE estimator was first applied to speech enhancement by McAulay and Malpass in 1980 [Ephraim, 1992]. The two states corresponded to speech presence and speech absence (silence) in the noisy observations. The estimator for the signal given that it is present in the noisy observations was implemented by the spectral subtraction approach. The estimator for the signal in the "silence state" is obviously equal to zero. This approach significantly improved the performance of the spectral subtraction approach.

Source Coding

An **encoder** for the clean signal maps vectors of that signal onto a finite set of representative signal vectors referred to as codewords. The mapping is performed by assigning each signal vector to its nearest neighbor codeword. The index of the chosen codeword is transmitted to the receiver in a signal communication system, and the signal is reconstructed using a copy of the chosen codeword. The codewords are designed to minimize the average distortion resulting from the nearest neighbor mapping. The codewords may simply represent waveform vectors of the signal. In another important application of low bit-rate speech coding, the codewords represent a set of parameter vectors of the AR model for the speech signal. Such coding systems synthesize the signal using the speech model in Fig. 15.2. The synthesis is performed using the encoded vector of AR coefficients as well as the parameters of the excitation signal. Reasonably good speech quality can be obtained using this coding approach at rates as low as 2400–4800 bits/sample [Gersho and Gray, 1991].

When only noisy signals are available for coding, the encoder operates on the noisy signal while representing the clean signal. In this case, the encoder is designed by minimizing the average distortion between the clean signal and the encoded signal. Specifically, let y denote the vector of clean signal to be encoded. Let z denote the corresponding given vector of the noisy signal. Let q denote the encoder. Let d denote a distortion measure. Then, the optimal encoder is designed by

$$\min_{q} E\{d(y, q(z))\} \tag{15.6}$$

When the clean signal is available for encoding the design problem is similarly defined, and it is obtained from Eq. (15.6) using $z = y$. The design problem in Eq. (15.6) is not standard since the encoder operates and represents different sources. The problem can be transformed into a standard coding problem by appropriately modifying the distortion measure. This was shown by Berger in 1971 and Ephraim and Gray in 1988 [Ephraim, 1992]. Specifically, define the modified distortion measure by

$$d'(z, q(z)) \triangleq E\{d(y, q(z)) \mid z\} \tag{15.7}$$

Then, by using iterated expectation in Eq. (15.8), the design problem becomes

$$\min_{q} E\{d'(z, q(z))\} \tag{15.8}$$

A useful class of encoders for speech signals are those obtained from vector quantization. Vector quantizers are designed using the Lloyd algorithm [Gersho and Gray, 1991]. This is an iterative algorithm in which the codewords and the nearest neighbor regions are alternatively optimized. This algorithm can be applied to design vector quantizers for clean and noisy signals using the modified distortion measure.

The problem of designing vector quantizers for noisy signals is related to the problem of estimating the clean signals from the noisy signals, as was shown by Wolf and Ziv in 1970 and Ephraim and Gray in 1988 [Ephraim, 1992]. Specifically, optimal waveform vector quantizers in the MMSE sense can be designed by first estimating the clean signal and then quantizing the estimated signal. Both estimation and quantization are performed in the MMSE sense. Similarly, optimal quantization of the vector of parameters of the AR model for the speech signal in the Itakura-Saito sense can be performed in two steps of estimation and quantization. Specifically, the autocorrelation function of the clean signal, which approximately constitutes the sufficient statistics of that signal for estimating the AR model, is first estimated in the MMSE sense. Then, optimal vector quantization in the Itakura-Saito sense is applied to the estimated autocorrelation.

The estimation-quantization approach has been most popular in designing encoders for speech signals given noisy signals. Since such design requires explicit knowledge of the statistics of the clean signal and the noise process, but this knowledge is not available as argued in the second section, a variety of suboptimal encoders were proposed. Most of the research in this area focused on designing encoders for the AR model of the signal due to the importance of such encoders in low bit-rate speech coding. The proposed encoders mainly differ in the estimators they used and the functionals of the speech signal these estimators have been applied to. Important examples of functionals which have commonly been estimated include the signal waveform, autocorrelation, and the spectral magnitude. The primarily set of estimators used for this application were obtained from the spectral subtraction approach and its derivatives. A version of the sample average estimator was also developed and applied to AR modeling by Juang and Rabiner in 1987 [Ephraim, 1992]. Recently, the HMM-based estimator of the autocorrelation function of the clean signal was used in AR model vector quantization [Ephraim, 1992].

Designing of AR model-based encoders from noisy signals has been a very successful application of speech enhancement. In this case both the quality and intelligibility of the encoded signal can be improved compared to the case where the encoder is designed for the clean signal and the input noise is simply ignored. The reason is that the input noise has devastating effects of the performance of AR model-based speech coders, and any "reasonable" estimation approach can significantly improve the performance of those coders in noisy environments.

Signal Classification

In recognition of clean speech signals a sample function of the signal is associated with one of the words in the vocabulary. The association or decision rule is designed to minimize the probability of classification error. When only noisy speech signals are available for recognition a very similar problem results. Specifically, a sample

function of the noisy signal is now associated with one of the words in the vocabulary in a way which minimizes the probability of classification error. The only difference between the two problems is that the sample functions of the clean and noisy signals from a given word have different statistics. The problem in both cases is that of partitioning the sample space of the given acoustic signals from all words in the vocabulary into L partition cells, where L is the number of words in the vocabulary.

Let $\{W_i,\ i = 1, \ldots, L\}$ denote a set of words in a given vocabulary. Let z denote the acoustic noisy signal from some word in the vocabulary. Let $\Omega \triangleq \{\omega_1, \ldots, \omega_L\}$ be a partition of the sample space of the noisy signals. The probability of error associated with this partition is given by

$$P_e(\Omega) = \sum_{i=1}^{L} P(W_i) \int_{z \notin \omega_i} p(z \mid W_i) dz \tag{15.9}$$

where $P(W_i)$ is the *a priori* probability of occurrence of the ith word, and $p(z \mid W_i)$ is the pdf of the noisy signal from the ith word. The minimization of $P_e(\Omega)$ is achieved by the well-known maximum *a posteriori* (MAP) decision rule. Specifically, z is associated with the word W_i for which $p(z \mid W_i)P(W_i) > p(z \mid W_j)P(W_j)$ for all $j = 1, \ldots, L$ and $j \neq i$. Ties are arbitrarily broken. In the absence of noise, the noisy signal z becomes a clean signal y, and the optimal recognizer is obtained by using the same decision rule with $z = y$. Hence, the only difference between recognition of clean signals and recognition of noisy signals is that in the first case the pdf's $\{p(y \mid W_i)\}$ are used in the decision rule, while in the second case the pdf's $\{p(z \mid W_i)\}$ are used in the same decision rule.

Note that optimal recognition of noisy signals requires explicit knowledge of the statistics of the clean signal and noise. Neither the clean signal nor any function of that signal needs to be estimated. Since, however, the statistics of the signal and noise are not explicitly available as argued in the second section, parametric models are usually assumed for these pdf's and their parameters are estimated from appropriate training data. Normally, HMMs with mixture of Gaussian pdf's at each state are attributed to both the clean signal and noise process. It can be shown (similarly to the case of classification of clean signals dealt with by Merhav and Ephraim in 1991 [Ephraim, 1992]) that if the pdf's of the signal and noise are precisely HMMs and the training sequences are significantly longer than the test data, then the MAP decision rule which uses estimates of the pdf's of the signal and noise is asymptotically optimal.

A key issue in applying hidden Markov modeling for recognition of speech signals is the matching of the energy contour of the signal to the energy contour of the model for that signal. Energy matching is required for two main reasons. First, speech signals are not strictly stationary and hence their energy contours cannot be reliably estimated from training data. Second, recording conditions during training and testing vary. An approach for gain adaptation was developed [Ephraim, 1992]. In this approach, HMMs for gain-normalized clean signals are designed and used together with gain contour estimates obtained from the noisy signals. The gain adaptation approach is implemented using the EM algorithm. This approach provides robust speech recognition at input SNRs greater than or equal to 10 dB.

The relation between signal classification and estimation was established in Kailath [1969] for *continuous* time signals contaminated by additive statistically independent Gaussian white noise. It was shown that minimum probability of error classification can be achieved by applying the MAP decision rule to the *causal* MMSE estimator of the clean signal. This interesting theoretical result provides the intuitive basis for a popular approach for recognition of noisy speech signals. In this approach, the clean signal or some feature vector of the signal is first estimated and then recognition is applied. In the statistical framework of hidden Markov modeling, however, the direct recognition approach presented earlier is significantly simpler since both the clean signal and the noisy signal are HMMs [Ephraim, 1992]. Hence, the complexity of recognizing the estimated signal is the same as that of recognizing the noisy signal directly.

Other commonly used approaches for recognition of noisy speech signals were developed for systems which are based on pattern recognition. When clean signals are available for recognition, these systems match the input signal to the nearest neighbor acoustic templet which represents some word in the vocabulary. The templets mainly comprise spectral prototypes of the clean signals. The matching is performed using a distance measure between the clean input signal and the templet. When only noisy signals are available for recognition,

several modifications of the pattern matching approach were proposed. Specifically, adapting the templets of the clean signal to reflect the presence of the noise was proposed by Roe in 1987 [Ephraim, 1992]; choosing templets for the noisy signal which are more robust than those obtained from adaptation of the templets for the clean signal was often proposed; and using distance measures which are robust to noise, such as the projection measure proposed by Mansour and Juang in 1989 [Ephraim, 1992]. These approaches along with the prefiltering approach in the sampled signal case are fairly intuitive and are relatively easy to implement. It is difficult, however, to establish their optimality in any well-defined sense. Another interesting approach based on robust statistics was developed by Merhav and Lee [Ephraim, 1992]. This approach was shown asymptotically optimal in the minimum probability of error sense within the hidden Markov modeling framework.

The speech recognition problem in noisy environments has also been a successful application of speech enhancement. Significant reduction in the error rate due to the noise presence was achieved by the various approaches mentioned above.

Comments

Three major aspects of speech enhancement were reviewed. These comprise improving the perception of speech signals in noisy environments and increasing the robustness of speech coders and recognition systems in noisy environments. The inherent difficulties associated with these problems were discussed, and the main solutions along with their strengths and weaknesses were presented. This section is an introductory presentation to the speech enhancement problem. A comprehensive treatment of the subject can be found in Lim [1979], Makhoul et al. [1989], Boll [1992], and Ephraim [1992]. Significant progress in understanding the problem and in developing new speech enhancement systems was made during the 1980s with the introduction of statistical model-based approaches. The speech enhancement problem, however, is far from being solved, and major progress is still needed. In particular, no speech enhancement system which is capable of simultaneously improving both the quality and intelligibility of the noisy signal is currently known. Progress in this direction can be made if more reliable statistical models for the speech signal and noise process as well as meaningful distortion measures can be found.

Defining Terms

Autoregressive model: Statistical model for resonant signals.
Classifier: Maps signal utterances into a finite set of word units, e.g., syllables.
Encoder: Maps signal vectors into a finite set of codewords. A vector quantizer is a particular type of encoder.
Hidden Markov model: Statistical model comprised of several subsources controlled by Markovian process.
Intelligibility: Objective quantitative measure of speech perception.
Noise: Any interfering signal adversely affecting the communication of the clean signal.
Quality: Subjective descriptive measure of speech perception.
Signal: Clean speech sample to be communicated with human or machine.
Signal-to-noise ratio: Ratio of the signal power to the noise power measured in decibels.
Speech enhancement: Improvement of perceptual aspects of speech signals.

Related Topics

48.1 Introduction • 73.2 Noise

References

S. F. Boll, "Speech enhancement in the 1980's: noise suppression with pattern matching," in *Advances in Speech Signal Processing,* S. Furui and M. M. Sonhdi, Eds., New York: Marcel Dekker, 1992.

H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. 16, 1968.

Y. Ephraim, "Statistical model based speech enhancement systems," *Proc. IEEE,* vol. 80, 1992.

Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing,* vol. 3, 251–316, 1995.
A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression,* Boston: Kluwer Academic Publishers, 1991.
T. Kailath, "A general likelihood-ratio formula for random signals in Gaussian noise," *IEEE Trans. on Inform Theory,* vol. 15, 1969.
J. S. Lim, Ed., *Speech Enhancement,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.
J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE,* vol. 67, 1979.
J. Makhoul, T. H. Crystal, D. M. Green, D. Hogan, R. J. McAulay, D. B. Pisoni, R. D. Sorkin, and T. G. Stockham, *Removal of Noise From Noise-Degraded Speech Signals,* Washington, D.C.: National Academy Press, 1989.
L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* vol. 77, 1989.
M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Processing speech signals to attenuate interference," in *IEEE Symp. on Speech Recognition,* Pittsburgh, 1974.

Further Information

A comprehensive treatment of the speech enhancement problem can be found in the four tutorial papers and book listed below.

J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE,* vol. 67, 1979.
J. Makhoul, T. H. Crystal, D. M. Green, D. Hogan, R. J. McAulay, D. B. Pisoni, R. D. Sorkin, and T. G. Stockham, *Removal of Noise From Noise-Degraded Speech Signals,* Washington, D.C.: National Academy Press, 1989.
S. F. Boll, "Speech enhancement in the 1980's: noise suppression with pattern matching," in *Advances in Speech Signal Processing,* S. Furui and M. M. Sonhdi, Eds., New York: Marcel Dekker, 1992.
Y. Ephraim, "Statistical model based speech enhancement systems," *Proc. IEEE,* vol. 80, 1992.
J. S. Lim, Ed., *Speech Enhancement,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.

15.3 Analysis and Synthesis

Jesse W. Fussell

After an acoustic speech signal is converted to an electrical signal by a microphone, it may be desirable to analyze the electrical signal to estimate some time-varying parameters which provide information about a model of the speech production mechanism. **Speech analysis** is the process of estimating such parameters. Similarly, given some parametric model of speech production and a sequence of parameters for that model, **speech synthesis** is the process of creating an electrical signal which approximates speech. While analysis and synthesis techniques may be done either on the continuous signal or on a sampled version of the signal, most modern analysis and synthesis methods are based on digital signal processing.

A typical speech production model is shown in Fig. 15.6. In this model the output of the excitation function is scaled by the gain parameter and then filtered to produce speech. All of these functions are time-varying.

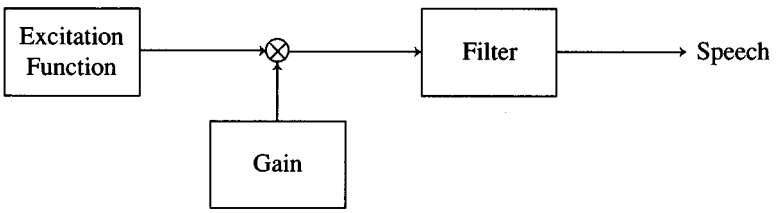


FIGURE 15.6 A general speech production model.

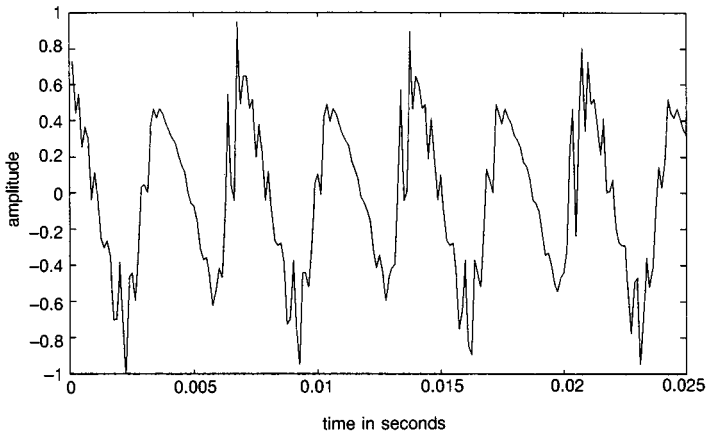


FIGURE 15.7 Waveform of a spoken phoneme /i/ as in beet.

For many models, the parameters are varied at a periodic rate, typically 50 to 100 times per second. Most speech information is contained in the portion of the signal below about 4 kHz.

The excitation is usually modeled as either a mixture or a choice of random noise and periodic waveform. For human speech, voiced excitation occurs when the vocal folds in the larynx vibrate; unvoiced excitation occurs at constrictions in the vocal tract which create turbulent air flow [Flanagan, 1965]. The relative mix of these two types of excitation is termed "**voicing**." In addition, the periodic excitation is characterized by a fundamental frequency, termed **pitch** or F0. The excitation is scaled by a factor designed to produce the proper amplitude or level of the speech signal. The scaled excitation function is then filtered to produce the proper spectral characteristics. While the filter may be nonlinear, it is usually modeled as a linear function.

Analysis of Excitation

In a simplified form, the excitation function may be considered to be purely periodic, for voiced speech, or purely random, for unvoiced. These two states correspond to voiced phonetic classes such as vowels and nasals and unvoiced sounds such as unvoiced fricatives. This binary voicing model is an oversimplification for sounds such as voiced fricatives, which consist of a mixture of periodic and random components. Figure 15.7 is an example of a time waveform of a spoken /i/ phoneme, which is well modeled by only periodic excitation.

Both time domain and frequency domain analysis techniques have been used to estimate the degree of voicing for a short segment or frame of speech. One time domain feature, termed the zero crossing rate, is the number of times the signal changes sign in a short interval. As shown in Fig. 15.7, the zero crossing rate for voiced sounds is relatively low. Since unvoiced speech typically has a larger proportion of high-frequency energy than voiced speech, the ratio of high-frequency to low-frequency energy is a frequency domain technique that provides information on voicing.

Another measure used to estimate the degree of voicing is the autocorrelation function, which is defined for a sampled speech segment, S, as

$$\mathrm{ACF}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} s(n)s(n-\tau) \tag{15.10}$$

where $s(n)$ is the value of the nth sample within the segment of length N. Since the autocorrelation function of a periodic function is itself periodic, voicing can be estimated from the degree of periodicity of the autocorrelation function. Figure 15.8 is a graph of the nonnegative terms of the autocorrelation function for a 64-ms frame of the waveform of Fig. 15.7. Except for the decrease in amplitude with increasing lag, which results from the rectangular window function which delimits the segment, the autocorrelation function is seen to be quite periodic for this voiced utterance.

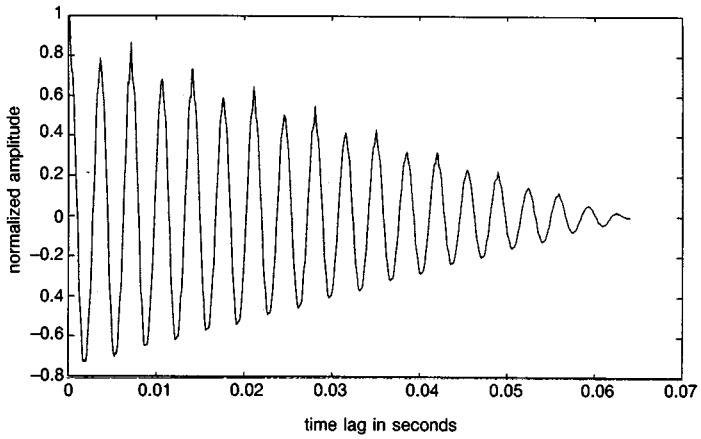


FIGURE 15.8 Autocorrelation function of one frame of /i/.

If an analysis of the voicing of the speech signal indicates a voiced or periodic component is present, another step in the analysis process may be to estimate the frequency (or period) of the voiced component. There are a number of ways in which this may be done. One is to measure the time lapse between peaks in the time domain signal. For example in Fig. 15.7 the major peaks are separated by about 0.0071 s, for a fundamental frequency of about 141 Hz. Note, it would be quite possible to err in the estimate of fundamental frequency by mistaking the smaller peaks that occur between the major peaks for the major peaks. These smaller peaks are produced by resonance in the vocal tract which, in this example, happen to be at about twice the excitation frequency. This type of error would result in an estimate of pitch approximately twice the correct frequency.

The distance between major peaks of the autocorrelation function is a closely related feature that is frequently used to estimate the pitch period. In Fig. 15.8, the distance between the major peaks in the autocorrelation function is about 0.0071 s. Estimates of pitch from the autocorrelation function are also susceptible to mistaking the first vocal track resonance for the glottal excitation frequency.

The absolute magnitude difference function (AMDF), defined as,

$$\text{AMDF}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}\left|s(n) - s(n-\tau)\right| \tag{15.11}$$

is another function which is often used in estimating the pitch of voiced speech. An example of the AMDF is shown in Fig. 15.9 for the same 64-ms frame of the /i/ phoneme. However, the minima of the AMDF is used as an indicator of the pitch period. The AMDF has been shown to be a good pitch period indicator [Ross et al., 1974] and does not require multiplications.

Fourier Analysis

One of the more common processes for estimating the spectrum of a segment of speech is the Fourier transform [Oppenheim and Schafer, 1975]. The Fourier transform of a sequence is mathematically defined as

$$S(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega n} \tag{15.12}$$

where $s(n)$ represents the terms of the sequence. The short-time Fourier transform of a sequence is a time-dependent function, defined as

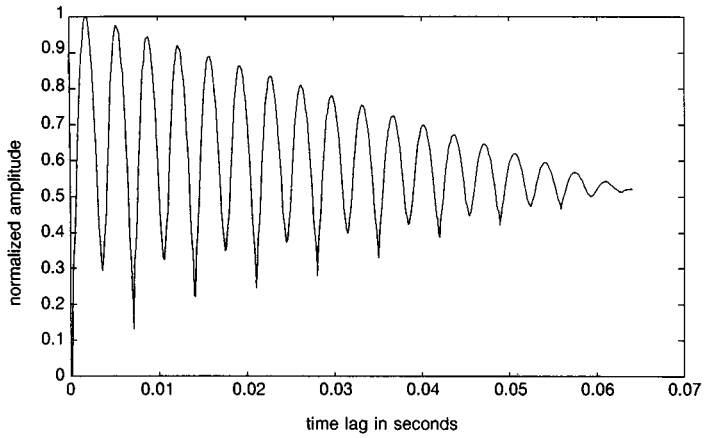


FIGURE 15.9 Absolute magnitude difference function of one frame of /i/.

$$S_m(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w(m-n)\, s(n) e^{-j\omega n} \tag{15.13}$$

where the window function $w(n)$ is usually zero except for some finite range, and the variable m is used to select the section of the sequence for analysis. The discrete Fourier transform (DFT) is obtained by uniformly sampling the short-time Fourier transform in the frequency dimension. Thus an N-point DFT is computed using Eq. (15.14),

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi nk/N} \tag{15.14}$$

where the set of N samples, $s(n)$, may have first been multiplied by a window function. An example of the magnitude of a 512-point DFT of the waveform of the /i/ from Fig. 15.10 is shown in Fig. 15.10. Note for this figure, the 512 points in the sequence have been multiplied by a Hamming window defined by

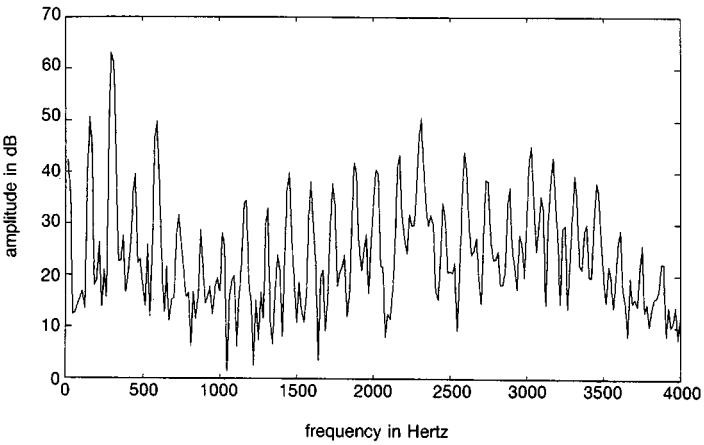


FIGURE 15.10 Magnitude of 512-point FFT of Hamming windowed /i/.

$$\begin{aligned} w(n) &= 0.54 - 0.46\ \cos\ (2\pi n\ /(N-1)) \quad 0 \le n \le N-1 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{15.15}$$

Since the spectral characteristics of speech may change dramatically in a few milliseconds, the length, type, and location of the window function are important considerations. If the window is too long, changing spectral characteristics may cause a blurred result; if the window is too short, spectral inaccuracies result. A Hamming window of 16 to 32 ms duration is commonly used for speech analysis.

Several characteristics of a speech utterance may be determined by examination of the DFT magnitude. In Fig. 15.10, the DFT of a voiced utterance contains a series of sharp peaks in the frequency domain. These peaks, caused by the periodic sampling action of the glottal excitation, are separated by the fundamental frequency which is about 141 Hz, in this example. In addition, broader peaks can be seen, for example at about 300 Hz and at about 2300 Hz. These broad peaks, called formants, result from resonances in the vocal tract.

Linear Predictive Analysis

Given a sampled (discrete-time) signal $s(n)$, a powerful and general parametric model for time series analysis is

$$s(n) = -\sum_{k=1}^{p} a(k)s(n-k) + G\sum_{l=0}^{q} b(l)\ u(n-l) \tag{15.16}$$

where $s(n)$ is the output and $u(n)$ is the input (perhaps unknown). The model parameters are $a(k)$ for $k = 1$, p, $b(l)$ for $l = 1$, q, and G. $b(0)$ is assumed to be unity. This model, described as an autoregressive moving average (ARMA) or pole-zero model, forms the foundation for the analysis method termed linear prediction. An autoregressive (AR) or all-pole model, for which all of the "b" coefficients except $b(0)$ are zero, is frequently used for speech analysis [Markel and Gray, 1976].

In the standard AR formulation of linear prediction, the model parameters are selected to minimize the mean-squared error between the model and the speech data. In one of the variants of linear prediction, the autocorrelation method, the minimization is carried out for a windowed segment of data. In the autocorrelation method, minimizing the mean-square error of the time domain samples is equivalent to minimizing the integrated ratio of the signal spectrum to the spectrum of the all-pole model. Thus, linear predictive analysis is a good method for spectral analysis whenever the signal is produced by an all-pole system. Most speech sounds fit this model well.

One key consideration for linear predictive analysis is the order of the model, p. For speech, if the order is too small, the formant structure is not well represented. If the order is too large, pitch pulses as well as formants begin to be represented. Tenth- or twelfth-order analysis is typical for speech. Figures 15.11 and 15.12 provide examples of the spectrum produced by eighth-order and sixteenth-order linear predictive analysis of the /i/ waveform of Fig. 15.7. Figure 15.11 shows there to be three formants at frequencies of about 300, 2300, and 3200 Hz, which are typical for an /i/.

Homomorphic (Cepstral) Analysis

For the speech model of Fig. 15.6, the excitation and filter impulse response are convolved to produce the speech. One of the problems of speech analysis is to separate or deconvolve the speech into these two components. One such technique is called homomorphic filtering [Oppenheim and Schafer, 1968]. The characteristic system for a system for homomorphic deconvolution converts a convolution operation to an addition operation. The output of such a characteristic system is called the complex **cepstrum.** The complex cepstrum is defined as the inverse Fourier transform of the complex logarithm of the Fourier transform of the input. If the input sequence is minimum phase (i.e., the z-transform of the input sequence has no poles or zeros outside the unit circle), the sequence can be represented by the real portion of the transforms. Thus, the real cepstrum can be computed by calculating the inverse Fourier transform of the log-spectrum of the input.

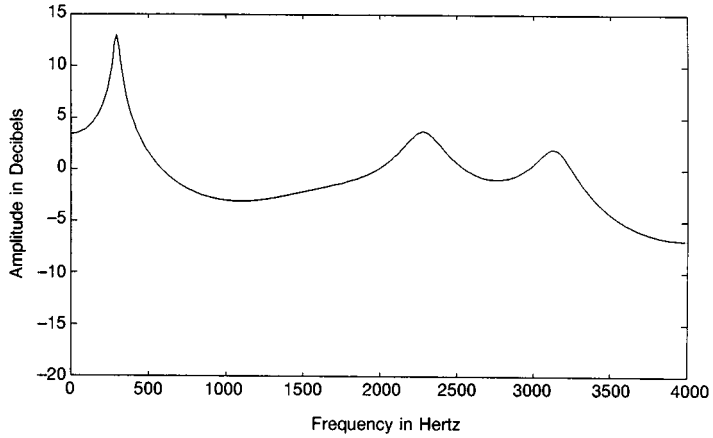


FIGURE 15.11 Eighth-order linear predictive analysis of an "i".

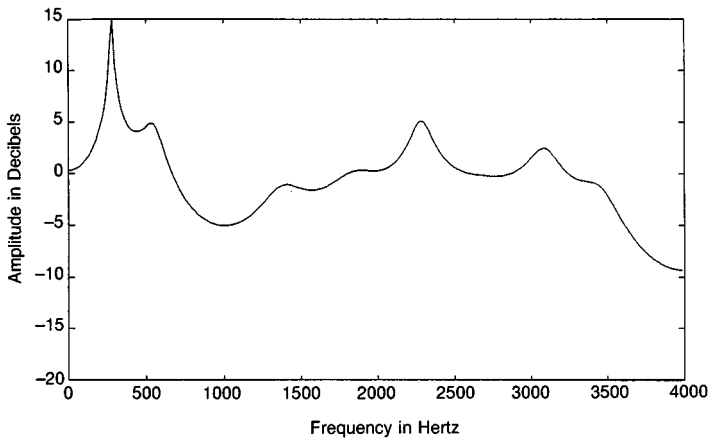


FIGURE 15.12 Sixteenth-order linear predictive analysis of an "i".

Figure 15.13 shows an example of the cepstrum for the voiced /i/ utterance from Fig. 15.7. The cepstrum of such a voiced utterance is characterized by relatively large values in the first one or two milliseconds as well as by pulses of decaying amplitudes at multiples of the pitch period. The first two of these pulses can clearly be seen in Fig. 15.13 at time lags of 7.1 and 14.2 ms. The location and amplitudes of these pulses may be used to estimate pitch and voicing [Rabiner and Schafer, 1978].

In addition to pitch and voicing estimation, a smooth log magnitude function may be obtained by windowing or "liftering" the cepstrum to eliminate the terms which contain the pitch information. Figure 15.14 is one such smoothed spectrum. It was obtained from the DFT of the cepstrum of Fig. 15.13 after first setting all terms of the cepstrum to zero except for the first 16.

Speech Synthesis

Speech synthesis is the creation of speech-like waveforms from textual words or symbols. In general, the speech synthesis process may be divided into three levels of processing [Klatt, 1982]. The first level transforms the text into a series of acoustic phonetic symbols, the second transforms those symbols to smoothed synthesis parameters, and the third level generates the speech waveform from the parameters. While speech synthesizers have

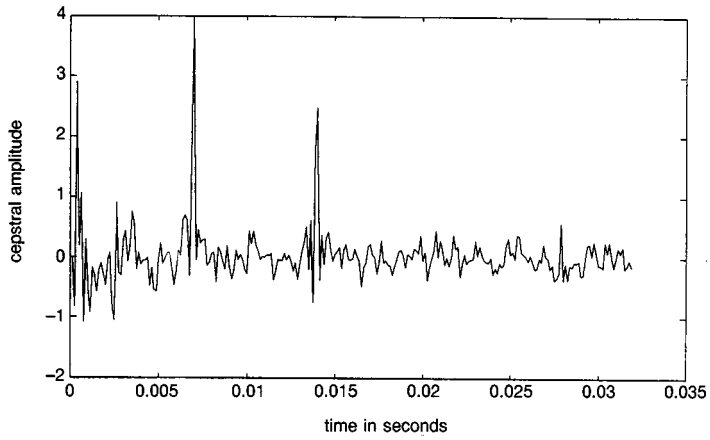


FIGURE 15.13 Real cepstrum of /i/.

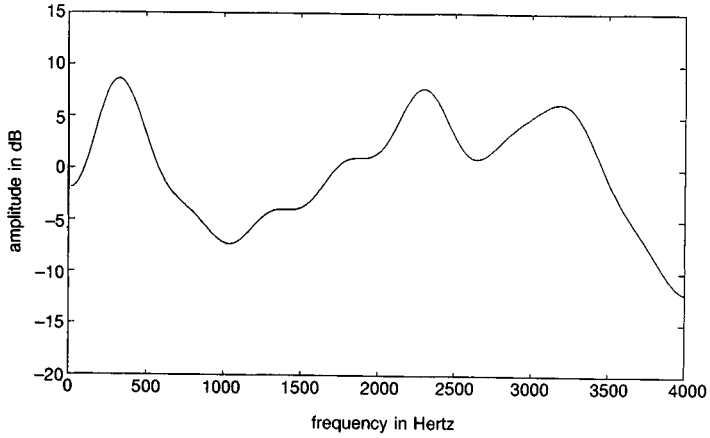


FIGURE 15.14 Smoothed spectrum of /i/ from 16 points of cepstrum.

been designed for a variety of languages and the processes described here apply to several languages, the examples given are for English text-to-speech.

In the first level of processing, abbreviations such as "Dr." (which could mean "doctor" or "drive"), numbers ("1492" could be a year or a quantity), special symbols such as "$", upper case acronyms (e.g., NASA), and nonspoken symbols such as "" (apostrophe) are converted to a standard form. Next prefixes and perhaps suffixes are removed from the body of words prior to searching for the root word in a lexicon, which defines the phonetic representation for the word. The lexicon includes words which do not obey the normal rules of pronunciation, such as "of". If the word is not contained in the lexicon, it is processed by an algorithm which contains a large set of rules of pronunciation.

In the second level, the sequences of words consisting of phrases or sentences are analyzed for grammar and syntax. This analysis provides information to another set of rules which determine the stress, duration, and pitch to be added to the phonemic representation. This level of processing may also alter the phonemic representation of individual words to account for coarticulation effects. Finally, the sequences of parameters which specify the pronunciation are smoothed in an attempt to mimic the smooth movements of the human articulators (lips, jaw, velum, and tongue).

The last processing level converts the smoothed parameters into a time waveform. Many varieties of waveform synthesizers have been used, including formant, linear predictive, and filter-bank versions. These waveform

synthesizers generally correspond to the synthesizers used in speech coding systems which are described in the first section of this chapter.

Defining Terms

Cepstrum: Inverse Fourier transform of the logarithm of the Fourier power spectrum of a signal. The complex cepstrum is the inverse Fourier transform of the complex logarithm of the Fourier tranform of the complex logarithm of the Fourier transform of the signal.

Pitch: Frequency of glottal vibration of a voiced utterance.

Spectrum or power density spectrum: Amplitude of a signal as a function of frequency, frequently defined as the Fourier transform of the autocovariance of the signal.

Speech analysis: Process of extracting time-varying parameters from the speech signal which represent a model for speech production.

Speech synthesis: Production of a speech signal from a model for speech production and a set of time-varying parameters of that model.

Voicing: Classification of a speech segment as being voiced (i.e., produced by glottal excitation), unvoiced (i.e., produced by turbulent air flow at a constriction) or some mix of those two.

Related Topic

14.1 Fourier Transforms

References

J. Allen, "Synthesis of speech from unrestricted text," *Proc. IEEE,* vol. 64, no. 4, pp. 433–442, 1976.

J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* Berlin: Springer-Verlag, 1965.

D. H. Klatt, "The Klattalk Text-to-Speech System" IEEE Int. Conf. on Acoustics, Speech and Signal Proc., pp. 1589–1592, Paris, 1982.

J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech,* Berlin: Springer-Verlag, 1976.

A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.,* pp. 221–226, 1968.

A. V. Oppenheim and R.W. Schafer, *Digital Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1975.

D. O'Shaughnessy, *Speech Communication,* Reading, Mass.: Addison-Wesley, 1987.

L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Englewood Cliffs, N.J.: Prentice-Hall, 1978.

M. J. Ross, H .L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoustics, Speech and Signal Proc.,* vol. ASSP-22, pp. 353–362, 1974.

R. W. Schafer and J. D. Markel, *Speech Analysis,* New York: IEEE Press, 1979.

Further Information

The monthly magazine *IEEE Transactions on Signal Processing,* formerly *IEEE Transactions on Acoustics, Speech and Signal Processing,* frequently contains articles on speech analysis and synthesis. In addition, the annual conference of the IEEE Signal Processing Society, the International Conference on Acoustics, Speech, and Signal Processing, is a rich source of papers on the subject.

15.4 Speech Recognition

Lynn D.Wilcox and Marcia A. Bush

Speech recognition is the process of translating an acoustic signal into a linguistic message. In certain applications, the desired form of the message is a verbatim transcription of a sequence of spoken words. For example, in using speech recognition technology to automate dictation or data entry to a computer, transcription accuracy is of prime importance. In other cases, such as when speech recognition is used as an interface to a database

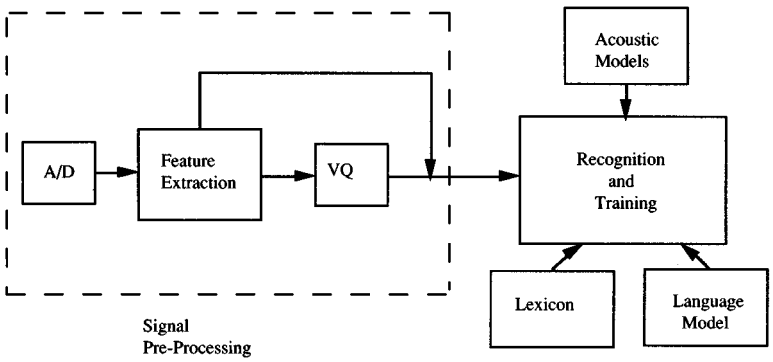


FIGURE 15.15 Architecture for a speech recognition system.

query system or to index by keyword into audio recordings, word-for-word transcription is less critical. Rather, the message must contain only enough information to reliably communicate the speaker's goal. The use of speech recognition technology to facilitate a dialog between person and computer is often referred to as "spoken language processing."

Speech recognition by machine has proven an extremely difficult task. One complicating factor is that, unlike written text, no clear spacing exists between spoken words; speakers typically utter full phrases or sentences without pause. Furthermore, acoustic variability in the speech signal typically precludes an unambiguous mapping to a sequence of words or subword units, such as phones.[1] One major source of variability in speech is coarticulation, or the tendency for the acoustic characteristics of a given speech sound or phone to differ depending upon the phonetic context in which it is produced. Other sources of acoustic variability include differences in vocal-tract size, dialect, speaking rate, speaking style, and communication channel.

Speech recognition systems can be constrained along a number of dimensions in order to make the recognition problem more tractable. Training the parameters of a recognizer to the speech of the user is one way of reducing variability and, thus, increasing recognition accuracy. Recognizers are categorized as speaker-dependent or speaker-independent, depending upon whether or not full training is required by each new user. Speaker-adaptive systems adjust automatically to the voice of a new talker, either on the basis of a relatively small amount of training data or on a continuing basis while the system is in use.

Recognizers can also be categorized by the speaking styles, vocabularies, and language models they accommodate. **Isolated word recognizers** require speakers to insert brief pauses between individual words. **Continuous speech recognizers** operate on fluent speech, but typically employ strict language models, or grammars, to limit the number of allowable word sequences. Wordspotters also accept fluent speech as input. However, rather than providing full transcription, wordspotters selectively locate relevant words or phrases in an utterance. **Wordspotting** is useful both in information-retrieval tasks based on keyword indexing and as an alternative to isolated word recogniton in voice command applications.

Speech Recognition System Architecture

Figure 15.15 shows a block diagram of a speech recognition system. Speech is typically input to the system using an analog transducer, such as a microphone, and converted to digital form. **Signal pre-processing** consists of computing a sequence of acoustic feature vectors by processing the speech samples in successive time intervals. In some systems, a clustering technique known as vector quantization is used to convert these continuous-valued features to a sequence of discrete codewords drawn from a codebook of acoustic prototypes. Recognition of an unknown utterance involves transforming the sequence of feature vectors, or codewords, into an appropriate message. The recognition process is typically constrained by a set of acoustic models which correspond to the basic units of speech employed in the recognizer, a lexicon which defines the vocabulary of the recognizer

[1]Phones correspond roughly to pronunciations of consonants and vowels.

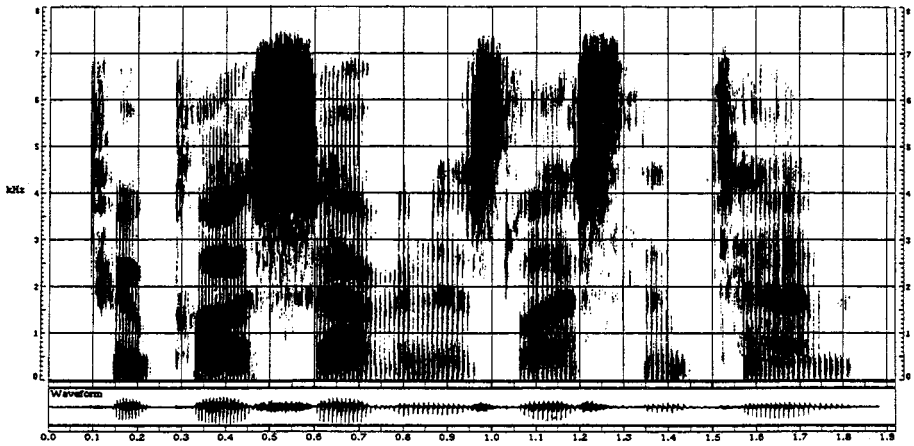


FIGURE 15.16 Speech spectrogram of the utterance "Two plus seven is less than ten." (*Source:* V.W. Zue, "The use of speech knowledge in automatic speech recognition," *Proc. IEEE,* vol. 73, no. 11, pp. 1602–1615, © 1985 IEEE. With permission.)

in terms of these basic units, and a language model which specifies allowable sequences of vocabulary items. The acoustic models, and in some cases the language model and lexicon, are learned from a set of representative training data. These components are discussed in greater detail in the remainder of this chapter, as are the two recognition paradigms most frequently employed in speech recognition: **dynamic time warping** and **hidden Markov models.**

Signal Pre-Processing

An amplitude waveform and speech spectrogram of the sentence "Two plus seven is less than ten" is shown in Fig. 15.16. The spectrogram represents the time evolution (horizontal axis) of the frequency spectrum (vertical axis) of the speech signal, with darkness corresponding to high energy. In this example, the speech has been digitized at a sampling rate of 16 kHz, or roughly twice the highest frequency of relevant energy in a high-quality speech signal. In general, the appropriate sampling rate is a function of the communication channel. In telecommunications, for example, a bandwidth of 4 kHz, and, thus, a Nyquist sampling rate of 8 kHz, is standard.

The speech spectrum can be viewed as the product of a source spectrum and the transfer function of a linear, time-varying filter which represents the changing configuration of the vocal tract. The transfer function determines the shape, or envelope, of the spectrum, which carries phonetic information in speech. When excited by a voicing source, the formants, or natural resonant frequencies of the vocal tract, appear as black bands running horizontally through regions of the speech spectrogram. These regions represent voiced segments of speech and correspond primarily to vowels. Regions characterized by broadband high-frequency energy, and by extremely low energy, result from noise excitation and vocal-tract closures, respectively, and are associated with the articulation of consonantal sounds.

Feature extraction for speech recognition involves computing sequences of numeric measurements, or feature vectors, which typically approximate the envelope of the speech spectrum. Spectral features can be extracted directly from the discrete Fourier transform (DFT) or computed using linear predictive coding (LPC) techniques. Cepstral analysis can also be used to deconvolve the spectral envelope and the periodic voicing source. Each feature vector is computed from a frame of speech data defined by windowing N samples of the signal. While a better spectral estimate can be obtained using more samples, the interval must be short enough so that the windowed signal is roughly stationary. For speech data, N is chosen such that the length of the interval covered by the window is approximately 25 to 30 msec. The feature vectors are typically computed at a frame rate of 10 to 20 msec by shifting the window forward in time. Tapered windowing functions, such as the Hamming window, are used to reduce dependence of the spectral estimate on the exact temporal position of

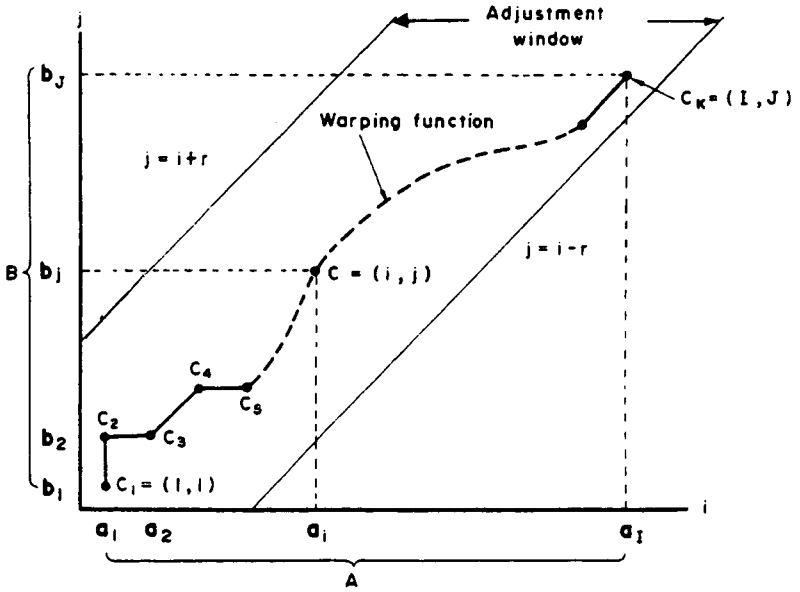


FIGURE 15.17 Dynamic time warping of utterances A and B. (*Source:* S. Furui, *Digital Speech Processing, Synthesis and Recognition,* New York: Marcel Dekker, 1989. With permission.)

the window. Spectral features are often augmented with a measure of the short time energy of the signal, as well as with measures of energy and spectral change over time [Lee, 1988].

For recognition systems which use discrete features, vector quantization can be used to quantize continuous-valued feature vectors into a set or codebook of K discrete symbols, or codewords [Gray, 1984]. The K codewords are characterized by prototypes $y^1 \ldots y^K$. A feature vector x is quantized to the kth codeword if the distance from x to y^k, or $d(x,y^k)$, is less than the distance from x to any other codeword. The distance $d(x,y)$ depends on the type of features being quantized. For features derived from the short-time spectrum and cepstrum, this distance is typically Euclidean or weighted Euclidean. For LPC-based features, the Itakura metric, which is based on spectral distortion, is typically used [Furui, 1989].

Dynamic Time Warping

Dynamic time warping (DTW) is a technique for nonlinear time alignment of pairs of spoken utterances. DTW-based speech recognition, often referred to as "template matching," involves aligning feature vectors extracted from an unknown utterance with those from a set of exemplars or templates obtained from training data. Nonlinear feature alignment is necessitated by nonlinear time-scale warping associated with variations in speaking rate.

Figure 15.17 illustrates the time correspondence between two utterances, A and B, represented as feature-vector sequences of unequal length. The time warping function consists of a sequence of points $F = c_1, \ldots, c_K$ in the plane spanned by A and B, where $c_k = (i_k, j_k)$. The local distance between the feature vectors a_i and b_j on the warping path at point $c = (i, j)$ is given as

$$d(c) = d(a_i, b_j) \tag{15.17}$$

The distance between A and B aligned with warping function F is a weighted sum of the local distances along the path,

$$D(F) = \frac{1}{N} \sum_{k=1}^{K} d(c_k) w_k \tag{15.18}$$

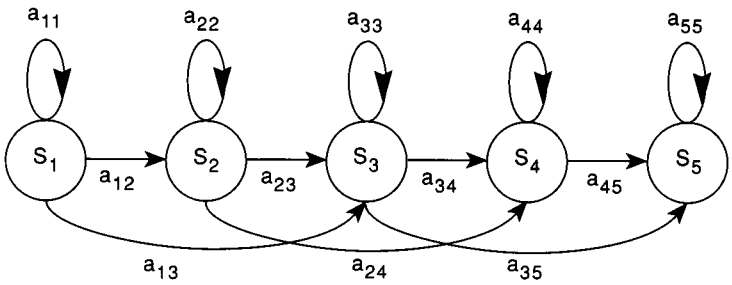


FIGURE 15.18 A typical HMM topology.

where w_k is a nonnegative weighting function and N is the sum of the weights. Path constraints and weighting functions are chosen to control whether or not the distance $D(F)$ is symmetric and the allowable degree of warping in each direction. Dynamic programming is used to efficiently determine the optimal time alignment between two feature-vector sequences [Sakoe and Chiba, 1978].

In DTW-based recognition, one or more templates are generated for each word in the recognition vocabulary. For speaker-dependent recognition tasks, templates are typically created by aligning and averaging the feature vectors corresponding to several repetitions of a word. For speaker-independent tasks, clustering techniques can be used to generate templates which better model pronunciation variability across talkers. In isolated word recognition, the distance $D(F)$ is computed between the feature-vector sequence for the unknown word and the templates corresponding to each vocabulary item. The unknown is recognized as that word for which $D(F)$ is a minimum. DTW can be extended to connected word recognition by aligning the input utterance to all possible concatenations of reference templates. Efficient algorithms for computing such alignments have been developed [Furui, 1989]; however, in general, DTW has proved most applicable to isolated word recognition tasks.

Hidden Markov Models[1]

Hidden Markov modeling is a probabilistic pattern matching technique which is more robust than DTW at modeling acoustic variability in speech and more readily extensible to continuous speech recognition. As shown in Fig. 15.18, hidden Markov models (HMMs) represent speech as a sequence of states, which are assumed to model intervals of speech with roughly stationary acoustic features. Each state is characterized by an output probability distribution which models variability in the spectral features or observations associated with that state. Transition probabilities between states model durational variability in the speech signal. The probabilities, or parameters, of an HMM are trained using observations (VQ codewords) extracted from a representative sample of speech data. Recognition of an unknown utterance is based on the probability that the speech was generated by the HMM.

More precisely, an HMM is defined by:

1. A set of N states $\{S_1 \ldots S_N\}$, where q_t is the state at time t.
2. A set of K observation symbols $\{v_1 \ldots v_K\}$, where O_t is the observation at time t.
3. A state transition probability matrix $A = \{a_{ij}\}$, where the probability of transitioning from state S_i at time t to state S_j at time $t + 1$ is $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$.
4. A set of output probability distributions B, where for each state j, $b_j(k) = P(O_t = v_k | q_t = S_j)$.
5. An initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i)$.

At each time t a transition to a new state is made, and an observation is generated. State transitions have the Markov property, in that the probability of transitioning to a state at time t depends only on the state at time

[1]Although the discussion here is limited to HMMs with discrete observations, output distributions such as Gaussians can be defined for continuous-valued features.

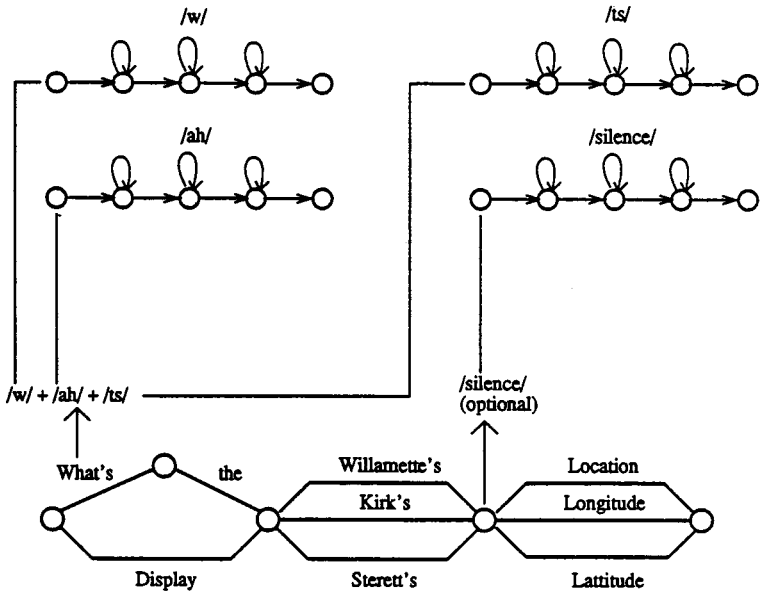


FIGURE 15.19 Language model, lexicon, and HMM phone models for a continuous speech recognition system. (*Source:* K.F. Lee, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System," Ph.D. Dissertation, Computer Science Dept., Carnegie Mellon, April 1988. With permission.)

$t-1$. The observations are conditionally independent given the state, and the transition probabilites are not dependent on time. The model is called hidden because the identity of the state at time t is unknown; only the output of the state is observed. It is common to specify an HMM by its parameters $\lambda = (A, B, \pi)$.

The basic acoustic unit modeled by the HMM can be either a word or a subword unit. For small recognition vocabularies, the lexicon typically consists of whole-word models similar to the model shown in Fig. 15.18. The number of states in such a model can either be fixed or be made to depend on word duration. For larger vocabularies, words are more often defined in the lexicon as concatenations of phone or triphone models. Triphones are phone models with left and right context specified [Lee, 1988]; they are used to model acoustic variability which results from the coarticulation of adjacent speech sounds.

In isolated word recognition tasks, an HMM is created for each word in the recognition vocabulary. In continuous speech recognition, on the other hand, a single HMM network is generated by expressing allowable word strings or sentences as concatenations of word models, as shown in Fig. 15.19. In wordspotting, the HMM network consists of a parallel connection of keyword models and a background model which represents the speech within which the keywords are embedded. Background models, in turn, typically consist of parallel connections of subword acoustic units such as phones [Wilcox and Bush, 1992].

The language model or grammar of a recognition system defines the sequences of vocabulary items which are allowed. For simple tasks, deterministic finite-state grammars can be used to define all allowable word sequences. Typically, however, recognizers make use of stochastic grammars based on n-gram statistics [Jelinek, 1985]. A bigram language model, for example, specifies the probability of a vocabulary item given the item which precedes it.

Isolated word recognition using HMMs involves computing, for each word in the recognition vocabulary, the probability $P(O|\lambda)$ of the observation sequence $O = O_1 \ldots O_T$. The unknown utterance is recognized as the word which maximizes this probability. The probability $P(O|\lambda)$ is the sum over all possible state sequences $Q = q_1 \ldots q_T$ of the probability of O and Q given λ, or

$$p(O|\lambda) = \sum_{Q} P(O, Q|\lambda) = \sum_{Q} P(O|Q, \lambda)P(Q|\lambda) = \sum_{q_1 \ldots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots \quad (15.19)$$

Direct computation of this sum is computationally infeasible for even a moderate number of states and observations. However, an iterative algorithm known as the **forward-backward** procedure [Rabiner, 1989] makes this computation possible. Defining the forward variable α as

$$\alpha_t(i) = P(O_1 \ldots O_t, q_t = S_i|\lambda) \tag{15.20}$$

and initializing $\alpha_1(i) = \pi_i b_i(O_1)$, subsequent $\alpha_t(i)$ are computed inductively as

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \tag{15.21}$$

By definition, the desired probability of the observation sequence given the model λ is

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{15.22}$$

Similarly, the backward variable β can be defined

$$\beta_t(i) = P(O_{t+1} \ldots O_T|q_t = S_i, \lambda) \tag{15.23}$$

The βs are computed inductively backward in time by first initializing $\beta_T(j) = 1$ and computing

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{15.24}$$

HMM-based continuous speech recognition involves determining an optimal word sequence using the **Viterbi** algorithm. This algorithm uses dynamic programming to find the optimal state sequence through an HMM network representing the recognizer vocabulary and grammar. The optimal state sequence $Q^* = (q_1^* \ldots q_T^*)$ is defined as the sequence which maximizes $P(Q|O,\lambda)$, or equivalently $P(Q,O|\lambda)$. Let $\delta_t(i)$ be the joint probability of the optimal state sequence and the observations up to time t, ending in state S_i at time t. Then

$$\delta_t(i) = \max P(q_1 \ldots q_{t-1}, q_t = S_i, O_1 \ldots O_t|\lambda) \tag{15.25}$$

where the maximum is over all state sequences $q_1 \ldots q_{t-1}$. This probability can be updated recursively by extending each partial optimal path using

$$\delta_{t+1}(j) = \max_i \delta_t(i) a_{ij} b_j(O_{t+1}) \tag{15.26}$$

At each time t, it is necessary to keep track of the optimal precursor to state j, that is, the state which maximized the above probability. Then, at the end of the utterance, the optimal state sequence can be retrieved by backtracking through the precursor list.

Training HMM-based recognizers involves estimating the parameters for the word or phone models used in the system. As with DTW, several repetitions of each word in the recognition vocabulary are used to train HMM-based isolated word recognizers. For continuous speech recognition, word or phone exemplars are typically extracted from word strings or sentences [Lee, 1988]. Parameters for the models are chosen based on a maximum likelihood criterion; that is, the parameters λ maximize the likelihood of the training data O, $P(O|\lambda)$. This maximization is performed using the **Baum-Welch** algorithm [Rabiner, 1989], a re-estimation

technique based on first aligning the training data O with the current models, and then updating the parameters of the models based on this alignment.

Let $\xi_t(i,j)$ be the probability of being in state S_i at time t and state S_j at time $t+1$ and observing the observation sequence O. Using the forward and backward variables $\alpha_t(i)$ and $\beta_t(j)$, $\xi_t(i,j)$ can be written as

$$\xi_t(i,\ j) = P(q_t = S_i,\ q_{t+1} = S_j \mid O, \lambda) = \frac{\alpha_t(i)a_{ij}\beta_{t+1}(j)b_j(O_{t+1})}{\sum_{ij=1}^{N} \alpha_t(i)a_{ij}\beta_{t+1}(j)b_j(O_{t+1})} \tag{15.27}$$

An estimate of a_{ij} is given by the expected number of transitions from state S_i to state S_j divided by the expected number of transitions from state S_i. Define $\gamma_t(i)$ as the probability of being in state S_i at time t, given the observation sequence O

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) = \sum_{j=1}^{N} \xi_t(i,\ j) \tag{15.28}$$

Summing $\gamma_t(i)$ over t yields a quantity which can be interpreted as the expected number of transitions from state S_i. Summing $\xi_t(i,j)$ over t gives the expected number of transitions from state i to state j. An estimate of a_{ij} can then be computed as the ratio of these two sums. Similarly, an estimate of $b_j(k)$ is obtained as the expected number of times being in state j and observing symbol v_k divided by the expected number of times in state j.

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,\ j)}{\sum_{t=1}^{T} \gamma_t(i)} \qquad \hat{b}_j(k) = \frac{\sum_{t:O_t=y_k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{15.29}$$

State-of-the-Art Recognition Systems

Dictation-oriented recognizers which accommodate isolated word vocabularies of many thousands of words in speaker-adaptive manner are currently available commercially. So too are speaker-independent, continuous speech recognizers for small vocabularies, such as the digits; similar products for larger (1000-word) vocabularies with constrained grammars are imminent. Speech recognition research is aimed, in part, at the development of more robust pattern classification techniques, including some based on neural networks [Lippmann, 1989] and on the development of systems which accommodate more natural spoken language dialogs between human and machine.

Defining Terms

Baum-Welch: A re-estimation technique for computing optimal values for HMM state transition and output probabilities.

Continuous speech recognition: Recognition of fluently spoken utterances.

Dynamic time warping (DTW): A recognition technique based on nonlinear time alignment of unknown utterances with reference templates.

Forward-backward: An efficient algorithm for computing the probability of an observation sequence from an HMM.

Hidden Markov model (HMM): A stochastic model which uses state transition and output probabilities to generate observation sequences.

Isolated word recognition: Recognition of words or short phrases preceded and followed by silence.

Signal pre-processing: Conversion of an analog speech signal into a sequence of numeric feature vectors or observations.

Viterbi: An algorithm for finding the optimal state sequence through an HMM given a particular observation sequence.

Wordspotting: Detection or location of keywords in the context of fluent speech.

References

S. Furui, *Digital Speech Processing, Synthesis, and Recognition,* New York: Marcel Dekker, 1989.

R. M. Gray, "Vector quantization," *IEEE ASSP Magazine,* vol. 1, no. 2, pp. 4–29, April 1984.

F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE,* vol. 73, no. 11, pp. 1616–1624, Nov. 1985.

K. F. Lee, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System," Ph.D. Dissertation, Computer Science Department, Carnegie Mellon University, April 1988.

R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Computation,* vol. 1, pp. 1–38, 1989.

L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 257–285, Feb. 1989.

H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, no. 1, pp. 43–49, Feb. 1978.

L. D. Wilcox and M. A. Bush, "Training and search algorithms for an interactive wordspotting system," in Proceedings, International Conference on Acoustics, Speech and Signal Processing, San Francisco, March 1992, pp. II-97–II-100.

V. W. Zue, "The use of speech knowledge in automatic speech recognition," *Proc. IEEE,* vol. 73, no. 11, pp. 1602–1615, Nov. 1985.

Further Information

Papers on speech recognition are regularly published in the *IEEE Speech and Audio Transactions* (formerly part of the *IEEE Transactions on Acoustics, Speech and Signal Processing*) and in the journal *Computer Speech and Language*. Speech recognition research and technical exhibits are presented at the annual IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), the biannual European Conference on Speech Communication and Technology (Eurospeech), and the biannual International Conference on Spoken Language Processing (ICSLP), all of which publish proceedings. Commercial applications of speech recognition technology are featured at annual American Voice Input-Output Society (AVIOS) and Speech Systems Worldwide meetings. A variety of standardized databases for speech recognition system development are available from the National Institute of Standards and Technology in Gaithersburg, MD.

15.5 Large Vocabulary Continuous Speech Recognition

Yuqing Gao, Bhuvana Ramabhadran, and Michael Picheny

Speech recognition is the process of converting an acoustic signal to a textual message. High recognition accuracy is of prime importance in order for a speech interface to be of any practical use in a dictation task, or any kind of intelligent human–machine interaction. Speech recognition is made extremely difficult by **co-articulation**, variations in speaking styles, rates, vocal-tract size across speakers, and communication channels. Speech research has been underway for over 4 decades, and many problems have been addressed and solved fully or partially. High performance can be achieved on tasks such as isolated word recognition, small and middle vocabulary recognition, and recognition of speech in nonadverse conditions. Large vocabulary (over 30K words), speaker-independent, continuous speech recognition has been one of the major research targets for years. Although for some large vocabulary tasks, high recognition accuracies have been achieved [7], significant challenges emerge as more and more applications make themselves viable for speech input.

Continuous Speech Recognition

Continuous speech recognition is significantly more difficult than isolated word recognition. Its complexity stems from the following three properties of continuous speech.

1. Word boundaries are unclear in continuous speech, whereas in isolated word recognition they are well-known and can be used to improve the accuracy and limit the search. For example, in the phrase "this ship," the /s/ of "this" is often omitted. Similarly, in "we were away a year," the whole sentence is one long vocalic segment, and the word boundaries are difficult to locate.
2. Co-articulatory effects are much stronger than in isolated speech. Although we try to pronounce words as concatenated sequences of individual speech sounds **(phones),** our articulators possess inertia which retards their motion from one position to another. As a result, a phone is strongly influenced by the previous and the following phones. This effect occurs both within single words and between words and is aggravated as the speaking rate increases.
3. Function words (articles, prepositions, pronouns, short verbs, etc.) tend to be poorly articulated. In particular, the phones are often shortened, skipped, or deleted.

As a result, speech recognition error rates increase drastically from isolated word to continuous speech. Moreover, the processing power needed to recognize continuous speech increases as well.

The primary advantages of continuous speech are two-fold. First, typical speaking rates for continuous speech are 140 to 170 words per minute, while isolated word mode speakers seldom exceed 70 words per minute. Secondly, continuous speech is a natural mode of human communication. Forcing pauses between words introduces artificiality and reduces user friendliness. The unnaturalness of isolated word speech breaks the speaker's train of thought.

Large Vocabulary

In the 1990s, the term "large vocabulary" has come to mean 30K words or more. Although the vocabulary size is certainly not the best measure of a task's difficulty, it does affect the severity of many problems such as the acoustic confusability of words, the degradation in performance due to using sub-word unit models, and the computational complexity of the **hypothesis search**.

Clearly, the number of confusable words grows substantially with the vocabulary size. As the vocabulary size increases, it becomes impractical to model each word individually, because neither the necessary training data nor the requisite storage is available. Instead, models must be based on sub-word units. These sub-word models usually lead to degraded performance because they fail to capture **co-articulation** effects as well as whole-word models. Additionally, the computational complexity of the search requires the introduction of efficient search methods such as "fast match" [26] which reject all but the most plausible word hypotheses to limit the computation effort. These word hypotheses which survive the "fast match" are then subjected to the full detailed analysis. Naturally, this process may introduce search errors, reducing the accuracy.

Some of the key engineering challenges in building speech recognition systems are selecting a proper set of sub-word units (e.g., phones), assembling units into words **(baseforms),** modeling co-articulating effects, accomodating the different stress patterns of different languages, and modeling pitch contours for tone-based languages such as Mandarin.

Overview of a Speech Recognition System

The general architecture of a typical speech recognition system is given in Fig. 15.20. The speech signal is typically input to the system via a microphone or a telephone. Signal preprocessing consists of computing a series of acoustic vectors by processing the speech signal at regular time intervals (frames), which are typically 10 ms long. These acoustic vectors are usually a set of parameters, such as LPC **cepstra** [23] or filter bank outputs (PLP [30], RASTA [28], etc.). In order to capture the change in these vectors over time, they have been augmented with their time derivatives or discriminant projection techniques (e.g., see LDA [10, 29]).

The recognizer consists of three parts: the **acoustic model,** the **language model,** and the hypothesis search. The recognition process involves the use of acoustic models over these feature vectors to label them with their

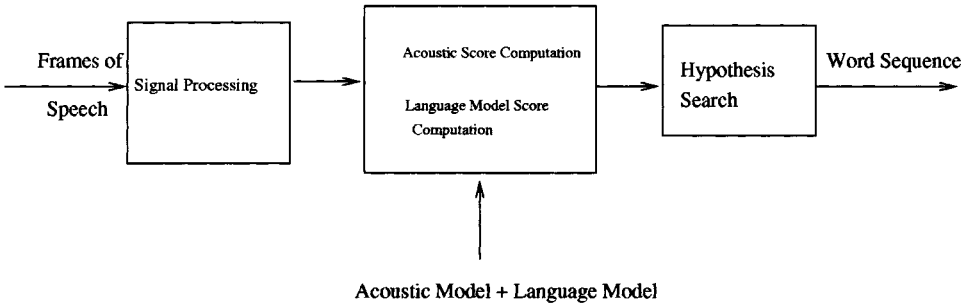


FIGURE 15.20 General architecture of a speech recognition system.

phonetic class. The **acoustic models** usually used are **Hidden Markov Models.** Artificial Neural Networks [16] or Dynamic Time Warping [17] based models have also been used, but will not be covered in this chapter section. Context-dependent acoustic models [9, 10] are obtained by querying the phonetic context using the concept of **tri-phones** or **decision trees** (networks) [2] that are constructed from a large amount of training data. A multidimensional **Gaussian mixture** model is used to model the feature vectors of the training data that have similar phonetic contexts. These models are then used as a set of observation densities in continuous Hidden Markov Models (HMMs). Each feature vector is labeled as a context-dependent phonetic class which is the closest acoustic class to the feature vector. A sequence of labels thus obtained is used to obtain a set of candidate words that are then pruned with the help of a **language model.** A language model bases its prediction of the next word on the history of the words preceding it. Finally, a **hypothesis search** is conducted through all possible sequences of hypothesized words to determine the optimal word sequence given the acoustic observations.

Several adaptation techniques have been proposed to derive speaker-dependent systems from the speaker-independent system described above. These techniques modify/tune the parameters of the acoustic models to the specific speaker.

Hidden Markov Models As Acoustic Models for Speech Recognition

There are many ways to characterize the temporal sequence of speech sounds as represented by a sequence of spectral observations. The most common way is to model the temporal sequence of spectra in terms of a Markov chain to describe the way one sound changes to another by imposing an explicitly probabilistic structure on the representation of the evolutional sequence.

If we denote the spectral vector at time t by O_t the observed spectral sequence, lasting from $t = 1$ to $t = T$, is then represented by

$$\left\{O_t\right\}_{t=1}^{T} = \left(O_1, O_2, \cdots, O_T\right)$$

Consider a first-order N-state Markov chain as illustrated for $N = 3$ in Fig. 15.21. Such a random process has the simplest memory: the value at time t depends only on the value at the preceding time and on nothing that went on before. However, it has a very useful property that leads to its application to speech recognition problem: the states of the chain generate observation sequences while the state sequence itself is hidden from the observer.

The system can be described as being one of the N distinct states, S_1, S_2, …, S_N, at any discrete time instant t. We use the state variable q_t as the state of the system at time t. Assume that the Markov chain is time invariant (homogeneous), so the transition probabilities do not depend on time. The Markov chain is then described by a state transition probability matrix $A = [a_{ij}]$, where

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i), \quad 1 \le i, j \le N \tag{15.30}$$

The transition probabilities satisfy the following constraints:

$$a_{ij} \ge 0 \tag{15.31}$$

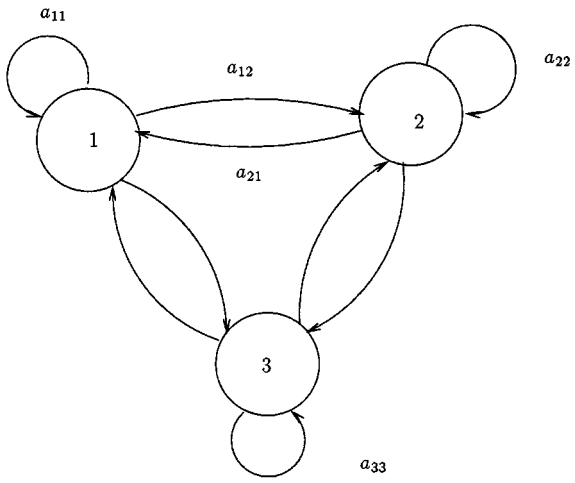


FIGURE 15.21 A first-order three-state hidden Markov model.

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i \tag{15.32}$$

Assume that at the initiated time, $t = 0$, the state of the system q_0 is specified by an initial state probability vector $\pi^T = [\pi_1, \pi_2, \ldots, \pi_N]$. Then for any state sequence $q = (q_0, q_1, q_2, \ldots, q_T)$, where $q_t \in \{S_1, S_2, \ldots, S_N\}$, the probability of q being generated by the Markov chain is

$$P(q|A, \pi) = \pi_{q0}, a_{q0q1} a_{q1q2} \ldots a_{qT-1qT} \tag{15.33}$$

Suppose now that the state sequence q is a sequence of speech sounds and cannot be observed directly. Instead, observation O_t is produced with the system in some unobserved state q_t (where $q_t \in \{S_1, S_2, \ldots, S_N\}$). Assume that the production of O_t in each possible S_i, $i = 1, 2, \ldots, N$ is stochastic and is characterized by a set of observation probability measures $B = \{b_i(O_t)_{i=1}^N\}$, where

$$b_i(O_t) = P(O_t|_{q_t} = S_i) \tag{15.34}$$

If the state sequence q that led to the observation sequence $O = (O_1, O_2, \ldots, O_T)$ is known, the probability of being generated by the system is assumed to be

$$P(O|q, B) = b_{q_1}(O_1) b_{q_2}(O_2) \ \ldots \ b_{q_T}(O_T) \tag{15.35}$$

Therefore, the joint probability of O and q being produced by the system is

$$P\left(O \mid \pi, A, b\right) = \pi_{q_0} \Pi_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}\left(O_t\right) \tag{15.36}$$

The probability of producing the observation sequence O by the random process without assuming knowledge of the state sequence is

$$P\left(O \mid \pi, A, B\right) = \sum_{q} P\left(O, q \mid \pi, A, B\right) = \sum_{q} \pi_{q_0} \Pi_{t-1}^{T} a_{q_t-1q_t} b_{q_t}\left(O_t\right) \tag{15.37}$$

Continuous Parameter Hidden Markow Models

The triple (π, A, B) defines a **Hidden Markov Model** (HMM). More specifically, a hidden Markov model is characterized by the following:

1. A state space $\{S_1, S_2, \ldots, S_N\}$. Although the states are not explicitly observed, for many applications there is often some physical significance attached to the states. In the case of speech recognition, this is often a phone or a portion–inital, middle, final–of a phone. We donote the state at time t as q_t.
2. A set of observations $O = (O_1, O_2, \ldots, O_T)$. The observations can be a set of discrete symbols chosen from a finite set, or continuous signals (or vectors). In speech recognition application, although it is possible to convert continuous speech representations into a sequence of discrete symbols via vector quantization codebooks and other methods, serious degradation tends to result from such discretization of the signal. In this article, we focus on HMMs with continuous observation output densities to model continuous signals.
3. The initial state distribution $\pi = \{\pi_i\}$ in which
4. The state transition probability distribution $A = \{a_{ij}\}$defined in Eq. (15.30).
5. The observations probability distribution, $B = \{b_j(O_t)\}$, defined in Eq. (15.34).

$$\pi_i = P(q_0 = S_i),\ 1 \le i \le N$$

Given the form of HMM, the following three basic problems of interest must be solved for the model to be useful in applications.

Task 1 (Evaluation): Given the observation sequence $O = (O_1, O_2, \ldots, O_T)$ and a model $\lambda = (\pi, A, B)$, how does one efficiently compute $P(O|\lambda)$?

Task 2 (Estimation): Given the observation sequence $O = (O_1, O_2, \ldots, O_T)$, how does one solve the inverse problem of estimating the parameters in λ?

Task 3 (Decoding): Given the observation sequence O and a model λ, how does we deduce the most likely state sequence q that is optimal in some sense or best explains the observations?

The Evaluation Problem

With unbounded computational power, Eq. (15.37) can be used to compute $P(O|\lambda)$. However, it involves on the order of $2TN^T$ calculations, because the summation in Eq. (15.37) has N^{T+1} possible sequences. This is computationally infeasible even for small values of N and T.

An iterative algorithm known as the **forward-backward** procedure makes this computation possible. Defining the forward variable α as

$$\alpha_t(i) = P(O_1, \ldots, O_t, q_t = S_i|\lambda) \tag{15.38}$$

and initializing $\alpha_1(i) = \pi_i b_i(O_1)$, subsequent $\alpha_t(i)$ are computed inductively as

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \tag{15.39}$$

By definition, the desired probability of the observation sequence given the model λ is

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{15.40}$$

Another alternative is to use the backward procedure by defining the backward variable β:

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T / q_t = S_i, \lambda) \tag{15.41}$$

$\beta_t(i)$ is the probability of the partial observation sequence from t + a to the end T, biven state S_i and model λ. The initial values are $\beta_T(i) = 1$ for all i. The values at time, $T - 1$, $T - 2$, …, 1, can be computed inductively:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{15.42}$$

The probability of the observation sequence given the model λ can be expressed in terms of the forward and backward probabilities:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) = \sum_{i=1}^{N} \alpha_T(i) \tag{15.43}$$

The forward and backward variables are very useful and will be used in the next section.

The Estimation Problem

Given an observation sequence or a set of sequences, (multiple utterances) the estimation problem is to find the "right" model parameter values that specify a model most likely to produce the given sequence. In speech recognition, this is called training. There is no known closed form analytic solution for the maximum likelihood model parameters. Nevertheless we can choose $\lambda = (\pi, A, B)$ such that its likelihood, $P(O|\lambda)$, is locally maximized using an iterative procedure such as the **Baum-Welch** re-estimation method (a form of the EM [expectation-maximization] method [4]. The method intoduces an auxiliary function $Q(\hat{\lambda}, \lambda)$ and maximizes it.

$$Q(\hat{\lambda}, \lambda) = \sum_{q} P(O, q | \hat{\lambda}) \log P(O, q | \lambda) \tag{15.44}$$

The re-estimation technique consists of first aligning the training data O with the current models, and then updating the parameters of the models based on the alignment to obtain a new estimate λ.

Let $\zeta_t(i, j)$ be the probability of being in state S_i at time t and state S_j at time $t + 1$, given the model λ and the observation sequence O:

$$\zeta_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) = \frac{P(q_t + S_i, q_{t+1} = S_j, O | \lambda)}{P(O|\lambda)} \tag{15.45}$$

Using the forward and backward variables defined in section 3.2, $\alpha_t(i)$ and $\beta_t(j)$, $\zeta_t(i, j)$ can be written as

$$\xi_t(i,j) = \frac{\alpha_t(i)\alpha_{ij}\beta_{t+1}(j)b_j(O_{t+1})}{\sum_{i,j=1}^{N}\alpha_t(i)\alpha_{ij}\beta_{t+1}(j)b_j(O_{t+1})} \tag{15.46}$$

An estimate of a_{ij} is given by the expected number of transitions from state S_i to state S_j, divided by the expected number of transitions from state S_i. Define $\gamma_t(i)$ as the probability of being in state S_i at time t, given the observation sequence O

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) = \sum_{j=1}^{N}\zeta_t(i,j) \tag{15.47}$$

Summing $\gamma_t(i)$ over t yields a quantity that can be interpreted as the expected number of transitions from state S_i. Summing $\zeta_t(i, j)$ over t gives the expected number of transitions from state S_i to S_j. An estimate of a_{ij} can be computed as the ratio of these two sums.

$$\hat{a}_{ij} = \frac{\Sigma_{t=1}^{T-1}\zeta_t(i,j)}{\Sigma_{t=1}^{T}\gamma_t(i)} \tag{15.48}$$

For the discrete observation case, an estimate of $b_j(k)$ is obtained as the expected number of times being in state S_i and observing symbol ν_k divided by the expected number of times in state S_j.

$$\hat{b}_j(k) = \frac{\Sigma_{t:O_t=\nu_k}\gamma_t(j)}{\Sigma_{t=1}^{T}\gamma_t(j)} \tag{15.49}$$

The most general representation of a continuous density of HMMs is a finite mixture of the form

$$b_j(O) = \sum_{k=1}^{M} c_{jk} N(O, \mu_{jk}, U_{jk}), \quad 1 \le j \le N$$

where O is the observation vector being modeled, c_{jk} is the mixture coefficient for the kth mixture in state j and N is any log-concave or elliptically symmetric density. Typically, we assume that N is Gaussian with mean vector μ_{jk} and covariance matrix U_{jk} for the kth mixture component in state j. The mixture weights c_{jk} satisfy the constraint:

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 1 \le j \le N \tag{15.50}$$

$$c_{jk} \ge 0, \quad 1 \le j \le N, \quad 1 \le k \le M \tag{15.51}$$

Let $\gamma_t(j,k)$ be the probability of being in state S_j at time t with k-th mixture component accounting for O_t:

$$\gamma_t(j,k) = P(q_t = S_j, k_t = k \mid O, \lambda) = \frac{P(q_t = S_j, k_t = k, O \mid \lambda)}{P(q_t = S_j, O \mid \lambda)} \tag{15.52}$$

The re-estimation formula for the coefficients of the mixture density are:

$$\hat{c}_{jk} = \frac{\Sigma_{t=1}^{T} \gamma_t(j,k)}{\Sigma_{t=1}^{T} \Sigma_{k=1}^{M} \gamma_t(j,k)} \tag{15.53}$$

$$\hat{\mu}_{jk} = \frac{\Sigma_{t=1}^{T} \gamma_t(j,k) o_t}{\Sigma_{t=1}^{T} \gamma_t(j,k)} \tag{15.54}$$

$$\hat{U}_{jk} = \frac{\Sigma_{t=1}^{T} \gamma_t(j,k)(o_t - \mu_{jk})(o_t - \mu_{jk})^T}{\Sigma_{t=1}^{T} \gamma_t(j,k)} \tag{15.55}$$

Details on how the re-estimation formulas are derived from the auxiliary function $Q(\hat{\lambda},\lambda)$ can be found in [25] and [23].

Viterbi Algorithm: One Solution for the Decoding Problem

We define the optimal state sequence $q^\star = (q_1^\star, \ldots, q_T^\star)$ is defined as the sequence which maximizes $P(q|O, \lambda)$, or equivalently $P(q, O|\lambda)$. Let $\delta_t(i)$ be the joint probability of the optimal state sequence and the observation up to time t, ending in state S_i at time t. Then,

$$\delta_t(i) = max\ P(q_1, \ldots\ q_{t-1}, q_t = S_i, O_1, \ldots, O_t|\lambda) \tag{15.56}$$

where the maximum is over all state sequences $q_1, \ldots, q_{t-1}$. This probability can be updated recursively by extending each partial optimal path using

$$\delta_t(i) = \max_i \delta_t(i) a_{ij} b_j(O_{t+1}) \tag{15.57}$$

At each time t, it is necessary to keep track of the optimal precursor t of state j, that is, the state that maximized the above probability. Then, at the end of the utterance, the optimal state sequence can be retrieved by backtracking through the precursor list [11].

Speaker Adaptation

In spite of recent progress in the design of speaker-independent (SI) systems, error rates are still typically two or three times higher than equivalent speaker-dependent (SD) systems. Variability in both anatomical and personal characteristics contribute to this effect. Anatomical differences include the length of the vocal tract, the size of the nasal cavity, etc. Similarly, there are variable speaking habits, such as accent, speed, and loudness. The straight-forward approach which blindly mixes the statistics for all speakers discards useful information.

The large amount of speaker-specific data required to train SD systems renders them impractical for many applications. However, it is possible to use a small amount of the new speaker's speech (adaptation data) to "tune" the SI models to the new speaker. Ideally, we would like to retain the robustness of well-trained SI models, yet improve the appropriateness of the models for the new speaker. Such methods are called speaker adaptation techniques. The adaptation is said to be *supervised* if the true text transcript of the adaptation data is known; otherwise, the adaption is said to be *unsupervised*.

Maximum *a posteriori* Estimation

A widely used speaker adaptation method maximizes the *posterior* estimation of HMMs [3]. The conventional maximum likelihood (ML) based algorithms assume the HMM parameters to be unknown but fixed, and the parameter estimators are derived entirely from the training observation sequence using the **Baum-Welch** algorithms. Sometimes, prior information about the HMM parameters is available, whether from subject matter

considerations or from previous experience. Designers may wish to use this prior information–in addition to the sample observations–to infer the HMM parameters.

The Maximum a Posteriori **(MAP)** framework naturally incorporates prior information into the estimation process, which is particularly useful for dealing with problems posed by sparse training data, where ML estimates become inaccurate. MAP parameter estimates approah the ML estimates when data is plentiful, but are governed by the prior information in the absence of data. If λ is the parameter vector to be estimated from the observation O with probability density function (pdf) $f(O|\lambda)$ and g is the prior pdf of λ, then the MAP estimate is defined as the maximum of the posterior pdf of λ, $g(\lambda|o)$.

Rather than maximizing the auxiliary function $Q(\hat{\lambda}, \lambda)$ as in Eq. 15.44, we instead maximize an auxiliary function that includes a contribution from the prior distribution of the model parameters.

$$R\left(\hat{\lambda}, \lambda\right) = Q\left(\hat{\lambda}, \lambda\right) + \log g\left(\hat{\lambda}\right) \tag{15.58}$$

The appropriate prior distributions are Gaussian distributions for the means, gamma distributions for the inverse variances, and Dirichlet distributions for the mixture weights [3].

The problem with MAP is that it adapts only parameters for which explicit training data is available, and it converges slowly for tasks where there are limited adaptation data and many parameters to be estimated. Many adaptation algorithms [15] [14] have been developed which attempt to generalize from "nearby" training data points to overcome this difficulty.

Transform-Based Adaptation

Another category of adaptation technique uses a set of regression-based transforms to tune the mean and variance of a **hidden Markov model** to the new speaker. Each of the transformations is applied to a number of HMMs and estimated from the corresponding data. Using this sharing of transformations and data, the method can produce improvements even if a small amount of adaptation data is available for the new speaker by using a global transform for all HMMs in the system. If more data is available, the number of transforms is increased.

Maximum Likelihood Linear Regression **(MLLR)**

The MLLR framework was first introduced in [27]. Consider the case of a continuous density HMM system with Gaussian output distributions. A particular Gaussian distribution, g, is characterized by a mean vector, μ_g, and a covariance matrix U_g. Given a speech vector o_t, the probability of that vector being generated by a Gaussian distribution g is $b_g(o_t)$:

$$b_g\left(o_t\right) = \frac{1}{\left(2\pi\right)^{d/2} |U_g|^{1/2}} e^{-1/2\left(o_t - \mu_g\right)^T U_g^{-1}\left(o_t - \mu_g\right)}$$

The adaptation of the mean vector is obtained by applying a transformation matrix W_g to the extended mean vector ζ_g to obtain an adapted mean vector $\hat{\mu}_g$

$$\hat{\mu}_g = W_g \zeta_g$$

where W_g is a d*(d + 1) matrix which maximizes the likelihood of the adaptation data, and the ζ_g is defined as

$$\zeta_g = [\Omega, \mu_1, \ldots, \mu_d]^T$$

where Ω is the offset term for the regression.

The probability for the adapted system becomes

$$b_g(o_t) = \frac{1}{(2\pi)^{d/2} |U_g|^{1/2}} e^{-1/2(o_t - W_g\mu_g)^T U_g^{-1}(o_t - W_g\mu_g)}$$

The auxiliary fuction in Eq. 15 can be used here to estimate W_g. It can be shown the W_g can be estimated using the equation below:

$$\sum_{t=1}^{T} \gamma_g(t) U_g^{-1} o_t \zeta_g^T = \sum_{t=1}^{T} \gamma_g(t) U_g^{-1} \hat{W}_g \zeta_g \zeta_g^T \tag{15.59}$$

where $\gamma_g(t)$ is the posterior probability of occupying state q at time t given that the observation sequence O is generated.

$$\gamma_g(t) = \frac{1}{P(O|\lambda)} \sum_{\phi \in \Phi} P(O, s_t = q|\lambda)$$

The **MLLR** algorithm can also be extended to transform the covariance matrices [38].

Cluster-Based Speaker Adaptation

Yet another category of speaker adaptation methodology is based on the fact that a speech training corpus contains a number of training speakers, some of whom are closer, acoustically, to the test speaker than others. Therefore, given a test speaker, if the **acoustic models** are re-estimated from a subset of the training speakers who are acoustically close to the test speaker, the system should be a better match to the test data of the speaker. A further improvement is obtained if the acoustic space of each of these selected training speakers is transformed, by using transform-based adaptation method, to come closer to the test speaker.

This scheme was shown to produce better speaker adaptation performance than other algorithms, for example MLLR [27] or **MAP** adaptation [3], when only a small amount of adaptation data was available.

However, the implementation of this method required the entire training corpus to be available online for the adaptation process, and this is not practical in many situations. This problem can be circumvented if a model is stored for each of the training speakers, and the transformation is applied to the model. The transformed models are then combined to produce the speaker-adapted model. However, due to the large number of training speakers, storing the models of each training speaker would require a prohibitively large amount of storage. Also, we may not have sufficient data from each training speaker to robustly estimate the parameters of the speaker-dependent model.

To solve this problem and retain the advantage of the method, a new algorithm is presented [21]. It is to precluster the training speakers acoustically into clusters. For each cluster, an HMM system (called a cluster-dependent system) is trained using speech data from the speakers who belong to the cluster. When a test speaker's data is available, we rank these cluster-dependent systems according to the distances between the test speaker and each cluster, and a subset of these clusters, acoustically closest to the test speaker, is chosen. Then the model for each of the selected clusters is transformed further to bring the model closer to the test speaker's acoustic space. Finally, these adapted cluster models are combined to form a speaker-adapted system. Hence, compared to [22], we now choose clusters that are acoustically close to the test speaker, rather than individual training speakers.

This method solves the problem of excessive storage for the training speaker models because the number of clusters is far fewer than the number of training speakers, and it is relatively inexpensive to store a model for each cluster. Also, as each cluster contains a number of speakers, we have enough data to robustly estimate the parameters of the model for the cluster.

Vocal Tract Length Normalization (VTL)

Several attempts have been made to model variations in vocal tract length across several speakers. The idea was originally introduced by Bamberg [42] and revived through a parametric approach in [39]. Assume a uniform tube with length L for the model of the vocal tract. Then each formant frequency will be proportional to $1/L$. The first-order effect of a difference in vocal tract length is the scaling of the frequency axis. The idea behind VTL is to rescale or warp the frequency axis during the signal processing step in a speech recognition system, to make speech from all speakers appear as if it was produced by a vocal tract of a single standard length. Such normalizations have led to significant gains in accuracy by reducing variability amongst speakers and allowing the pooling of training data for the construction of sharper models. Three VTL methods have been recently proposed. In [39], a parametric method of normalization which counteracts the effect of varied vocal tract length is presented. This method is particularly useful when only a small amount of training data is available and requires the determination of the formant frequencies. In [40], an automated method is presented that uses a simple generic voiced speech model to rapidly select appropriate frequency scales. This generic model is a mixture of 256 multiversity Gaussians with diagonal covariances trained on the unwarped data. Different warp scales are selected to linearly transform the frequency axis of the speaker's data. The resulting warped features are scored against the generic model. The warp scale that scores the best is selected as the warp scale for that speaker. An iterative process updates the generic model with the new features obtained after warping each speaker with the best warp scale. Once the best warp scales for each speaker have been determined, SI models are built with the appropriately warped feature vectors. This warp selection method allows data from all speakers to be merged into one set of cannonical models. In [41], a class of transforms are proposed which achieve a remapping of the frequency axis much like the conventional VTL methods. These mappings known as all-pass transforms, are linear in the cepstral domain which makes speaker normalization simple to implement. The parameters of these transforms are computed using conjugate gradient methods.

Modeling Context in Continuous Speech

Speech recognition cannot be accurately modeled by a concatenation of elementary HMMs corresponding to individual **phones** of a word baseform. A phone is a sub-word acoustic unit of speech. The realizations of the phones depend on their context. This is especially true for continuous speech where the phenomenon called *co-articulation* is observed. Co-articulation is when the pronunciation of a phoneme is affected by the phones preceding and following it, such as, the *t* in *top* and *pot*. This section discussed several methods that will yield HMM building blocks that take into account phonetic context. A word is specified by its phonetic baseform, the phones are transformed into their appropriate allophones according to the context in which they appear, and a concatenation of the HMMs of these allophones results in the word HMM. Two standard approaches are used to make use of contextual information:

1. **Tri-phones** as building blocks
2. **Decision trees** that lead to general contextual building blocks

Tri-Phones

In order to take into account the influence of context on pronunciation, many speech recognizers base their modeling on the tri-phone concept. The tri-phone concept was first introduced by Chow et al. [5, 24] and Lee et al. [7, 8] in the 1980s. In this concept, the pronunciation of a phone is influenced by the preceding and following phone (i.e., the triplet is used to model the realization of the phone). The phone p embedded in the context p_1 and p_2 is specified by the tri-phone p_1pp_2, where p_1 and p_2 are the previous and the following **phones**. Different such realizations of p are called allophones. This amounts to saying that the contextual influence of the preceding and following phone is most important. If this solution were carried out literally, the resulting allophone alphabet would be too large to be useful. For example, a phonetic alphabet of size M would produce M^3 allophones. Even though in practice, not all M^3 allophones occur, the number of possible allophones is still large.

So the tri-phone method relies on an equivalence classification of the context of phones. One simple scheme involves the clustering of tri-phones into distinct categories to characterize them. The criterion used for

clustering can be an information theoretic measure such as likelihood or entropy. This is the concept behind *generalized tri-phones* [8]. Decision trees (described in the next section) can also be used to determine the distinct categories. Another scheme is to tie the HMM distributions of these tri-phones. More recently, methods that cluster individual states of the HMM [1] have been introduced. A drawback in using tri-phones is that wider contexts (i.e., three, four, or five phones to the left) may be important.

Decision Trees

The purpose of the decision tree is to map a large number of conditions (i.e., phonetic contexts) to a small manageable number of equivalence classes. Each terminal node of the tree represents a set of phonetic contexts. The aim in decision tree construction is to find the best possible definition of equivalence classes [2].

During decoding, the acoustic models to be used for a given observation string are chosen based on the current acoustic context — that is, by pouring the data down the decision tree until a terminal node is reached and using the models at that terminal node to compute the likelihood of the data.

Decision Tree Construction

Maximum Likelihood (ML) estimation is one common technique used for constructing a decision tree; i.e., the aim is to find the different classes that maximize the likelihood of the given training data. A binary decision tree is grown in the following fashion.

1. From among a set of questions (at the phonetic or lexeme level), the best question for partitioning the data into two classes (i.e., the question that maximizes the likelihood) is found.
2. The above step is repeated recursively on each of the two classes until either there is insufficient data to continue or the best question is not sufficiently helpful.

The easiest way to construct a decision tree is to create — in advance — a list of possible questions for each variable that may be tested. Finding the best question at any given node consists of subjecting all the relevant variables to each of the questions on the corresponding list and picking the best combination of the variable and the question.

In building an acoustic decision tree using phonetic context, at least 10 variables may be interrogated, 5 preceding phones and 5 following phones. Since all of these variables belong to the same phonetic alphabet, only one set of questions needs to be prepared, where each question is a subset of the phonetic alphabet.

Typically, trees are less than 20 layers deep, as beyond that the algorithm runs out of data.

Let $X_1 \ldots X_n$ denote n discrete random variables whose values may be tested. Let Q_{ij} denote the jth predetermined question for X_i

1. Starting at the root node, try splitting each node into two subnodes.
2. For each variable X_i, evaluate questions Q_{i1}, Q_{i2}, etc. Let Q_b denote the best question estimated using any one of the criteria described earlier. The best question at a node is the question that maximizes the likelihood of the training data at that node after applying the question.
3. Find the best pair X_i, Q_b denoted as X, Q_b.
4. If the selected question is not sufficiently helpful (gain in likelihood due to the split is not significant) or does not have sufficient data points, make the current node a leaf.
5. Otherwise, split the current node into two new subnodes according to the answer of question Q_b on variable X_b.

The algorithm stops when all nodes are either too small to split further or have been marked as leaves. Overtraining is prevented by limiting the number of asked questions.

Questions

A decision tree has a question associated with every non-terminal node. These can be grouped into continuous and discrete questions.

Discrete questions. If X is a discrete random variable that takes values in some finite alphabet R, then a question about X has the form: Is X an element of S where S is a subset of R? Typically, questions are of the form "Is the preceding phone a vowel?" or "Is the following phone an unvoiced stop?"

Continuous questions. If X is a continuous random variable that takes real values, a question about X has the form: X ∈ t where t is some real value. Instead of limiting the questions to a predefined set, we could search for the best subset of values taken by the random variable at any node and use the best question found. This implies that we generate questions on the fly during tree construction. The disadvantages of this approach include too much CPU time to search for the best subset and because there are so many subsets, there is too much freedom in the tree-growing algorithm, resulting in over-training or spurious questions that do not generalize very well.

All of these questions can be constructed in advance by experts. For example, phonetic questions can be generated by linguists or by algorithms [9].

Language Modeling

Humans recognize words based not only on what they hear, but also on what they have heard in the past, as well as what they anticipate to hear in the future. It is this capability that make humans the best speech recognition systems. Modern speech recognition systems attempt to achieve this human capability through *language modeling*. Language modeling is the art and science of anticipating or predicting words or word sequences from nonacoustic sources of information such as context, structure, and grammar of the particular language, previously heard word sequences.

In large vocabulary speech recognition, in which word sequences W are uttered to convey some message, the language model $P(W)$ is of critical importance to the recognition accuracy. In most cases, the language model must be estimated from a large text corpus.

For practical reasons, the word sequence probability $P(W)$ is approximated by

$$P_N(W) = \prod_{i=1}^{Q} P(w_i \mid w_{i-1}, w_{i-2}, \ldots w_{i-N+1}) \tag{15.60}$$

This is called an N-gram language model, where N is the number of words from the history that are used in the computation. Typically, $N = 3$ and these are referred to as trigram language models, Q is the number of words in the sequence being decoded.

The conditional probabilities in Eq. (15.60) are estimated by the simple relative frequency approach described in [23].

The maximum entropy approach is a method of estimating the conditional probability distributions (discribed below). In cases when the text corpus is not large enough to reliably estimate the probabilities, smoothing techniques such as linear smoothing (deleted interpolation) are applied (described below).

Perplexity is a measure of performance of language models. Perplexity, defined in Eq. (15.61) is the average word branching factor of the language model.

$$B = 2^H = P(w_1, w_2, \ldots, w_Q)^{-1/Q} \tag{15.61}$$

Perplexity is an important parameter is specifying the degree of sophistication in a recognition task, from the source uncertainty to the quality of the language model.

Other techniques that have been used in language modeling include desision tree models [43] and automatically inferred linked grammars to model long range correlations [44].

Smoothing

In computing the language model probabilities, we desire the following: fewer parameters to estimate; the available data is sufficient for the estimation of parameters; and that the probability can be constructed at recognition time from the parameter values while occupying limited storage. Several smoothing techniques have been proposed to handle the scarcity of data [25]. These are essential in the construction of n-gram language models. They include linear smoothing, also known as deleted interpolation, backing-off, bucketing

and equivalence classification techniques. An extensive empirical study of these techniques for language modeling is given in [35]. A brief description of two smoothing techniques is covered in this section. Linear smoothing, is due to Jelinek and Mercer [37], where a class of smoothing models that involve linear interpolation are presented. The maximum-likelihood estimate is interpolated with the smoothed lower-order distribution defined analogously, i.e.

$$P_{\text{int}\,erp}\left(\omega_i \mid \omega_{i-n+1}^{i-1}\right) = \lambda_{\omega_{i-n+1}^{i-1}} P_{ML}\left(\omega_i \mid \omega_{i-n+1}^{i-1}\right) + \left(1 - \lambda_{\omega_{i-n+1}^{i-1}}\right) P_{\text{int}\,erp}\left(\omega_i \mid \omega_{i-n+2}^{i-1}\right) \tag{15.62}$$

To yield meaningful results, the training data used to estimate $\lambda_{\omega_{i-n+1}^{i-1}}$ need to be distinct from the data used to estimate P_{ML}. In *held-out interpolation*, a section of training data is reserved for this purpose. In[37], a technique called deleted interpolation is described where different parts of the training data rotate in train either P_{ML} or the $\lambda_{\omega_{i-n+1}^{i-1}}$ and the results are then averaged. The other widely used smoothing technique in speech recognition is the backing-off technique described by Katz [36]. Here, the Good-Turing estimate [36] is extended by adding the interpolation of higher-order models with lower-order models. This technique performs best for bigram models estimated from small training sets. The trigram **language model** probability is defined by,

$$p(\omega_3|\omega_1,\omega_2) = \lambda_3 f(\omega_3|\omega_1,\omega_2) + \lambda_2 f(\omega_3|\omega_2) + \lambda_1 f(\omega_3) \tag{15.63}$$

The backing-off technique suggests that if the counts $C(\omega_1, \omega_2)$ is sufficiently large, the $f(\omega_3|\omega_1,\omega_2)$ by itself is a better estimate of $p(\omega_3|\omega_1,\omega_2)$. Hence, for different values of counts, a different estimate of $p(\omega_3|\omega_1,\omega_2)$ is used. Several variations on the choice of this threshold and the Good-Turing type function are described in [25].

Maximum Entropy based Language Models

A maximum-likelihood approach for automatically constructing maximum entropy models is presented in [34]. The maximum entropy method finds a model that simultaneously satifies a set of constraints Such a model is a method of estimating the conditional probability distributions. The principle is simple: model all that is known and assume nothing about that which is not known. Let x and y be a set of random variables such that $P(y|\varkappa)$ is the probability that the model assigns an output y given $\varkappa$. Let $f(\varkappa, y)$ be the indicator function (the expected value of this function is the feature function) that takes a binary value of 1 or 0 to fit the training data. If $P(\varkappa)$ satisfies $\sum_{\varkappa,y} P(\varkappa)P(y|\varkappa)f(\varkappa,y) = d(i)$ where $d(i)$ are the constraints then there must be a probability $P(\varkappa)$ that satisfies all the constraints uniformly. A mathematical measure of uniformity of conditional distributions is the conditional entropy, H. The solution to this problem is obtained by selecting the model with maximum entropy from the set of possible models, C, i.e.,

$$\mathrm{p}^{\star} = \underset{p \epsilon C}{arg\max} H \tag{15.64}$$

It can be shown that p* is well-defined and that there is always a model with maximum entropy in any constrained set C. For simple cases, the above equation can be solved mathematically, but for a more general case, the use of Lagrange multipliers from constrained optimization theory is employed. This approach leads to the following statement. The maximum entropy model subject to a set of constraints, has the parametric form given by Eq. 15.65,

$$P\left(y \mid \aleph\right) = Q_{\lambda}\left(\aleph\right) \exp \sum_{i} \lambda_i f\left(\aleph, y\right) \tag{15.65}$$

where the Lagrangian multipliers, $\lambda_i s$ can be determined by maximizing the Lagrangian, $\lambda = H + \sum_i \lambda_i (p(f_i) - \hat{p}(f_i))$. $Q_{\lambda}(\varkappa)$ is the normalization constant and $p(f_i)$ and $\hat{p}(f_i)$ are the emperical and expected distributions. Since the Lagrangian is the log-likelihood for the exponential model, P, the solution states that the model with the maximum entropy is one that maximizes the likelihood of the training data.

Details on the construction of maximum entropy models, techniques for the selection of features to be included in models and the computations of the parameters of these models are addressed in [34]. Techniques for computing the parameters of such models such as, hill climbing, iterative projection and iterative scaling algorithms are described in [25].

Hypothesis Search

It is the aim of the speech recognizer to determine the best possible word sequence given the acoustic observation; that is, the word string $\hat{W}$ that satisfies

$$\hat{W} = \underset{w}{argmax}\, P(W)P(A \mid W) \tag{15.66}$$

Since the word string is made up of a sequence of words, the search for $\hat{W}$ is done from several possible hypotheses. **Viterbi** Search, a time-synchronous search strategy, and Tree Search, a time-asynchronous search strategy, are presented here.

Viterbi Search

The Viterbi algorithm [11] introduced previously finds the most likely path through an HMM. Equation (15.66) demands that we find for each candidate word string W, the probability of the set of paths that corresponds to the word string W and then identify the word string whose set of paths has the highest probability.

Section 2 described the hidden Markev Model concept. The Viterbi algorithm that finds the maximizing state sequence for successive levels i (there may be several), and deciding at the final level k, from among the competing sequences. At each level i, the paths whose probabilities fall below a threshold are purged. A traceback from the final state for which the probability is maximized to the start state in the purged trellis yields the most likely state sequence.

Each word is represented as a concatenation of several HMMs, one corresponding to each of the phones that make up the word. If we are using a bigram **language model**, then $P(W) = P(w_i)\Pi_{i=2}^{n}\, P(w_i|w_{i-1}) = 0$, and $\hat{W}$ the most likely path through these HMMs. The number of states is proportional to the vocabulary size, V. If we are using the trigram language model, then, $P(W) = P(w_1)P(w_2|w_1)\Pi_{i=3}^{n}\, P(w_i|w_{i-2}, w_{i-1})$, and the graph becomes more complicated with the number of states being proportional to $|V|^2$. No practical algorithms exist for finding the exact solution, but the Viterbi algorithm will find the most likely path through these HMMs whose identity can then determine the recognized word string.

One drawback of the Viterbi algorithm is the number of states that have to be evaluated for a bigram language model, even for a practical vocabulary size of 60,000 words. A shortcut that is commonly used is the *beam search*. Here, the maximal probability P_{i-1}^{m} of the states at stage $i - 1$, i.e., $\max_{s1, \ldots, sl} P(s_1, s_2, \ldots, s_{i-1}, s_{l=i}, y_1, \ldots, y_i|s_0)$ is computed and used as the basis for computing a dynamic threshold to prune out all states in the trellis whose path probabilities fall below this threshold. Multi-pass search strategies have been proposed over the thresholding used in the beam search to handle more complex language models [6].

Tree Search

The search for the most likely word sequence can be thought of as searching for a path in a tree whose branches are labeled with the various words of the vocabulary V such that there are $|V|$ branches leaving each node, one for each word (i.e., the size of the vocabulary). Typically, in large vocabulary continuous speech recognition, this search from a tree of possible hypotheses turns out to be a very large computational effort. Hence, the search is limited by a fast match approach [26] that will reject from consideration several branches of the tree without subjecting them to a detailed analysis. The Viterbi algorithm achieves the same kind of pruning using the beam search approach and multi-pass strategies.

Stack Search

Stack search algorithms for speech recognition have been used at IBM [19] and MIT Lincoln Labs [20]. This heuristic search algorithm helps to reduce computational and storage needs without sacrificing accuracy. Any

tree search must be based on some evaluation criterion related to the search's purpose. The algorithm below is a popular algorithm used for the heuristic determination of minimum-cost paths

1. Insert into a stack all the single-branch paths corresponding to the words of the vocabulary.
2. Sort these entries in descending order of a function $F(w_i)$, where $w_i \in$ vocabulary V.
3. If the top entry in the stack is the end of the utterance, the search ends; otherwise, each entry in the stack is extended using $F(\cdot)$ for all possible words in the vocabulary and inserted into the stack while maintaining the stack order.
   $F(\cdot)$ is

$$F\left(w_1^k\right) = \max_{w^r} P\left(a_1^m, w_1^k\right) \tag{15.67}$$

   where w^r denotes a word string of length r and a_1^m denotes the acoustic data to be recognized.

The methods described in [14] incorporate the definition of an envelope that is used to mark partial paths in the stack as *alive* or *dead*; these considerably speed up the search. In [13], a tree search strategy called the *envelope search* is presented. This is a time-asynchronous search that combines aspects of the A* search with the time-synchronous **Viterbi** search. Several bi-directional search strategies have been proposed by Li et al. [18]. Kenny et al. discuss several aspects of A* algorithms in [12]. A different approach involving majority decisions on observed discrete acoustic output strings leading to a polling fast match is introduced by Bahl et al. in [15]. The use of multiple stacks is yet another way to control the search procedure and is presented in [13].

Tree Search vs. Viterbi Search

A Viterbi search of a trellis finds the most likely succession of transitions through a composite HMM composed of word HMMs. The number of states in a trellis stage (determined by the end states of the word HMMs) must be limited to keep the search's storage and computational requirements feasible. The tree search imposed no such constraints on the number of end states as long as this search does not prune out the correct path. Both algorithms are suboptimal in the sense that they do not guarantee to find the most probable word string.

State-of-the-Art Systems

In the 1998 DARPA Hub-4E English Broadcast News Benchmark Test, an overall recognition error rate of 13.5% was achieved. This test includes recognition of baseline broadcast speech, spontaneous broadcast speech, speech over telephone channels, speech in the presence of background music, speech under degraded acoustic conditions, speech from non-native speakers, and all other kinds of speech. Details can be obtained from the NIST Web site. Another benchmark test is the Switchboard Task, which is the transcription of conversations between two people over the telephone. Error rates for this task are approximately 37%. In the Airline Travel Information System (ATIS) speech recognition evaluation conducted by DARPA, error rates close to 2% have been obtained. High recognition accuracies have been obtained for digit recognition, with error rates under 1% (TIMIT database [31]).

Challenges in Speech Recognition

Some of the issues that still arise in speech recognition, and make interesting research problems for the present and the future, include:

1. Accurate transcription of spontaneous speech compared to read speech is still a major challenge because of its inherent casual and incoherent nature, embedded disfluencies, and incomplete voicing of several **phones** or words.
2. Recognizing speech between individuals and/or multiple speakers in a conference.
3. Robustness of recognition to different kinds of channels, background noise in the form of music, speech over speech, variation in distances between the speaker and the microphone, etc.
4. Recognition across age groups, speaking rates, and accents.
5. Building of a **language model** for an unknown domain and the addition of out-of-vocabulary (oov) words.

6. In order to dynamically adapt to speakers (i.e., make use of their speech when no transcription is available), unsupervised adaptation is necessary. To do this accurately, we need a confidence measure on the decoded script.
7. Speech recognition systems do not have any understanding of decoded speech. To move toward understanding/machine translation, we need some post-processing of the transcription that could lead to intelligent conversational systems.

Applications

The use of speech as a means of input, in a fashion similar to the use of the keyboard and the mouse, has resulted in the application of speech recognition to a wide set of fields. These can be broadly divided into three segments: desktop, telephony, and embedded systems.

1. In the desktop areas, continuous speech recognition has been used for dictation of text documents, for commands to navigate the desktop environment, and Internet surfing. Dictation accuracies of the order of 96% and greater have been achieved. The main players in this field include IBM with their ViaVoice series of products. Dragon Systems, L&H, Philips, and Microsoft.[1] Software tailored to dictation in specialized fields, such as radiology, general medicine, and the legal domain, have also been put out by some of these companies. Recorded speech using hand-held digital recorders can also be transcribed subsequently by the same software.
2. Telephony is an emerging field for applications of speech recognition. These include repertory dialing, automated call type recognition, credit card validation, directory listening retrieval, speaker identification, financial applications such as trading of stocks and mutual funds, banking, voice mail transcription,

   companies have their own specialized products for telephony.
3. The use of speech input for embedded systems is a relatively new field because only recently handheld systems have adequate CPV and memory for accurate speech recognition.

Defining Terms

Acoustic Model: Any statistical or syntactic model that represents a speech signal.
Baseforms: Representation of a word as a sequence of phones.
Baum-Welch algorithm: A form of EM (expectation-maximization) is an iterative procedure to estimate the parameters of a stochastic model by maximizing the likelihood of the data.
Cepstra: The Fourier Transform of the logarithm of the power spectrum sampled at regular intervals.
Co-articulation: Pronunciation of **phones** being influenced by the previous and following phones.
Decision trees: A technique used to group several conditions into classes.
Forward-backward: A recursive algorithm for computing the posterior probability of a HMM using forward and backward variables.
Gaussian mixtures: Convex combination of Gaussian (a kind of probability distribution function) functions.
Hidden Markov Model (HMM): A stochastic model that uses state transition and output probabilities to generate observation sequences.
Hypothesis search: Search through a large set of hypotheses of word sequences to find the optimal word sequence.
Language Model: Language Models predict words or word sequences from nonacoustic sources of information, such as context, structure, and grammar of the particular language.
Linear Prediction Coefficients (LPC): A representation of an analog signal using an Auto Regressive model.
MAP: Maximum *a posteriori*. Technique for speaker adaptation.
MLLR: Maximum Likelihood Linear Regression. Technique for speaker adaptation.
Phones: Sub-word acoustic unit.

[1]While this is not an exhaustive list of companies with continuous speech recognition software products, they are the leaders in the field to date.

Signal pre-processing: Conversion of an analog speech signal into a sequence of numeric feature vectors or observations.

Speaker adaptation: The process of using a small amount of data from a new speaker to tune a set of speaker-independent acoustic models to a new speaker.

Supervised and Unsupervised Adaptation: In speaker adaptation, the procedure is said to be supervised if the true transcription of the adaptation data is known and is unsupervised otherwise.

Tri-phone: Context-dependent model of a phone as a function of its previous and succeeding phones.

Viterbi: An algorithm for finding the optimal state sequence through an HMM, given a particular observation sequence.

References

1. Young, S. J. and Woodland, P. C., State clustering in HMM-based continuous speech recognition, *Computer Speech and Language*, 8, 369, 1994.
2. Bahl, L. R. et al., Decision trees for phonological rules in continuous speech, *ICASSP*, 1991.
3. Gauvain, Jean-Luc and Lee, Chin-Hui, Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Transactions on Speech and Audio Processing*, 2, 1994.
4. Baum, L. E., An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes, *Inequalities*, 3, 1, 1972.
5. Chow, Y. et al., BYBLOS: The BBN continuous speech recognition system, *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 89, 1987.
6. Lee, C. H. , Soong, F. K., Paliwal, K. K., Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, 1996.
7. Lee, K. and Hon, H., Large vocabulary speaker independent continuous speech recognition, *IEEE International Conference on Acoustics Speech and Signal Processing*, 1988.
8. Lee, K., Hon, H., Hwang, M., Mahajan, S., and Reddy, R., The Sphinx speech recognition system, *IEEE International Conference on Acoustics Speech and Signal Processing*, 1989.
9. Bahl, L., de Souza, P., Gopalakrishman, P. S., and Picheny, M., Context-dependent vector quantization for continuous speech recognition, *ICASSP*, 1993.
10. Bahl, L., de Souza, P., Gopalakrishman, P. S., Nahamoo, D., and Picheny, M., Robust methods for using context dependent features and models in a continuous speech recognizer, *ICASSP*, I, 533, 1994.
11. Viterbi, A. J., Error bounds for convolution codes and an asymmetrically optimum decoding algorithm, *IEEE Transactions on Information Theory*, IT-13, 260, 1967.
12. Kenny, P. et al., A new fast match for very large vocabulary continuous speech recognition, *ICASSP*, II, 656, 1993.
13. Bahl, L. R., Gopalakrishman, P. S., and Mercer, R. L., Search issues in large vocabulary speech recognition, *Proceedings of the 1993 Workshop on Automatic Speech Recognition*, 1993.
14. Gopalakrishman, P. S., Bahl, L. R., and Mercer, R. L., A tree search strategy for large-vocabulary continuous speech recognition, *ICASSP*, I, 572, 1995.
15. Bahl, L. R., Bakis, R., de Souza, P. V., and Mercer, R. L., Obtaining candidate words by polling in a large vocabulary speech recognition system, *ICASSP*, I, 489, 1998.
16. Lippman, R. P., Review of neural networks for speech recognition, in *Readings in Speech Recognition*, Waibel, A. and Lee, K. F., Eds., Morgan Kaufmann, San Mateo, CA, 1990.
17. Rabiner, L. R. and Levinson, S. E., Isolated and connected word recognition — Theory and selected applications, *IEEE Transactions on Communications*, COM-29, 621, 1981.
18. Li, Z., Boulianne, G., Laboute, P., Barszcz, M., Garudadri, H., and Kenny, P., Bi-directional graph search strategies for speech recognition, *Computer Speech and Language*, 10, 295, 1996.
19. Bahl, L. R., Jelinek, F., and Mercer, R. L., A maximum likelihood approach to continuous speech recognition, *IEEE Trans. Pat. Anal. and Mach. Int.*, PAMI-5, 179, 1983.
20. Paul, D., An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model, *Proc. DARPA Workshop on Speech and Natural Language*, pp. 405, 1992.
21. Gao et al., Speaker adaptation based on pre-clustering training speakers, *Eurospeech-97*, pp. 2091, 1997.

22. Padmanabhan et al., Speaker clustering transformation for speaker adaptation in large vocabulary speech recognition systems, *ICASSP*, 1996.
23. Rabiner, L. and Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice-Hall Signal Process Series, 1993.
24. Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., and Makhoul, J., Context-dependent modeling for acoustic-phonetic recognition of continuous speech, *ICASSP*, 1985.
25. Jelinek, F., *Statistical Methods for Speech Recognition*, MIT Press, 1997.
26. Bahl, L. R. et al., A fast approximate match for large vocabulary speech recognition, *IEEE Transactions on Speech and Audio Processing*, 1, 59, 1993.
27. Leggetter, C. J. and Woodland, P. C., Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech and Language*, Vol. 9, pp. 171, 1995.
28. Hermansky, H. and Morgan, N., RASTA processing of speech, *IEEE Transactions on Speech and Audio Processing*, 2, 587, 1994.
29. Hunt, M. J., A statistical approach to metrics for word and syllable recognition, *Journal of Acoustic Society of America*, 66(S1), S35(A), 1979.
30. Hermansky, H., Perceptual linear predictive (PLP) analysis of speech, *Journal of Acoustic Society of America*, 87(4), 1748, 1990.
31. Lamel, L., Speech database development: Design and analysis of the acoustic-phonetic corpus, *Proceedings of the DARPA Speech Recognition Workshop*, pp. 100, 1986.
32. Lee, K., *Automatic Speech Recognition*, Kluwer Academic, 1989.
33. Pallett, D. S., Fiscus, J. G., Fisher, J. S., Garafolo, W. M., Lund, B. A., Martin, A., and Brzybocki, M. A., 1994 Benchmark Tests for the ARPA Spoken Language Program, *Proceedings of the spoken language systems technology workshop*, Austin, TX, Jan. 22–25 1995.
34. Berger, A. L., Pietra, S. A., V. J., A maximum entropy approach to natural language processing, *Computational Linguistics*, Vol. 22, Np. 1, pp. 1, pp. 39–73, Mar. 1996.
35. Chen, S. A., Goodman, J., An empirical study of smoothing techniques for language modeling, *Technical Report*, TR-10-98, Havard University, August 1998.
36. Katz, S. M., Estimation of probabilities from Sse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-35(3):400–401, Mar. 1987.
37. Jelinek, F. and Mercer, R. L., Interpolated estimation of Markov source parameters from sparse data, *Proceedings of the Workshop on Pattern Recognition in Practice*, May 1980.
38. Gales, M., Maximum likelihood linear transformations for HMM-based speech recognition, *Computer Speech and Language*, Vol. 12, pp 75–98, 1998.
39. Eide, E., et al., A parametric approach to vocal tract normalization, P*roceedings of the 15th Annual Speech Research Symposium*, CLSP, Baltimore, pp. 161-167, June 1995.
40. Wegmann, S. et al., Speaker normalization on conversational telephone speech, *ICASSP-96*, Vol. 1, pp. 339–341, May 1996.
41. McDonough, J. et al., Speaker adaptation with all-pass transforms, *ICASSP-99*, Vol. II, pp. 7575-760, Phoneix, May 1999.
42. Bamberg, P., Vocal tract normalization, *Verbex Internal Technical Report*, 1981.
43. Bahl, L. R. et al., A tree based statistical language model for natural language speech, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37(7), 1989.
44. Berger, A. et al., The Candide System for Machine Translation, *Proceedings of the ARPA Conference on Human Language Technology*, New Jersey, 1994.
45. Chen, S., Adaptation by correlation, *Proceedings of the DARPA Speech Recognition Workshop*, Virginia, 1997.
46. Shinoda, K., Lee, C.-H., Structural MAP speaker adaptation using hierarchical priors, *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp 381–388, 1997.

For Further Information

There are three major speech-related conferences each year, namely, *International Conference on Acoustics, Speech and Signal Processing (ICASSP), International Conference on Spoken Language Processing (ICSLP), and European Conference on Speech Communication and Technology (EUROSPEECH)*. Besides this, Defense Advanced Research Projects Agency (DARPA) conducts workshops on Broadcast News Transcription (transcription of live television broadcasts) and Switchboard (conversations between individuals over the telephone) tasks. Also, there are several conferences addressing specific issues such as phonetic sciences, robust methods for speech recognition in adverse conditions, etc. Journals related to speech include *IEEE Transactions on Speech and Audio Processing, IEEE Transactions on Signal Processing, Computer and Speech Language, Speech Communications* and *IEEE Transactions on Information Theory.* Additional details on the statistical techniques used in speech recognition can be found in several books [23, 25, 32]. A good review of current techniques can also be found in [6].

Acknowledgements

The authors wish to thank Dr. Harry Printz and Dr. R. T. Ward of IBM T. J. Watson Research Center for their careful review of this manuscript and many useful suggestions.

Pillai, S.U., Shim, T.I., Batalama, S.N., Kazakos, D., Daum, F. "Spectral Estimation and Modeling"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

16

Spectral Estimation and Modeling

S. Unnikrishna Pillai
Polytechnic University

Theodore I. Shim
Polytechnic University

Stella N. Batalama
State University of New York at Buffalo

Dimitri Kazakos
University of Southwestern Louisiana

Fred Daum
Raytheon Company

16.1 Spectral Analysis
Historical Perspective • Modern Spectral Analysis
16.2 Parameter Estimation
Bayesian Estimation • Mean-Square Estimation • Minimax Estimation • Maximum Likelihood Estimation • Other Parameter Estimation Schemes
16.3 Kalman Filtering
Kalman Filter Equations • Kalman Filter Examples • Extended Kalman Filter • Nonlinear Filters • Practical Issues

16.1 Spectral Analysis

S. Unnikrishna Pillai and Theodore I. Shim

Historical Perspective

Modern spectral analysis dates back at least to Sir Isaac Newton [Newton, 1671], whose prism experiments with sunlight led him to discover that each color represented a particular wavelength of light and that the sunlight contained all wavelengths. Newton used the word *spectrum,* a variant of the Latin word *specter,* to describe the band of visible light colors.

In the early eighteenth century, Bernoulli discovered that the solution to the wave equation describing a vibrating string can be expressed as an infinite sum containing weighted sine and cosine terms. Later, the French engineer Joseph Fourier in his *Analytical Theory of Heat* [Fourier, 1822] extended Bernoulli's wave equation results to arbitrary periodic functions that might contain a finite number of jump discontinuities. Thus, for some $T_0 > 0$, if $f(t) = f(t + T_0)$ for all t, then $f(t)$ represents a periodic signal with period T_0 and in the case of real signals, it has the Fourier series representation

$$f(t) = A_0 + 2\sum_{k=1}^{\infty} (A_k \cos k\omega_0 t + B_k \sin k\omega_0 t)$$

where $\omega_0 = 2\pi/T_0$, and

$$A_k = \frac{1}{T_0}\int_0^{T_0} f(t)\cos k\omega_0 t\, dt, \qquad B_k = \frac{1}{T_0}\int_0^{T_0} f(t)\sin k\omega_0 t\, dt$$

with A_0 representing the dc term ($k = 0$). Moreover, the infinite sum on the right-hand side of the above expression converges to $[f(t_{-0}) + f(t_{+0})]/2$. The total power P of the periodic function satisfies the relation

$$P = \frac{1}{T_0}\int_0^{T_0} |f(t)|^2\, dt = A_0^2 + 2\sum_{k=1}^{\infty}(A_k^2 + B_k^2)$$

implying that the total power is distributed only among the dc term, the fundamental frequency $\omega_0 = 2\pi/T_0$ and its harmonics $k\omega_0$, $k \geq 1$, with $2(A_k^2 + B_k^2)$ representing the power contained at the harmonic $k\omega_0$. For every periodic signal with finite power, since $A_k \to 0$, $B_k \to 0$, eventually the overharmonics become of decreasing importance.

The British physicist Schuster [Schuster, 1898] used this observation to suggest that the partial power $P_k = 2(A_k^2 + B_k^2)$ at frequency $k\omega_0$, $k = 0 \to \infty$, be displayed as the spectrum. Schuster termed this method the *periodogram*, and information over a multiple of periods was used to compute the Fourier coefficients and/or to smooth the periodogram, since depending on the starting time, the periodogram may contain irregular and spurious peaks. A notable exception to periodogram was the linear regression analysis introduced by the British statistician Yule [Yule, 1927] to obtain a more accurate description of the periodicities in noisy data. Because the sampled periodic process $x(k) = \cos k\omega_0 T$ containing a single harmonic component satisfies the recursive relation

$$x(k) = ax(k-1) - x(k-2)$$

where $a = 2\cos\omega_0 T$ represents the harmonic component, its noisy version $x(k) + n(k)$ satisfies the recursion

$$x(k) = ax(k-1) - x(k-2) + n(k)$$

Yule interpreted this time series model as a recursive harmonic process driven by a noise process and used this form to determine the periodicity in the sequence of sunspot numbers. Yule further generalized the above recursion to

$$x(k) = ax(k-1) + bx(k-2) + n(k)$$

where a and b are arbitrary, to describe a truly autoregressive process and since for the right choice of a, b the least-square solution to the above autoregressive equation is a damped sinusoid, this generalization forms the basis for the modern day parametric methods.

Modern Spectral Analysis

Norbert Wiener's classic work on Generalized Harmonic Analysis [Wiener, 1930] gave random processes a firm statistical foundation, and with the notion of ensemble average several key concepts were then introduced. The formalization of modern day probability theory by Kolmogorov and his school also played an indispensable part in this development. Thus, if $x(t)$ represents a continuous-time stochastic (random) process, then for every fixed t, it behaves like a **random variable** with some **probability density function** $f_x(x,t)$. The ensemble average or **expected value** of the process is given by

$$\mu_x(t) = E[x(t)] = \int_{-\infty}^{\infty} x f_x(x,t)\, dx$$

and the statistical correlation between two time instants t_1 and t_2 of the random process is described through its **autocorrelation function**

$$R_{xx}(t_1,t_2) = E[x(t_1)x^*(t_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{-\infty} x_1 x_2^* f_{x_1x_2}(x_1,x_2,t_1,t_2)dx_1dx_2 = R_{xx}^*(t_2,t_1)$$

where $f_{x1\,x2}(x_1, x_2, t_1, t_2)$ represents the joint probability density function of the random variable $x_1 = x(t_1)$ and $x_2 = x(t_2)$ and * denotes the complex conjugate transpose in general. Processes with autocorrelation functions that depend only upon the difference of the time intervals t_1 and t_2 are known as wide sense stationary processes. Thus, if $x(t)$ is wide sense stationary, then

$$E[x(t+\tau)x^*(t)] = R_{xx}(\tau) = R_{xx}^*(-\tau)$$

To obtain the distribution of power versus frequency in the case of a **stochastic process,** one can make use of the Fourier transform based on a finite segment of the data. Letting

$$P_T(\omega) = \frac{1}{2T}\left|\int_{-T}^{T} x(t)e^{-j\omega t}dt\right|^2$$

represent the power contained in a typical realization over the interval $(-T, T)$, its ensemble average value as $T \to \infty$ represents the true power contained at frequency ω. Thus, for wide sense stationary processes

$$\begin{aligned} S(\omega) &= \lim_{T\to\infty} E[P_T(\omega)] = \lim_{T\to\infty}\int_{-T}^{T}\int_{-T}^{T} R_{xx}(t_1 - t_2)e^{-j\omega(t_1-t_2)}dt_1dt_2 \\ &= \lim_{T\to\infty}\int_{-2T}^{2T} R_{xx}(\tau)\left(1 - \frac{|\tau|}{2T}\right)e^{-j\omega\tau}d\tau = \int_{-\infty}^{\infty} R_{xx}(\tau)e^{-j\omega\tau}d\tau \geq 0 \end{aligned} \tag{16.1}$$

Moreover, the inverse relation gives

$$R_{xx}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)e^{j\omega\tau}d\omega \tag{16.2}$$

and hence

$$R_{xx}(0) = E\left[|x(t)|^2\right] = P = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\,d\omega$$

Thus $S(\omega)$ represents the power spectral density and from Eqs. (16.1) and (16.2), the power spectral density and the autocorrelation function form a Fourier transform pair, the well-known Wiener–Khinchin theorem.

If $x(kT)$ represents a discrete-time wide sense stationary stochastic process, then

$$r_k = E\{x\big((n+k)T\big)x^*(nT)\} = r_{-k}^*$$

and the power spectral density is given by

$$S(\omega) = \sum_{k=-\infty}^{\infty} r_k e^{-jk\omega T}$$

or in terms of the normalized variable $\theta = \omega T$,

$$S(\theta) = \sum_{k=-\infty}^{\infty} r_k e^{-jk\theta} = S(\theta + 2\pi k) \geq 0 \tag{16.3}$$

and the inverse relation gives the autocorrelations to be

$$r_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\theta)e^{jk\theta}d\theta = r_{-k}^{*}$$

Thus, the power spectral density of a discrete-time process is periodic. Such a process can be obtained by sampling a continuous-time process at $t = kT$, $|k| = 0 \rightarrow \infty$, and if the original continuous-time process is band-limited with a two-sided bandwidth equal to $2\omega_b = 2\pi/T$, then the set of discrete samples so obtained is equivalent to the original process in a mean-square sense.

As Schur [Schur, 1917] has observed, for discrete-time stationary processes the nonnegativity of the **power spectrum** is equivalent to the nonnegative definiteness of the Hermitian Toeplitz matrices, i.e.,

$$S(\theta) \geq 0 \Leftrightarrow \mathbf{T}_k = \begin{pmatrix} r_0 & r_1 & \cdots & r_k \\ r_1^{*} & r_0 & \cdots & r_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_k^{*} & r_{k-1} & \cdots & r_0 \end{pmatrix} = \mathbf{T}_k^{\star} \geq 0, \qquad k = 0 \rightarrow \infty \tag{16.4}$$

If $x(nT)$ is the output of a discrete-time linear time-invariant causal system driven by $w(nT)$, then we have the following representation:

$$w(nT) \rightarrow \boxed{H(z) = \sum_{k=0}^{\infty} h(kT)z^k} \rightarrow x(nT) = \sum_{k=0}^{\infty} h(kT)w\Big((n-k)T\Big) \tag{16.5}$$

In the case of a stationary input, the output is also stationary, and its power spectral density is given by

$$S_x(\theta) = \left|H(e^{j\theta})\right|^2 S_w(\theta) \tag{16.6}$$

where $S_w(\theta)$ represents the power spectral density of the input process. If the input is a white noise process, then $S_w(\theta) = \sigma^2$ and

$$S_x(\theta) = \sigma^2\left|H(e^{j\theta})\right|^2$$

Clearly if $H(z)$ is rational, so is $S_x(\theta)$. Conversely, given a power spectral density

$$S_x(\theta) = \sum_{k=-\infty}^{\infty} r_k e^{jk\theta} \geq 0 \tag{16.7}$$

that satisfies the integrability condition

$$\int_{-\pi}^{\pi} S_x(\theta)\, d\theta < \infty \tag{16.8}$$

and the physical realizability (Paley–Wiener) criterion

$$\int_{-\pi}^{\pi} \ln S_x(\theta)\, d\theta > -\infty \tag{16.9}$$

there exists a unique function $H(z)$ that is analytic together with its inverse in $|z| < 1$ (minimum phase factor) such that

$$H(z) = \sum_{k=0}^{\infty} b_k z^k, \qquad |z| < 1 \tag{16.10}$$

and

$$S_x(\theta) = \lim_{r \to 1-0} |H(re^{j\theta})|^2 = |H(e^{j\theta})|^2, \; a.e.$$

$H(z)$ is known as the Wiener factor associated with $S_x(\theta)$ and as Eq. (16.5) shows, when driven by white noise, it generates a stochastic process $x(nT)$ from past samples and its power spectral density matches with the given $S_x(\theta)$.

In this context, given a finite set of autocorrelations $r_0, r_1, \ldots, r_n$, the spectral extension problem is to obtain the class of all extensions that match the given data, i.e., such an extension $K(\theta)$ must automatically satisfy

$$K(\theta) \geq 0$$

and

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} K(\theta) e^{jk\theta} d\theta = r_k, \quad k = 0 \to n$$

in addition to satisfying Eqs. (16.8) and (16.9).

The solution to this problem is closely related to the trigonometric moment problem, and it has a long and continuing history through the works of Schur [1917]; Nevanlinna, Akheizer and Krein [Akheizer and Krein, 1962]; Geronimus [1954]; and Shohat and Tamarkin [1970], to name a few. If the given autocorrelations are such that the matrix $\mathbf{T}_n$ in Eq. (16.4) is singular, then there exists an $m \leq n$ such that $\mathbf{T}_{m-1}$ is positive definite ($\mathbf{T}_{m-1} > 0$) and $\mathbf{T}_m$ is singular [$\det \mathbf{T}_m = 0$, $\det(.)$ representing the determinant of (.)]. In that case, there exists a unique vector $\mathbf{X} = (x_0, x_1, \ldots, x_m)^T$ such that $\mathbf{T}_m \mathbf{X} = 0$ and further, the autocorrelations have a unique extension given by

$$c_k = \sum_{i=1}^{m} P_i e^{jk\theta_i}, \qquad |k| = 0 \to \infty \tag{16.11}$$

where $e^{j\theta i}$, $i = 1 \to m$ are the m zeros of the polynomial $x_0 + x_1 z + \ldots + x_m z^m$ and $P_i > 0$. This gives

$$\mathbf{T}_{m-1} = \mathbf{A} \begin{pmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_m \end{pmatrix} \mathbf{A}^* \tag{16.12}$$

where **A** is an $m \times m$ Vandermonde matrix given by

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_m \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_m^2 \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_1^{m-1} & \lambda_2^{m-1} & \cdots & \lambda_m^{m-1} \end{pmatrix}, \qquad \lambda_i = e^{j\theta_i}, \qquad i = 1 \to m$$

and Eq. (16.12) can be used to determine $P_k > 0$, $k = 1 \to m$. The power spectrum associated with Eq. (16.11) is given by

$$S(\theta) = \sum_{k=1}^{m} P_k \delta(\theta - \theta_k)$$

and it represents a discrete spectrum that corresponds to pure uncorrelated sinusoids with signal powers P_1, P_2, …, P_m.

If the given autocorrelations satisfy $\mathbf{T}_n > 0$, from Eq. (16.4), every unknown r_k, $k \geq n+1$, must be selected so as to satisfy $\mathbf{T}_k > 0$, and this gives

$$\left| r_{k+1} - \zeta_k \right|^2 \leq \mathrm{R}^2{}_k \tag{16.13}$$

where $\zeta_k = \mathbf{f}_k^T \mathbf{T}_k^{-1} \mathbf{b}_k$, $\mathbf{f}_k = (r_1, r_2, \ldots, r_k)^T$, $\mathbf{b}_k = (r_k, r_{k-1}, \ldots, r_1)$ and $R_k = \det \mathbf{T}_k / \det \mathbf{T}_{k-1}$.

From Eq. (16.13), the unknowns could be anywhere inside a sequence of circles with center ζ_k and radius R_k, and as a result, there are an infinite number of solutions to this problem. Schur and Nevanlinna have given an analytic characterization to these solutions in terms of bounded function extensions. A bounded function $\rho(z)$ is analytic in $|z| < 1$ and satisfies the inequality $|\rho(z)| \leq 1$ everywhere in $|z| < 1$.

In a network theory context, Youla [1980] has also given a closed form parametrization to this class of solutions. In that case, given $r_0, r_1, \ldots, r_n$, the minimum phase transfer functions satisfying Eqs. (16.8) and (16.9) are given by

$$H_\rho(z) = \frac{\Gamma(z)}{P_n(z) - z\rho(z)\tilde{P}_n(z)} \tag{16.14}$$

where $\rho(z)$ is an *arbitrary* bounded function that satisfies the inequality (Paley-Wiener criterion)

$$\int_{-\pi}^{\pi} \ln\left[1 - \left|\rho(e^{j\theta})\right|^2\right] d\theta > -\infty$$

and $\Gamma(z)$ is the minimum phase factor obtained from the factorization

$$1 - \left|\rho(e^{j\theta})\right|^2 = \left|\Gamma(e^{j\theta})\right|^2$$

Further, $P_n(z)$ represents the Levinson polynomial generated from $r_0 \to r_n$ through the recursion

$$\sqrt{1 - |s_n|^2}\, P_n(z) = P_{n-1}(z) - z s_n \tilde{P}_{n-1}(z)$$

that starts with $P_0(z) = 1/\sqrt{r_0}$, where

$$s_n = \left\{ P_{n-1}(z) \sum_{k=1}^{n} r_k z^k \right\}_n P_{n-1}(0) \tag{16.15}$$

represents the reflection coefficient at stage n. Here, $\{\ \}_n$ denotes the coefficient of z^n in the expansion $\{\ \}$, and $\tilde{P}_n(z) \triangleq z^n P_n^*(1/z^*)$ represents the polynomial reciprocal to $P_n(z)$. Notice that the given information $r_0 \rightarrow r_n$ enters $P_n(z)$ through Eq. (16.5). The power spectral density

$$K(\theta) = \left| H_\rho(e^{j\theta}) \right|^2$$

associated with Eq. (16.14) satisfies all the interpolation properties described before. In Eq. (16.14), the solution $\rho(z) \equiv 0$ gives $H(z) = 1/P_n(z)$, a pure AR(n) system that coincides with Burg's maximum entropy extension. Clearly, if $H_\rho(z)$ is rational, then $\rho(z)$ must be rational and, more interestingly, every rational system must follow from Eq. (16.14) for a specific rational bounded function $\rho(z)$. Of course, the choice of $\rho(z)$ brings in extra freedom, and this can be profitably used for system identification as well as rational and stable approximation of nonrational systems [Pillai and Shim, 1993].

Defining Terms

Autocorrelation function: The expected value of the product of two random variables generated from a random process for two time instants; it represents their interdependence.

Expected value (or mean) of a random variable: Ensemble average value of a random variable that is given by integrating the random variable after scaling by its probability density function (weighted average) over the entire range.

Power spectrum: A nonnegative function that describes the distribution of power versus frequency. For wide sense stationary processes, the power spectrum and the autocorrelation function form a Fourier transform pair.

Probability density function: The probability of the random variable taking values between two real numbers x_1 and x_2 is given by the area under the nonnegative probability density function between those two points.

Random variable: A continuous or discrete valued variable that maps the set of all outcomes of an experiment into the real line (or complex plane). Because the outcomes of an experiment are inherently random, the final value of the variable cannot be predetermined.

Stochastic process: A real valued function of time t, which for every fixed t behaves like a random variable.

Related Topics

14.1 Fourier Transforms • 40.2 Spectrum, Specifications, and Measurement Techniques • 73.3 Stochastic Processes

References

N.I. Akheizer and M. Krein, *Some Questions in the Theory of Moments,* American Math. Soc. Monogr., 2, 1962.

J.B.J. Fourier, *Theorie Analytique de la Chaleur (Analytical Theory of Heat),* Paris, 1822.

L. Y. Geronimus, *Polynomials Orthogonal on a Circle and Their Applications,* American Math. Soc., Translation, 104, 1954.

I. Newton, *Philos. Trans.,* vol. IV, p. 3076, 1671.

S.U. Pillai and T.I. Shim, *Spectrum Estimation and System Identification,* New York: Springer-Verlag, 1993.

I. Schur, "Uber Potenzreihen, die im Innern des Einheitzkreises Beschrankt Sind," *Journal fur Reine und Angewandte Mathematik,* vol. 147, pp. 205–232, 1917.

J.A. Shohat and J.D. Tamarkin, *The Problem of Moments,* American Math. Soc., Math. Surveys, 1, 1970.

N. Wiener "Generalized harmonic analysis," *Acta Math.,* vol. 55, pp. 117–258, 1930.
D.C. Youla, "The FEE: A New Tunable High-Resolution Spectral Estimator: Part I," Technical note, no. 3, Dept. of Electrical Engineering, Polytechnic Institute of New York, Brooklyn, New York; also RADC Report, RADC-TR-81-397, AD A114996, 1982, 1980.
G.U. Yule, "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers," *Philos. Trans. R. Soc. London, Ser. A,* vol. 226, pp. 267–298, 1927.

16.2 Parameter Estimation

Stella N. Batalama and Dimitri Kazakos

Parameter estimation is the operation of assigning a value in a continuum of alternatives to an unknown parameter based on a set of observations involving some function of the parameter. **Estimate** is the value assigned to the parameter and **estimator** is the function of the observations that yields the estimate.

The basic elements in the parameter estimation are a vector parameter θ^m, a vector space $\mathcal{E}^m$ where θ^m takes its values, a stochastic process $X(t)$ parameterized by θ^m and a performance criterion or cost function. The estimate $\hat{\theta}^m(x^n)$ based on the observation vector $x^n = [x_1, x_2, ..., x_n]$ is a solution of some optimization problem according to the performance criterion. In the following, the function $f(x^n|\theta^m)$ will denote the conditional joint probability density function of the random variables $x_1, ..., x_n$.

There are several parameter estimation schemes. If the process $X(t)$ is parametrically known, i.e., if its conditional joint probability density functions are known for each fixed value θ^m of the vector parameter θ^m, then the corresponding parameter estimation scheme is called **parametric.** If the statistics of the process $X(t)$ are nonparametrically described, i.e., given $\theta^m \in \mathcal{E}^m$ any joint probability density function of the process is a member of some nonparametric class of probability density functions, then the **nonparametric estimation** schemes arise.

Let Γ^n denote the n-dimensional observation space. Then an estimator $\hat{\theta}(x^n)$ of a vector parameter θ^m is a function from the observation space, Γ^n, to the parameter space, $\mathcal{E}^m$. Since this is a function of random variables, it is itself a random variable (or random vector).

There are certain stochastic properties of estimators that quantify somehow their quality. In this sense an estimator is said to be **unbiased** if its expected value is the true parameter value, i.e., if

$$E_\theta\{\hat{\theta}^m(\mathbf{x}^n)\} = \theta^m$$

where the subscript θ on the expectation symbol denotes that the expectation is taken according to the probability density function $f(x^n|\theta^m)$. In the case where the observation space is the $\Re^n$ and the parameter is a scalar, it is

$$E_\theta\{\hat{\theta}(\mathbf{x}^n)\} = \int_{R^n} \hat{\theta}(x^n) f(x^n|\theta^m) dx^n$$

The **bias** of the estimate is the Euclidean norm $\|\theta^m - E_\theta\{\theta^m(x^n)\}\|^{1/2}$. Thus, the bias measures the distance between the expected value of the estimate and the true value of the parameter. Clearly, the estimator is unbiased when the bias is zero.

Usually it is of interest to know the conditional variance of an unbiased estimate. The bias of the estimate $\hat{\theta}^m(x^n)$ and the conditional variance

$$E_\theta\{\|\hat{\theta}^m(\mathbf{x}^n) - E_\theta\{\hat{\theta}^m(\mathbf{x}^n)\}\|^2|\theta^m\}$$

generally represent a trade-off. Indeed, an unbiased estimate may induce relatively large variance. On the other hand, the introduction of some low-level bias may then result in a significant reduction of the induced variance.

In general, the bias versus variance trade-off should be studied carefully for the correct evaluation of any given parameter estimate. A parameter estimate is called **efficient** if the conditional variance equals a lower bound known as the Rao-Cramèr bound.

It will be useful to present briefly this bound.

The Rao-Cramèr bound gives a theoretical minimum for the variance of any estimate. More specifically, let $\hat{\theta}(x^n)$ be the estimate of a scalar parameter θ given the observation vector x^n. Let $f(x^n|\theta)$ be given, twice continuously differentiable with respect to θ, and satisfy also some other mild regularity conditions. Then,

$$E_\theta\left\{\left[\hat{\theta}(\mathbf{x}^n) - \theta\right]^2\right\} \geq E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log f(x^n|\theta)\right]^2\right\}^{-1}.$$

Sometimes we need to consider the case where the sample size n increases to infinity. In such a case, an estimator is said to be **consistent** if

$$\hat{\theta}^m(\mathbf{x}^n) \to \theta^m \text{ as } n \to \infty$$

Since the estimate $\hat{\theta}^m(\mathbf{x}^n)$ is a random variable, we have to specify in what sense the above holds. Thus, if the above limit holds w.p. 1, we say that $\hat{\theta}^m(x^n)$ is *strongly consistent* or *consistent w.p. 1*. In a similar way we can define a *weakly consistent* estimator.

As far as the asymptotic distribution of $\theta(x^n)$ as $n \to \infty$ is concerned, it turns out that the central limit theorem can often be applied to $\hat{\theta}(x^n)$ to infer that $\sqrt{n}[\hat{\theta}(x^n) - \theta]$ is asymptotically normal with zero mean as $n \to \infty$.

In order to examine certain parameter estimation schemes we need first to present the definition of some related functions. **Penalty or cost function** $c[\theta^m, \hat{\theta}^m(x^n)]$ is a scalar, nonnegative function whose values vary as θ^m varies in the parameter space $\mathscr{E}^m$ and as the sequence x^n takes different values in the observation space, Γ^n. *Conditional expected penalty* $c(\theta^m, \hat{\theta}^m)$ induced by the parameter estimate and the penalty function is a function defined as follows:

$$c(\theta^m, \hat{\theta}^m) = E_\theta\{c[\theta^m, \hat{\theta}^m(\mathbf{x}^n)]\}$$

If an *a priori* density function $p(\theta^m)$ is available, then the expected penalty $c(\hat{\theta}^m, p)$ can be evaluated.

The various existing parameter estimation schemes evolve as the solutions of optimization problems whose objective function is either the conditional expected penalty or the conditional density function $f(x^n|\theta^m)$.

Bayesian Estimation Scheme

In the **Bayesian estimation** scheme the available assets are:

1. A parametrically known stochastic process parameterized by θ^m, in other words, a given conditional joint density function $f(x^n|\theta^m)$ defined on the observation space Γ^n, where θ^m is a well-defined parameter vector.
2. A realization x^n from the underlying active process, where the implied assumption is that the process remains unchanged throughout the whole observation period.
3. A density function $p(\theta^m)$ defined on the parameter space $\mathscr{E}^m$.
4. For each data sequence x^n, parameter vector θ^m and parameter estimate $\hat{\theta}^m(x^n)$, a penalty scalar function $c[\theta^m, \hat{\theta}^m(x^n)]$ is given.
5. A performance criterion which is the minimization of the expected penalty $c(\theta^m, p)$.

The estimate $\hat{\theta}_0^m$ that minimizes the expected penalty is called *optimal Bayesian estimate at p*. Under some mild conditions the optimal Bayesian estimate $\hat{\theta}_0^m(x^n)$ is the conditional expectation $E\{\theta^m|x^n\}$.

If the penalty function has the form $c[\theta^m,\hat{\theta}^m] = 1 - \delta(\|\theta^m - \hat{\theta}^m\|)$, where $\delta()$ is the delta function, then the optimal Bayesian estimate is called maximum a posteriori estimate since it happens to maximize the a *posteriori* density $p(\theta^m|x^n)$.

Another special case of penalty function is the function $\|\theta^m - \hat{\theta}^m\|^2$. In this case the Bayesian estimate is called *minimum mean-square estimate* and equals the conditional expectation $E\{\theta^m|x^n\}$.

In the following we present some more details about **mean-square estimation** since it is one of the most popular schemes.

Mean-Square Estimation

For the simplicity of our discussion we consider the case of estimating a single continuous type random variable θ with density $p(\theta)$ instead of estimating a random vector. We also reduce the dimensionality of the observation space to one. In this framework the penalty function will be the square of the estimation error $(\theta - \hat{\theta})^2$ and the performance or optimality criterion will be the minimization of the mean (expected) square value of the estimation error.

We will first consider the case of estimating a random variable θ by a constant $\hat{\theta}$. This means that we wish to find a constant $\hat{\theta}$ such that the mean-square (MS) error

$$e = E\{(\theta - \hat{\theta})^2\} = \int_{\infty}^{\infty} (\theta - \hat{\theta})^2 p(\theta) d\theta$$

is minimum. Since e depends on $\hat{\theta}$, it is minimum if

$$\frac{de}{d\theta} = \int_{-\infty}^{\infty} 2(\theta - \hat{\theta}) p(\theta) d\theta = 0$$

i.e., if

$$\hat{\theta} = E\{\theta\} = \int_{-\infty}^{\infty} \theta p(\theta) d\theta$$

The case where θ is to be estimated by a function $\hat{\theta}(\mathbf{x})$ of the random variable (observation) $\mathbf{x}$, and not by a constant, is examined next. In this case the MS error takes the form:

$$e = E\{[\theta - \hat{\theta}(\mathbf{x})]^2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\theta - \hat{\theta}(x)]^2 p(\theta, x) dx dy$$

where $p(\theta, x)$ is the joint density of the random variables $\boldsymbol{\theta}$ and x. In this case we need to find that function $\hat{\theta}(x)$ which minimizes the MS error. It can be proved that the function that minimizes the MS error is

$$\hat{\theta}(x) = E\{\theta \mid x\} = \int_{-\infty}^{\infty} \theta p(\theta \mid x) d\theta$$

The function $\hat{\theta}(\mathbf{x})$ is called **nonlinear MS estimate.**

As we have seen, when the penalty function is the quadratic function $(\theta - \hat{\theta})^2$, then the optimal Bayesian estimate is the conditional expectation $E\{\boldsymbol{\theta}|x\}$. If $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian, then the above conditional expectation is a linear function of x. But when the above statistics are not Gaussian, then the optimal Bayesian estimate is generally a nonlinear function of x. Thus, to resolve this problem we introduce suboptimal Bayesian schemes for this quadratic penalty function. In particular we consider only the class of linear parameter estimates and we try to find that estimate which minimizes the expected quadratic penalty. This estimate is called linear MS estimate and it is used in many applications because of the simplicity of the solution.

The linear estimation problem is the estimation of a random variable $\boldsymbol{\theta}$ in terms of a linear function $Ax + B$ of $\mathbf{x}$, i.e., $\hat{\theta}(x) = Ax + B$. In this case we need to find the constants A and B in order to minimize the MS error

$$e = E\{[\theta - (A\mathbf{x} + B)]^2\}$$

A fundamental principle in the MS estimation is the **orthogonality principle.** This principle states that the optimum linear MS estimate $A\mathbf{x} + B$ of $\boldsymbol{\theta}$ is such that the estimation error $\boldsymbol{\theta} - (A\mathbf{x} + B)$ is orthogonal to the data $\mathbf{x}$, i.e.,

$$E\{[\boldsymbol{\theta} - (A\mathbf{x} + B)]\mathbf{x}\} = 0$$

Using the above principle, it can be proved that e is minimum if

$$A = \frac{r\sigma_\theta}{\sigma_x} \text{ and } B = \eta_\theta - A\eta_x$$

$$\eta_x = E\{\mathbf{x}\}, \qquad \eta_\theta = E\{\boldsymbol{\theta}\}$$

$$\sigma_x^2 = E\{(\mathbf{x} - \eta_x)^2\}, \; \sigma_\theta^2 = E\{(\theta - \eta_\theta)^2\}$$

$$r = \frac{E\{(\mathbf{x} - \eta_x)(\theta - \eta_\theta)\}}{\sigma_x \sigma_\theta}$$

i.e., η_x, η_θ are the means of $\mathbf{x}$ and $\boldsymbol{\theta}$; σ_x^2, σ_θ^2 are the corresponding variances; σ_x, σ_θ is the standard deviation of x and θ; and r is the correlation coefficient of $\mathbf{x}$ and $\boldsymbol{\theta}$. Thus the MS error takes the form $e = \sigma_\theta^2 (1 - r^2)$.

The estimate

$$\hat{\theta}(\mathbf{x}) = A\mathbf{x} + B$$

is called the **nonhomogeneous linear estimate** of $\boldsymbol{\theta}$ in terms of $\mathbf{x}$. If $\boldsymbol{\theta}$ is estimated by a function $\hat{\theta}(\mathbf{x}) = \alpha\mathbf{x}$, the estimate is called homogeneous. It can be also shown by the orthogonality principle that for the homogeneous estimate

$$\alpha = \frac{E\{\mathbf{x}\boldsymbol{\theta}\}}{E\{\mathbf{x}^2\}}$$

Using the orthogonality principle it can be shown that if the random variables $\boldsymbol{\theta}$ and $\mathbf{x}$ are Gaussian zero mean, then the optimum nonlinear estimate of θ equals the linear estimate. In other words if $\hat{\theta}(\mathbf{x}) = E\{\boldsymbol{\theta}|x\}$ is the optimum nonlinear estimate of $\boldsymbol{\theta}$ and $\hat{\theta} = \alpha\mathbf{x}$ is the optimum linear estimate, then $\hat{\theta}(\mathbf{x}) = E\{\hat{\theta}|x\} = \hat{\theta} = \alpha x$.

This is true since the random variables $\boldsymbol{\theta}$ and $\mathbf{x}$ have zero mean, $E\{\boldsymbol{\theta}\} = E\{\mathbf{x}\} = 0$, and thus the linear estimate $\hat{\theta}$ has zero mean too, $E\{\hat{\theta}\} = 0$. This implies that the linear estimation error $\varepsilon = \boldsymbol{\theta} - \hat{\theta}$ also has zero mean, $E\{\varepsilon\} = E\{\boldsymbol{\theta} - \hat{\theta}\} = 0$.

On the other hand, the orthogonality principle can be applied, which implies that the linear estimation error ε is orthogonal to the data, $E\{\varepsilon \mathbf{x}\} = 0$. Since ε is Gaussian, it is independent of $\mathbf{x}$ and thus $E\{\varepsilon | x\} = E\{\varepsilon\} = 0$, which is equivalent to the following:

$$\mathrm{E}\{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} | x\} = 0 \Rightarrow E\{\boldsymbol{\theta} | x\} - E\{\hat{\boldsymbol{\theta}} | x\} = 0$$

$$\Rightarrow E\{\boldsymbol{\theta} | x\} = E\{\hat{\boldsymbol{\theta}} | x\} \Rightarrow \hat{\theta}(\mathbf{x}) = \alpha \mathrm{x} \Rightarrow \hat{\theta}(\mathbf{x}) = \hat{\theta}$$

i.e., the nonlinear and the linear estimates coincide.

In addition, since the linear estimation error $\varepsilon = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ is independent of the data x, so is the square error, i.e.,

$$E\{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2 | \mathrm{x}\} = \mathrm{E}\{(\theta - \hat{\theta}^2\} = V$$

Thus, the conditional mean of $\boldsymbol{\theta}$ assuming the data x equals its MS estimate and the conditional variance the MS error. That simplifies the evaluation of conditional densities when Gaussian random variables are involved because since $f(\theta | x)$ is Gaussian, it has the form

$$f(\theta \,|\, X) = \frac{1}{\sqrt{2\pi V}} \exp\left\{\frac{-[\theta - \alpha x]^2}{2V}\right\}$$

Minimax Estimation Scheme

In the **minimax estimation** scheme the available assets are:

1. A parametrically known stochastic process parameterized by θ^m.
2. A realization x^n from the underlying active process.
3. A scalar penalty function $c[\theta^m, \hat{\theta}^m(x^n)]$ for each data sequence x^n, parameter vector θ^m, and parameter estimate $\hat{\theta}^m(x^n)$.

The minimax schemes are solutions of saddle-point game formalizations, with payoff function the expected penalty $c(\hat{\theta}^m, p)$ and with variables the parameter estimate $\hat{\theta}^m$ and the *a priori* parameter density function p. If a minimax estimate $\hat{\theta}_0^m$ exists, it is an optimal Bayesian estimate, at some least favorable *a priori* distribution p_0.

Maximum Likelihood Estimation Scheme

Maximum likelihood estimation was first introduced by Fisher. It is a very powerful estimation procedure that yields many of the well-known estimation methods as special cases.

The essential difference between Bayesian and maximum likelihood parameter estimation is that in Bayesian Estimation the parameter θ^m is considered to be random with a given density function, while in the maximum likelihood framework it is unknown but fixed.

Consider a random process $X(t)$ parameterized by θ^m, where θ^m is an unknown fixed parameter vector of finite dimensionality m (e.g., $\theta^m \in \Re^m$). More specifically the conditional joint probability density function $f(x_1, \ldots, x_n | \theta^m)$ is well known for every θ^m, where $x^n = [x_1, \ldots, x_n]$ is a realization (or observation vector or sample vector) of the process $X(t)$.

The problem is to find an estimate of the parameter vector θ^m based on the realization of $X(t)$. (Note that the dimensionality of the parameter vector θ^m in the joint probability density function is assumed to be fixed.)

The intuition behind the maximum likelihood method is that we choose those parameters $[\theta_1, \ldots, \theta_m]$ from which the actually observed sample vector is most likely to have come. This means that the estimator of θ^m is selected so that the observed sample vector becomes as "likely as possible."

In this sense we call the conditional joint probability density function $f(x^n|\theta^m)$ as likelihood function $\ell(\theta^m)$. The likelihood function $\ell(\theta^m)$ is a deterministic function of the parameter vector θ^m once the observed variables $x_1, \ldots, x_n$ are inserted. This means that θ^m is variable and the sample vector x^n is fixed, while the conditional joint probability density function is considered as a function of the observation vector x^n with θ^m fixed. The maximum likelihood estimator of θ^m is that value of the parameter vector for which the likelihood function is maximized.

In many cases it is more convenient to work with another function called log-likelihood function, $L(\theta^m)$, rather than the likelihood function. The log-likelihood function is the natural logarithm of $\ell(\theta^m)$. Since the logarithm is a monotone function, it follows that whenever $L(\theta)$ achieves its maximum value, $\ell(\theta^m)$ is maximized too, for the same value of the parameter vector θ^m. Thus the log-likelihood function is maximized for that value of the vector parameter θ^m for which the first partial derivatives with respect to θ_i, $i = 1, \ldots, m$ are equal to zero, i.e.,

$$\hat{\theta}_{ML} \; : \; \frac{\partial L(\theta^m)}{\partial \theta_i} = 0$$

where $\hat{\theta}_{ML}$ denotes the maximum likelihood estimate of the vector parameter θ^m.

It can be shown that when the process $X(t)$ is memoryless and stationary (i.e., when $x_1, \ldots, x_n$ are independent, identically distributed) then the ML estimators are consistent, asymptotically efficient, and asymptotically Gaussian.

Example: Let $\mathbf{x}_i$, $i = 1, \ldots, n$, be Gaussian independent random variables with mean θ and variance σ_i^2: $\mathbf{x}_i \in N(\theta, \sigma_i^2)$. The mean θ is to be estimated and the Rao-Cramèr bound is to be evaluated. Since θ is unknown but fixed, we will use the maximum likelihood estimation scheme. The random variable $\mathbf{x}_i$ has the probability density function

$$\frac{1}{\sqrt{2\pi\sigma_i}} \exp\left\{-\frac{(x_i - \theta)^2}{2\sigma_i^2}\right\}$$

Since $\mathbf{x}_i$, $i = 1, \ldots, n$, are independent, the joint density function is

$$f(x_i, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left\{-\frac{(x_i - \theta)^2}{2\sigma_i^2}\right\}$$

which is exactly the likelihood function for this estimation problem. The log-likelihood function is

$$\log f(x_1, \ldots, x_n|\theta) = -\frac{n}{2} \log(2\pi) - \sum_{i=1}^{n} \log \sigma_i - \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{\sigma_i^2}$$

We can maximize the log-likelihood function with respect to θ and find the maximizing value to be equal to

$$\hat{\theta}_{ML}(x^n) = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}$$

Note that for equal variances the maximum likelihood estimate coincides with the commonly used sample mean.

The Rao-Cramèr bound can be found as follows:

$$E_\theta\left\{\left[\frac{d}{d\theta}\log f(x^n|\theta)\right]^2\right\}^{-1} = -E_\theta\left\{\frac{d^2}{d\theta^2}\log f(x^n|\theta)\right\} = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}$$

In conclusion, we see that for Gaussian data the sample mean estimate is efficient because it coincides with the maximum likelihood estimate.

When the data are contaminated with a fraction of data coming from an unknown probability density function, the so called *outliers*, the sample mean performs poorly even when the fraction of outliers is small. This observation gave birth to the branch of statistics called robust statistics.

Other Parameter Estimation Schemes

The Bayesian, minimax, and maximum likelihood estimation schemes described above make up the class of parametric parameter estimation procedures. The common characteristic of those procedures is the availability of some parametrically known stochastic process that generates the observation sequence x^n. When for every given parameter value θ^m the stochastic process that generates x^n is nonparametrically described, the nonparametric estimation schemes arise. The latter schemes may evolve as the solutions of certain saddle-point games, whose payoff function originates from the parametric maximum likelihood formalizations. It is assumed that, in addition to the nonparametrically described data-generating process, the only assets available are a realization x^n from the underlying active process and the parameter space $\mathcal{C}^m$. The **qualitative robustness** in parameter estimation corresponds to local performance stability for small deviations from a nominal, parametrically known, data-generating process.

Defining Terms

Bayesian estimation: An estimation scheme in which the parameter to be estimated is modeled as a random variable with known probability density function.

Bias: The norm of the difference between the true value of the estimate and its mean value.

Consistent estimator: An estimator whose value converges to the true parameter value as the sample size tends to infinity. If the convergence holds w.p. 1, then the estimator is called *strongly consistent* or *consistent w.p. 1.*

Efficient estimator: An estimator whose variance achieves the Rao-Cramèr bound.

Estimate: Our best guess of the parameter of interest based on a set of observations.

Estimator: A mapping from the data space to the parameter space that yields the estimate.

Homogeneous linear estimator: An estimator which is a homogeneous linear function of the data.

Maximum likelihood estimate: An estimate that maximizes the probability density function of the data conditioned on the parameter.

Mean-square estimation: An estimation scheme in which the cost function is the mean-square error.

Minimax estimate: The optimum estimate for the least favorable prior distribution.

Nonhomogeneous linear estimator: An estimator which is a nonhomogeneous linear function of the data.

Nonlinear MS estimate: The optimum estimate under the mean-square performance criterion.

Nonparametric estimation: An estimation scheme in which no parametric description of the statistical model is available.

Orthogonality principle: The fundamental principle for MS estimates. It states that the estimation error is orthogonal to the data.

Parameter estimation: The procedure by which we combine all available data to obtain our best guess about a parameter of interest.

Parametric estimation: An estimation scheme in which the statistical description of the data is given according to a parametric family of statistical models.

Penalty or cost function: A nonnegative scalar function which represents the cost incurred by an inaccurate value of the estimate.

Qualitative robustness: A geometric formulation of robust estimation.

Robust estimation: An estimation scheme in which we optimize performance for the least favorable statistical environment among a specified statistical class.

Unbiased estimator: An estimator whose mean value is equal to the true parameter value.

Related Topics

73.1 Signal Detection • 73.3 Stochastic Processes

References

S. Haykin, *Adaptive Filter Theory,* Englewood Cliffs, N.J.: Prentice-Hall, 1991.

D. Kazakos and P. Papantoni-Kazakos, *Detection and Estimation,* New York: Computer Science Press, 1990.

L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification,* Cambridge, Mass.: The MIT Press, 1983.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* New York: McGraw-Hill, 1984.

Further Information

IEEE Transactions on InformationTheory is a bimonthly journal that publishes papers on theoretical aspects of estimation theory and in particular on transmission, processing, and utilization of information.

IEEE Transactions on Signal Processing is a monthly journal which presents applications of estimation theory to speech recognition and processing, acoustical signal processing, and communication.

IEEE Transactions on Communications is a monthly journal presenting applications of estimation theory to data communication problems, synchronization of communication systems, channel equalization, and image processing.

16.3 Kalman Filtering

Fred Daum

The **Kalman filter** is a linear recursive algorithm that solves the least squares problem for time-varying linear systems with non-stationary noise. It estimates the state of a linear dynamical system given linear measurements corrupted by additive noise. It is an optimal estimator, assuming that the measurement noise is Gaussian, and assuming that all other relevant probability densities are also Gaussian.

For example, the location of your car can be estimated using a Kalman filter to combine noisy measurements of distance from four or more satellites. As a second example, the position and velocity of an airplane can be estimated by a Kalman filter using noisy measurements of range, azimuth, and elevation from one or more radars. As a third example, the future price of IBM stock can be predicted using a Kalman filter to combine noisy data on thousands of relevant economic variables, using a dynamic model of the stock market and the overall economy.

The Kalman filter has been applied to solve many diverse real-world engineering problems, including spacecraft navigation, GPS navigation, robotics, air traffic control, missile guidance, chemical plant control, stock market prediction, weather prediction, speech recognition, speech encoding and compression, radar target tracking, satellite orbit estimation, and inertial navigation. See [Sorenson, 1985] for other applications.

Most real-world engineering problems have measurement equations or dynamical system equations that are nonlinear in the **state vector.** Therefore, the Kalman filter equations cannot be applied directly; rather, the problem must be approximated using linear equations. This linear approximation is very straightforward, and it is called the **"extended Kalman filter" (EKF)**. One of the main reasons for the wide application ofthe Kalman filter is the ease with which a nonlinear system can be approximated by a linear system. The resulting approximation is often very good, resulting in good EKF performance. Unfortunately, the EKF performance is sometimes poor, in which case a plethora of alternative approximations can be attempted.

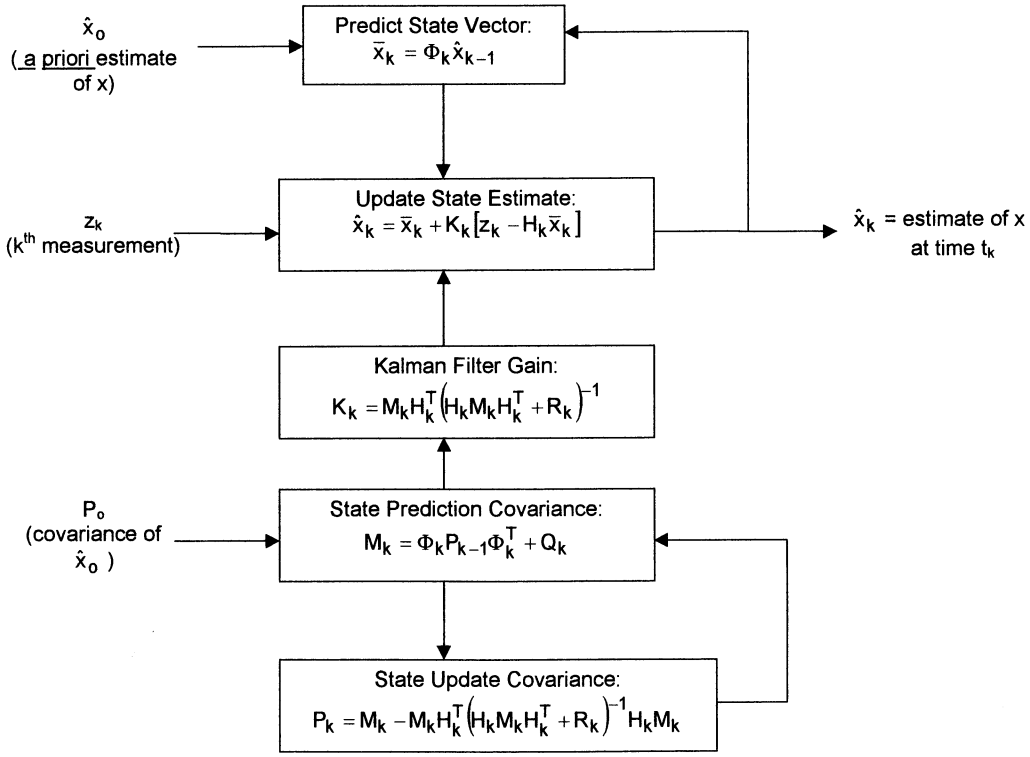


FIGURE 16.1 Block diagram of Kalman filter.

Kalman Filter Equations

The **Kalman filter** algorithm is shown as a block diagram in Fig. 16.1. The estimate of x is updated recursively as new measurements become available. Each measurement update corresponds to one iteration of Fig. 16.1. The symbols in Fig. 16.1 are defined in Table 16.1.

The Kalman filter uses both a model of the system dynamics

$$x_k = \Phi_k x_{k-1} + w_k \tag{16.16}$$

as well as a model of the measurements

$$z_k = H_k x_k + v_k \tag{16.17}$$

These models are a combination of deterministic and random effects. One can think of the true state vector, x_k, evolving in time according to a deterministic linear dynamical system:

$$x_k = \Phi_k x_{k-1} \tag{16.18}$$

with a random perturbation modeled by w_k. Likewise, the measurement model consists of a deterministic linear part:

$$z_k = H_k x_k \tag{16.19}$$

with an additive random perturbation of v_k. As shown in Fig. 16.1, the Kalman filter predicts the state vector from one time (t_{k-1}) to the next (t_k) using the deterministic part of the linear dynamical system. This is the

TABLE 16.1 Definition of Symbols

Symbol	Meaning	Mathematical Definition	Value
x_k	State vector of a linear dynamical system at time t_k	$x_k = \Phi_k x_{k-1} + w_k$	Vector of dimension n
t_k	Time of the k^{th} measurement	—	Scalar
k	Index of discrete time measurements	—	Integer
z_k	k^{th} measurement	$z_k = H_k x_k + v_k$	Vector of dimension m
H_k	Measurement matrix at time t_k	See above	m × n matrix
R_k	Covariance matrix of measurement noise at time t_k	$R_k = E(v_k v_k^T)$	m × m matrix
v_k	Measurement noise at time t_k	Gaussian zero mean random variable with covariance matrix R_k, statistically independent from sample to sample, and statistically independent of x_k	Vector of dimension m
Φ_k	Transition matrix of linear dynamical system from time t_{k-1} to t_k	See above	n × n matrix
w_k	Process noise	Gaussian zero mean random variable with covariance matrix Q_k, statistically independent from sample to sample, and statistically independent of x_k	Vector of dimension n
Q_k	Covariance matrix of process noise at time t_k	$Q_k = E(w_k w_k^T)$	n × n matrix
P_k	Error covariance matrix of x_k conditioned on Z_k	$P_k = E\left[\left(x_k - \hat{x}_k\right)\left(x_k - \hat{x}_k\right)^T \middle| Z_k\right]$	n × n matrix
M_k	Error covariance matrix of x_k conditioned on Z_{k-1}	$M_k = E\left[\left(x_k - \bar{x}_k\right)\left(x_k - \bar{x}_k\right)^T \middle| Z_{k-1}\right]$	n × n matrix
$\hat{x}_k$	Estimate of x at time t_k conditioned on Z_k	$\hat{x}_k = E\left(x_k \middle| Z_k\right)$	Vector of dimension n
$\bar{x}_k$	Estimate of x at time t_k conditioned on Z_{k-1}	$\bar{x}_k = E\left(x_k \middle| Z_{k-1}\right)$	Vector of dimension n
Z_k	Set of measurements up to and including time t_k	$Z_k = \left\{z_1, z_2, \ldots, z_k\right\}$	Set of m-dimensional vectors
$\hat{x}_o$	Initial estimate of x at time t_o, prior to any measurements	$\hat{x}_o = E\left(x_o\right)$	Vector of dimension n
P_o	Initial error covariance matrix of $\hat{x}_o$, prior to any measurements	$P_o = E\left[\left(x_o - \hat{x}_o\right)\left(x_o - \hat{x}_o\right)^T\right]$	n × n matrix
$E(\cdot)$	Expected value of $(\cdot)$	$E(A) = \int A\, p(A) dA$	—
$p(A)$	Probability density of A	Gaussian probability density	Function
$p(A\|B)$	Probability density of A conditioned on B	Gaussian probability density	Function
$(\cdot)^T$	Transpose of $(\cdot)$	$(A^T)_{ij} = A_{ji}$ for a matrix A with ij^{th} element A_{ij}	Operation
$(\cdot)^{-1}$	Inverse of matrix $(\cdot)$	A^{-1} is the inverse of matrix A if and only if $A^{-1}A = AA^{-1} = I$	Operation
I	Identity matrix	$I_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases}$	n × n matrix

best prediction of $x(t_k)$ given $x(t_{k-1})$, assuming that w_k is a zero-mean random variable uncorrelated with $x(t_{k-1})$. Hence, the **state vector** prediction in the **Kalman filter** is intuitively exactly what one would expect.The state vector update is also intuitively reasonable. In particular, it is a linear combination of the predicted value of x and the latest measurement z_k. This linear combination of $\bar{x}_k$ and z_k is computed as a compromise between prediction errors in $\bar{x}_k$ and the measurement noise errors in z_k. That is, the Kalman filter gain is computed to optimally combine $\bar{x}_k$ and z_k, using the known models of system dynamics and measurements. More specifically, the Kalman filter gain is computed to minimize the following error criterion at the time of the k^{th} measurement:

$$J = \left(\bar{x}_k - \hat{x}_k\right)^T M_k^{-1} \left(\bar{x}_k - \hat{x}_k\right) + \left(z_k - H_k \hat{x}_k\right)^T R_k^{-1} \left(z_k - H_k \hat{x}_k\right) \tag{16.20}$$

The **covariance matrix** of the prediction errors is M_k, and R_k is the measurement error covariance matrix. If M_k is large, then the first term in J is weighted less because the prediction, $\bar{x}_k$, is relatively inaccurate. Similarly, if R_k is large, then the second term in J is weighted less because the measurement, z_k, is relatively inaccurate. The weighted least squares criterion, J, strikes a balance between prediction errors and measurement errors. To find the value of $\hat{x}_k$ that minimizes J, we can set the derivative of J with respect to $\hat{x}_k$ equal to zero:

$$\frac{\partial J}{\partial \hat{x}_k} = 2\left(\hat{x}_k - \bar{x}_k\right)^T M_k^{-1} + 2\left(z_k - H_k \hat{x}_k\right)^T R_k^{-1} \left(-H_k\right) = 0 \tag{16.21}$$

Using the fact that covariance matrices are symmetric, and rearranging terms, we get:

$$\left(M_k^{-1} + H_k^T R_k^{-1} H_k\right)\hat{x}_k = M_k^{-1} \bar{x}_k + H_k^T R_k^{-1} z_k \tag{16.22}$$

and solving for $\hat{x}_k$:

$$\hat{x}_k = \left(M_k^{-1} + H_k^T R_k^{-1} H_k\right)^{-1} \left[M_k^{-1} \bar{x}_k + H_k^T R_k^{-1} z_k\right] \tag{16.23}$$

This can be put into the form:

$$\hat{x}_k = \bar{x}_k + K_k \left[z_k - H_k \bar{x}_k\right] \tag{16.24}$$

where:

$$K_k = \left(M_k^{-1} + H_k^T R_k^{-1} H_k\right)^{-1} H_k^T R_k^{-1} \tag{16.25}$$

Further matrix algebra can be used to represent the Kalman filter gain in two other forms:

$$K_k = M_k H_k^T \left(H_k M_k H_k^T + R_k\right)^{-1} \tag{16.26}$$

$$K_k = P_k H_k^T R_k^{-1} \tag{16.27}$$

in which P_k is the error covariance matrix of $\hat{x}_k$.

The above calculation of the **Kalman filter** gain is by no means a derivation of the Kalman filter, but rather it is merely a simple heuristic calculation that gives insight into these equations. The error criterion, J, is the logarithm of a Gaussian probability density of x conditioned on Z_k. A lucid derivation of the Kalman filter is given in Ho and Lee [1964], from a Bayesian viewpoint. In contrast, Kalman's original derivation does not use Bayesian methods; see Kalman [1960]. Other derivations of the Kalman filter are in Gelb [1974] and Jazwinski [1970]. The Kalman filter is "optimal" in a number of ways, under different assumptions, which are discussed in these references.

The Kalman filter is stable under rather mild assumptions, which can always be achieved in practice, by using a **state vector** with minimum dimension to realize the linear dynamical model and the linear measurement model. This corresponds to a *completely controllable* dynamical system and a *completely observable* measurement model. Stability of the Kalman filter under these conditions was proved in a beautiful paper by Kalman (1963). Kalman's stability theory has great practical value because engineers do not need to check for Kalman filter stability as a separate issue. This result is all the more impressive when we consider that the Kalman filter is a time-varying linear dynamical system that can be very complex and have very high dimension. Nevertheless, the Kalman filter is automatically stable under the simple minimality assumption given above.

It is important to emphasize that both the linear dynamical system and the measurement model can be time-varying, and both the process noise and measurement noise can be nonstationary. That is, all of the following matrices can vary with time: Φ_k, H_k, Q_k, and R_k. Also, the discrete times at which measurements are made (t_k for $k = 1, 2, \ldots$) are completely arbitrary; that is, the sample rate can be nonuniform.

Kalman Filter Examples

A good way to understand the Kalman filter equations is to look at some simple low-dimensional examples.

Example 1

Consider the problem of estimating the value of an unknown constant scalar, x, given a sequence of noisy measurements:

$$z_k = x_k + v_k \tag{16.28}$$

where the variance of measurement noise is constant, and the variance of the *a priori* estimate of x is infinite. For this example, using the Kalman filter notation:

$$\Phi_k = 1$$

$$Q_k = 0$$

$$H_k = 1$$

$$R_k = \text{constant} = c$$

$$P_o = \infty$$

the corresponding Kalman filter equations given in Fig. 16.1 are

$$\bar{x}_k = \hat{x}_{k-1} \tag{16.29}$$

$$\hat{x}_k = \bar{x}_k + K_k\left(z_k - \bar{x}_k\right) \tag{16.30}$$

$$K_k = M_k \big/ \left(M_k + c\right) \tag{16.31}$$

$$M_k = P_{k-1} \tag{16.32}$$

$$P_k = M_k - M_k^2 / \left(M_k + c\right) \tag{16.33}$$

$$P_o = \infty \tag{16.34}$$

which simplifies to

$$\hat{x}_k = \hat{x}_{k-1} + K_k \left(z_k - \hat{x}_{k-1}\right) \tag{16.35}$$

$$K_k = P_{k-1} / \left(P_{k-1} + c\right) \tag{16.36}$$

$$P_k = P_{k-1} - P_{k-1}^2 / \left(P_{k-1} + c\right) \tag{16.37}$$

After some more algebra, it turns out that

$$P_k^{-1} = P_{k-1}^{-1} + \frac{1}{c} \tag{16.38}$$

where $P_o^{-1} = 0$, and hence

$$P_k^{-1} = \sum_{j=1}^{k} 1/c \tag{16.39}$$

$$P_k^{-1} = k/c \tag{16.40}$$

Therefore, the variance of estimation error after k measurements is

$$P_k = c/k \tag{16.41}$$

which is exactly what we should expect for this example. Also, the Kalman filter gain is

$$K_k = P_{k-1} / \left(P_{k-1} + c\right) \tag{16.42}$$

$$K_k = \frac{1}{1 + c/P_{k-1}} \tag{16.43}$$

$$K_k = \frac{1}{1 + k - 1} \tag{16.44}$$

$$K_k = 1/k \tag{16.45}$$

which is intuitively very reasonable. Furthermore, the Kalman filter can now be written as:

$$\hat{x}_k = \hat{x}_{k-1} + \frac{1}{k}\left(z_k - \hat{x}_{k-1}\right) \tag{16.46}$$

which has the solution

$$\hat{x}_k = \frac{1}{k}\sum_{j=1}^{k} z_j \tag{16.47}$$

The Kalman filter for this simple example is nothing more than our old friend, the arithmetic average.

Example 2

Consider the same problem as in Example 1, but with R_k not constant. It is easy to show that the Kalman filter in this case is

$$\hat{x}_k = \hat{x}_{k-1} + \left(P_k / R_k\right)\left(z_k - \hat{x}_{k-1}\right) \tag{16.48}$$

where the estimation error variance after k measurements is given by

$$P_k = 1\Big/\sum_{j=1}^{k}\left(1/R_j\right) \tag{16.49}$$

and the Kalman filter estimate of x after k measurements is:

$$\hat{x}_k = \frac{\sum_{j=1}^{k}\left(z_j/R_j\right)}{\sum_{j=1}^{k}\left(1/R_j\right)} \tag{16.50}$$

This result is intuitively very reasonable. In particular, the more accurate measurements (corresponding to small R_j) are weighted more heavily in estimating x; conversely, relatively inaccurate measurements (with large R_j) are given little weight.

Example 3

Consider the problem of estimating the value of a quantity, y, that changes linearly with time with an unknown rate of change, given a sequence of measurements of y corrupted by additive noise that is statistically independent from sample to sample. In the Kalman filter setup, we could model this problem as follows. Let the state vector be:

$$x = \begin{bmatrix} y \\ \dot{y} \end{bmatrix} \tag{16.51}$$

The transition matrix would be:

$$\Phi_k = \begin{bmatrix} 1 & \Delta t_k \\ 0 & 1 \end{bmatrix} \tag{16.52}$$

where $\Delta t_k = t_k - t_{k-1}$. Furthermore,

$$H_k = [1 \quad 0]$$

$$Q_k = 0$$

$$R_k = \text{constant}$$

$$P_o^{-1} = 0$$

Assuming a constant value of $\Delta t_k = T$, it turns out that the error **covariance matrix** is:

$$M_k = \begin{bmatrix} 2(2k-1) & 6/T \\ 6/T & 12/T^2(k-1) \end{bmatrix} \frac{R}{k(k+1)} \tag{16.53}$$

See [Sorenson, 1967] for details.

Extended Kalman Filter

In practical applications, it is rather rare to find a problem with dynamical equations and measurement equations that are linear in x. Nevertheless, engineers use the **Kalman filter** theory applied to a linear approximation of the actual nonlinear dynamics and measurements. This approximation is called the **extended Kalman filter (EKF)**; it is very straightforward and popular. The Kalman filter itself is almost never used in real-world applications, but rather the EKF is essentially ubiquitous.

Figure 16.2 shows a block diagram of the EKF. Note that the EKF uses the nonlinear dynamics and nonlinear measurement equations to predict $\bar{x}_k$ and $\bar{z}_k$, rather than using a linear approximation. In contrast, the EKF uses linear approximations of f(x) and h(x) to compute the covariance matrices and the Kalman filter gain. The nonlinear dynamical model for x is:

$$x_k = f(x_{k-1}) + w_k \tag{16.54}$$

and the nonlinear measurement model is:

$$z_k = h(x_k) + v_k \tag{16.55}$$

Also, note in Fig. 16.2 that the estimate of x is used to compute the Kalman filter gain, unlike the Kalman filter, in which the filter gain and the error covariance matrices do not depend on x (see Fig. 16.1). Unlike the Kalman filter, there is no guarantee that the EKF is stable. Moreover, there is no reason to suppose that the EKF will give optimal performance. Although in many applications the EKF performance is good, it is well known that the EKF performance is often poor or far from optimal. Unfortunately, there is no theory that predicts when the EKF will give good performance, but rather engineers use Monte Carlo simulations to evaluate EKF performance.

There is a vast literature on methods to improve the EKF performance, including second-order Taylor series, iteration of Fig. 16.2 to improve the linearization, tuning the process noise covariance matrix, decoupling the

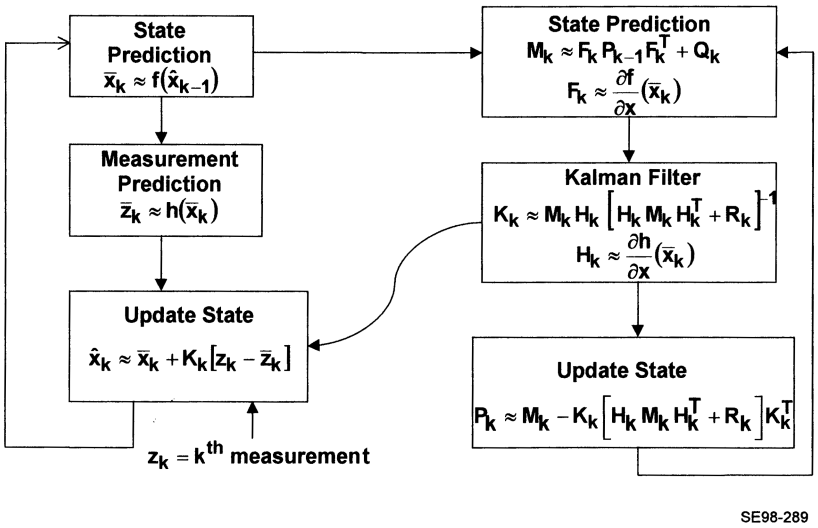


FIGURE 16.2 The extended Kalman (EKF) is a linear approximation.

error **covariance matrix**, preferred order of processing the components of a vector-valued measurement,careful choice of coordinates (e.g., polar vs. Cartesian), hybrid coordinate systems, etc. There is no guarantee that any of these methods will improve EKF performance; in some cases, second-order corrections and/or iteration actually make EKF performance worse, contrary to intuition. Reviews of these techniques are given in Tanizaki [1996], as well as Wishner et al. [1969], Mehra [1971], Leskiw et al. [1982], Gelb [1974], Bellaire et al. [1995], Henriksen [1982], Fang [1976], Daum et al. [1983], and Jazwinski [1970].

Nonlinear Filters

Considering the frequently disappointing performance of the EKF noted in the previous section, there has been intense research to develop better nonlinear filters. An exact nonlinear recursive filter was derived by Beneš (for a certain class of nonlinear problems) in a seminal paper [1981]. The Beneš filter is "exact" in the sense that it computes an optimal estimate of x, without any approximations, in contrast to the EKF, which uses linear approximations. A generalization of the Beneš filter and the **Kalman filter** was developed by Daum [1986a; 1986b]. Figure 16.3 shows the superior performance of this new nonlinear filter compared to the EKF for certain practical applications; see Schmidt [1993] for details. A more general class of exact nonlinear recursive filters is based on the exponential family of probability densities. The Kalman filter theory is based on a Gaussian density, which is a special case of the exponential family; see Daum [1988; 1997a, b] for a development of this theory, which is summarized in Table 16.2.

Another alternative to the EKF, reported in Julier et al. [1995], is called the *unscented filter*, and in contrast to the EKF, does not use Jacobians, but rather evaluates multidimensional integrals by sampling at carefully selected points much like Gauss-Hermite quadrature formulas. The unscented filter shows much better performance than the EKF in certain applications, with less computational complexity than the EKF.

Exact recursive filters for nonlinear estimation problems generally do not exist. This is not surprising, considering that the existence of an exact recursive filter corresponds to the following minor miracle:

$$\underset{\substack{\uparrow \\ \text{growing} \\ \text{dimension} \\ \text{with k}}}{p\left(x,\ t\middle|Z_k\right)} = \underset{\substack{\uparrow \\ \text{fixed} \\ \text{dimension} \\ \text{for all k}}}{p\left(x,\ t\middle|\psi_k\right)} \tag{16.56}$$

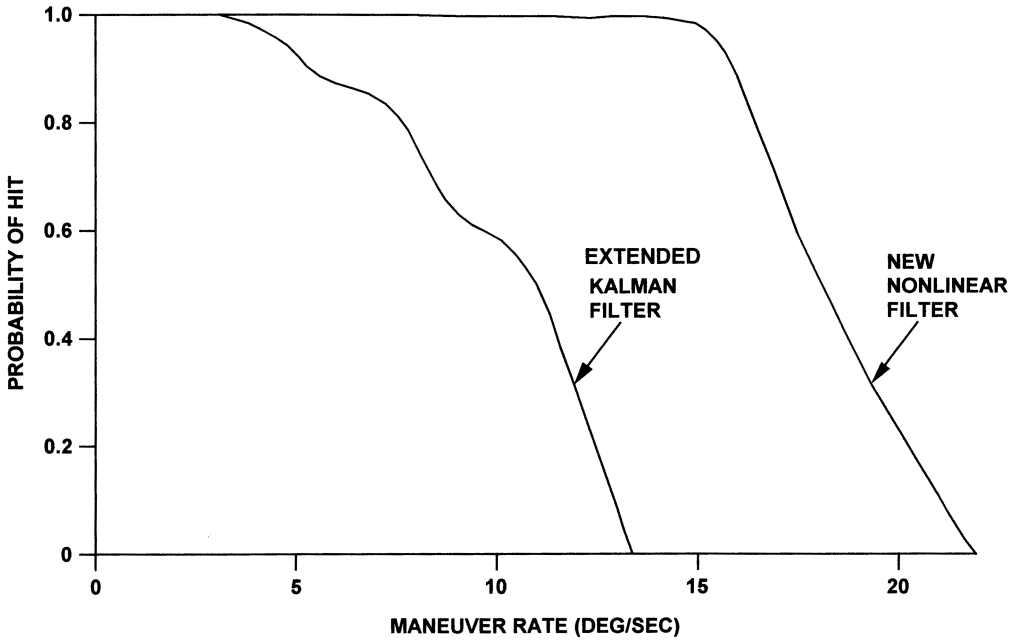


FIGURE 16.3 New nonlinear filter vs. extended Kalman filter [See Schmidt (1993)].

in which ψ_k is a sufficient statistic for x. A *recursive* filter exists when there is a sufficient statistic with fixed finite dimension. In classical statistics, for parameter estimation, it is well known that this will happen (assuming certain regularity conditions) if and only if the conditional density is from an exponential family; see Daum [1988]. The theory of fixed finite dimensional filters has also been developed from a completely different perspective, using Lie algebras; see Beneš [1987].

Non-recursive filters generally have superior performance compared with the EKF, at the cost of higher computational complexity. For parameter estimation problems (corresponding to zero process noise, $Q_k = 0$), these non-recursive filters are popular in practical applications despite the increased computational complexity compared with the EKF. Gauss invented this type of non-recursive nonlinear filter over 200 years ago; see Sorenson [1980]. On the other hand, non-recursive filters for estimating x, where x is a Markov process with non-zero process noise ($Q_k \neq 0$), generally have much greater computational complexity. Nevertheless, with a sufficiently fast computer, it would be practical to implement such a non-recursive algorithm. The theory to design such algorithms is well known, and some Monte Carlo simulations have shown excellent performance relative to the EKF; see Sorenson [1988] and Kastella et al. [1997]. Presumably, with computers getting ten times faster (at fixed cost) every 5 years, the application of such non-recursive filters will become common in the future, despite very high computational complexity relative to the EKF.

Practical Issues

Data Association

One of the best ways to ruin the performance of a **Kalman filter** is to put the wrong data into it. In a dense multiple-target environment, for applications with sensors such as radar, passive infrared, sonar, acoustic, or optical, the question of which measurement originated from which target is a very serious issue. In addition, the measurements could be unresolved mixtures from two or more objects. There is a plethora of algorithms to mitigate these problems, as well as a vast literature on this subject, including Blackman [1986]; Blackman et al. [1999], Bar-Shalom [1995], and Daum [1992].

TABLE 16.2 Exact Recursive Filters

Filter	Conditional Density $p(x, t \mid Z_k)$	Class of Dynamics	Propagation Equations
1. Kalman (1960)	η = Gaussian	$\frac{\partial f}{\partial x} = A(t)$	$\dot{m} = Am$ $\dot{P} = AP + PA^T + GG^T$
2. Beneš (1981)	$\eta \exp\left[\int^x f(x)dx\right]$	$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x}\right)^T$ and $\left\|f(x)\right\|^2 + \mathrm{tr}\left(\frac{\partial f}{\partial x}\right) = x^T Ax + b^T x + c$	$\dot{m} = -P\ Am - \frac{1}{2}Pb$ $\dot{P} = I - PAP$
3. Daum (1986)	$\eta\, P_{ss}^{\alpha}(x)$	$f - \alpha Q r^T = Dx + E$ and $\mathrm{tr}\left(\frac{\partial f}{\partial x}\right) + \frac{\alpha}{2} r Q r^T = x^T Ax + b^T x + c$ where $r = \frac{\partial}{\partial x} \log P_{ss}(x)$	$\dot{m} = 2(\alpha - 1)PAm + Dm + (\alpha - 1)Pb + E$ $\dot{P} = 2(\alpha - 1)PAP + DP + PD^T + Q$
4. Daum (1986)	$\eta\, q^{\alpha}(x, t)$	Same as filter 3, but with $r = \frac{\partial}{\partial x} \log q(x, t)$	Same as filter 3
5. Daum (1986)	$\eta Q(x, t)$	$\frac{\partial f}{\partial x} - \left(\frac{\partial f}{\partial x}\right)^T = D^T - D$ and $\frac{\partial f}{\partial t} + \dot{D}x + \dot{E} = -\frac{\partial f^T}{\partial x} f - \frac{1}{2}\left[\frac{\partial}{\partial x}\mathrm{tr}\left(\frac{\partial f}{\partial x}\right)\right]^T$ $+ (2A + D^T D)x + D^T E + b$	$\dot{m} = -(2PA + D)m - E - Pb$ $\dot{P} = -2PAP - PD^T - DP + I$
6. Daum (1988)	$p(x, t)\exp\left[\theta^T(x, t)\psi(Z_k, t)\right]$	Solution of PDE for $\theta(x, t)$	$\frac{d\psi}{dt} = A^T \psi + \Gamma$ where $\Gamma = (\Gamma_1, \Gamma_2, \ldots, \Gamma_M)^T$ with $\Gamma_j = \psi^T B_j \psi$

Ill-Conditioning

The **Kalman filter** covariance matrix can be extremely ill-conditioned in certain applications, as analyzed in Daum et al. [1983]. A Kalman filter is ill-conditioned if its performance is significantly degraded by numerical errors. Special factorizations of the covariance matrices have been developed to mitigate this problem; the standard reference is Bierman [1977]. There are many other methods to mitigate ill-conditioning, as discussed by Daum et al. [1983].

Adaptive Kalman Filters

The Kalman filter was derived assuming that the process noise covariance matrix, Q_k, as well as all other matrices (Φ_k, H_k, R_k) are known exactly *a priori*. In practice, however, these assumptions may be inappropriate. To mitigate this uncertainty a number of adaptive Kalman filter algorithms can be used. Most of these algorithms consist of a bank of Kalman filters combined adaptively using Bayes' rule. This basic structure was invented by Magill [1965], and it has now evolved into a more sophisticated algorithm, called *interacting multiple models* invented by Blom [1984]. A recent survey of this topic is given by Bar-Shalom et al. [1995].

Measurement Models

The Kalman filter theory assumes Gaussian measurement errors that are statistically independent from sample to sample, with zero mean and exactly known covariance matrix (R_k). However, in many practical applications, these are poor approximations of reality. For example, in radar applications, the measurments of range, azimuth, and elevations are often biased, non-Gaussian, and correlated with time, owing to diverse physical effects including multipath, tropospheric refraction, ducting, ionospheric refraction, glint, RV wake, rocket exhaust plume, RFI, ECM, unresolved measurements, bias errors in time, location and angular orientation for the radar itself, radar hardware errors, etc. Gauss himself cautioned against naïve least squares fitting of data with bias and drift; see Gauss [1995].

Performance Evaluation

As shown in Fig. 16.1, the Kalman filter computes the covariance matrix of the estimation error (P_k). However, in practical applications, this theoretical covariance matrix may be extremely optimistic, owing to the effects noted earlier (nonlinearity, ill-conditioning, data association errors, unresolved data, errors in modeling both measurement errors and target dynamics) as well as bugs in the software itself. Therefore, the standard approach to evaluate Kalman filter performance is Monte Carlo simulation. However, no one in their right mind would believe the results of a complex Monte Carlo simulation without a back-of-the-envelope calculation that is in rough agreement with the simulation results. A good source of such simple formulas is Brookner [1998].

Obviously, the very best way to evaluate Kalman filter performance is to conduct extensive real-world testing. Unfortunately, the cost and practicality of this approach is often prohibitive or is deemed to be not cost-effective. A judicious combination of extensive Monte Carlo simulation and limited real-world testing is often the most practical approach to performance evaluation.

The best possible performance for EKFs can be computed using theoretical lower bounds on the error covariance matrix, such as the Cramér-Rao bound (CRB) for parameter estimation. For the standard Kalman filter setup with zero process noise ($Q_k = 0$), it turns out that the CRB is simply the Kalman filter error covariance matrix itself; see Taylor [1979]. On the other hand, for non-zero process noise, the available bounds are much more complex to compute and they are generally not tight; see Kerr [1989] for a detailed review of the state of the art. More generally, the theory of error bounds when data association errors are considered is developed in Daum [1997a, b].

Digital Realization

All of the algorithms discussed here are always implemented using digital computers, owing to their superior accuracy, flexibility, and dynamic range, as compared to currently available analog devices. Generally, 32-bit or 64-bit floating point arithmetic is required for most practical applications, although extra precision may be required for extremely ill-conditioned problems.

The idea of using analog computers to implement Kalman filters (which is sometimes suggested in academic circles) is rather naïve, owing to the limited accuracy, limited dynamic range, and inflexibility of analog

computers. Likewise, the filtering theory for continuous time measurements (which dominates the academic literature on nonlinear filtering) is also impractical because measurements must be made in discrete time to accommodate digital computers. The naïve approximation of discrete time measurements by continuous time data generally results in poor performance, and it is not used by practical engineers. The academic literature is out of touch with such practical issues; for example, see Hazewinkel et al. [1981].

Defining Terms

Kalman filter: A recursive algorithm that estimates the state vector of a linear dynamical system given noisy measurements that are linear in the state vector. This algorithm was invented by Rudolf E. Kalman, and it was published in 1960.

Extended Kalman filter: A recursive algorithm for estimating the state of nonlinear dynamical systems that uses the Kalman filter equations based on a linear approximation to the nonlinear dynamical system and/or nonlinear measurement equations.

State vector: A vector that specifies the state of a dynamical system. For example, the position and velocity of a leaf falling to the ground could be the state vector for the leaf. For deterministic systems, this corresponds to the initial conditions for a system of ordinary differential equations. For a special type of random process, called a Markov process, the future state is statistically independent of the past, conditioned on knowledge of the state at the present time.

Covariance matrix: A matrix that gages the uncertainty in a vector using second moments. The diagonal elements of this matrix are the variances of uncertainty in the components of the vector.

References

Y. Bar-Shalom and X. Li. *Multitarget-Multisensor Tracking*, YBS, 1995.

R. Bellaire, E. W. Kamen, and S. M. Zabin, A new nonlinear iterated filter with applications to target tracking, *SPIE Proceedings*, San Diego, 1995.

V. E. Beneš, Nonlinear filtering: problems, examples, applications, *Advances in Statistical Signal Processing,* Vol. 1, pp. 1–14, JAI Press, 1987.

V. E. Beneš, Exact finite-dimensional filters for certain diffusions with nonlinear drift, *Stochastics,* 5, 65–92, 1981.

G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, New York: Academic, 1977.

S. S. Blackman and R. F. Popoli. *Design and Analysis of Modern Tracking Systems*, Artech House, 1999.

S. S. Blackman. *Multi-Target Tracking with Radar Applications*, Artech House Inc., 1986.

H. A. P. Blom. A sophisticated tracking algorithm for ATC surveillance data, *Proceedings of International Radar Conference*, Paris, 1984.

E. Brookner. *Tracking and Kalman Filtering Made Easy*, John Wiley & Sons, 1998.

R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, third edition, John Wiley & Sons, 1997.

A.E. Bryson and Y. C. Ho, *Applied Optimal Control*, Blaisdell Publishing, 1969.

R. S. Bucy, Linear and nonlinear filtering, *Proc. IEEE,* 58, 854–864, 1970.

F. E. Daum. (1997a), Virtual measurements for nonlinear filters, *Proceedings of IEEE Control and Decision Conference*, San Diego, pp. 1657–1662, 1997.

F. E. Daum. (1997b) Cramér-Rao type bounds for random set problems, pp. 165–183 in *Random Sets*, Ed. By J. Goutsias, R. Mahler, and H. T. Nguyen, Springer-Verlag, 1997.

F. E. Daum. Beyond Kalman filters: practical design of nonlinear filters, *Proceedings of the SPIE Conference on Signal and Data Processing of Small Targets*, pp. 252–262, Orlando, FL, April 1995.

F. E. Daum. A system approach to multiple target tracking, *Multitarget-Multisensor Tracking*, Volume II (Y. Bar-Shalom, ed.), pp. 149–181, Artech House, 1992.

F. E. Daum. New exact nonlinear filters, *Bayesian Analysis of Time Series and Dynamic Models* (J.C. Spall, ed.), pp. 199–226, Marcel Dekker, New York, 1988.

F. E. Daum. (1986a). Exact finite dimensional nonlinear filters, *IEEE Trans. Autom. Control* AC-31(7), 616–622, 1986.

F. E. Daum. (1986b). New nonlinear filters and exact solutions of the Fokker-Planck equations, in *Proceedings of the American Control Conference,* pp. 884–888, 1986.
F. E. Daum and R. J. Fitzgerald. Decoupled Kalman filters for phased array radar tracking, *IEEE Trans. Autom. Control,* AC-28, 269–283, 1983.
B. T. Fang. A nonlinear counterexample for batch and extended sequential estimation algorithms, *IEEE Trans. Autom. Control*, AC-21, 138–139, 1976.
C. F. Gauss. *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, translated by G. W. Stewart, SIAM, 1995.
A. Gelb (Editor). *Applied Optimal Estimation*, MIT Press, 1974.
M. Hazewinkel and J. C. Willems, Eds. *Stochastic Systems: The Mathematics of Filtering and Identification and Applications,* D. Reidel, Dordrecht, The Netherlands, 1981.
R. Henriksen. The truncated second-order nonlinear filter revisited, *IEEE Trans. Autom. Control,* AC-27, 247–251, 1982.
Y. C. Ho and R. C. K. Lee. A Bayesian approach to problems in stochastic estimation and control, *IEEE Trans. Autom. Control*, AC-9, 333–339, 1964.
C. E. Hutchinson. The Kalman filter applied to aerospace and electronic systems, *IEEE Trans. Aerosp. Electron. Syst.* 500–504, 1984.
A. H. Jazwinski. *Stochastic Processes and Filtering Theory,* Academic Press, New York, 1970.
S. Julier, J. Uhlmann, and H. Durrant-Whyte. A new approach to filtering nonlinear systems, *Proceedings of American Control Conference*, June 1995.
R. E. Kalman. New methods in Wiener filtering theory, in *Proc. Symp. Eng. Appl. of Random Function Theory and Probability,* F. Kozin and J. L. Bogdanoff, Eds. New York: Wiley, 1963.
R. E. Kalman. A new approach to linear filtering and prediction problems, *Trans. ASME J. Basic Eng.*, 82D, 35–45, 1960.
K. Kastella and A. Zatezalo. Nonlinear filtering for detection, tracking and ID algorithms, *ONR/NSWC Workshop on Filtering and Tracking*, May 1997.
T. H. Kerr. Status of CR-like lower bounds for nonlinear filtering, *IEEE Trans. Aerosp. Electron. Syst.*, 25, 590–601, 1989.
D. M. Leskiw and K. S. Miller. Nonlinear estimation with radar observations, *IEEE Trans. Aerosp. Electron. Syst.,* AES-18, 192–200, 1982.
D. T. Magill. Optimal adaptive estimation of sampled stochastic processes, *IEEE Trans. Autom. Control*, AC-10, 434–439, 1965.
R. K. Mehra. A comparison of several nonlinear filters for reentry vehicle tracking, *IEEE Trans. Autom. Control,* AC-16, 307–319, 1971.
G. C. Schmidt. Designing nonlinear filters based on Daum's Theory, *AIAA Journal of Guidance, Control and Dynamics,* 16, 371–376, 1993.
B. E. Schutz, J. D. McMillan, and B. D. Tapley. Comparison of statistical orbit determination methods, *AIAA J.*, Nov. 1974.
H. W. Sorenson. Recursive estimation for nonlinear dynamic systems, *Bayesian Analysis of Time Series and Dynamic Models,* J.C. Spall, Ed., pp. 127–165, Marcel Dekker, 1988.
H. W. Sorenson. *Kalman Filtering: Theory and Applications*, IEEE Press, New York, 1985.
H. W. Sorenson. *Parameter Estimation*, Marcel Dekker, 1980.
H. W. Sorenson. On the development of practical nonlinear filters, *Inf. Sci.* 7, 253–270, 1974.
H. W. Sorenson. On the error behavior in linear minimum variance estimation problems, *IEEE Trans. Autom. Control*, AC-12, 557–562, 1967.
H. Tanizaki. *Nonlinear Filters*, 2nd ed., Springer-Verlag, 1996.
J. H. Taylor. Cramér-Rao estimation error lower bound analysis for nonlinear systems with unknown deterministic variables, *IEEE Trans. Autom. Control*, April 1979.
R. P. Wishner, R. E. Larson, and M. Athans. Status of radar tracking algorithms, *Symp. on Nonlinear Estimation Theory and Appl.*, 1970.
R. P. Wishner, J. A. Tabaczynski, and M. Athans. A comparison of three non-linear filters, *Automatica* 5, 487–496, 1969.

Further Information

The best concise introduction to Kalman filtering is Chapter 12 in Bryson and Ho [1969]. The three best books on Kalman filters are Gelb [1974], Sorenson [1985], and Brown et al. [1997]. The standard reference on nonlinear filters is Jazwinski [1970]. The best journal on Kalman filter applications is the *IEEE Transactions on Aerospace and Electronic Systems*, which typically has several practical papers on Kalman filters each issue. Two good conferences with many papers on Kalman filters are the *IEEE Conference on Decision and Control* (mid-December annually) and the *SPIE Conference on Signal and Data Processing* (April each year).

Delp, E.J. Allebach, J., Bouman, C.A., Rajala, S.A., Bose, N.K., Sibul, L.H., Wolf, W., Zhang, Y-Q. "Multidimensional Signal Processing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

17

Multidimensional Signal Processing

Edward J. Delp
Purdue University

Jan Allebach
Purdue University

Charles A. Bouman
Purdue University

Sarah A. Rajala
North Carolina State University

N. K. Bose
Pennsylvania State University

L. H. Sibul
Pennsylvania State University

Wayne Wolf
Princeton University

Ya-Qin Zhang
Microsoft Research, China

17.1 Digital Image Processing
Image Capture • Point Operations • Image Enhancement • Digital Image Compression • Reconstruction • Edge Detection • Analysis and Computer Vision

17.2 Video Signal Processing
Sampling • Quantization • Vector Quantization • Video Compression • Information-Preserving Coders • Predictive Coding • Motion-Compensated Predictive Coding • Transform Coding • Subband Coding • HDTV • Motion Estimation Techniques • Token Matching Methods • Image Quality and Visual Perception • Visual Perception

17.3 Sensor Array Processing
Spatial Arrays, Beamformers, and FIR Filters • Discrete Arrays for Beamforming • Discrete Arrays and Polynomials • Velocity Filtering

17.4 Video Processing Architectures
Computational Techniques • Heterogeneous Multiprocessors • Video Signal Processors • Instruction Set Extensions

17.5 MPEG-4 Based Multimedia Information System
MPEG-4 Multimedia System

17.1 Digital Image Processing

Edward J. Delp, Jan Allebach, and Charles A. Bouman

What is a **digital image**? What is digital image processing? Why does the use of computers to process pictures seem to be everywhere? The space program, robots, and even people with personal computers are using digital image processing techniques. In this section we shall describe what a digital image is, how one obtains digital images, what the problems with digital images are (they are not trouble-free), and finally how these images are used by computers. A discussion of processing the images is presented later in the section. At the end of this section is a bibliography of selected references on digital image processing.

The use of computers to process pictures is about 30 years old. While some work was done more than 50 years ago, the year 1960 is usually the accepted date when serious work was started in such areas as optical character recognition, **image coding**, and the space program. NASA's Ranger moon mission was one of the first programs to return digital images from space. The Jet Propulsion Laboratory (JPL) established one of the early general-purpose image processing facilities using second-generation computer technology.

The early attempts at digital image processing were hampered because of the relatively slow computers used, i.e., the IBM 7094, the fact that computer time itself was expensive, and that image digitizers had to be built by the research centers. It was not until the late 1960s that image processing hardware was generally available (although expensive). Today it is possible to put together a small laboratory system for less than $60,000; a system based on a popular home computer can be assembled for about $5,000. As the cost of computer hardware

decreases, more uses of digital image processing will appear in all facets of life. Some people have predicted that by the turn of the century at least 50% of the images we handle in our private and professional lives will have been processed on a computer.

Image Capture

A digital image is nothing more than a matrix of numbers. The question is how does this matrix represent a real image that one sees on a computer screen?

Like all imaging processes, whether they are analog or digital, one first starts with a sensor (or transducer) that converts the original imaging energy into an electrical signal. These sensors, for instance, could be the photomultiplier tubes used in an x-ray system that converts the x-ray energy into a *known* electrical voltage. The transducer system used in ultrasound imaging is an example where sound pressure is converted to electrical energy; a simple TV camera is perhaps the most ubiquitous example. An important fact to note is that the process of conversion from one energy form to an electrical signal is not necessarily a *linear* process. In other words, a proportional charge in the input energy to the sensor will not always cause the same proportional charge in the output electrical signal. In many cases calibration data are obtained in the laboratory so that the relationship between the input energy and output electrical signal is known. These data are necessary because some transducer performance characteristics change with age and other usage factors.

The sensor is not the only thing needed to form an image in an imaging system. The sensor must have some spatial extent before an image is formed. By spatial extent we mean that the sensor must not be a simple point source examining only one location of energy output. To explain this further, let us examine two types of imaging sensors used in imaging: a CCD video camera and the ultrasound transducer used in many medical imaging applications.

The CCD camera consists of an *array* of light sensors known as charge-coupled devices. The image is formed by examining the output of each sensor in a preset order for a finite time. The electronics of the system then forms an electrical signal which produces an image that is shown on a cathode-ray tube (CRT) display. The image is formed because there is an array of sensors, each one examining only one spatial location of the region to be sensed.

The process of **sampling** the output of the sensor array in a particular order is known as *scanning*. Scanning is the typical method used to convert a two-dimensional energy signal or image to a one-dimensional electrical signal that can be handled by the computer. (An image can be thought of as an energy field with spatial extent.) Another form of scanning is used in ultrasonic imaging. In this application there is *only* one sensor instead of an array of sensors. The ultrasound transducer is moved or steered (either mechanically or electrically) to various spatial locations on the patient's chest or stomach. As the sensor is moved to each location, the output electrical signal of the sensor is sampled and the electronics of the system then form a television-like signal which is displayed. Nearly all the transducers used in imaging form an image by either using an array of sensors or a single sensor that is moved to each spatial location.

One immediately observes that both of the approaches discussed above are equivalent in that the energy is sensed at various spatial locations of the object to be imaged. This energy is then converted to an electrical signal by the transducer. The image formation processes just described are classical analog image formation, with the distance between the sensor locations limiting the spatial resolution in the system. In the array sensors, resolution is determined by how close the sensors are located in the array. In the single-sensor approach, the spatial resolution is limited by how far the sensor is moved. In an actual system spatial resolution is also determined by the performance characteristics of the sensor. Here we are assuming for our purposes *perfect* sensors.

In digital image formation one is concerned about two processes: *spatial sampling* and **quantization**. Sampling is quite similar to scanning in analog image formation. The second process is known as *quantization* or *analog-to-digital conversion,* whereby at each spatial location a *number* is assigned to the amount of energy the transducer observes at that location. This number is usually proportional to the electrical signal at the output of the transducer. The overall process of sampling and quantization is known as *digitization*. Sometimes the digitization process is just referred to as analog-to-digital conversion, or A/D conversion; however, the reader should remember that digitization also includes spatial sampling.

The digital image formulation process is summarized in Fig. 17.1. The spatial sampling process can be considered as overlaying a grid on the object, with the sensor examining the energy output from each grid box

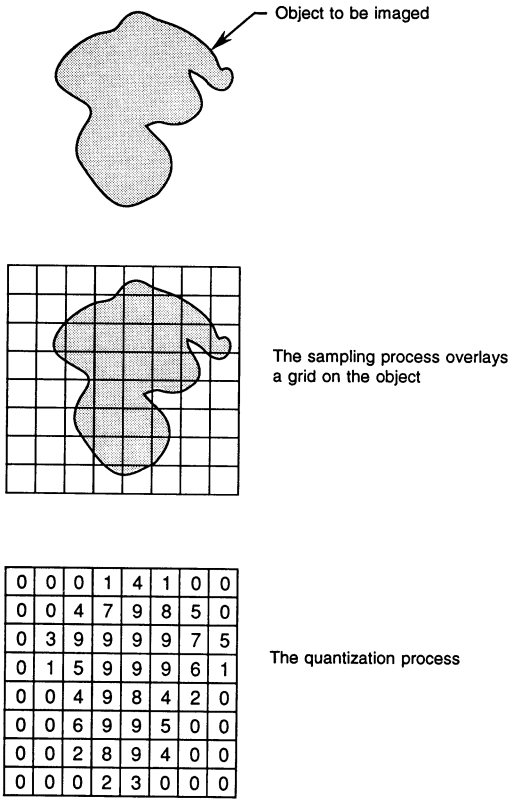


FIGURE 17.1 Digital image formation: sampling and quantization.

and converting it to an electrical signal. The quantization process then assigns a number to the electrical signal; the result, which is a *matrix* of numbers, is the digital representation of the image. Each spatial location in the image (or grid) to which a number is assigned is known as a *picture element* or **pixel** (or pel). The size of the sampling grid is usually given by the number of pixels on each side of the grid, e.g., 256 × 256, 512 × 512, 488 × 380.

The quantization process is necessary because all information to be processed using computers must be represented by numbers. The quantization process can be thought of as one where the input energy to the transducer is represented by a finite number of energy values. If the energy at a particular pixel location does not take on one of the finite energy values, it is assigned to the closest value. For instance, suppose that we assume *a priori* that only energy values of 10, 20, 50, and 110 will be represented (the units are of no concern in this example). Suppose at one pixel an energy of 23.5 was observed by the transducer. The A/D converter would then assign this pixel the energy value of 20 (the closest one). Notice that the quantization process makes mistakes; this error in assignment is known as *quantization* error or *quantization noise.*

In our example, each pixel is represented by one of four possible values. For ease of representation of the data, it would be simpler to assign to each pixel the index value 0, 1, 2, 3, instead of 10, 20, 50, 110. In fact, this is typically done by the quantization process. One needs a simple table to know that a pixel assigned the value 2 corresponds to an energy of 50. Also, the number of possible energy levels is typically some integer power of two to also aid in representation. This power is known as the number of *bits* needed to represent the energy of each pixel. In our example each pixel is represented by two bits.

One question that immediately arises is how accurate the digital representation of the image is when one compares the digital image with a corresponding analog image. It should first be pointed out that after the digital image is obtained one requires special hardware to convert the matrix of pixels back to an image that can be viewed on a CRT display. The process of converting the digital image back to an image that can be viewed is known as *digital-to-analog conversion,* or *D/A conversion.*

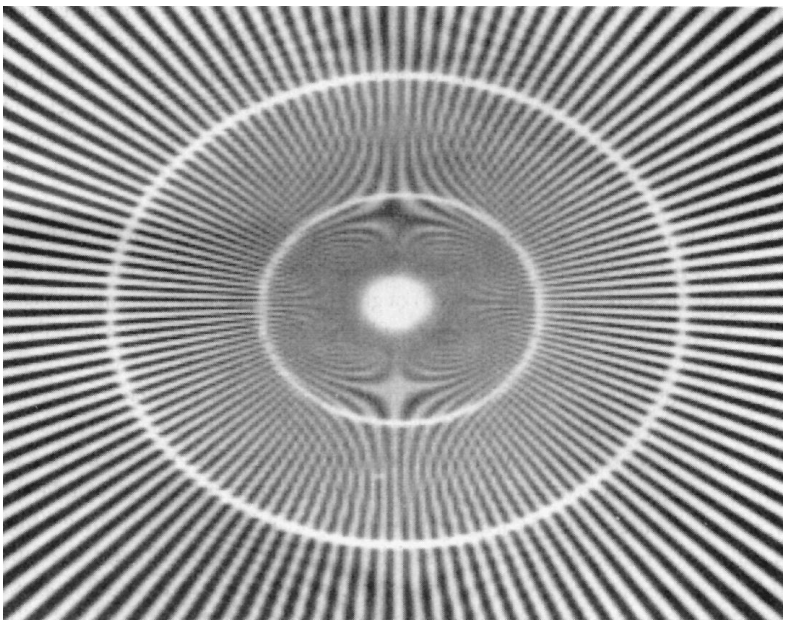


FIGURE 17.2 This image shows the effects of aliasing due to sampling the image at too low a rate. The image should be straight lines converging at a point. Because of undersampling, it appears as if there are patterns in the lines at various angles. These are known as moiré patterns.

The quality of representation of the image is determined by how close spatially the pixels are located and how many levels or numbers are used in the quantization, i.e., how coarse or fine is the quantization. The sampling accuracy is usually measured in how many pixels there are in a given area and is cited in pixels/unit length, i.e., pixels/cm. This is known as the *spatial sampling rate*. One would desire to use the lowest rate possible to minimize the number of pixels needed to represent the object. If the sampling rate is too low, then obviously some details of the object to be imaged will not be represented very well. In fact, there is a mathematical theorem which determines the lowest sampling rate possible to preserve details in the object. This rate is known as the *Nyquist* sampling rate (named after the late Bell Laboratories engineer Harry Nyquist). The theorem states that the sampling rate must be *twice* the highest possible detail one expects to image in the object. If the object has details closer than, say 1 mm, one must take at least 2 pixels/mm. (The Nyquist theorem actually says more than this, but a discussion of the entire theorem is beyond the scope of this section.) If we sample at a lower rate than the theoretical lowest limit, the resulting digital representation of the object will be distorted. This type of distortion or sampling error is known as *aliasing* errors. Aliasing errors usually manifest themselves in the image as moiré patterns (Fig. 17.2). The important point to remember is that there is a *lower limit* to the spatial sampling rate such that object detail can be maintained. The sampling rate can also be stated as the total number of pixels needed to represent the digital image, i.e., the matrix size (or grid size). One often sees these sampling rates cited as 256×256, 512×512, and so on. If the same object is imaged with a large matrix size, the sampling rate has obviously increased. Typically, images are sampled on 256×256, 512×512, or 1024×1024 grids, depending on the application and type of modality. One immediately observes an important issue in digital representation of images: that of the large number of pixels needed to represent the image. A 256×256 image has 65,536 pixels and a 512×512 image has 262,144 pixels! We shall return to this point later when we discuss processing or storage of these images.

The quality of the representation of the digital image is also determined by the number of levels or shades of gray that are used in the quantization. If one has more levels, then fewer mistakes will be made in assigning values at the output of the transducer. Figure 17.3 demonstrates how the number of gray levels affects the digital representation of an artery. When a small number of levels are used, the quantization is coarse and the quantization error is large. The quantization error usually manifests itself in the digital image by the appearance

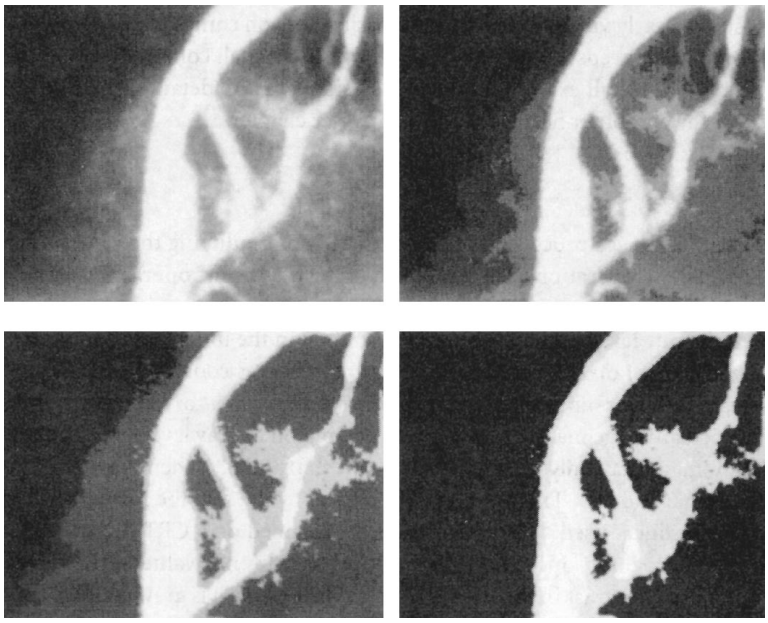


FIGURE 17.3 This image demonstrates the effects of quantization error. The upper left image is a coronary artery image with 8 bits (256 levels or shades of gray) per pixel. The upper right image has 4 bits/pixel (16 levels). The lower left image has 3 bits/pixel (8 levels). The lower right image has 2 bits/pixel (4 levels). Note the false contouring in the images as the number of possible levels in the pixel representation is reduced. This false contouring is the quantization error, and as the number of levels increases the quantization error decreases because fewer mistakes are being made in the representation.

of *false contouring* in the picture. One usually needs at least 6 bits or 64 gray levels to represent an image adequately. Higher-quality imaging systems use 8 bits (256 levels) or even as many as 10 bits (1024 levels) per pixel. In most applications, the human observer cannot distinguish quantization error when there are more than 256 levels. (Many times the number of gray levels is cited in bytes. One byte is 8 bits, i.e., high-quality monochrome digital imaging systems use one byte per pixel.)

One of the problems briefly mentioned previously is the large number of pixels needed to represent an image, which translates into a large amount of digital data needed for the representation. A 512×512 image with 8 bits/pixel (1 byte/pixel) of gray level representation requires 2,097,152 bits of computer data to describe it. A typical computer file that contains 1000 words usually requires only about 56,000 bits to describe it. The 512×512 image is 37 times larger! (A picture is truly worth more than 1000 words.) This data requirement is one of the major problems with digital imaging, given that the storage of digital images in a computer file system is expensive. Perhaps another example will demonstrate this problem. Many computers and word processing systems have the capability of transmitting information over telephone lines to other systems at data rates of 2400 bits per second. At this speed it would require nearly 15 minutes to transmit a 512×512 image! Moving objects are imaged digitally by taking *digital snapshots* of them, i.e., digital video. True digital imaging would acquire about 30 images/s to capture all the important motion in a scene. At 30 images/s, with each image sampled at 512×512 and with 8 bits/pixel, the system must handle 62,914,560 bits/s. Only very expensive acquisition systems are capable of handling these large data rates.

The greatest advantage of digital images is that they can be processed on a computer. Any type of operation that one can do on a computer can be done to a digital image. Recall that a digital image is just a (huge) matrix of numbers. Digital image processing is the process of using a computer to extract useful information from this matrix. Processing that cannot be done optically or with analog systems (such as early video systems) can be easily done on computers. The disadvantage is that a large amount of data needs to be processed and on some small computer systems this can take a long time (hours). We shall examine image processing in more detail in the next subsection and discuss some of the computer hardware issues in a later chapter.

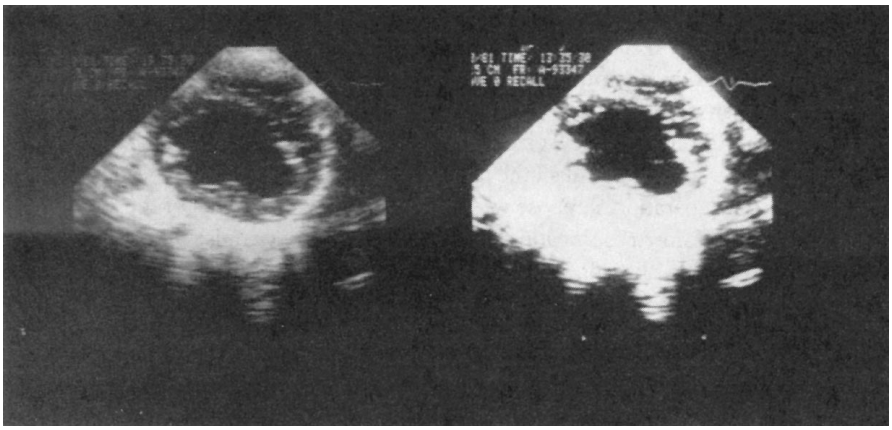


FIGURE 17.4 Contrast stretching. The image on the right has gray values between 0 and 63, causing the contrast to look washed out. The image on the right has been contrast enhanced by multiplying the gray levels by four.

Point Operations

Perhaps the simplest image processing operation is that of modifying the values of individual pixels in an image. These operations are commonly known as **point operations**. A point operation might be used to highlight certain regions in an image. Suppose one wished to know where all the pixels in a certain gray level region were *spatially* located in the image. One would modify all those pixel values to 0 (black) or 255 (white) such that the observer could see where they were located.

Another example of a point operation is *contrast enhancement* or *contrast stretching*. The pixel values in a particular image may occupy only a small region of gray level distribution. For instance, the pixels in an image may only take on values between 0 and 63, when they could nominally take on values between 0 and 255. This is sometimes caused by the way the image was digitized and/or by the type of transducer used. When this image is examined on a CRT display the contrast looks washed out. A simple point operation that multiplies each pixel value in the image by four will increase the apparent contrast in the image; the new image now has gray values between 0 and 252. This operation is shown in Fig. 17.4. Possibly the most widely used point operation in medical imaging is *pseudo-coloring*. In this point operation all the pixels in the image with a particular gray value are assigned a *color*. Various schemes have been proposed for appropriate pseudo-color tables that assign the gray values to colors. It should be mentioned that point operations are often cascaded, i.e., an image undergoes contrast enhancement and then pseudo-coloring.

The operations described above can be thought of as operations (or *algorithms*) that modify the range of the gray levels of the pixels. An important feature that describes a great deal about an image is the *histogram* of the pixel values. A histogram is a table that lists how many pixels in an image take on a particular gray value. These data are often plotted as a function of the gray value. Point operations are also known as *histogram modification* or *histogram stretching*. The contrast enhancement operation shown in Fig. 17.4 modifies the histogram of the resultant image by stretching the gray values from a range of 0–63 to a range of 0–252. Some point operations are such that the resulting histogram of the processed image has a particular shape. A popular form of histogram modification is known as *histogram equalization*, whereby the pixels are modified such that the histogram of the processed image is almost flat, i.e., all the pixel values occur equally.

It is impossible to list all possible types of point operations; however, the important thing to remember is that these operations process one pixel at a time by modifying the pixel based *only* on its gray level value and *not* where it is distributed spatially (i.e., location in the pixel matrix). These operations are performed to enhance the image, make it easier to see certain structures or regions in the image, or to force a particular shape to the histogram of the image. They are also used as initial operations in a more complicated image processing algorithm.

Image Enhancement

Image enhancement is the use of image processing algorithms to remove certain types of distortion in an image. The image is enhanced by removing noise, making the edge structures in the image stand out, or any other operation that makes the image *look* better.[1] Point operations discussed above are generally considered to be enhancement operations. Enhancement also includes operations that use groups of pixels and the spatial location of the pixels in the image.

The most widely used algorithms for enhancement are based on pixel functions that are known as **window operations**. A window operation performed on an image is nothing more than the process of examining the pixels in a certain region of the image, called the window region, and computing some type of mathematical function derived from the pixels in the window. In most cases the windows are square or rectangle, although other shapes have been used. After the operation is performed, the result of the computation is placed in the center pixel of the window where a 3×3 pixel window has been extracted from the image. The values of the pixels in the window, labeled $a_1, a_2, \ldots, a_9$, are used to compute a new pixel value which replaces the value of a_5, and the window is moved to a new center location until all the pixels in the original image have been processed. As an example of a window operation, suppose we computed the average value of the pixels in the window. This operation is known as *smoothing* and will tend to reduce noise in the image, but unfortunately it will also tend to blur edge structures in the image.

Another window operation often used is the computation of a linear weighted sum of the pixel values. Let a'_5 be the new pixel value that will replace a_5 in the original image. We then form

$$a'_5 = \sum_{i=1}^{9} \alpha_i a_i \tag{17.1}$$

where the α_i's are any real numbers. For the simple smoothing operation described above we set $\alpha_i = 1/9$ for all *i*. By changing the values of the α_i weights, one can perform different types of enhancement operations to an image. Any window operation that can be described by Eq. 17.1 is known as a *linear window operation* or *convolution* operator. If some of the α_i coefficients take on negative values, one can enhance the appearance of edge structures in the image.

It is possible to compute a nonlinear function of the pixels in the window. One of the more powerful nonlinear window operations is that of *median filtering*. In this operation all the pixels in the window are listed in descending magnitude and the middle, or *median,* pixel is obtained. The median pixel then is used to replace a_5. The median filter is used to remove noise from an image and at the same time preserve the edge structure in the image. More recently there has been a great deal of interest in *morphological operators*. These are also nonlinear window operations that can be used to extract or enhance shape information in an image.

In the preceding discussion, all of the window operations were described on 3×3 windows. The current research in window operations is directed at using large window sizes, i.e., 9×9, 13×13, or 21×21. The philosophy in this work is that small window sizes only use local information and what one really needs to use is information that is more global in nature.

Digital Image Compression

Image compression refers to the task of reducing the amount of data required to store or transmit a digital image. As discussed earlier, in its natural form, a digital image comprises an array of numbers. Each such

[1]Image enhancement is often confused with *image restoration*. Image enhancement is the ad hoc application of various processing algorithms to enhance the appearance of the image. Image restoration is the application of algorithms that use knowledge of the degradation process to enhance or restore the image, i.e., deconvolution algorithms used to remove the effect of the aperture point spread function in blurred images. A discussion of image restoration is beyond the scope of this section.

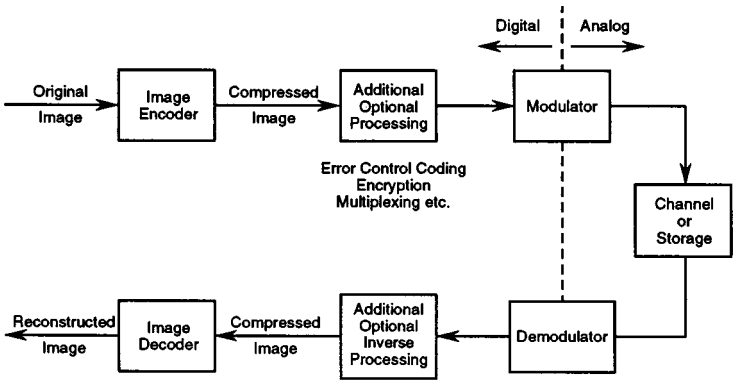


FIGURE 17.5 Overview of an image compression system.

number is the sampled value of the image at a pixel (picture element) location. These numbers are represented with finite precision using a fixed number of bits. Until recently, the dominant image size was 512×512 pixels with 8 bits or 1 byte per pixel. The total storage size for such an image is $512^2 \approx 0.25 \times 10^6$ bytes or 0.25 Mbytes. When digital image processing first emerged in the 1960s, this was considered to be a formidable amount of data, and so interest in developing ways to reduce this storage requirement arose immediately. Since that time, image compression has continued to be an active area of research. The recent emergence of standards for image coding algorithms and the commercial availability of very large scale integration (VLSI) chips that implement image coding algorithms is indicative of the present maturity of the field, although research activity continues apace.

With declining memory costs and increasing transmission bandwidths, 0.25 Mbytes is no longer considered to be the large amount of data that it once was. This might suggest that the need for image compression is not as great as previously. Unfortunately (or fortunately, depending on one's point of view), this is not the case because our appetite for image data has also grown enormously over the years. The old 512×512 pixels $\times$ 1 byte per pixel "standard" was a consequence of the spatial and gray scale resolution of sensors and displays that were commonly available until recently. At this time, displays with more than $10^3 \times 10^3$ pixels and 24 bits/pixel to allow full color representation (8 bits each for red, green, and blue) are becoming commonplace. Thus, our 0.25-Mbyte standard image size has grown to 3 Mbytes. This is just the tip of the iceberg, however. For example, in desktop printing applications, a 4-color (cyan, magenta, yellow, and black) image of an 8.5×11 in.^{2} page sampled at 600 dots per in. requires 134 Mbytes. In remote sensing applications, a typical hyperspectral image contains terrain irradiance measurements in each of 200 10-nm-wide spectral bands at 25-m intervals on the ground. Each measurement is recorded with 12-bit precision. Such data are acquired from aircraft or satellite and are used in agriculture, forestry, and other fields concerned with management of natural resources. Storage of these data from just a 10×10 km^2 area requires 4800 Mbytes.

Figure 17.5 shows the essential components of an image compression system. At the system input, the image is encoded into its compressed form by the image coder. The compressed image may then be subjected to further digital processing, such as error control coding, encryption, or multiplexing with other data sources, before being used to modulate the analog signal that is actually transmitted through the channel or stored in a storage medium. At the system output, the image is processed step by step to undo each of the operations that was performed on it at the system input. At the final step, the image is decoded into its original uncompressed form by the image decoder. Because of the role of the image encoder and decoder in an image compression system, **image coding** is often used as a synonym for image compression. If the reconstructed image is identical to the original image, the compression is said to be **lossless**. Otherwise, it is **lossy**.

Image compression algorithms depend for their success on two separate factors: redundancy and irrelevancy. *Redundancy* refers to the fact that each pixel in an image does not take on all possible values with equal probability, and the value that it does take on is not independent of that of the other pixels in the image. If this were not true, the image would appear as a white noise pattern such as that seen when a television receiver is tuned to an unused channel. From an information-theoretic point of view, such an image contains the

FIGURE 17.6 Key elements of an image encoder.

maximum amount of information. From the point of view of a human or machine interpreter, however, it contains no information at all. *Irrelevancy* refers to the fact that not all the information in the image is required for its intended application. First, under typical viewing conditions, it is possible to remove some of the information in an image without producing a change that is perceptible to a human observer. This is because of the limited ability of the human viewer to detect small changes in luminance over a large area or larger changes in luminance over a very small area, especially in the presence of detail that may mask these changes. Second, even though some degradation in image quality may be observed as a result of image compression, the degradation may not be objectionable for a particular application, such as teleconferencing. Third, the degradation introduced by image compression may not interfere with the ability of a human or machine to extract the information from the image that is important for a particular application. Lossless compression algorithms can only exploit redundancy, whereas lossy methods may exploit both redundancy and irrelevancy.

A myriad of approaches have been proposed for image compression. To bring some semblance of order to the field, it is helpful to identify those key elements that provide a reasonably accurate description of most encoding algorithms. These are shown in Fig. 17.6. The first step is *feature extraction*. Here the image is partitioned into $N \times N$ blocks of pixels. Within each block, a feature vector is computed which is used to represent all the pixels within that block. If the feature vector provides a complete description of the block, i.e., the block of pixel values can be determined exactly from the feature vector, then the feature is suitable for use in a lossless compression algorithm. Otherwise, the algorithm will be lossy. For the simplest feature vector, we let the block size $N = 1$ and take the pixel values to be the features. Another important example for $N = 1$ is to let the feature be the error in the prediction of the pixel value based on the values of neighboring pixels which have already been encoded and, hence, whose values would be known as the decoder. This feature forms the basis for *predictive encoding*, of which differential pulse-code modulation (DPCM) is a special case. For larger size blocks, the most important example is to compute a two-dimensional (2-D) Fourier-like transform of the block of pixels and to use the N^2 transform coefficients as the feature vector. The widely used Joint Photographic Experts Group (JPEG) standard image coder is based on the discrete cosine transform (DCT) with a block size of $N = 8$. In all of the foregoing examples, the block of pixel values can be reconstructed exactly from the feature vector. In the last example, the inverse DCT is used. Hence, all these features may form the basis for a lossless compression algorithm. A feature vector that does not provide a complete description of the pixel block is a vector consisting of the mean and variance of the pixels within the block and an $N \times N$ binary mask indicating whether or not each pixel exceeds the mean. From this vector, we can only reconstruct an approximation to the original pixel block which has the same mean and variance as the original. This feature is the basis for the lossy block truncation coding algorithm. Ideally, the feature vector should be chosen to provide as nonredundant as possible a representation of the image and to separate those aspects of the image that are relevant to the viewer from those that are irrelevant.

The second step in image encoding is **vector quantization**. This is essentially a clustering step in which we partition the feature space into cells, each of which will be represented by a single prototype feature vector. Since all feature vectors belonging to a given cell are mapped to the same prototype, the quantization process is irreversible and, hence, cannot be used as part of a lossless compression algorithm. Figure 17.7 shows an example for a two-dimensional feature space. Each dot corresponds to one feature vector from the image. The X's signify the prototypes used to represent all the feature vectors contained within its quantization cell, the boundary of which is indicated by the dashed lines. Despite the simplicity with which vector quantization may be described, the implementation of a vector quantizer is a computationally complex task unless some structure is imposed on it. The clustering is based on minimizing the distortion between the original and quantized feature vectors, averaged over the entire image. The distortion measure can be chosen to account for the relative sensitivity of the human viewer to different kinds of degradation. In one dimension, the vector quantizer reduces to the Lloyd-Max scalar quantizer.

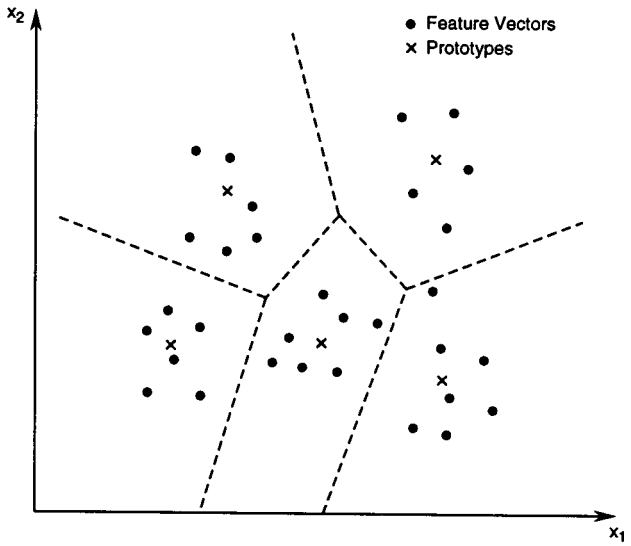


FIGURE 17.7 Vector quantization of a 2-D feature space.

The final step in image encoding is ***entropy*** *coding*. Here we convert the stream of prototype feature vectors to a binary stream of 0's and 1's. Ideally, we would like to perform this conversion in a manner that yields the minimum average number of binary digits per prototype feature vector.

In 1948, Claude Shannon proved that it is possible to code a discrete memoryless source using on the average as few binary digits per source symbol as the *source entropy* defined as

$$H = -\sum_n p_n \log_2 p_n$$

Here p_n denotes the probability or relative frequency of occurrence of the nth symbol in the source alphabet, and $\log_2(x) = \ln(x)/\ln(2)$ is the base 2 logarithm of x. The units of H are bits/source symbol. The proof of Shannon's theorem is based on grouping the source symbols into large blocks and assigning binary code words of varying length to each block of source symbols. More probable blocks of source symbols are assigned shorter code words, whereas less probable blocks are assigned longer code words. As the block length approaches infinity, the bit rate tends to H. Huffman determined the *optimum* variable-length coding scheme for a discrete memoryless source using blocks of any finite length.

Table 17.1 provides an example illustrating the concept of source coding. The source alphabet contains eight symbols with the probabilities indicated. For convenience, these symbols have been labeled in order of decreasing probability. In the context of image encoding, the source alphabet would simply consist of the prototype feature vectors generated by the vector quantizer. The entropy of this source is 2.31 bits/source symbol. If we were to use a fixed-length code for this source, we would need to use three binary digits for each source symbol as shown in Table 17.1. On the other hand, the code words for the Huffman code contain from 1 to 4 code letters (binary digits). In this case, the average code word length

$$\bar{l} = \sum_n p_n l_n$$

is $\bar{l} = 2.31$ binary digits. Here l_n is the number of code letters in the code word for the source symbol a_n. This is the average number of binary digits per source symbol that would be needed to encode the source, and it is equal to the entropy. Thus, for this particular source, the Huffman code achieves the lower bound. It can be shown that in general the rate for the Huffman code will always be within 1 binary digit of the source entropy. By grouping source symbols into blocks of length L and assigning code words to each block, this maximum

TABLE 17.1 A Discrete Source with an Eight-Symbol Alphabet and Two Schemes for Encoding It

Source Symbol	Probability p_n	Fixed-Length Code	Huffman Code
a_1	1/2	000	0
a_2	1/8	001	100
a_3	1/8	010	101
a_4	1/16	011	1100
a_5	1/16	100	1101
a_6	1/16	101	1110
a_7	1/32	110	11110
a_8	1/32	111	11111
	$H = 2.31$ bits/source symbol	$\bar{l} = 3$ binary digits/source symbol	$\bar{l} = 2.31$ binary digits/source symbol

distance can be decreased to $1/L$ binary digits. Note the subtle distinction here between *bits*, which are units of information, a property of the source alone, and *binary digits*, which are units of code word length, and hence only a property of the code used to represent the source. Also note that the Huffman code satisfies the *prefix condition,* i.e., no code word is the prefix of another longer code word. This means that a stream of 0's and 1's may be uniquely decoded into the corresponding sequence of source symbols without the need for markers to delineate boundaries between code words.

The Huffman code is determined from the *binary tree* shown in Fig. 17.8. This tree is constructed recursively by combining the two least probable symbols in the alphabet into one composite symbol whose probability of occurrence is the sum of the probabilities of the two symbols that it represents. The code words for these two symbols are the same as that of the composite symbol with a 0 or a 1 appended at the end to distinguish between them. This procedure is repeated until the reduced alphabet contains only a single code word. Then the code word for a particular source symbol is determined by traversing the tree from its root to the leaf node for that source symbol.

Reconstruction

The objective of **image reconstruction** is to compute an unknown image from many complex measurements of the image. Usually, each measurement depends on many pixels in the image which may be spatially distant from one another.

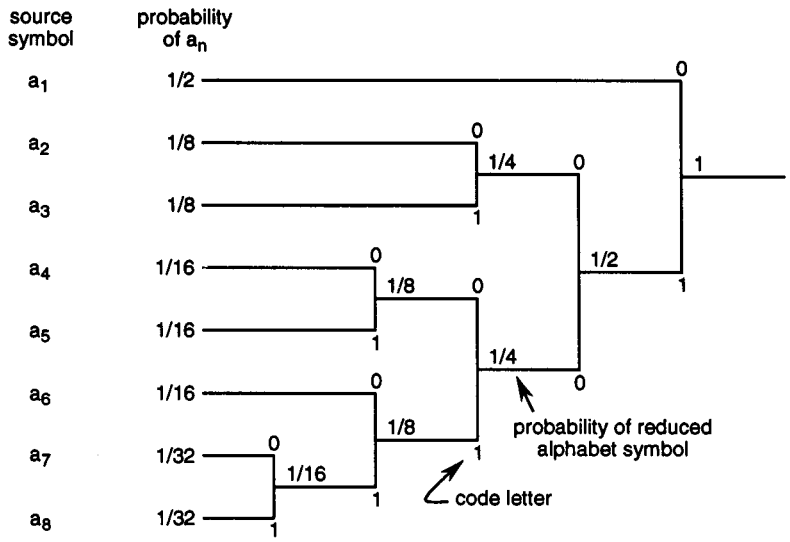


FIGURE 17.8 Binary tree used to generate the Huffman code for the source shown in Table 17.1.

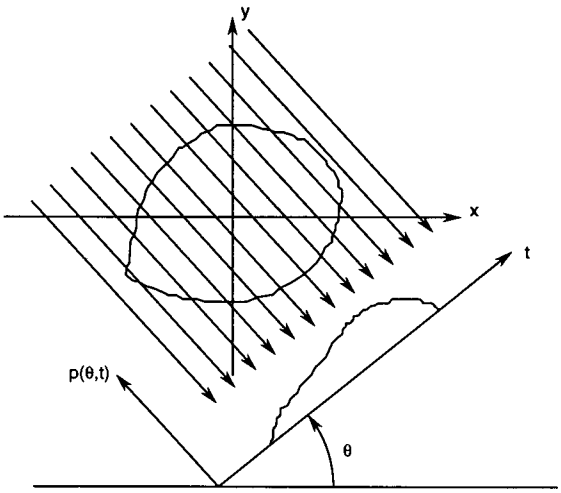


FIGURE 17.9 Projection data for angle θ, resulting in the one-dimensional function $p(\theta,t)$.

A typical reconstruction problem is **tomography**, in which each measurement is obtained by integrating the pixel values along a ray through the image. Figure 17.9 illustrates the measurement of these ray integrals in the **projection** process. For each angle θ a set of ray integrals is computed by varying the position t at which the ray passes through the image. The points along a ray are given by all the solutions (x,y) to the equation

$$t = x \cos\theta + y \sin\theta$$

We may therefore compute the ideal projection integrals by the following expression known as the Radon transform

$$p(\theta, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(t - x\cos\theta - y\sin\theta)dxdy \tag{17.2}$$

where $\delta(t - x\cos\theta - y\sin\theta)$ is an impulse function that is nonzero along the projection ray.

In practice, these projection integrals may be measured using a variety of physical techniques. In transmission tomography, λ_T photons are emitted into an object under test. A detector then counts the number of photons, $\lambda(\theta,t)$, which pass through the object without being absorbed. Collimators are used to ensure the detected energy passes straight through the object along the desired path. Since the attenuation of energy as it passes through the object is exponentially related to the integral of the object's density, the projection integral may be computed from the formula

$$p(\theta,t) = -\log\left(\frac{\lambda(\theta,t)}{\lambda_T}\right)$$

In emission tomography, one wishes to measure the rate of photon emission at each pixel. In this case, various methods may be used to collect and count all the photons emitted along a ray passing through the object.

Once the projections $p(\theta,t)$ have been measured, the objective is to compute the unknown cross section $f(x,y)$. The image and projections may be related by first computing the Fourier transform of the 2-D image

$$F(\omega_x, \omega_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)e^{j(\omega_x x + \omega_y y)}dxdy$$

and the 1-D projection for each angle

$$P(\theta, \omega) = \int_{-\infty}^{\infty} p(\theta, t) e^{j\omega t} dt$$

These two transforms are then related by the Fourier slice theorem.

$$F(\omega \cos \theta, \omega \sin \theta) = P(\theta, \omega)$$

In words, $P(\theta,\omega)$ corresponds to the value of the 2-D Fourier transform $F(\omega_x, \omega_y)$ along a 1-D line at an angle of θ passing through the origin.

The Fourier slice theorem may be used to develop two methods for inverting the Radon transform and thereby computing the image f. The first method, known as filtered back projection, computes the inverse Fourier transform in polar coordinates using the transformed projection data.

$$f(x, y) = \frac{1}{2\pi} \int_0^{\pi} \int_{-\infty}^{\infty} P(\theta, \omega) |\omega| e^{j\omega(x \cos \theta + y \sin \theta)} d\omega d\theta$$

Notice that the $|\omega|$ term accounts for the integration in polar coordinates.

A second inversion method results from performing all computations in the space domain rather than first transforming to the frequency domain ω. This can be done by expressing the inner integral of filtered back projection as a convolution in the space domain.

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} P(\theta, \omega) |\omega| e^{j\omega s} d\omega = \int_{-\infty}^{\infty} p(\theta, t) h(s - t) dt$$

Here $h(t)$ is the inverse Fourier transform of $|\omega|$. This results in the inversion formula known as convolution back projection

$$f(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} p(\theta, t) h(x \cos \theta + y \sin \theta - t) dt d\theta$$

In practice, h must be a low-pass approximation to the true inverse Fourier transform of $|\omega|$. This is necessary to suppress noise in the projection data. In practice, the choice of h is the most important element in the design of the reconstruction algorithm.

Edge Detection

The ability to find gray level edge structures in images is an important image processing operation. We shall define an **edge** to be regions in the image where there is a large change in gray level over a relatively small spatial region. The process of finding edge locations in digital images is known as edge detection. Most edge detection operators, also known as edge operators, use a window operator to first enhance the edges in the image, followed by thresholding the enhanced image.

There has been a great deal of research performed in the area of edge detection. Some of the research issues include robust threshold selection, window size selection, noise response, edge linking, and the detection of edges in moving objects. While it is beyond the scope of this section to discuss these issues in detail, it is obvious that such things as threshold selection will greatly affect the performance of the edge detection algorithm. If the threshold is set too high, then many edge points will be missed; if set too low, then many "false" edge points will be obtained because of the inherent noise in the image. The investigation of the "optimal" choice of the threshold is an important research area. Selection of the particular window operation to enhance the edges of an image, as an initial step in edge detection, has recently been based on using models of the performance of the human visual system in detecting edges.

Analysis and Computer Vision

The process of extracting useful measurements from an image or sequence of images is known as *image analysis* or *computer vision*. Before analysis can be performed one must first determine pertinent features or attributes of the object in the scene and extract information about these features. The selection of which features in the image to measure must be chosen *a priori*, based on empirical results. Most features used consist of shape properties, shape change properties, shading, texture, motion, depth, and color. After the features are extracted, one must then use the feature measurements to determine scene characteristics such as object identification. In the past, simple pattern recognition algorithms, i.e., nearest-neighbor classification, have been used to compare the feature measurements of an image to a set of feature measurements that correspond to a known object. A decision is then made as to whether or not the features of the image match those of the known type.

Recently, there has been work in the application of *artificial intelligence* techniques to image analysis. These approaches are very much different from classical statistical pattern recognition in that the feature measurements are used in a different manner as part of a larger system that attempts to model the scene and then determine what is in it based on the model.

Defining Terms

Digital image: An array of numbers representing the spatial distribution of energy in a scene which is obtained by a process of sampling and quantization.

Edge: A localized region of rapid change in gray level in the image.

Entropy: A measure of the minimum amount of information required on the average to store or transmit each quantized feature vector.

Image compression or coding: The process of reducing the number of binary digits or bits required to represent the image.

Image enhancement: An image processing operation that is intended to improve the visual quality of the image or to emphasize certain features.

Image feature: An attribute of a block of image pixels.

Image reconstruction: The process of obtaining an image from nonimage data that characterizes that image.

Lossless vs. lossy compression: If the reconstructed or decoded image is identical to the original, the compression scheme is lossless. Otherwise, it is lossy.

Pixel: A single sample or picture element in the digital image which is located at specific spatial coordinates.

Point operation: An image processing operation in which individual pixels are mapped to new values irrespective of the values of any neighboring pixels.

Projection: A set of parallel line integrals across the image oriented at a particular angle.

Quantization: The process of converting from a continuous-amplitude image to an image that takes on only a finite number of different amplitude values.

Sampling: The process of converting from a continuous-parameter image to a discrete-parameter image by discretizing the spatial coordinate.

Tomography: The process of reconstructing an image from projection data.

Vector quantization: The process of replacing an exact vector of features by a prototype vector that is used to represent all feature vectors contained within a cluster.

Window operation: An image processing operation in which the new value assigned to a given pixel depends on all the pixels within a window centered at that pixel location.

Related Topics

15.1 Coding, Transmission, and Storage • 73.6 Data Compression

References

H. C. Andrews and B.R. Hunt, *Digital Image Restoration,* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

D. H. Ballard and C. M. Brown, *Computer Vision,* Englewood Cliffs, N.J.: Prentice-Hall, 1982.

H. Barrow and J. Tenenbaum, "Computational vision," *Proc. IEEE,* vol. 69, pp. 572–595, May 1981.
A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression,* Norwell, Mass.: Kluwer Academic Publishers, 1991.
R. C. Gonzalez and P. Wintz, *Digital Image Processing,* Reading, Mass.: Addison-Wesley, 1991.
G.T. Herman, *Image Reconstruction from Projections,* New York: Springer-Verlag, 1979.
T. S. Huang, *Image Sequence Analysis,* New York: Springer-Verlag, 1981.
A. K. Jain, *Fundamentals of Digital Image Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.
A. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging,* New York: IEEE Press, 1988.
A. Macovski, *Medical Imaging Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.
M. D. McFarlane, "Digital pictures fifty years ago," *Proc. IEEE,* pp. 768–770, July 1972.
W. K. Pratt, *Digital Image Processing,* New York: Wiley, 1991.
A. Rosenfeld and A. Kak, *Digital Picture Processing,* vols. 1 and 2, San Diego: Academic Press, 1982.
J. Serra, *Image Analysis and Mathematical Morphology,* vols. 1 and 2, San Diego: Academic Press, 1982 and 1988.

Further Information

A number of textbooks are available that cover the broad area of image processing and several that focus on more specialized topics within this field. The texts by Gonzalez and Wintz [1991], Jain [1989], Pratt [1991], and Rosenfeld and Kak (Vol. 1) [1982] are quite broad in their scope. Gonzalez and Wintz's treatment is written at a somewhat lower level than that of the other texts. For a more detailed treatment of computed tomography and other medical imaging modalities, the reader may consult the texts by Herman [1979], Macovski [1983], and Kak and Slaney [1988]. To explore the field of computer vision, the reader is advised to consult the text by Ballard and Brown [1982]. Current research and applications of image processing are reported in a number of journals. Of particular note are the *IEEE Transactions on Image Processing;* the *IEEE Transactions on Pattern Analysis and Machine Intelligence;* the *IEEE Transactions on Geoscience and Remote Sensing;* the *IEEE Transactions on Medical Imaging;* the *Journal of the Optical Society of America, A: Optical Engineering;* the *Journal of Electronic Imaging;* and *Computer Vision, Graphics, and Image Processing.*

17.2 Video Signal Processing

Sarah A. Rajala

Video signal processing is the area of specialization concerned with the processing of time sequences of image data, i.e., video. Because of the significant advances in computing power and increases in available transmission bandwidth, there has been a proliferation of potential applications in the area of video signal processing. Applications such as high-definition television, digital video, multimedia, video phone, interactive video, medical imaging, and information processing are the driving forces in the field today. As diverse as the applications may seem, it is possible to specify a set of fundamental principles and methods that can be used to develop the applications.

Considerable understanding of a video signal processing system can be gained by representing the system with the block diagram given in Fig. 17.10. Light from a real-world scene is captured by a **scanning system** and causes an image frame $f(x,y,t_0)$ to be formed on a focal plane. A video signal is a sequence of image frames that are created when a scanning system captures a new image frame at periodic intervals in time. In general, each frame of the video sequence is a function of two spatial variables x and y and one temporal variable t. An integral part of the scanning system is the process of converting the original analog signal into an appropriate digital representation. The conversion process includes the operations of sampling and quantization. **Sampling**

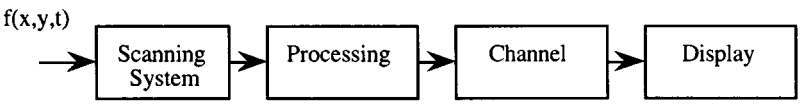


FIGURE 17.10 Video signal processing system block diagram.

is the process of converting a continuous-time/space signal into a discrete-time/space signal. **Quantization** is the process of converting a continuous-valued signal into a discrete-valued signal.

Once the video signal has been sampled and quantized, it can be processed digitally. Processing can be performed on special-purpose hardware or general-purpose computers. The type of processing performed depends on the particular application. For example, if the objective is to generate high-definition television, the processing would typically include compression and motion estimation. In fact, in most of the applications listed above these are the fundamental operations. **Compression** is the process of compactly representing the information contained in an image or video signal. **Motion estimation** is the process of estimating the displacement of the moving objects in a video sequence. The displacement information can then be used to interpolate missing frame data or to improve the performance of compression algorithms.

After the processing is complete, a video signal is ready for transmission over some channel or storage on some medium. If the signal is transmitted, the type of channel will vary depending on the application. For example, today analog television signals are transmitted one of three ways: via satellite, terrestrially, or by cable. All three channels have limited transmission bandwidths and can adversely affect the signals because of the imperfect frequency responses of the channels. Alternatively, with a digital channel, the primary limitation will be the bandwidth.

The final stage of the block diagram shown in Fig. 17.10 is the display. Of critical importance at this stage is the human observer. Understanding how humans respond to visual stimuli, i.e., the psychophysics of vision, will not only allow for better evaluation of the processed video signals but will also permit the design of better systems.

Sampling

If a continuous-time video signal satisfies certain conditions, it can be exactly represented by and be reconstructed from its sample values. The conditions which must be satisfied are specified in the *sampling theorem.* The sampling theorem can be stated as follows:

Sampling Theorem:

Let $f(x,y,t)$ be a bandlimited signal with $F(\omega_x,\omega_y,\omega_t) = 0$ for $|\omega_x| > \omega_{xM}$, $|\omega_y| > \omega_{yM}$, and $|\omega_t| > \omega_{tM}$. Then $f(x,y,t)$ is uniquely determined by its samples $f(jX_S, kY_S, lT_S) = f(j,k,l)$, where $j,k,l = 0, \pm 1, \pm 2, \ldots$ if

$$\omega_{sx} > 2\omega_{xM},\ \omega_{sy} > 2\omega_{yM},\ \text{and}\ \omega_{st} > 2\omega_{tM}$$

and

$$\omega_{sx} = 2\pi/X_S,\ \omega_{sy} = 2\pi/Y_S,\ \text{and}\ \omega_{st} = 2\pi/T_S$$

X_S is the sampling period along the x direction, $\omega_x = 2\pi/X_S$ is the spatial sampling frequency along the x direction, Y_S is the sampling period along the y direction, $\omega_y = 2\pi/Y_S$ is the spatial sampling frequency along the y direction, T_S is the sampling period along the temporal direction, and $\omega_t = 2\pi/T_S$ is the temporal sampling frequency.

Given these samples, $f(x,y,t)$ can be reconstructed by generating a periodic impulse train in which successive impulses have amplitudes that are successive sample values. This impulse train is then processed through an ideal low-pass filter with appropriate gain and cut-off frequencies. The resulting output signal will be exactly equal to $f(x,y,t)$. (*Source:* Oppenheim et al., 1983, p. 519.)

If the sampling theorem is not satisfied, **aliasing** will occur. Aliasing occurs when the signal is undersampled and therefore no longer recoverable by low-pass filtering. Figure 17.11(a) shows the frequency spectrum of a sampled bandlimited signal with no aliasing. Figure 17.11(b) shows the frequency response of the same signal with aliasing. The aliasing occurs at the points where there is overlap in the diamond-shaped regions. For video signals aliasing in the temporal direction will give rise to flicker on the display. For television systems, the standard temporal sampling rate is 30 frames per second in the United States and Japan and 25 frames per second in Europe. However, these rates would be insufficient without the use of interlacing.

If the sampling rate (spatial and/or temporal) of a system is fixed, a standard approach for minimizing the effects of aliasing for signals that do not satisfy the sampling theorem is to use a presampling filter. Presampling

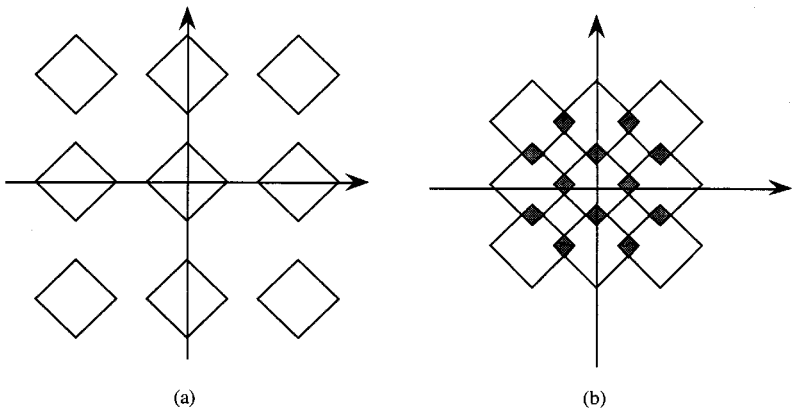


FIGURE 17.11 (a) Frequency spectrum of a sampled signal with no aliasing. (b) Frequency spectrum of a sampled signal with aliasing.

filters are low-pass filters whose cut-off frequencies are chosen to be less than $\omega_M, \omega_M, \omega_M$. Although the signal will still not be able to be reconstructed exactly, the degradations are less annoying. Another problem in a real system is the need for an ideal low-pass filter to reconstruct an analog signal. An ideal filter is not physically realizable, so in practice an approximation must be made. Several very simple filter structures are common in video systems: sample and hold, bilinear, and raised cosine.

Quantization

Quantization is the process of converting the continuous-valued amplitude of the video signal into a discrete-valued representation, i.e., a finite set of numbers. The output of the quantizer is characterized by quantities that are limited to a finite number of values. The process is a many-to-one mapping, and thus there is a loss of information. The quantized signal can be modeled as

$$f_q(j,k,l) = f(j,k,l) - e(j,k,l)$$

where $f_q(j,k,l)$ is the quantized video signal and $e(j,k,l)$ is the quantization noise. If too few bits per sample are used, the quantization noise will produce visible false contours in the image data.

The quantizer is a mapping operation which generally takes the form of a staircase function (see Fig. 17.12). A rule for quantization can be defined as follows: Let $\{d_k, k = 1, 2, \ldots, N+1\}$ be the set of decision levels with d_1 the minimum amplitude value and d_N the maximum amplitude value of $f(j,k,l)$. If $f(j,k,l)$ is contained in the interval (d_k, d_{k+1}), then it is mapped to the kth reconstruction level r. Methods for designing quantizers can be broken into two categories: uniform and nonuniform. The input-output function for a typical uniform quantizer is shown in Fig. 17.12. The mean square value of the quantizing noise can be easily calculated if it is assumed that the amplitude probability distribution is constant within each quantization step. The quantization step size for a uniform quantizer is

$$q = \frac{d_{N+1} - d_1}{N}$$

and all errors between $q/2$ and $-q/2$ are equally likely. The mean square quantization error is given by:

$$\left\langle e^2(j,k,l) \right\rangle = \int_{-q/2}^{q/2} \frac{f^2}{q}\, df = \frac{q^2}{12}$$

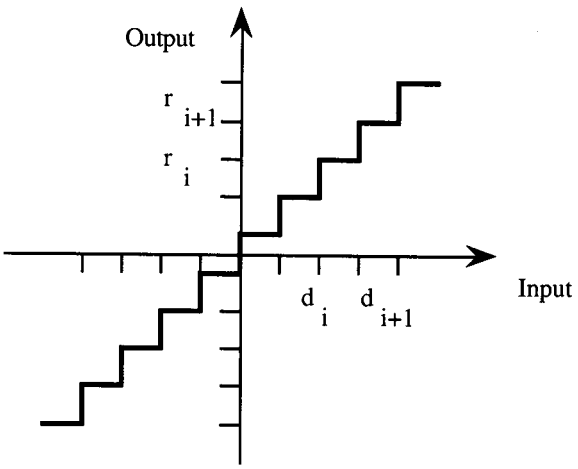


FIGURE 17.12 Characteristics of a uniform quantizer.

If one takes into account the exact amplitude probability distribution, an optimal quantizer can be designed. Here the objective is to choose a set of decision levels and reconstruction levels that will yield the minimum quantization error. If f has a probability density function $p_f(f)$, the mean square quantization error is

$$\left\langle e^2(j,k,l)\right\rangle = \sum_{i=1}^{N} \int_{di}^{di+1} (f - r_i)^2 p_f(f) df$$

where N is the number of quantization levels. To minimize, the mean square quantization error is differentiated with respect to d_i and r_i. This results in the Max quantizer:

$$d_i = \frac{r_i + r_{i-1}}{2}$$

and

$$r_i = \frac{\int_{di}^{di+1} f p_f(f) df}{\int_{di}^{di+1} p_f(f) df}$$

Thus, the quantization levels need to be midway between the reconstruction levels, and the reconstruction levels are at the centroid of that portion of $p_f(f)$ between d_i and f_{d+1}. Unfortunately these requirements do not lead to an easy solution. Max used an iterative numerical technique to obtain solutions for various quantization levels assuming a zero-mean Gaussian input signal. These results and the quantization levels for other standard amplitude distributions can be found in Jain [1989].

A more common and less computationally intense approach to nonuniform quantization is to use a compandor (compressor–expander). The input signal is passed through a nonlinear compressor before being quantized uniformly. The output of the quantizer must then be expanded to the original dynamic range (see Fig. 17.13). The compression and expansion functions can be determined so that the compandor approximates a Max quantizer.

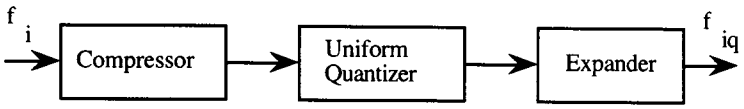


FIGURE 17.13 Nonuniform quantization using a compandor.

Vector Quantization

Quantization does not have to be done on a single pixel at a time. In fact, better results can be achieved if the video data are quantized on a vector (block) basis. In vector quantization, the image data are first processed into a set of vectors. A code book (set of code words or templates) that best matches the data to be quantized is then generated. Each input vector is then quantized to the closest code word. Compression is achieved by transmitting only the indices for the code words. At the receiver, the images are reconstructed using a table look-up procedure. Two areas of ongoing research are finding better methods for designing the code books and developing better search and update techniques for matching the input vectors to the code words.

Video Compression

Digital representations of video signals typically require a very large number of bits. If the video signal is to be transmitted and/or stored, compression is often required. Applications include conventional and high-definition television, video phone, video conferencing, multi-media, remote-sensed imaging, and magnetic resonance imaging. The objective of compression (source encoding) is to find a representation that maximizes picture quality while minimizing the data per picture element (pixel). A wealth of compression algorithms have been developed during the past 30 years for both image and video compression. However, the ultimate choice of an appropriate algorithm is application dependent. The following summary will provide some guidance in that selection process.

Compression algorithms can be divided into two major categories: information-preserving, or lossless, and lossy techniques. Information-preserving techniques introduce no errors in the encoding/decoding process; thus, the original signal can be reconstructed exactly. Unfortunately, the achievable compression rate, i.e., the reduction in bit rate, is quite small, typically on the order of 3:1. On the other hand, lossy techniques introduce errors in the coding/decoding process; thus, the received signal cannot be reconstructed exactly. The advantage of the lossy techniques is the ability to achieve much higher compression ratios. The limiting factor on the compression ratio is the required quality of the video signal in a specific application.

One approach to compression is to reduce the spatial and/or temporal sampling rate and the number of quantization levels. Unfortunately, if the sampling is too low and the quantization too coarse, aliasing, contouring, and flickering will occur. These distortions are often much greater than the distortions introduced by more sophisticated techniques at the same compression rate. Compression systems can generally be modeled by the block diagram shown in Fig. 17.14. The first stage of the compression system is the mapper. This is an operation in which the input pixels are mapped into a representation that can be more effectively encoded. This stage is generally reversible. The second stage is the quantizer and performs the same type of operation as described earlier. This stage is not reversible. The final stage attempts to remove any remaining statistical redundancy. This stage is reversible and is typically achieved with one of the information-preserving coders.

Information-Preserving Coders

The data rate required for an original digital video signal may not represent its average information rate. If the original signal is represented by M possible independent symbols with probabilities p_i, $i = 0, 1,..., M-1$, then the information rate as given by the first-order entropy of the signal H is

$$H = -\sum_{i=1}^{M-1} p_i \log_2 p_i \text{ bits per sample}$$

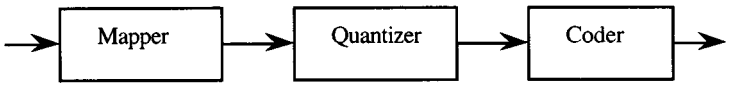


FIGURE 17.14 Three-stage model of an encoder.

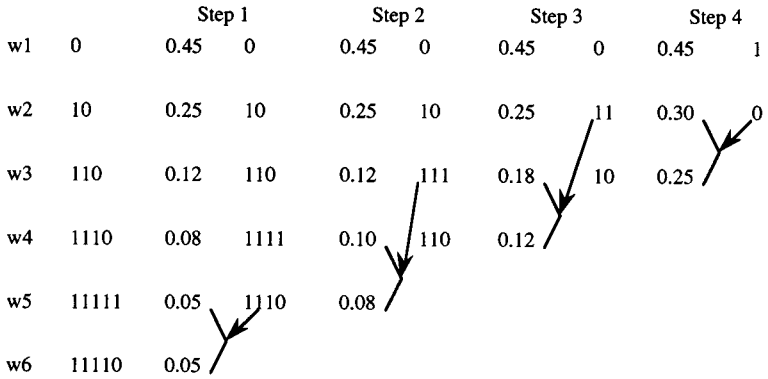


FIGURE 17.15 An example of constructing a Huffman code.

According to Shannon's coding theorem [see Jain, 1989], it is possible to perform lossless coding of a source with entropy H bits per symbol using $H + \varepsilon$ bits per symbol. ε is a small positive quantity. The maximum obtainable compression rate C is then given by:

$$C = \frac{\text{average bit rate of the original data}}{\text{average bit rate of the encoded data}}$$

Huffman Coding

One of the most efficient information-preserving (entropy) coding methods is Huffman coding. Construction of a Huffman code involves arranging the symbol probabilities in decreasing order and considering them as leaf nodes of a tree. The tree is constructed by merging the two nodes with the smallest probability to form a new node. The probability of the new node is the sum of the two merged nodes. This process is continued until only two nodes remain. At this point, 1 and 0 are arbitrarily assigned to the two remaining nodes. The process now moves down the tree, decomposing probabilities and assigning 1's and 0's to each new pair. The process continues until all symbols have been assigned a code word (string of 1's and 0's). An example is given in Fig. 17.15. Many other types of information-preserving compression schemes exist (see, for example, Gonzalez and Wintz [1987]), including arithmetic coding, Lempel-Ziv algorithm, shift coding, and run-length coding.

Predictive Coding

Traditionally one of the most popular methods for reducing the bit rate has been predictive coding. In this class, differential pulse-code modulation (DPCM) has been used extensively. A block diagram for a basic DPCM system is shown in Fig. 17.16. In such a system the difference between the current pixel and a predicted version of that pixel gets quantized, coded, and transmitted to the receiver. This difference is referred to as the prediction error and is given by

$$e_i = f_i - \hat{f}_i$$

The prediction is based on previously transmitted and decoded spatial and/or temporal information and can be linear or nonlinear, fixed or adaptive. The difference signal e_i is then passed through a quantizer. The signal

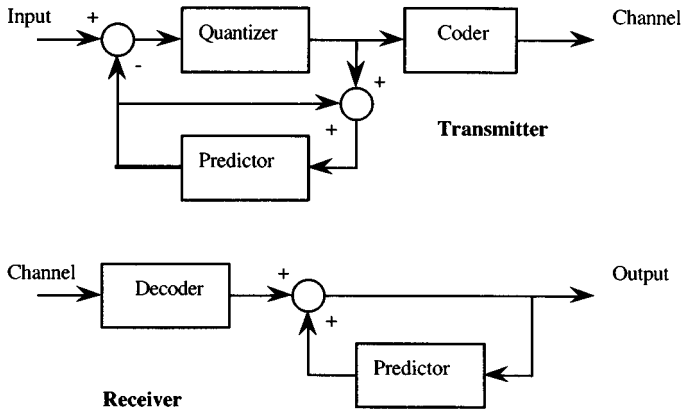


FIGURE 17.16 Block diagram of a basic DPCM system.

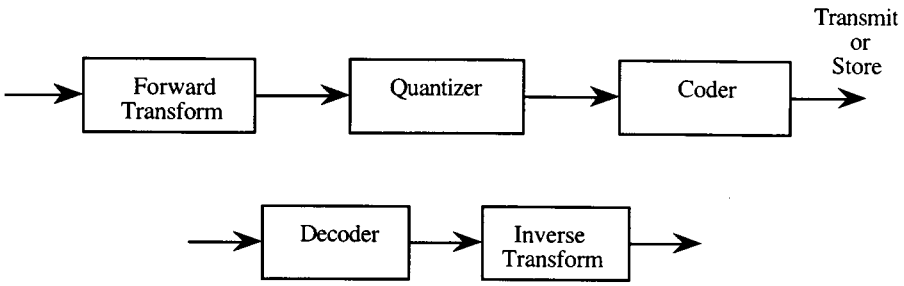


FIGURE 17.17 Transform coding system.

at the output of the quantizer is the quantized prediction error e_{iq}, which is entropy encoded transmission. The first step at the receiver is to decode the quantized prediction error. After decoding, d_{iq} is added to the predicted value of the current pixel $\hat{f}_i$ to yield the reconstructed pixel value. Note that as long as a quantizer is included in the system, the output signal will not exactly equal the input signal.

The predictors can include pixels from the present frame as well as those from previous frames (see Fig. 17.17). If the motion and the spatial detail are not too high, frame (or field) prediction works well. If the motion is high and/or the spatial detail is high, intrafield prediction generally works better. A primary reason is that there is less correlation between frames and fields when the motion is high.

For more information on predictive coding, see Musmann et al. [1985] or Jain [1989].

Motion-Compensated Predictive Coding

Significant improvements in image quality, at a fixed compression rate, can be obtained when adaptive prediction algorithms take into account the frame-to-frame displacement of moving objects in the sequence. Alternatively, one could increase the compression rate for a fixed level of image quality. The amount of increase in performance will depend on one's ability to estimate the motion in the scene. Techniques for estimating the motion are described in a later subsection.

Motion-compensated prediction algorithms can be divided into two categories. One category estimates the motion on a block-by-block basis and the other estimates the motion one pixel at a time. For the block-based methods an estimate of the displacement is obtained for each block in the image. The block matching is achieved by finding the maximum correlation between a block in the current frame and a somewhat larger search area in the previous frame. A number of researchers have proposed ways to reduce the computational complexity,

including using a simple matching criterion and using logarithmic searches for finding the peak value of the correlation.

The second category obtains a displacement estimate at each pixel in a frame. These techniques are referred to as pel recursive methods. They tend to provide more accurate estimates of the displacement but at the expense of higher complexity. Both categories of techniques have been applied to video data; however, block matching is used more often in real systems. The primary reason is that more efficient implementations have been feasible. It should be noted, however, that every pixel in a block will be assigned the same displacement estimate. Thus, the larger the block size the greater the potential for errors in the displacement estimate for a given pixel. More details can be found in Musmann et al. [1985].

Transform Coding

In transform coding, the video signal $f(x,y,t)$ is subjected to an invertible transform, then quantized and encoded (see Fig. 17.17). The purpose of the transformation is to convert statistically dependent picture elements into a set of statistically independent coefficients. In practice, one of the separable fast transforms in the class of unitary transforms is used, e.g., cosine, Fourier, or Hadamard. In general, the transform coding algorithms can be implemented in 2-D or 3-D. However, because of the real-time constraints of many video signal processing applications, it is typically more efficient to combine a 2-D transform with a predictive algorithm in the temporal direction, e.g., motion compensation.

For 2-D transform coding the image data are first subdivided into blocks. Typical block sizes are 8×8 or 16×16. The transform independently maps each image block into a block of transform coefficients; thus, the processing of each block can be done in parallel. At this stage the data have been mapped into a new representation, but no compression has occurred. In fact, with the Fourier transform there is an expansion in the amount of data. This occurs because the transform generates coefficients that are complex-valued. To achieve compression the transform coefficients must be quantized and then coded to remove any remaining redundancy.

Two important issues in transform coding are the choice of transformation and the allocation of bits in the quantizer. The most commonly used transform is the discrete cosine transform (DCT). In fact, many of the proposed image and video standards utilize the DCT. The reasons for choosing a DCT include: its performance is superior to the other fast transforms and is very close to the optimal Karhunen-Loeve transform, it produces real-valued transform coefficients, and it has good symmetry properties, thus reducing the blocking artifacts inherent in block-based algorithms. One way to reduce these artifacts is by using a transform whose basis functions are even, i.e., the DCT, and another is to use overlapping blocks. For bit allocation, one can determine the variance of the transform coefficients and then assign the bits so the distortion is minimized. An example of a typical bit allocation map is shown in Fig. 17.18.

8	7	6	5	3	3	4	4	4	1	1	1	1	1	0	0
7	6	5	4	3	3	2	2	1	1	1	1	1	0	0	0
6	5	4	3	3	2	2	2	1	1	1	1	1	0	0	0
5	4	3	3	3	2	2	2	1	1	1	1	1	0	0	0
3	3	3	3	2	2	2	1	1	1	1	1	0	0	0	0
3	3	2	2	2	2	2	1	1	1	1	1	0	0	0	0
2	2	2	2	2	2	1	1	1	1	1	0	0	0	0	0
2	2	2	2	1	1	1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 17.18 A typical bit allocation for 16×16 block coding of an image using the DCT.

Subband Coding

Recently, subband coding has proved to be an effective technique for image compression. Here, the original video signal is filtered into a set of bandpass signals (subbands), each sampled at successively lower rates. This process is known as the subband analysis stage. Each of the bandpass images is then quantized and encoded for transmission/storage. At the receiver, the signals must be decoded and then an image reconstructed from the subbands. The process at the receiver is referred to as the subband synthesis stage. A one-level subband

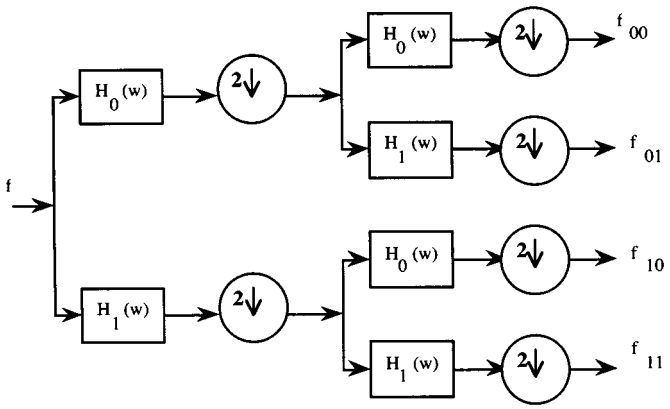


FIGURE 17.19 A two-dimensional subband analysis system for generating four equal subbands.

analysis results in 4 subbands and a 2-level analysis in 16 equal subbands or 7 unequal subbands. A block diagram for a separable two-dimensional subband analysis system is shown in Fig. 17.19.

HDTV

High-definition television (HDTV) has received much attention in the past few years. With the recent push for all digital implementations of HDTV, the need for video signal processing techniques has become more obvious. In order for the digital HDTV signal to fit in the transmission bandwidth, there is a need for a compression ratio of approximately 10:1, with little or no degradation introduced. The goal of HDTV is to produce high-quality video signals by enhancing the detail, improving the aspect ratio and the viewing distance. The detail is enhanced by increasing the video bandwidth. The proposed aspect ratio of 16/9 will allow for a wide-screen format which is more consistent with the formats used in the motion-picture industry. The eye's ability to resolve fine detail is limited. To achieve full resolution of the detail, the HDTV image should be viewed at a distance of approximately three times the picture height. To accommodate typical home viewing environments, larger displays are needed.

Motion Estimation Techniques

Frame-to-frame changes in luminance are generated when objects move in video sequences. The luminance changes can be used to estimate the displacement of the moving objects if an appropriate model of the motion is specified. A variety of motion models have been developed for dynamic scene analysis in machine vision and for video communications applications. In fact, motion estimates were first used as a control mechanism for the efficient coding of a sequence of images in an effort to reduce the temporal redundancy. Motion estimation algorithms can be classified in two broad categories: gradient or differential-based methods and token matching or correspondence methods. The gradient methods can be further divided into pel recursive, block matching, and optical flow methods.

Pel Recursive Methods

Netravali and Robbins [1979] developed the first pel recursive method for television signal compression. The algorithm begins with an initial estimate of the displacement, then iterates recursively to update the estimate. The iterations can be performed at a single pixel or at successive pixels along a scan line. The true displacement **D** at each pixel is estimated by

$$\hat{\mathbf{D}}^i = \hat{\mathbf{D}}^{i-1} + \mathbf{U}^i$$

where $\hat{\mathbf{D}}^i$ is the displacement estimate at the ith iteration and $\mathbf{U}^i$ is the update term. $\mathbf{U}^i$ is an estimate of $\mathbf{D} - \hat{\mathbf{D}}^{i-1}$. They then used the displaced frame difference (DFD):

$$DFD(x, y, \hat{\mathbf{D}}^{i-1}) = I(x, y, t) - I(x - \hat{\mathbf{D}}^{i-1}, t - T_S)$$

to obtain a relationship for the update term $\mathbf{U}^i$. In the previous equation, T_S is the temporal sample spacing. If the displacement estimate is updated from sample to sample using a steepest-descent algorithm to minimize the weighted sum of the squared displaced frame differences over a neighborhood, then $\hat{\mathbf{D}}^i$ becomes

$$\hat{\mathbf{D}}^i = \hat{\mathbf{D}}^{i-1} - \frac{\varepsilon}{2} \nabla \hat{\mathbf{D}}^i \left[\sum_j W_j [DFD(x_{k-j}, \hat{\mathbf{D}}^{i-1})]^2 \right]$$

where $W_j \geq 0$ and

$$\sum_j W_j = 1$$

A graphical representation of pel recursive motion estimation is shown in Fig. 17.20.

A variety of methods to calculate the update term have been reported. The advantage of one method over another is mainly in the improvement in compression. It should be noted that pel recursive algorithms assume that the displacement to be estimated is small. If the displacement is large, the estimates will be poor. Noise can also affect the accuracy of the estimate.

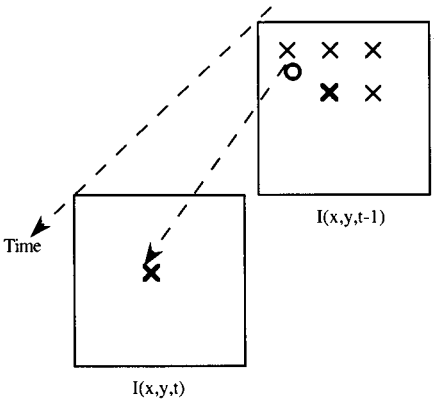


FIGURE 17.20 A graphical illustration of pel recursive motion estimation. The distance between the x and o pixels in the frame at $t - 1$ is $\hat{\mathbf{D}}^{i-1}$.

Block Matching

Block matching methods estimate the displacement within an $M \times N$ block in an image frame. The estimate is determined by finding the best match between the $M \times N$ block in a frame at time t and its best match from frame at $t - T_S$. An underlying assumption in the block matching techniques is that each pixel within a block has the same displacement. A general block matching algorithm is given as follows:

1. Segment the image frame at time t into a fixed number of blocks of size $M \times N$.
2. Specify the size of the search area in the frame at time $t - 1$. This depends on the maximum expected displacement. If $D_{\max}$ is the maximum displacement in either the horizontal or vertical direction, then the size of the search area, SA, is

$$\text{SA} = (M + 2D_{\max}) \times (N + 2D_{\max})$$

   Figure 17.21 illustrates the search area in the frame at time $t - 1$ for an $M \times N$ block at time t.
3. Using an appropriately defined matching criterion, e.g., mean-squared error or sum of absolute difference, find the best match for the $M \times N$ block.
4. Proceed to the next block in frame t and repeat step 3 until displacement estimates have been determined for all blocks in the image.

Optical Flow Methods

The optical flow is defined as the apparent motion of the brightness patterns from one frame to the next. The optical flow is an estimate of the velocity field and hence requires two equations to solve for it. Typically a

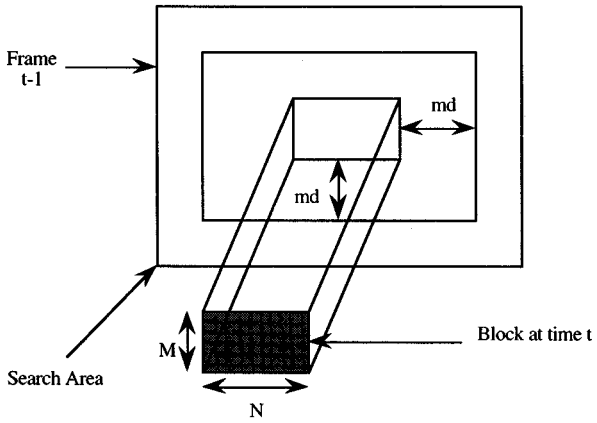


FIGURE 17.21 An illustration of block matching.

constraint is imposed on the motion model to provide the necessary equations. Optical flow can give useful information about the spatial arrangement of the objects in a scene, as well as the rate of change of those objects. Horn [1986] also defines a motion field, which is a two-dimensional velocity field resulting from the projection of the three-dimensional velocity field of an object in the scene onto the image plane. The motion field and the optical flow are not the same.

In general, the optical flow has been found difficult to compute because of the algorithm sensitivity to noise. Also, the estimates may not be accurate at scene discontinuities. However, because of its importance in assigning a velocity vector at each pixel, there continues to be research in the field.

The optical flow equation is based on the assumption that the brightness of a pixel at location (x,y) is constant over time; thus,

$$I_x \frac{dx}{dt} + I_y \frac{dy}{dt} + I_t = 0$$

where dx/dt and dy/dt are the components of the optical flow. Several different constraints have been used with the optical flow equation to solve for dx/dt and dy/dt. A common constraint to impose is that the velocity field is smooth.

Token Matching Methods

Token matching methods are often referred to as discrete methods since the goal is to estimate the motion only at distinct image features (tokens). The result is a sparse velocity field. The algorithms attempt to match the set of discrete features in the frame at time $t-1$ with a set that best resembles them in the frame at time t. Most of the algorithms in this group assume that the estimation will be achieved in a two-step process. In the first step, the features are identified. The features could be points, corners, centers of mass, lines, or edges. This step typically requires segmentation and/or feature extraction. The second step determines the various velocity parameters. The velocity parameters include a translation component, a rotation component, and the rotation axis. The token matching algorithms fail if there are no distinct features to use.

All of the methods described in this subsection assume that the intensity at a given pixel location is reasonably constant over time. In addition, the gradient methods assume that the size of the displacements is small. Block matching algorithms have been used extensively in real systems, because the computational complexity is not too great. The one disadvantage is that there is only one displacement estimate per block. To date, optical flow algorithms have found limited use because of their sensitivity to noise. Token matching methods work well for applications in which the features are well defined and easily extracted. They are probably not suitable for most video communications applications.

TABLE 17.2 Quality and Impairment Ratings

5 Excellent	5 Imperceptible	3 Much better
4 Good	4 Perceptible but not annoying	2 Better
3 Fair	3 Slightly annoying	1 Slightly better
2 Poor	2 Annoying	0 Same
1 Bad	1 Very annoying	–1 Slightly worse
		–2 Worse
		–3 Much worse

Image Quality and Visual Perception

An important factor in designing video signal processing algorithms is that the final receiver of the video information is typically a human observer. This has an impact on how the quality of the final signal is assessed and how the processing should be performed. If our objective is video transmission over a limited bandwidth channel, we do not want to waste unnecessary bits on information that cannot be seen by the human observer. In addition, it is undesirable to introduce artifacts that are particularly annoying to the human viewer. Unfortunately, there are no perfect quantitative measures of visual perception. The human visual system is quite complicated. In spite of the advances that have been made, no complete model of human perception exists. Therefore, we often have to rely on subjective testing to evaluate picture quality. Although no comprehensive model of human vision exists, certain functions can be characterized and then used in designing improved solutions. For more information, see Netravali and Haskell [1988].

Subjective Quality Ratings

There are two primary categories of subjective testing: *category-judgment* (*rating-scale*) methods and *comparison* methods. Category-judgment methods ask the subjects to view a sequence of pictures and assign each picture (video sequence) to one of several categories. Categories may be based on overall quality or on visibility of impairment (see Table 17.2).

Comparison methods require the subjects to compare a distorted test picture with a reference picture. Distortion is added to the test picture until both pictures appear of the same quality to the subject. Viewing conditions can have a great impact on the results of such tests. Care must be taken in the experimental design to avoid biases in the results.

Visual Perception

In this subsection, a review of the major aspects of human psychophysics that have an impact in video signal processing is given. The phenomena of interest include light adaptation, visual thresholding and contrast sensitivity, masking, and temporal phenomena.

Light Adaptation

The human visual system (HVS) has two major classes of photoreceptors, the rods and the cones. Because these two types of receptors adapt to light differently, two different adaptation time constants exist. Furthermore, these receptors respond at different rates going from dark to light than from light to dark. It should also be noted that although the HVS has an ability to adapt to an enormous range of light intensity levels, on the order of 10^{10} in millilamberts, it does so adaptively. The simultaneous range is on the order of 10^3.

Visual Thresholding and Contrast Sensitivity

Determining how sensitive an observer is to small changes in luminance is important in the design of video systems. One's sensitivity will determine how visible noise will be and how accurately the luminance must be represented. The contrast sensitivity is determined by measuring the just-noticeable difference (JND) as a

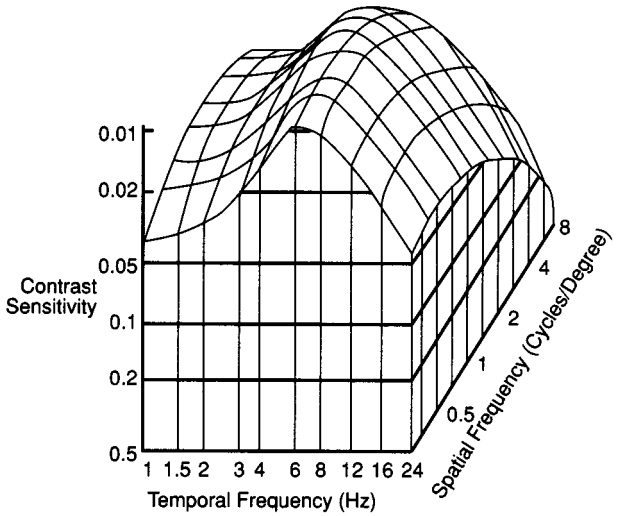


FIGURE 17.22 A perspective view of the spatio-temporal threshold surface.

function of the brightness. The JND is the amount of additional brightness needed to distinguish a patch from the background. It is a visibility threshold. What is significant is that the JND is dependent on the background and surrounding luminances, the size of the background and surrounding areas, and the size of the patch, with the primary dependence being on the luminance of the background.

Masking

The response to visual stimuli is greatly affected by what other visual stimuli are in the immediate neighborhood (spatially and temporally). An example is the reduced sensitivity of the HVS to noise in areas of high spatial activity. Another example is the masking of details in a new scene by what was present in the previous scene. In both cases, the masking phenomenon can be used to improve the quality of image compression systems.

Temporal Effects

One relevant temporal phenomenon is the flicker fusion frequency. This is a temporal threshold which determines the point at which the HVS fuses the motion in a sequence of frames. Unfortunately this frequency varies as a function of the average luminance. The HVS is more sensitive to flicker at high luminances than at low luminances. The spatial-temporal frequency response of the HVS is important in determining the sensitivity to small-amplitude stimuli. In both the temporal and spatial directions, the HVS responds as a bandpass filter (see Fig. 17.22). Also significant is the fact that the spatial and temporal properties are not independent of one another, especially at low frequencies.

For more details on image quality and visual perception see Schreiber [1991] and Netravali and Haskell [1988].

Defining Terms

Aliasing: Distortion introduced in a digital signal when it is undersampled.
Compression: Process of compactly representing the information contained in a signal.
Motion estimation: Process of estimating the displacement of moving objects in a scene.
Quantization: Process of converting a continuous-valued signal into a discrete-valued signal.
Sampling: Process of converting a continuous-time/space signal into a discrete-time/space signal.
Scanning system: System used to capture a new image at periodic intervals in time and to convert the image into a digital representation.

Related Topics

8.5 Sampled Data • 15.1 Coding, Transmission, and Storage

References

R. C. Gonzalez and P. Wintz, *Digital Image Processing,* Reading, Mass.: Addison-Wesley, 1987.

R. A. Haddad and T. W. Parsons, *Digital Signal Processing: Theory, Applications, and Hardware,* New York: Computer Science Press, 1991.

B. P. Horn, *Robot Vision,* Cambridge, Mass.: The MIT Press, 1986.

A. K. Jain, *Fundamentals of Digital Image Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

N. Jayant, "Signal compression: Technology targets and research directions," *IEEE Journal on Selected Areas in Communications,* vol. 10, no. 5, pp. 796–818, 1992.

H. G. Musmann, P. Pirsch, and H.-J. Grallert, "Advances in picture coding," *Proc. IEEE,* vol. 73, no. 4, pp. 523–548, 1985.

A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression,* New York: Plenum Press, 1988.

A. N. Netravali and J. D. Robbins, "Motion-compensated television coding: Part I," *Bell Syst. Tech. J.*, vol. 58, no. 3, pp. 631–670, 1979.

A. V. Oppenheim, A. S. Willsky, and I. T. Young, *Signals and Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.

W. F. Schreiber, *Fundamentals of Electronic Imaging Systems,* Berlin: Springer-Verlag, 1991.

Further Information

Other recommended sources of information include *IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing,* and the *Proceedings of the IEEE,* April 1985, vol. 73, and *Multidimensional Systems and Signal Processing Journal,* 1992, vol. 3.

17.3 Sensor Array Processing

N. K. Bose and L. H. Sibul

Multidimensional signal processing tools apply to aperture and sensor array processing. Planar sensor arrays can be considered to be sampled apertures. Three-dimensional or volumetric arrays can be viewed as multidimensional spatial filters. Therefore, the topics of sensor array processing, aperture processing, and multidimensional signal processing can be studied under a unified format. The basic function of the receiving array is transduction of propagating waves in the medium into electrical signals. Propagating waves are fundamental in radar, communication, optics, sonar, and geophysics. In electromagnetic applications, basic transducers are antennas and arrays of antennas. A large body of literature that exists on antennas and antenna arrays can be exploited in the areas of aperture and sensor array processing. Much of the antenna literature deals with transmitting antennas and their radiation patterns. Because of the reciprocity of transmitting and receiving transducers, key results that have been developed for transmitters can be used for analysis of receiver aperture and/or array processing. Transmitting transducers radiate energy in desired directions, whereas receiving apertures/arrays act as spatial filters that emphasize signals from a desired look direction while discriminating against interferences from other directions. The spatial filter **wavenumber** response is called the receiver beam pattern. Transmitting apertures are characterized by their radiation patterns.

Conventional beamforming deals with the design of fixed beam patterns for given specifications. Optimum beamforming is the design of beam patterns to meet a specified optimization criterion. It can be compared to optimum filtering, detection, and estimation. Adaptive **beamformers** sense their operating environment (for example, noise covariance matrix) and adjust beamformer parameters so that their performance is optimized [Monzingo and Miller, 1980]. Adaptive beamformers can be compared with adaptive filters.

Multidimensional signal processing techniques have found wide application in seismology—where a group of identical seismometers, called seismic arrays, are used for event location, studies of the earth's sedimentation structure, and separation of coherent signals from noise, which sometimes may also propagate coherently across the array but with different horizontal velocities—by employing **velocity filtering** [Claerbout, 1976]. Velocity filtering is performed by multidimensional filters and allows also for the enhancement of signals which may occupy the same wavenumber range as noise or undesired signals do. In a broader context, beamforming can be used to separate signals received by sensor arrays based on frequency, wavenumber, and velocity (speed as well as direction) of propagation. Both the transfer and unit impulse-response functions of a velocity filter are two-dimensional functions in the case of one-dimensional arrays. The transfer function involves frequency and wavenumber (due to spatial sampling by equally spaced sensors) as independent variables, whereas the unit impulse response depends upon time and location within the array. Two-dimensional filtering is not limited to velocity filtering by means of seismic array. Two-dimensional spatial filters are frequently used, for example, in the interpretation of gravity and magnetic maps to differentiate between regional and local features. Input data for these filters may be observations in the survey of an area conducted over a planar grid over the earth's surface. Two-dimensional wavenumber digital filtering principles are useful for this purpose. Velocity filtering by means of two-dimensional arrays may be accomplished by properly shaping a three-dimensional response function $H(k_1,k_2,\omega)$. Velocity filtering by three-dimensional arrays may be accomplished through a four-dimensional function $H(k_1,k_2,k_3,\omega)$ as explained in the following subsection.

Spatial Arrays, Beamformers, and FIR Filters

A propagating plane wave, $s(\mathbf{x},t)$, is, in general, a function of the three-dimensional space variables and the time variable $(x_1, x_2, x_3) \Delta = \mathbf{x}$ and the time variable t. The 4-D Fourier transform of the stationary signal $s(\mathbf{x},t)$ is

$$S(\mathbf{k},\omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(\mathbf{x},t)e^{-j(\omega t-\sum_{i=1}^{3}k_i x_i)}dx_1dx_2dx_3dt \tag{17.3}$$

which is referred to as the wavenumber–frequency spectrum of $s(\mathbf{x},t)$, and $(k_1,k_2,k_3) \underline{\underline{\Delta}}\ \mathbf{k}$ denotes the wavenumber variables in radians per unit distance and ω is the frequency variable in radians per second. If c denotes the velocity of propagation of the plane wave, the following constraint must be satisfied

$$k_1^2 + k_2^2 + k_3^2 = \frac{\omega^2}{c^2}$$

If the 4-D Fourier transform of the unit impulse response $h(\mathbf{x},t)$ of a 4-D linear shift-invariant (LSI) filter is denoted by $H(k,\omega)$, then the response $y(\mathbf{x},t)$ of the filter to $s(\mathbf{x},t)$ is the 4-D linear convolution of $h(\mathbf{x},t)$ and $s(\mathbf{x},t)$, which is, uniquely, characterized by its 4-D Fourier transform

$$Y(\mathbf{k},\omega) = H(\mathbf{k},\omega)S(\mathbf{k},\omega) \tag{17.4}$$

The inverse 4-D Fourier transform, which forms a 4-D Fourier transform pair with Eq. (17.3), is

$$s(\mathbf{x},t) = \frac{1}{(2\pi)^4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(\mathbf{k},\omega)e^{j(\omega t-\sum_{i=1}^{3}k_i x_i)}dk_1dk_2dk_3d\omega \tag{17.5}$$

It is noted that $S(\mathbf{k},\omega)$ in Eq. (17.3) is product separable, i.e., expressible in the form

$$S(\mathbf{k},\omega) = S_1(k_1)\,S_2(k_2)\,S_3(k_3)\,S_4(\omega) \tag{17.6}$$

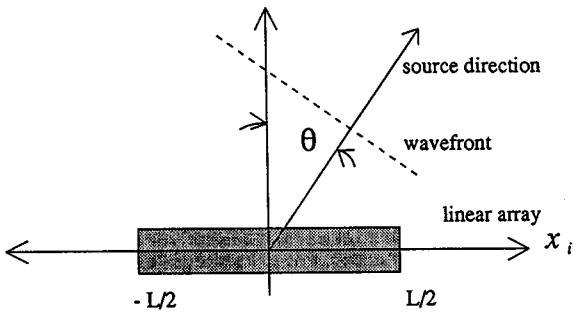


FIGURE 17.23 Uniformly weighted linear array.

where each function on the right-hand side is a univariate function of the respective independent variable, if and only if $s(\mathbf{x},t)$ in Eq. (17.3) is also product separable. In beamforming, $S_i(k_i)$ in Eq. (17.6) would be the far-field beam pattern of a linear array along the x_i-axis. For example, the normalized beam pattern of a uniformly weighted (shaded) linear array of length L is

$$S(k,\theta) = \frac{\sin\left(\frac{kL\sin\theta}{2}\right)}{\left(\frac{kL}{2}\sin\theta\right)}$$

where $\lambda = (2\pi/k)$ is the wavelength of the propagating plane wave and θ is the angle of arrival at array site as shown in Fig. 17.23. Note that θ is explicitly admitted as a variable in $S(k,\theta)$ to allow for the possibility that for a fixed wavenumber, the beam pattern could be plotted as a function of the angle of arrival. In that case, when θ is zero, the wave impinges the array broadside and the normalized beam pattern evaluates to unity.

The counterpart, in aperture and sensor array processing, of the use of window functions in spectral analysis for reduction of sidelobes is the use of aperture shading. In aperture shading, one simply multiplies a uniformly weighted aperture by the shading function. The resulting beam pattern is, then, simply the convolution of the beam pattern of the uniformly shaded volumetric array and the beam pattern of the shading function. Fourier transform relationship between the stationary signal $s(\mathbf{x},t)$ and the wavenumber frequency spectrum $S(\mathbf{k},\omega)$ allows one to exploit high-resolution spectral analysis techniques for the high-resolution estimation of the direction of arrival [Pillai, 1989]. The superscript *, t, and H denote, respectively, complex conjugate, transpose, and conjugate transpose.

Discrete Arrays for Beamforming

An array of sensors could be distributed at distinct points in space in various ways. Line arrays, planar arrays, and volumetric arrays could be either uniformly spaced or nonuniformly spaced, including the possibility of placing sensors randomly according to some probability distribution function. Uniform spacing along each coordinate axis permits one to exploit the well-developed multidimensional signal processing techniques concerned with filter design, DFT computation via FFT, and high-resolution spectral analysis of sampled signals [Dudgeon, 1977]. Nonuniform spacing sometimes might be useful for reducing the number of sensors, which otherwise might be constrained to satisfy a maximum spacing between uniformly placed sensors to avoid **grating lobes** due to aliasing, as explained later. A discrete array, uniformly spaced, is convenient for the synthesis of a digital filter or beamformer by the performing of digital signal processing operations (namely delay, sum, and multiplication or weighting) on the signal received by a collection of sensors distributed in space. The sequence of the nature of operations dictates the types of beamformer. Common beamforming systems are of

the straight summation, delay-and-sum, and weighted delay-and-sum types. The geometrical distribution of sensors and the weights w_i associated with each sensor are crucial factors in the shaping of the filter characteristics. In the case of a linear array of N equispaced sensors, which are spaced D units apart, starting at the origin $x_1 = 0$, the function

$$W(k_1) = \frac{1}{N} \sum_{n=0}^{N-1} w_n e^{-jk_1 nD} \tag{17.8}$$

becomes the **array pattern**, which may be viewed as the frequency response function for a finite impulse response (FIR) filter, characterized by the unit impulse response sequence $\{w_n\}$. In the case when $w_n = 1$, Eq. (17.8) simplifies to

$$W(k_1) = \frac{1}{N} \frac{\sin\left(\dfrac{k_1 ND}{2}\right)}{\sin\left(\dfrac{k_1 D}{2}\right)} \exp\left\{-j \frac{(N-1)k_1 D}{2}\right\}$$

If the N sensors are symmetrically placed on both sides of the origin, including one at the origin, and the sensor weights are $w_n = 1$, then the linear array pattern becomes

$$W(k_1) = \frac{1}{N} \frac{\sin\left(\dfrac{k_1 ND}{2}\right)}{\sin\left(\dfrac{k_1 D}{2}\right)}$$

For planar arrays, direct generalizations of the preceding linear array results can be obtained. To wit, if the sensors with unity weights are located at coordinates (kD, lD), where $k = 0, \pm 1, \pm 2, \ldots, \pm[(N-1)/2]$, and $l = 0, \pm 1, \pm 2, \ldots, \pm[(M-1)/2]$, for odd integer values of N and M, then the array pattern function becomes

$$\begin{aligned} W(k_1, k_2) &= \frac{1}{NM} \sum_{k=-\left(\frac{N-1}{2}\right)}^{\left(\frac{N-1}{2}\right)} \sum_{l=-\left(\frac{M-1}{2}\right)}^{\left(\frac{M-1}{2}\right)} \exp\{-j(k_1 kD + k_2 lD\} \\ &= \frac{1}{NM} \frac{\sin\left(\dfrac{k_1 ND}{2}\right)}{\sin\left(\dfrac{k_1 D}{2}\right)} \frac{\sin\left(\dfrac{k_2 MD}{2}\right)}{\sin\left(\dfrac{k_2 D}{2}\right)} \end{aligned} \tag{17.10}$$

Routine generalizations to 3-D spatial arrays are also possible. The array pattern functions for other geometrical distributions may also be routinely generated. For example, if unit weight sensors are located at the six vertices and the center of a regular hexagon, each of whose sides is D units long, then the array pattern function can be shown to be

$$W(k_1, k_2) = \frac{1}{7}\left[1 + 2\cos k_1 D + 4\cos\frac{k_1 D}{2}\cos\frac{\sqrt{3}k_2 D}{2}\right] \tag{17.11}$$

The array pattern function reveals how selective a particular beamforming system is. In the case of a typical array function shown in Eq. (17.9), the beamwidth, which is the width of the main lobe of the array pattern, is inversely proportional to the array aperture. Because of the periodicity of the array pattern function, the main lobe is repeated at intervals of $2\pi/D$. These repetitive lobes are called grating lobes, whose existence may be interpreted in terms of spatial frequency aliasing resulting from a sampling interval D due to the N receiving sensors located at discrete points in space. If the spacing D between sensors satisfies

$$D \leq \frac{\lambda}{2} \tag{17.12}$$

where λ is the smallest wavelength component in the signal received by the array of sensors, then the grating lobes have no effect on the received signal. A plane wave of unit amplitude which is incident upon the array at bearing angle θ degrees, as shown in Fig. 17.23, produces outputs at the sensors given by the vector

$$\mathbf{s}(\theta) \underline{\underline{\Delta}} \; \mathbf{s}_\theta = [\exp(j0)\ \exp(jk_1 D\sin\theta)\ \ldots\ \exp(jk_1(N-1)D\sin\theta)]^t \tag{17.13}$$

where $k_1 = 2\pi/\lambda$ is the wavenumber. In array processing, the array output y_θ may be viewed as the inner product of an array weight vector $\mathbf{w}$ and the steering vector $\mathbf{s}_\theta$. Thus, the beamformer response along a direction characterized by the angle θ is, treating $\mathbf{w}$ as complex,

$$y_\theta = \left\langle \mathbf{w}(\theta), \mathbf{s}_\theta \right\rangle = \sum_{k=0}^{N-1} w_k^\star \exp(jk_1 kD\sin\theta) \tag{17.14}$$

The beamforming system is said to be robust if it performs satisfactorily despite certain perturbations [Ahmed and Evans, 1982]. It is possible for each component $s_{k\theta}$ of $\mathbf{s}_\theta$ to belong to an interval $[s_{k\theta} - \phi_{k\theta}, s_{k\theta} + \phi_{k\theta}]$, and a robust beamformer will require the existence of at least one weight vector $\mathbf{w}$ which will guarantee the output y_θ to belong to an output envelope for each $\mathbf{s}_\theta$ in the input envelope. The robust beamforming problem can be translated into an optimization problem, which may be tackled by minimizing the value of the array output power

$$P(\theta) = \mathbf{w}^H(\theta) R \mathbf{w}(\theta) \tag{17.15}$$

when the response to a unit amplitude plane wave incident at the steering direction θ is constrained to be unity, i.e., $\mathbf{w}^H(\theta)\mathbf{s}(\theta) = 1$, and R is the additive noise-corrupted signal autocorrelation matrix. The solution is called the minimum variance beamformer and is given by

$$\mathbf{w}_{\mathrm{MV}}(\theta) = \frac{R^{-1}\mathbf{s}(\theta)}{\mathbf{s}^H(\theta)R^{-1}\mathbf{s}(\theta)} \tag{17.16}$$

and the corresponding power output is

$$P_{\mathrm{MV}}(\theta) = \frac{1}{\mathbf{s}^H(\theta)R^{-1}\mathbf{s}(\theta)} \tag{17.17}$$

The minimum variance power as a function of θ can be used as a form of the data-adaptive estimate of the directional power spectrum. However, in this mode of solution, the coefficient vector is unconstrained except

at the steering direction. Consequently, a signal tends to be regarded as an unwanted interference and is, therefore, suppressed in the beamformed output unless it is almost exactly aligned with the steering direction. Therefore, it is desirable to broaden the signal acceptance angle while at the same time preserving the optimum beamformer's ability to reject noise and interference outside this region of angles. One way of achieving this is by the application of the principle of superdirectivity.

Discrete Arrays and Polynomials

It is common practice to relate discrete arrays to polynomials for array synthesis purposes [Steinberg, 1976]. For volumetric equispaced arrays (it is only necessary that the spacing be uniform along each coordinate axis so that the spatial sampling periods D_i and D_j along, respectively, the ith and jth coordinate axes could be different for $i \neq j$), the weight associated with sensors located at coordinate $(i_1 D_1, i_2 D_2, i_3 D_3)$ is denoted by $w(i_1, i_2, i_3)$. The function in the complex variables (z_1, z_2, and z_3) that is associated with the sequence $\{w(i_1, i_2, i_3)\}$ is the generating function for the sequence and is denoted by

$$W(z_1, z_2, z_3) = \sum_{i_1} \sum_{i_2} \sum_{i_3} w(i_1, i_2, i_3) z_1^{i_1} z_2^{i_2} z_3^{i_3} \tag{17.18}$$

In the electrical engineering and geophysics literature, the generating function $W(z_1, z_2, z_3)$ is sometimes called the z-transform of the sequence $\{w(i_1, i_2, i_3)\}$. When there are a finite number of sensors, a realistic assumption for any physical discrete array, $W(z_1, z_2, z_3)$ becomes a trivariate polynomial. In the special case when $w(i_1, i_2, i_3)$ is product separable, the polynomial $W(z_1, z_2, z_3)$ is also product separable. Particularly, this separability property holds when the shading is uniform, i.e., $w(i_1, i_2, i_3) = 1$. When the support of the uniform shading function is defined by $i_1 = 0,1, \ldots, N_1 - 1$, $i_2 = 0,1, \ldots, N_2 - 1$, and $i_3 = 0,1, \ldots, N_3 - 1$, the associated polynomial becomes

$$W(z_1, z_2, z_3) = \sum_{i_1=0}^{N_1-1} \sum_{i_2=0}^{N_2-1} \sum_{i_3=0}^{N_3-1} z_1^{i_1} z_2^{i_2} z_3^{i_3} = \prod_{i=1}^{3} \frac{z_i^{N_i} - 1}{z_i - 1} \tag{17.19}$$

In this case, all results developed for the synthesis of linear arrays become directly applicable to the synthesis of volumetric arrays. For a linear uniform discrete array composed of N sensors with intersensor spacing D_1 starting at the origin and receiving a signal at a known fixed wavenumber k_1 at a receiving angle θ, the far-field beam pattern

$$S(k_1, \theta) \triangleq S(\theta) = \sum_{r=0}^{N-1} e^{jk_1 r D_1 \sin\theta}$$

may be associated with a polynomial $\Sigma_{r=0}^{N-1} z_1^r$, by setting $z_1 = e^{jk1D1\sin\theta}$. This polynomial has all its zeros on the unit circle in the z_1-plane. If the array just considered is not uniform but has a weighting factor w_r, for $r = 0,1, \ldots, N_1 - 1$, the space factor,

$$Q(\theta) \triangleq \sum_{r=0}^{N_1-1} w_r e^{jk_1 D_1 r \sin\theta}$$

may again be associated with a polynomial $\Sigma_{r=0}^{N_1-1} w_r z_1^r$. By the pattern multiplication theorem, it is possible to get the polynomial associated with the total beam pattern of an array with weighted sensors by multiplying the polynomials associated with the array element pattern and the polynomial associated with the space factor $Q(\theta)$. The array factor $|Q(\theta)|^2$ may also be associated with the polynomial spectral factor

$$\left|Q(\theta)\right|^2 \leftrightarrow \sum_{r=0}^{N_1-1} w_r z_1^r \sum_{r=0}^{N_1-1} w_r^{\star} (z_1^{\star})^r \tag{17.20}$$

where the weighting (shading) factor is allowed to be complex. Uniformly distributed apertures and uniformly spaced volumetric arrays which admit product separable sensor weightings can be treated by using the well-developed theory of linear discrete arrays and their associated polynomial. When the product separability property does not hold, scopes exist for applying results from multidimensional systems theory [Bose, 1982] concerning multivariate polynomials to the synthesis problem of volumetric arrays.

Velocity Filtering

Combination of individual sensor outputs in a more sophisticated way than the delay-and-sum technique leads to the design of multichannel velocity filters for linear and planar as well as spatial arrays. Consider, first, a linear (1-D) array of sensors, which will be used to implement velocity discrimination. The pass and rejection zones are defined by straight lines in the (k_1,ω)-plane, where

$$k_1 = \frac{\omega}{V} = \frac{\omega}{(v/\sin\theta)}$$

is the wavenumber, ω the angular frequency in radians/second, V the apparent velocity on the earth's surface along the array line, v the velocity of wave propagation, and θ the horizontal arrival direction. The transfer function

$$H(\omega, k_1) = \begin{cases} 1, & -\dfrac{|\omega|}{V} \le k_1 \le \dfrac{|\omega|}{V} \\ 0, & \text{otherwise} \end{cases}$$

of a "pie-slice" or "fan" velocity filter [Bose, 1985] rejects totally wavenumbers outside the range $-|\omega|/V \le k_1 \le |\omega|/V$ and passes completely wavenumbers defined within that range. Thus, the transfer function defines a high-pass filter which passes signals with apparent velocities of magnitude greater than V at a fixed frequency ω. If the equispaced sensors are D units apart, the spatial sampling results in a periodic wavenumber response with period $k_1 = 1/(2D)$. Therefore, for a specified apparent velocity V, the resolvable wavenumber and frequency bands are, respectively, $-1/(2D) \le k_1 \le 1/(2D)$ and $-V/(2D) \le \omega \le V/(2D)$ where $\omega/(2D)$ represents the folding frequency in radians/second.

Linear arrays are subject to the limitation that the source is required to be located on the extended line of sensors so that plane wavefronts approaching the array site at a particular velocity excite the individual sensors, assumed equispaced, at arrival times which are also equispaced. In seismology, the equispaced interval between successive sensor arrival times is called a move-out or step-out and equals $(D \sin\theta)/v = D/V$. However, when the sensor-to-source azimuth varies, two or more independent signal move-outs may be present. Planar (2-D) arrays are then required to discriminate between velocities as well as azimuth. Spatial (3-D) arrays provide additional scope to the enhancement of discriminating capabilities when sensor/source locations are arbitrary. In such cases, an array origin is chosen and the mth sensor location is denoted by a vector $(x_{1m}\, x_{2m}\, x_{3m})^t$ and the frequency wavenumber response of an array of sensors is given by

$$H(\omega, k_1, k_2, k_3) = \frac{1}{N} \sum_{m=1}^{N} H_m(\omega) \exp\left[\sum_{i=1}^{3} -j2\pi k_i x_{im}\right]$$

where $H_m(\omega)$ denotes the frequency response of a filter associated with the mth recording device (sensor). The sum of all N filters provides flat frequency response so that waveforms arriving from the estimated directions of arrival at estimated velocities are passed undistorted and other waveforms are suppressed. In the planar

specialization, the 2-D array of sensors leads to the theory of 3-D filtering involving a transfer function in the frequency wavenumber variables f, k_1, and k_2. The basic design equations for the optimum, in the least-mean-square error sense, frequency wavenumber filters have been developed [Burg, 1964]. This procedure of Burg can be routinely generalized to the 4-D filtering problem mentioned above.

Acknowledgment

N.K. Bose and L.H. Sibul acknowledge the support provided by the Office of Naval Research under, respectively, Contract N00014-92-J-1755 and the Fundamental Research Initiatives Program.

Defining Terms

Array pattern: Fourier transform of the receiver weighting function taking into account the positions of the receivers.

Beamformers: Systems commonly used for detecting and isolating signals that are propagating in a particular direction.

Grating lobes: Repeated main lobes in the array pattern interpretable in terms of spatial frequency aliasing.

Velocity filtering: Means for discriminating signals from noise or other undesired signals because of their different apparent velocities.

Wavenumber: 2π (spatial frequency in cycles per unit distance).

Related Topic

14.3 Design and Implementation of Digital Filters

References

K.M. Ahmed and R.J. Evans, "Robust signal and array processing," *IEE Proceedings, F: Communications, Radar, and Signal Processing,* vol. 129, no. 4, pp. 297–302, 1982.

N.K. Bose, *Applied Multidimensional Systems Theory,* New York: Van Nostrand Reinhold, 1982.

N.K. Bose, *Digital Filters,* New York: Elsevier Science North-Holland, 1985. Reprint ed., Malabar, Fla.: Krieger Publishing, 1993.

J.P. Burg, "Three-dimensional filtering with an array of seismometers," *Geophysics,* vol. 23, no. 5, pp. 693–713, 1964.

J.F. Claerbout, *Fundamentals of Geophysical Data Processing,* New York: McGraw-Hill, 1976.

D.E. Dudgeon, "Fundamentals of digital array processing," *Proc. IEEE,* vol. 65, pp. 898–904, 1977.

R.A. Monzingo and T.W. Miller, *Introduction to Adaptive Arrays,* New York: Wiley, 1980.

S.M. Pillai, *Array Signal Processing,* New York: Springer-Verlag, 1989.

B.D. Steinberg, *Principles of Aperture and Array System Design,* New York: Wiley, 1976.

Further Information

Adaptive Signal Processing, edited by Leon H. Sibul, includes papers on adaptive arrays, adaptive algorithms and their properties, as well as other applications of adaptive signal processing techniques (IEEE Press, New York, 1987).

Adaptive Antennas:Concepts and Applications, by R. T. Compton, Jr., emphasizes adaptive antennas for electromagnetic wave propagation applications (Prentice-Hall, Englewood-Cliffs, N.J., 1988).

Array Signal Processing: Concepts and Techniques, by D. H. Johnson and D. E. Dudgeon, incorporates results from discrete-time signal processing into array processing applications such as signal detection, estimation of direction of propagation, and frequency content of signals (Prentice-Hall, Englewood Cliffs, N.J., 1993).

Neural Network Fundamentals with Graphs, Algorithms, and Applications, by N. K. Bose and P. Liang, contains the latest information on adaptive-structure networks, growth algorithms, and adaptive techniques for learning and capability for generalization (McGraw-Hill, New York, N.Y., 1996).

17.4 Video Processing Architectures

Wayne Wolf

Video processing has become a major application of computing: personal computers display multimedia data, digital television provides more channels, etc. The characteristics of video algorithms are very different from traditional applications of computers; these demands require new architectures.

Two fundamental characteristics of video processing make it challenging and different than applications like database processing. First, the video processor must handle streaming data that arrives constantly. Traditional applications assume that data has a known, fixed location. In video processing, not only are new input samples always arriving, but our time reference in the stream is constantly changing. At one time instant, we may consider a sample x_t, but at the next sampling interval that sample becomes x_{t-1}. The need to sweep through the data stream puts additional demands on the memory system. Since streaming data must be processed in realtime. If the deadline for completing an output is missed, the results will be visible on the screen. When designing realtime systems, it is not sufficient to look at aggregate throughput because data can become backed up for a period and still meet some long-term timing requirements. Processing must complete every realtime result by the appointed deadline. Architectures must provide underlying support for predictable computation times.

The challenges of processing streaming data in realtime are made greater by the fact that video processing algorithms are becoming very complex. Video compression algorithms make use of several different techniques and complex search algorithms to maximize their ability to compress the video data; video display systems provide much more sophisticated controls to the user; content analysis systems make use of multiple complex algorithms working together; mixed computer graphics-video systems combine geometric algorithms with traditional video algorithms. Expect video processing algorithms to become more complex in the future. This complexity puts greater demands on the realtime nature of the video architecture: more complex algorithms generally have less predictable execution times. The architecture should be designed so that algorithms can take advantage of idle hardware caused by early completions of functions, rather than letting hardware sit idle while it waits for other operations to complete.

Luckily, VLSI technology is also advancing rapidly and allows us to build ever more sophisticated video processing architectures. The state of video processing architectures will continue to advance as VLSI allows us to integrate more transistors on a chip; in particular, the ability to integrate a significant amount of memory along with multiple processing elements will provide great strides in video processing performance over the next several years. However, the basic techniques for video processing used today [Pir98] will continue to be the basis for video architectures in the long run.

This chapter section first reviews two basic techniques for performing video operations: single instruction multiple data (SIMD) and vectorization; and then looks at the three major styles of video architectures: heterogeneous multiprocessors, video signal processors, and microprocessor instruction set extensions.

Computational Techniques

Many of the fundamental operations in video processing are filters that can be described as linear equations; for example,

$$\sum_{1 \le i \le n} c_i x_i$$

There are two techniques for implementing such equations: single-instruction multiple data (SIMD) processing and vector processing. The two are similar in underlying hardware structure; the most important differences come in how they relate to the overall computer architecture of which they are a part.

SIMD

The term **SIMD** comes from Flynn's classification of computer architectures, based on the number of data elements they processed simultaneously and the number of instructions used to control the operations on those

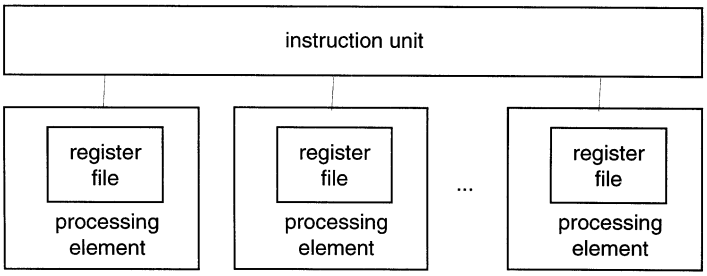


FIGURE 17.24 A SIMD architecture.

data. In a SIMD machine, a single instruction is used to control the operation performed on many data elements. Thus, the same operation is performed simultaneously on all that data. Figure 17.24 shows a SIMD structure: several function units, each with its own register file, has an **ALU** for performing operations on data; the controller sends identical signals to all function units so that the same operation is performed on all function units at the same; there is also a network that allows processing elements to pass data among themselves.

Consider how to use a SIMD machine to perform the filtering operation given at the beginning of this section. The multiplications are all independent, so we can perform N multiplications in parallel on the N processing elements. We need to perform N – 1 additions on the multiplication results; by properly arranging the computation in a tree, many of those operations can be performed in parallel as well. We will need to use the data transfer network in two ways: to transfer x values between processing elements for the data streaming time shift; and to transfer the partial addition results in the addition tree. SIMD architectures can of course be used to implement multidimensional functions as well. For example, two-dimensional correlation is used in video compression, image recognition, etc., and can easily be mapped onto a SIMD machine.

SIMD architectures provide a high degree of parallelism at high speeds. Instruction distribution and decoding is not a bottleneck. Furthermore, each **processing element** has its own data registers and the communication network between the processing elements can be designed to be fast. However, not all algorithms can be efficiently mapped onto SIMD architectures. Global computation is difficult in SIMD machines. Operations that cause global changes to the machine state also create problems.

Vectorization

Vector instructions were originally invented for supercomputers to improve the performance of scientific calculations. Although video operations are generally done in fixed-point rather than floating-point arithmetic, vector instructions are well-suited to the many video operations that can be expressed in linear algebra. Vectorization was used in many early video processors. More recently, SIMD has become more popular, but with vectorization becoming more popular in general-purpose microprocessors, there may be a resurgence of vector units for multimedia computation.

A vector is a data structure supported by hardware. The vector is stored in memory as a set of memory locations; special vector registers are also provided to hold the vectors for arithmetic operations. Our filter example could be implemented as a single vector instruction (after loading the vector registers with the c and x vectors): a vector multiply-accumulate instruction, similar to scalar multiply-accumulate instructions in DSPs, could multiply the x_i's by the c_i's and accumulate the result.

The motivation for supporting vector instructions is pipelining the arithmetic operations. If an arithmetic operation takes several clock cycles, pipelining allows high throughput at a high clock rate at the cost of latency. As shown in Fig. 17.25 vectors are well-suited to pipelined execution because all the operations in the vector are known to be independent in advance.

Vector units allow linear algebra to be performed at very high speeds with high hardware utilization. Furthermore, because they have a long history in scientific computing, compiling high-level languages into vector instructions is well understood. However, the latencies of integer arithmetic operations for video operations is smaller than that for the floating-point operations typically used in scientific **vector processors**.

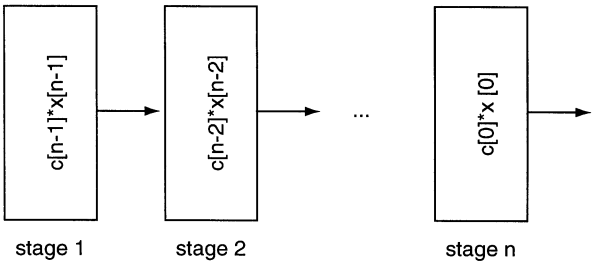


FIGURE 17.25 Pipelining to support vector operations.

Heterogeneous Multiprocessors

The earliest style of video processor is the **heterogeneous multiprocessor**. These machines cannot execute arbitrary programs — they are restricted to a single algorithm or variations on that algorithm. The microarchitecture of the machine is tuned to the target application. In the early days of digital video, special-purpose heterogeneous multiprocessors were the only way to implement VLSI video processing because chips were not large enough to support the hardware required for instruction-set processors. Today, heterogeneous multiprocessors are used to implement low-cost video systems, since by specializing the hardware for a particular application, less hardware is generally required, resulting in smaller, less-expensive chips.

A simple heterogeneous architecture is shown in Fig. 17.26. This machine implements a sum-of-absolute-differences correlation in two dimensions for block motion estimation. The architecture of this machine is derived from the data flow of the computation, where for each offset (r, s), the sum-of-absolute differences between a $n \times n$ macroblock and a $T \times T$ reference area can be computed:

$$\sum_{1 \le i \le n} \sum_{1 \le i \le n} | M(i, j) - R(i + r, j + s) |$$

The machine executes one column of the computation per clock cycle: n absolute differences are formed and then passed onto a summation unit. This machine is not a SIMD architecture because it does not execute instructions — it is designed to perform one algorithm.

Heterogeneous architectures can also be used for more complex algorithms. Figure 17.27 shows a sketch for a possible architecture for MPEG-style video compression [MPE]. The unit has separate blocks for the major operations: block motion estimation, discrete cosine transform (DCT) calculation, and channel coding; it also has a processor used for overall control.

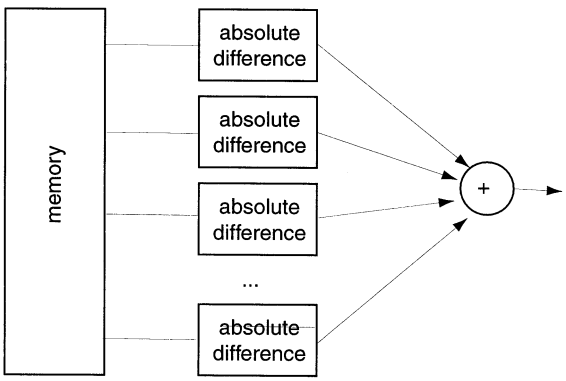


FIGURE 17.26 A heterogeneous multiprocessor.

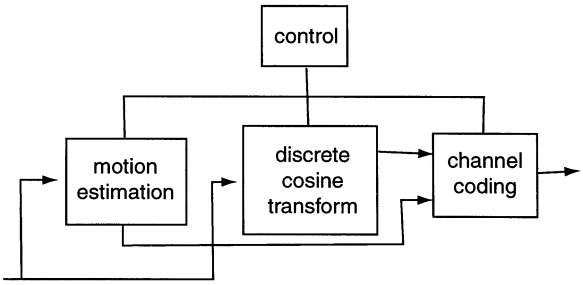


FIGURE 17.27 A heterogeneous architecture for MPEG-style compression.

Heterogeneous architectures are designed by careful examination of the algorithm to be implemented. The most time-critical functions must be identified early. Those operations are typically implemented as special-purpose function units. For example, the block motion estimation engine of Fig. 17.26 can be used as a special-purpose function unit in a more complex application like an MPEG video compressor. Communication links must be provided between the function units to provide adequate bandwidth for the data transfers. In structured communication architectures, data transfers can be organized around buses or more general communication networks like crossbars. Heterogeneous communication systems make specialized connections as required by the algorithm. Many modern heterogeneous video processors use as much structured communication as possible but add specialized communication links as required to meet performance requirements. Many modern heterogeneous processors are at least somewhat programmable. Basic architectures may use registers to control certain parameters of the algorithm. More complex algorithms may use general-purpose microprocessors as elements of the architecture. Small microcontrollers are frequently used for system interfacing, such as talking to a keyboard or other controlling device. Larger microprocessors can be used to run algorithms that do not benefit from special-purpose function units.

Heterogeneous multiprocessors will continue to dominate high-volume, low-cost markets for video and multimedia functions. When an application is well-defined, it is often possible to design a special-purpose architecture that performs only that operation but is significantly cheaper than a system built from a programmable processor. Furthermore, heterogeneous multiprocessors may require significantly less power than programmable solutions and, therefore, an increasing number of battery-operated multimedia devices. However, heterogeneous multiprocessors are not well-suited to other application areas. If the algorithm is not well-defined, if the system must be able to execute a variety of algorithms, or if the size of the market will not support the cost of designing an application-specific solution, heterogeneous multiprocessors are not appropriate.

Video Signal Processors

The term digital signal processor (DSP) is generally reserved for microprocessors optimized for signal processing algorithms and which run at audio rates. A video signal processor (VSP) is a DSP that is capable of running at video rates. Using separate names for audio and video rate processors is reasonable because VSPs provide much greater parallelism and significantly different microarchitectures.

Many early video processors were vector machines because vector units provide high throughput with relatively small amounts of hardware. Today, most VSPs today make use of very-long instruction word (VLIW) processor technology, as shown in Fig. 17.28. The architecture has several function units connected to a single register file. The operations on all the function units are controlled by the instruction decoder based on the current instruction. A **VLIW** machine differs from a SIMD machine in two important ways. First, the VLIW machine connects all function units to the same register file, while the SIMD machine uses separate registers for the function units. The common register file gives the VLIW machine much more flexibility; for example, a data value can be used on one function unit on one cycle and on another function unit on the next cycle without having to copy the value. Second, the function units in the VLIW machine need not perform the same operation. The instruction is divided into fields, one for each unit. Under control of its instruction field, each instruction unit can request data from the register file and perform operations as required.

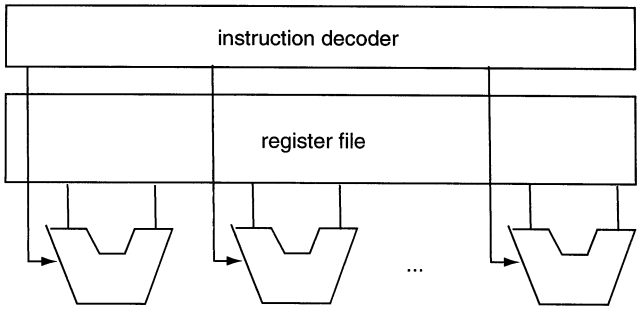


FIGURE 17.28 A simple VLIW machine.

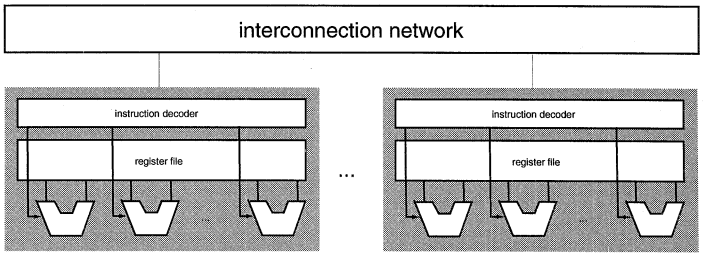


FIGURE 17.29 A clustered VLIW machine.

Although having a common register file is very flexible, there are physical limitations on the number of function units that can be connected to a single register file. A single addition requires three ports to the register file: one to read each operand and a third to write the result back to the register file. Register files are built from static random access memory (SRAMs) and slow down as the number of read/write ports grows. As a result, VLIW machines are typically built in clusters as shown in Fig. 17.29. Each cluster has its own register file and function units, with three or four function units per cluster typical in today's technology. A separate interconnection network allows data transfers between the clusters. When data held in one register file is needed in a different cluster, an instruction must be executed to transfer the data over the interconnection network to the other register file.

The major difference between VLIW architectures and the superscalar architectures found in modern microprocessors is that VLIW machines have statically scheduled operations. A superscalar machine has hardware that examines the instruction stream to determine what operations can be performed in parallel; for example, when two independent operations appear in consecutive instructions, those instructions can be executed in parallel. A VLIW machine relies on a compiler to identify parallelism in the program and to pack those operations into instruction words. This requires sophisticated compilers that can extract parallelism and effectively make use of it when generating instructions. Video is especially well-suited to VLIW because video programs have a great deal of parallelism that is relatively easy to identify and take advantage of in a VLIW machine.

VLIW has potential performance advantages because its control unit is relatively simple. Because the work of finding parallelism is performed by the compiler, a VLIW machine does not require the sophisticated execution unit of a superscalar processor. This allows a VLIW video processor to run at high clock rates. However, it does rely on the compiler's ability to find enough parallelism to keep the function units busy. Furthermore, complex algorithms may have some sections that are not highly parallel and therefore will not be sped up by the VLIW mechanism. If one is not careful, these sequential sections of code can come to limit the overall performance of the application.

Practical video signal processors are not pure VLIW machines, however. In general, they are in fact hybrid machines that use VLIW processing for some operations and heterogeneous multiprocessing techniques for others. This is necessary to meet the high performance demand on video processing; certain critical operations can be sped up with special-purpose function units, leaving the VLIW processor to perform the rest. An example

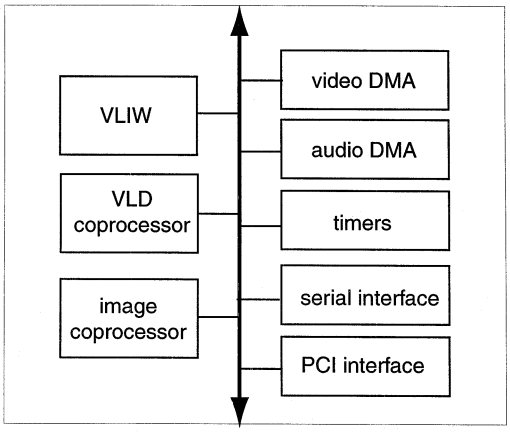


FIGURE 17.30 The Trimedia TM-1 video signal processor.

of this technique is the Trimedia TM-1 processor [Rat96] shown in Fig. 17.30. This machine has a VLIW processor. It also has several function units for specialized video operations, principal among these being a variable-length decoding for channel coding, an image coprocessor. The TM-1 also supports multiple DMA channels to speed up data transfers as well as timers to support realtime operation.

VLIW VSPs represent one end of the programmable video processor spectrum. These machines are designed from the ground up to execute video algorithms. The VLIW architecture is very well-suited to video applications due to the embarrassing levels of parallelism available in video programs. Special-purpose function units can be used to speed up certain key operations. However, VLIW VSPs may not be as well-suited to executing code that is more typically found on workstation microprocessors, such as error checking, bit-level operations, etc. As a result, VLIW VSPs can be used in conjunction with standard microprocessors to implement a complex video application, with the VSP performing traditional parallel video sections of the code and the microprocessor performing the less regular computations.

Instruction Set Extensions

Both heterogeneous multiprocessors and VSPs are specialized architectures for video. However, there are many applications in which it is desirable to execute video programs directly on a workstation or PC: programs that are closely tied to the operating system, mixed video/graphics applications, etc. Traditional microprocessors are fast but are not especially well-utilized by video programs. For these applications, microprocessor instruction set extensions have been developed to allow video algorithms to be executed more efficiently on traditional microprocessors.

The basic principle of instruction set extensions is subword parallelism [Lee95], as illustrated in Fig. 17.31. This technique takes advantage of the fact that modern microprocessors support native 32- or 64-bit operations while most video algorithms require much smaller data accuracy, such as 16 bits or even 8 bits. One can divide the microprocessor data path, on which the instructions are executed, into subwords. This is a relatively simple operation, mainly entailing adding a small amount of logic to cut the ALU's carry chain at the appropriate points when subword operations are performed. When a 64-bit data path is divided for use by 16-bit subwords, the machine can support four simultaneous subword operations. Subword parallelism is often referred to as SIMD because a single microprocessor instruction causes the same operation to be performed on all the subwords in parallel. However, there is no separate SIMD instruction unit — all the work is done by adding a small amount of hardware to the microprocessor data path. Subword parallelism is powerful because it has a very small cost in the microprocessor (both in terms of chip area and performance) and because it provides substantial speedups on parallel code.

A typical instruction set extension will of course support logical and arithmetic operations on subwords. They may support saturation arithmetic as well as two's-complement arithmetic. Saturation arithmetic generates the maximum value on overflow, more closely approximating physical devices. They may also support

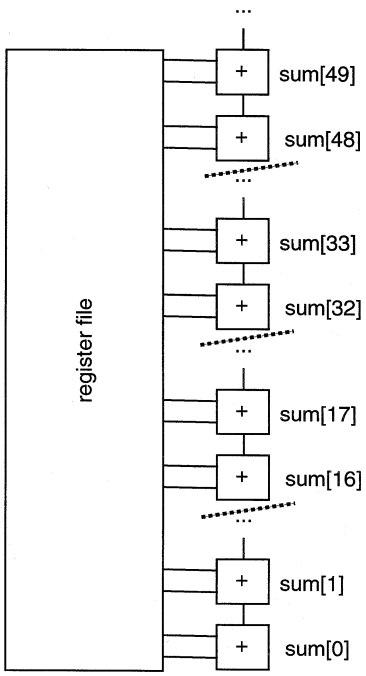


FIGURE 17.31 Implementing subword parallelism on a microprocessor.

permutation operations so that the order of subwords in a word can be shuffled. Loads and stores are performed on words — not subwords.

ISA extensions have been defined for the major microprocessor architectures. The MAX extension for the HP PA-RISC architecture [Lee96] was the first ISA extension and introduced the notion of subword parallelism. The VIS (Visual Instruction Set) extension [Tre96] has been added to the Sun SPARC architecture. The Intel x86 architecture has been extended with the MMX instructions [Pel96].

The MMX extension is based on the well-known Intel architecture. It supports operations on 8-bit bytes, 16-bit words, 32-bit doublewords, and 64-bit quadwords. All these data types are packed into 64-bit words. All MMX operations are performed in the floating-point registers; this means that the floating-point registers must be saved at the beginning of MMX code and restored at its end. (Although floating-point operations access these registers as a stack, MMX instructions can arbitrarily address the registers.) MMX supports addition, subtraction, comparison, multiplication, shifts, and logical operations. Arithmetic can optionally be performed in saturation mode. There are also instructions for packing and unpacking subwords into words. Some conversion operations are provided so that intermediate calculations can be performed at higher precisions and then converted to a smaller format.

The Sun VIS extension also uses the floating-point registers. The MAX-2 extension is the latest extension to the HP architecture. It uses integer registers rather than floating-point registers. It does not directly implement multiplication, but instead provides a shift-and-add operation for software-driven multiplication. MAX-2 also supports a permutation operation to allow subwords to be rearranged in a word.

The ability to mix multimedia instructions with other instructions on a standard microprocessor is the clear advantage of instruction set extensions. These extensions are very well-suited to the implementation of complex algorithms because the microprocessor can efficiently execute the nonlinear algebra operations as well as the highly parallel video operations. Furthermore, instruction set extensions take advantage of the huge resources available to microprocessor manufacturers to build high-performance chips.

The main disadvantages of instruction set extensions are related to the tight coupling of the video and non-video instructions. First, the memory system is not changed to fit the characteristics of the video application.

The streaming data typical of video is not very well-suited to the caches used in microprocessors. Caches rely on temporal and spatial locality; they assume that a variable is used many times after its initial use. In fact, streaming data will be used a certain number of times and then be discarded, to be replaced with a new datum. Second, the available parallelism is limited by the width of the data path. A 64-bit data path can exhibit at most four-way parallelism when subdivided into 16-bit subwords. Other architectures can be more easily extended for greater parallelism when technology and cost permit.

Summary

There is no one best way to design a video processing architecture. The structure of the architecture depends on the intended application environment, algorithms to be run, performance requirements, cost constraints, and other factors. Computer architects have developed a range of techniques that span a wide range of this design space: heterogeneous multiprocessors handle low-cost applications effectively; VLIW video signal processors provide specialized video processing; instruction set extensions to microprocessors enhance video performance on traditional microprocessors. As VLSI technology improves further, these techniques will be extended to create machines that hold significant amounts of video memory on-chip with the processing elements that operate on the video data.

Defining Terms

ALU: Arithmetic/logic unit.
MPEG: A set of standards for video compression.
Processing element: A computational unit in a parallel architecture.
SIMD (single-instruction multiple data): An architecture in which a single instruction controls the operation of many separate processing elements.
Heterogeneous multiprocessors: An architecture in which several dissimilar processing units are connected together to perform a particular computation.
Vector processor: A machine that operates on vector and matrix quantities in a pipelined fashion.
VLIW (very-long instruction word): An architecture in which several ALUs are connected to a common register file, under the control of an instruction word that allows the ALU operations to be determined separately.

References

[Lee95] R. B. Lee, Accelerating multimedia with enhanced microprocessor, *IEEE Micro*, April 1995, pp. 22–32.
[Lee96] R. B. Lee, Subword parallelism with MAX-2, *IEEE Micro*, August 1996, pp. 51–59.
[MPE] MPEG Web site, http://www.mpeg.org.
[Pel96] A. Peleg and U. Weiser, MMX technology extension to the Intel architecture, *IEEE Micro*, August 1996, pp. 42–50.
[Pir98] P. Pirsch and J.-J. Stolberg, VLSI implementations of image and video multimedia processing systems, *IEEE Transactions on Circuits and Systems for Video Technology*, 8(7), November 1998, pp. 878–891.
[Rat96] S. Rathnam and G. Slavenburg, An architectural overview of the programmable media processor, TM-1, in *Proc. Compcon*, IEEE Computer Society Press, 1996, pp. 319–326.
[Tre96] M. Tremblay, J. M. O'Connor, Ventatesh Narayanan, and Liang He, VIS speeds new media processing, *IEEE Micro*, August 1996, pp. 10–20.

Further Reading

Two journals, *IEEE Transactions on Circuits and Systems for Video Technology* and *IEEE Micro* — provide up-to-date information on developments in video processing. A number of conferences cover this area, including the International Solid State Circuits Conference (ISCCC) and the Silicon Signal Processing (SiSP) Workshop.

17.5 MPEG-4 Based Multimedia Information System

Ya-Qin Zhang[1]

Recent creation and finalization of the Motion-Picture Expert Group (**MPEG-4**) international standard has provided a common platform and unified framework for multimedia information representation. In addition to providing highly efficient compression of both natural and synthetic audio-visual (AV) contents such as video, audio, sound, texture maps, graphics, still images, MIDI, and animated structure, MPEG-4 enables greater capabilities for manipulating AV contents in the compressed domain with object-based representation. MPEG-4 is a natural migration of the technological convergence of several fields: digital television, computer graphics, interactive multimedia, and Internet. This tutorial chapter briefly discusses some example features and applications enabled by the MPEG-4 standard.

During the last decade, a spectrum of standards in digital video and multimedia has emerged for different applications. These standards include: the ISO JPEG for still images [1]; ITU-T H.261 for video conferencing from 64 kilobits per second (kbps) to 2 megabits per second (Mbps) [2]; ITU-T H.263 for PSTN-based video telephony [3]; ISO MPEG-1 for CD-ROM and storage at VHS quality [4]; the ISO MPEG-2 standard for digital television [5]; and the recently completed ISO/MPEG-4 international standard for multimedia representation and integration [6]. Two new ISO standards are under development to address the next-generation still image coding (JPEG-2000) and content-based multimedia information description (MPEG-7). Several special issues of *IEEE* journals have been devoted to summarizing recent advances in digital image, video compression, and advanced television in terms of standards, algorithms, implementations, and applications [7–11].

The successful convergence and implementation of MPEG-1 and MPEG-2 have become a catalyst for propelling the new digital consumer markets such as Video CD, Digital TV, DVD, and DBS. While the MPEG-1 and MPEG-2 standards were primarily targeted at providing high compression efficiency for storage and transmission of pixel-based video and audio, MPEG-4 envisions to support a wide variety of multimedia applications and new functionalities of object-based audio-visual (AV) contents. The recent completion of MPEG-4 Version 1 is expected to provide a stimulus to the emerging multimedia applications in wireless networks, Internet, and content creation.

The MPEG-4 effort was originally conceived in late 1992 to address very low bit rate (VLBR) video applications at below 64 kbps such as PSTN-based videophone, video e-mail, security applications, and video over cellular networks. The main motivations for focusing MPEG-4 at VLBR applications were:

- Applications such as PSTN videophone and remote monitoring were important, but not adequately addressed by established or emerging standards. In fact, new products were introduced to the market with proprietary schemes. The need for a standard at rates below 64 kbps was imminent.
- Research activities had intensified in VLBR video coding, some of which have gone beyond the boundary of the traditional statistical-based and pixel-oriented methodology.

It was felt that a new breakthrough in video compression was possible within a five-year time window. This "quantum leap" would likely make compressed-video quality at below 64 kbps, adequate for many applications such as videophone.

Based on the above assumptions, a workplan was generated to have the MPEG-4 Committee Draft (CD) completed in 1997 to provide a generic audio visual coding standard at very low bit rates. Several MPEG-4 seminars were held in parallel with the WG11 meetings, many workshops and special sessions have been organized, and several special issues have been devoted to such topics. However, as of July 1994 in the Norway WG11 meeting, there was still no clear evidence that a "quantum leap" in compression technology was going to happen within the MPEG-4 timeframe. On the other hand, ITU-T has embarked on an effort to define the H.263 standard for videophone applications in PSTN and mobile networks. The need for defining a pure compression standard at very low bit rates was, therefore, not entirely justified.

[1]The author was the director of Multimedia Technology Laboratory at Sarnoff Corporation in Princeton, New Jersey when this work was performed.

In light of the situation, a change of direction was called to refocus on new or improved functionalities and applications that are not addressed by existing and emerging standards. Examples include object-oriented features for content-based multimedia database, error-robust communications in wireless networks, hybrid nature and synthetic image authoring and rendering. With the technological convergence of digital video, computer graphics, and Internet, MPEG-4 aims at providing an audiovisual coding standard allowing for interactivity, high compression, and/or universal accessibility, with a high degree of flexibility and extensibility.

In particular, MPEG-4 intends to establish a flexible content-based audio-visual environment that can be customized for specific applications and that can be adapted in the future to take advantage of new technological advances. It is foreseen that this environment will be capable of addressing new application areas ranging from conventional storage and transmission of audio and video to truly interactive AV services requiring content-based AV database access, e.g., video games or AV content creation. Efficient coding, manipulation, and delivery of AV information over Internet will be key features of the standard.

MPEG-4 Multimedia System

Figure 17.32 shows an architectural overview of MPEG-4. The standard defines a set of syntax to represent individual *audiovisual objects*, with both natural and synthetic contents. These objects are first encoded independently into their own elementary streams. Scene description information is provided separately, defining the location of these objects in space and time that are composed into the final scene presented to the user. This representation includes support for user interaction and manipulation. The scene description uses a tree-based structure, following the Virtual Reality Modeling Language (VRML) design. Moving far beyond the capabilities of VRML, MPEG-4 scene descriptions can be dynamically constructed and updated, enabling much higher levels of interactivity. Object descriptors are used to associate scene description components that relate digital video to the actual elementary streams that contain the corresponding coded data. As shown in Fig. 17.32, these components are encoded separately and transmitted to the receiver. The receiving terminal then has the responsibility of composing the individual objects for presentation and for managing user interaction.

Following are eight MPEG-4 functionalities, defined and clustered into three classes:

- Content-based interactivity: Content-based manipulation and bit stream editing; content-based multimedia data access tools; hybrid natural and synthetic data coding; improved temporal access.
- Compression: Improved coding efficiency; coding of multiple concurrent data streams.
- Universal access: Robustness in error-prone environments; content-based scalability.

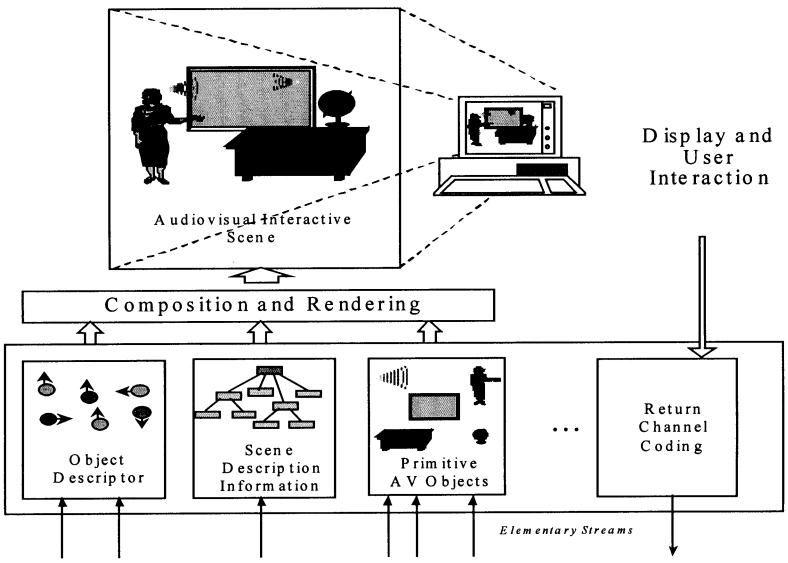


FIGURE 17.32 MPEG-4 Overview. Audio-visual objects, natural audio, as well as synthetic media are independently coded and then combined according to scene description information (courtesy of the ISO/MPEG-4 committee).

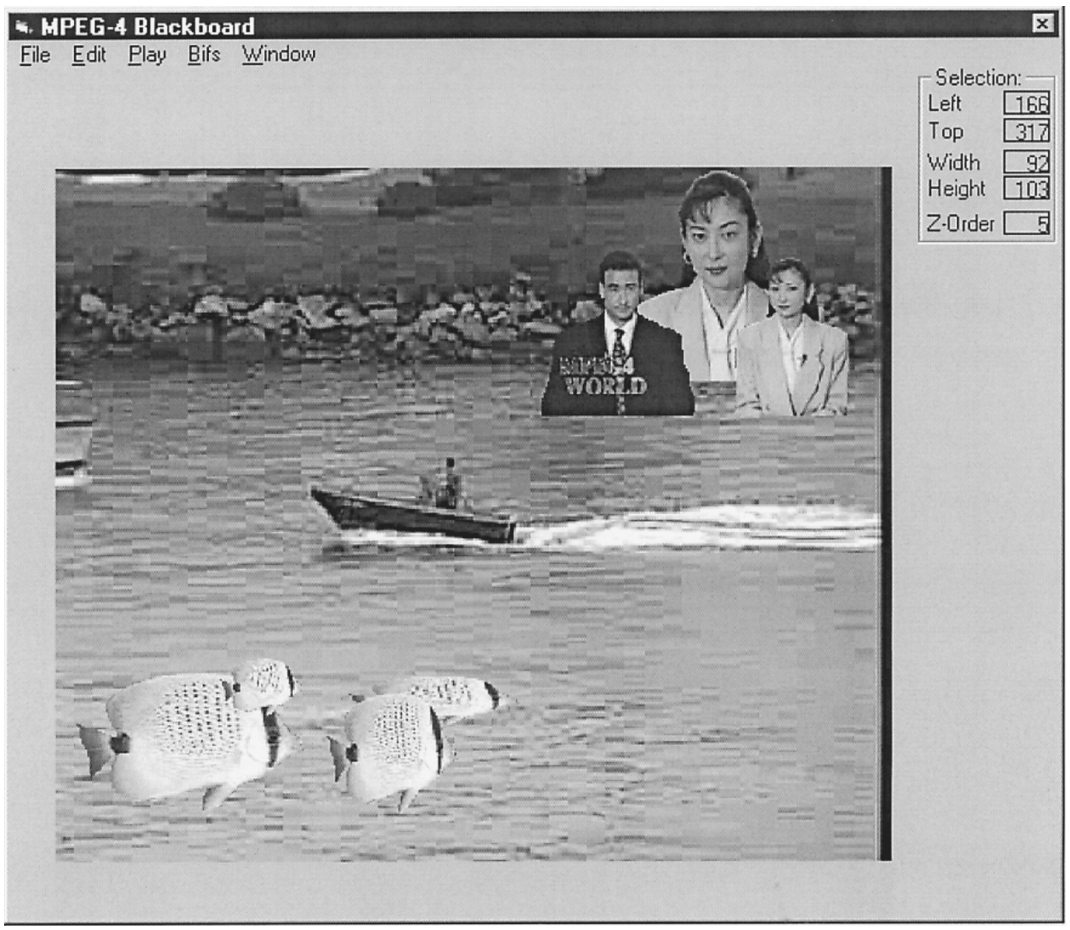


FIGURE 17.33 An example of a multimedia authoring system using MPEG-4 tools and functionalities (courtesy of Sarnoff Corporation).

Some of the applications enabled by these functionalities include:

- Video streaming over Internet.
- Multimedia authoring and presentations.
- View of the contents of video data in different resolutions, speeds, angles, and quality levels.
- Storage and retrieval of multimedia database in mobile links with high error rates and low channel capacity (e.g., Personal Digital Assistant).
- Multipoint teleconference with selective transmission, decoding, and display of "interesting" parties.
- Interactive home shopping with customers' selection from a video catalogue.
- Stereo-vision and multiview of video contents, e.g., sports.
- "Virtual" conference and classroom.
- Video email, agents, and answering machines.

Object-based Authoring Tool Example

Figure 17.33 shows an example of an object-based authoring tool for MPEG-4 AV contents, recently developed by the Multimedia Technology Laboratory at Sarnoff Corporation in Princeton, New Jersey. This tool has the following features:

- Compression/decompression of different visual objects into MPEG-4-compliant bitstreams.
- Drag-and-drop of video objects into a window while resizing the objects or adapting them to different frame rates, speeds, transparencies, and layers.

- Substitution of different backgrounds.
- Mixing natural image and video objects with computer-generated, synthetic texture and animated objects.
- Creating metadata information for each visual objects.

This set of authoring tools can be used for interactive Web design, digital studio, and multimedia presentation. It empowers users to compose and interact with digital video on a higher semantic level.

References

1. JPEG Still Image Coding Standard, ISO/IEC 10918–1, 1990.
2. Video Code for Audiovisual Services at 64 to1920 kbps, CCITT Recommendation H.261, 1990.
3. Recommendation H.263P Video Coding for Narrow Telecommunication Channels at below 64 kbps, ITU-T/SG15/LBC, May 1995.
4. Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbps, ISO/IEC 11172, 1992.
5. Generic Coding of Moving Pictures and Associated Audio, ISO/IEC 13818, 1994.
6. MPEG-4 Draft International Standard, ISO/IEC JTC1/SC29/WG11, October, 1998.
7. Y.-Q. Zhang, W. Li, and M. Liou, eds., Advances in Digital Image and Video Compression, *Special Issue, Proceedings of IEEE,* Feb. 1995.
8. M. Kunt, ed. Digital Television, *Special Issue, Proceedings of IEEE,* July 1995.
9. Y.-Q. Zhang, F. Pereria, T. Sikora, and C. Reader, eds., MPEG-4, *Special Issue, IEEE Transactions on Circuits and Systems for Video Technology,* Feb. 1997.
10. T. Chen, R. Liu, and A. Tekalp, eds., Multimedia Signal Processing, *Special Issue on Proceedings of IEEE,* May 1998.
11. M.T. Sun, K. Ngan, T. Sikora, and S. Panchnatham, eds., Representation and Coding of Images and Video, *IEEE Transactions on Circuits and Systems for Video Technology,* November 1998.
12. MPEG-4 Requirements Ad-Hoc Group, MPEG-4 Requirements, ISO/IEC JTC1/SC29/WG11/MPEG-4, Maceio, Nov. 1996.

Parhi, K.K., Chassaing, R., Bitler, B. "VLSI for Signal Processing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

18

VLSI for Signal Processing

18.1 Special Architectures
Pipelining • Parallel Processing • Retiming • Unfolding • Folding Transformation • Look-Ahead Technique • Associativity Transformation • Distributivity • Arithmetic Processor Architectures • Computer-Aided Design • Future VLSI DSP Systems

18.2 Signal Processing Chips and Applications
DSP Processors • Fixed-Point TMS320C25-Based Development System • Implementation of a Finite Impulse Response Filter with the TMS320C25 • Floating-Point TMS320C30-Based Development System • EVM Tools • Implementation of a Finite Impulse Response Filter with the TMS320C30 • FIR and IIR Implementation Using C and Assembly Code • Real-Time Applications • Conclusions and Future Directions

Keshab K. Parhi
University of Minnesota

Rulph Chassaing
Roger Williams University

Bill Bitler
InfiMed

18.1 Special Architectures

Keshab K. Parhi

Digital signal processing (DSP) is used in numerous applications. These applications include telephony, mobile radio, satellite communications, speech processing, video and image processing, biomedical applications, radar, and sonar. Real-time implementations of DSP systems require design of hardware that can match the application sample rate to the hardware processing rate (which is related to the clock rate and the implementation style). Thus, real-time does not always mean high speed. Real-time architectures are capable of processing samples as they are received from the signal source, as opposed to storing them in buffers for later processing as done in batch processing. Furthermore, real-time architectures operate on an infinite time series (since the number of the samples of the signal source is so large that it can be considered infinite). While speech and sonar applications require lower sample rates, radar and video image processing applications require much higher sample rates. The sample rate information alone cannot be used to choose the architecture. The algorithm complexity is also an important consideration. For example, a very complex and computationally intensive algorithm for a low-sample-rate application and a computationally simple algorithm for a high-sample-rate application may require similar hardware speed and complexity. These ranges of algorithms and applications motivate us to study a wide variety of architecture styles.

Using very large scale integration (VLSI) technology, DSP algorithms can be prototyped in many ways. These options include (1) single or multiprocessor programmable digital signal processors, (2) the use of core programmable digital signal processor with customized interface logic, (3) semicustom gate-array implementations, and (4) full-custom dedicated hardware implementation. The DSP algorithms are implemented in the programmable processors by translating the algorithm to the processor assembly code. This can require an extensive amount of time. On the other hand, high-level compilers for DSP can be used to generate the assembly code. Although this is currently feasible, the code generated by the compiler is not as efficient as hand-optimized code. Design of DSP compilers for generation of efficient code is still an active research topic. In the case of

dedicated designs, the challenge lies in a thorough understanding of the DSP algorithms and theory of architectures. For example, just minimizing the number of multipliers in an algorithm may not lead to a better dedicated design. The area saved by the number of multipliers may be offset by the increase in control, routing, and placement costs.

Off-the-shelf programmable digital signal processors can lead to faster prototyping. These prototyped systems can prove very effective in fast simulation of computation-intensive algorithms (such as those encountered in speech recognition, video compression, and seismic signal processing) or in benchmarking and standardization. After standards are determined, it is more useful to implement the algorithms using dedicated circuits.

Design of dedicated circuits is not a simple task. Dedicated circuits provide limited or no programming flexibility. They require less silicon area and consume less power. However, the low production volume, high design cost, and long turnaround time are some of the difficulties associated with the design of dedicated systems. Another difficulty is the availability of appropriate computer-aided design (CAD) tools for DSP systems. As time progresses, however, the architectural design techniques will be better understood and can be incorporated into CAD tools, thus making the design of dedicated circuits easier. Hierarchical CAD tools can integrate the design at various levels in an automatic and efficient manner. Implementation of standards for signal and image processing using dedicated circuits will lead to higher volume production. As time progresses, dedicated designs will be more acceptable to customers of DSP.

Successful design of dedicated circuits requires careful algorithm and architecture considerations. For example, for a filtering application, different equivalent realizations may possess different levels of concurrency. Thus, some of these realizations may be suitable for a particular application while other realizations may not be able to meet the sample rate requirements of the application. The lower-level architecture may be implemented in a word-serial or word-parallel manner. The arithmetic functional units may be implemented in bit-serial or digit-serial or bit-parallel manner. The synthesized architecture may be implemented with a dedicated data path or shared data path. The architecture may be systolic or nonsystolic.

Algorithm transformations play an important role in the design of dedicated architectures [Parhi, 1989]. This is because the transformed algorithms can be made to operate with better performance (where the performance may be measured in terms of speed, area, or power). Examples of these transformations include pipelining, parallel processing, retiming, unfolding, folding, look-ahead, associativity, and distributivity. These transformations and other architectural concepts are described in detail in subsequent sections.

Pipelining

Pipelining can increase the amount of concurrency (or the number of activities performed simultaneously) in an algorithm. Pipelining is accomplished by placing latches at appropriate intermediate points in a data flow graph that describes the algorithm. Each latch also refers to a storage unit or buffer or register. The latches can be placed at *feed-forward cutsets* of the data flow graph. In synchronous hardware implementations, pipelining can increase the clock rate of the system (and therefore the sample rate). The drawbacks associated with pipelining are the increase in system latency and the increase in the number of registers. To illustrate the speed increase using pipelining, consider the second-order three-tap finite impulse response (FIR) filter shown in Fig. 18.1(a). The signal $x(n)$ in this system can be sampled at a rate limited by the throughput of one multiplication and two additions. For simplicity, if we assume the multiplication time to be two times the addition time (T_{add}), the effective sample or clock rate of this system is $1/4T_{add}$. By placing latches as shown in Fig. 18.1(b) at the cutset shown in the dashed line, the sample rate can be improved to the rate of one multiplication or two additions. While pipelining can be easily applied to all algorithms with no feedback loops by the appropriate placement of latches, it cannot easily be applied to algorithms with feedback loops. This is because the cutsets in feedback algorithms contain feed-forward and feedback data flow and cannot be considered as feed-forward cutsets.

Pipelining can also be used to improve the performance in software programmable multiprocessor systems. Most software programmable DSP processors are programmed using assembly code. The assembly code is generated by high-level compilers that perform scheduling. Schedulers typically use the acyclic precedence graph to construct schedules. The removal of all edges in the signal (or data) flow graph containing delay

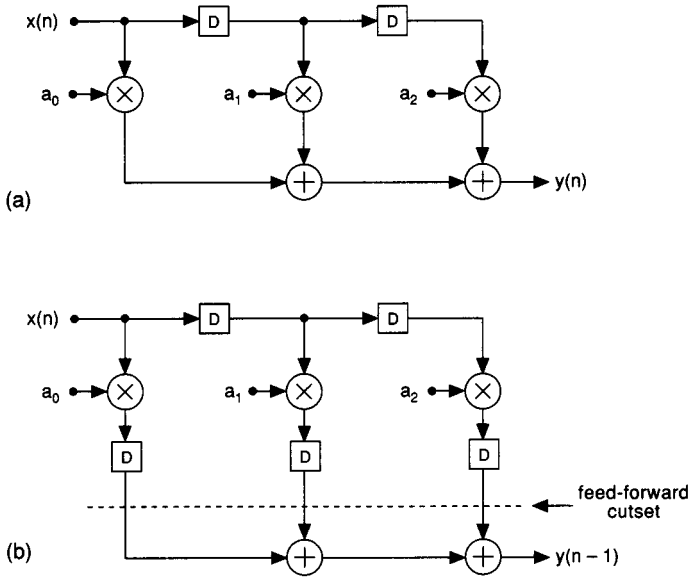


FIGURE 18.1 (a) A three-tap second-order nonrecursive digital filter; (b) the equivalent pipelined digital filter obtained by placing storage units at the intersection of the signal wires and the feed-forward cutset. If the multiplication and addition operations require 2 and 1 unit of time, respectively, then the maximum achievable sampling rates for the original and the pipelined architectures are 1/4 and 1/2 units, respectively.

elements converts the signal flow graph to an acyclic precedence graph. By placing latches to pipeline a data flow graph, we can alter the acyclic precedence graph. In particular, the critical path of the acyclic precedence graph can be reduced. The new precedence graph can be used to construct schedules with lower iteration periods (although this may often require an increase in the number of processors).

Pipelining of algorithms can increase the sample rate of the system. Sometimes, for a constant sample rate, pipelining can also reduce the power consumed by the system. This is because the data paths in the pipelined system can be charged or discharged with lower supply voltage. Since the capacitance remains almost constant, the power can be reduced. Achieving low power can be important in many battery-powered applications [Chandrakasan et al., 1992].

Parallel Processing

Parallel processing is related to pipelining but requires replication of hardware units. Pipelining exploits concurrency by breaking a large task into multiple smaller tasks and by separating these smaller tasks by storage units. On the other hand, parallelism exploits concurrency by performing multiple larger tasks simultaneously in separate hardware units.

To illustrate the speed increase due to parallelism, consider the parallel implementation of the second-order three-tap FIR filter of Fig. 18.1(a) shown in Fig. 18.2. In the architecture of Fig. 18.2, two input samples are processed and two output samples are generated in each clock cycle period of four addition times. Because each clock cycle processes two samples, however, the effective sample rate is $1/2T_{\text{add}}$ which is the same as that of Fig. 18.1(b). The parallel architecture leads to the speed increase with significant hardware overhead. The entire data flow graph needs to be replicated with an increase in the amount of parallelism. Thus, it is more desirable to use pipelining as opposed to parallelism. However, parallelism may be useful if pipelining alone cannot meet the speed demand of the application or if the technology constraints (such as limitations on the clock rate by the I/O technology) limit the use of pipelining. In obvious ways, pipelining and parallelism can be combined also. Parallelism, like pipelining, can also lead to power reduction but with significant overhead in hardware requirements. Achieving pipelining and parallelism can be difficult for systems with feedback loops. Concurrency may be created in these systems by using the look-ahead transformation.

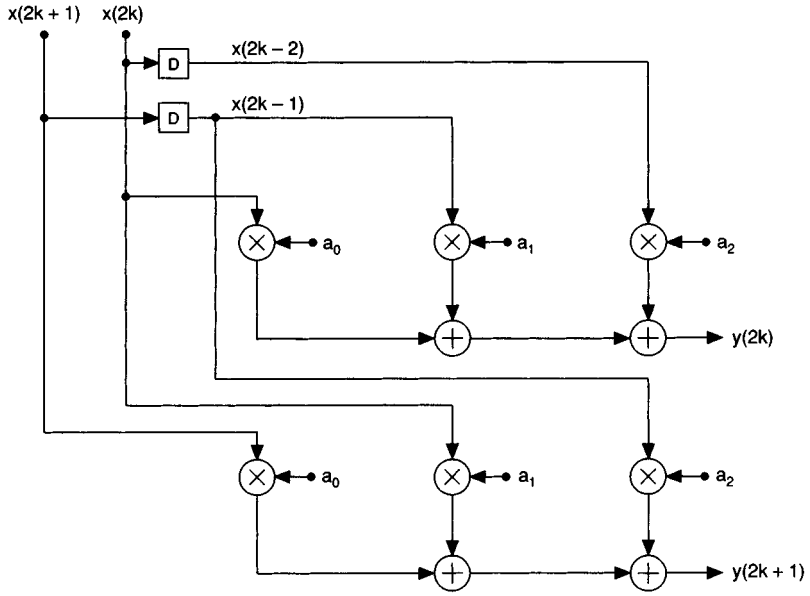


FIGURE 18.2 Twofold parallel realization of the three-tap filter of Fig. 18.1(a).

Retiming

Retiming is similar to pipelining but yet different in some ways [Leiserson et al., 1983]. Retiming is the process of moving the delays around in the data flow graph. Removal of one delay from all input edges of a node and insertion of one delay to each outgoing edge of the same node is the simplest example of retiming. Unlike pipelining, retiming does not increase the latency of the system. However, retiming alters the number of delay elements in the system. Retiming can reduce the critical path of the data flow graph. As a result, it can lead to clock period reduction in hardware implementations or critical path of the acyclic precedence graph or the iteration period in programmable software system implementations.

The single host formulation of the retiming transformation preserves the latency of the algorithm. The retiming formulation with no constraints on latency (i.e., with separate input and output hosts) can also achieve *pipelining with no retiming* or *pipelining with retiming*. Pipelining with retiming is the most desirable transformation in DSP architecture design. Pipelining with retiming can be interpreted to be identical to retiming of the original algorithm with a large number of delays at the input edges. Thus, we can increase the system latency arbitrarily and remove the appropriate number of delays from the inputs after the transformation.

The retiming formulation assigns retiming variables $r(.)$ to each node in the data flow graph. If $i(U \rightarrow V)$ is the number of delays associated with the edge $U \rightarrow V$ in the original data flow graph and $r(V)$ and $r(U)$, respectively, represent the retiming variable value of the nodes V and U, then the number of delays associated with the edge $U \rightarrow V$ in the retimed data flow graph is given by

$$i_r(U \rightarrow V) = i(U \rightarrow V) + r(V) - r(U)$$

For the data flow graph to be realizable, $i_r(U \rightarrow V) \geq 0$ must be satisfied. The retiming transformation formulates the problem by calculating path lengths and by imposing constraints on certain path lengths. These constraints are solved as a shortest-path problem.

To illustrate the usefulness of retiming, consider the data flow graph of a two-stage pipelined lattice digital filter graph shown in Fig. 18.3(a) and its equivalent pipelined-retimed data flow graph shown in Fig. 18.3(b). If the multiply time is two units and the add time is one unit, the architecture in Fig. 18.3(a) can be clocked with period 10 units whereas the architecture in Fig. 18.3(b) can be clocked with period 2 units.

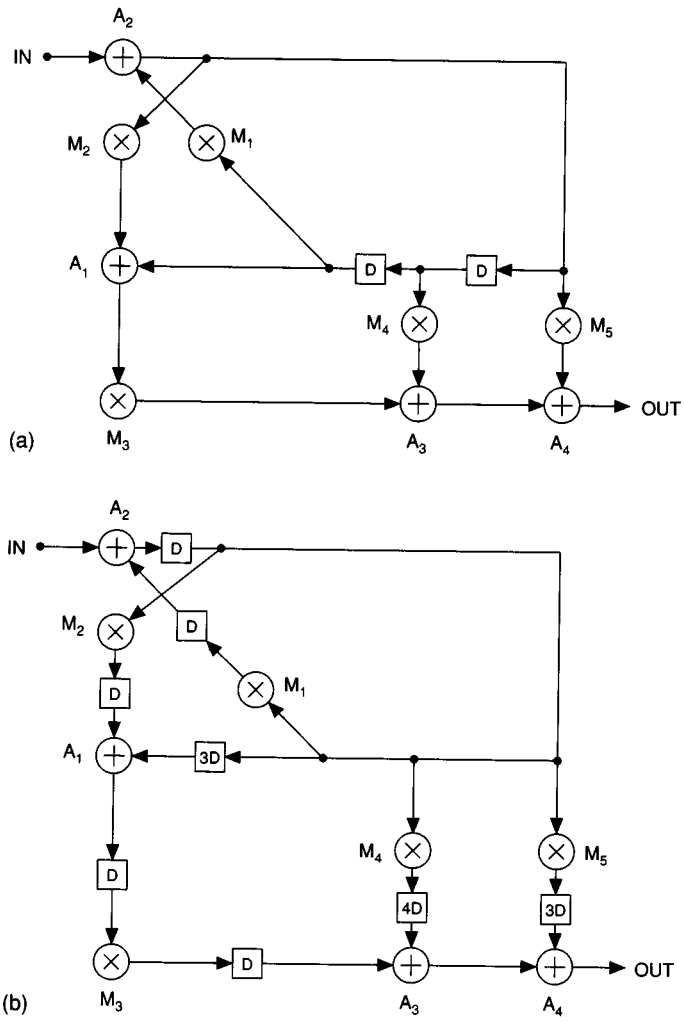


FIGURE 18.3 (a) A two-stage pipelinable time-invariant lattice digital filter. If multiplication and addition operations require 2 and 1 time units, respectively, then this data flow graph can achieve a sampling period of 10 time units (which corresponds to the critical path $M_1 \rightarrow A_2 \rightarrow M_2 \rightarrow A_1 \rightarrow M_3 \rightarrow A_3 \rightarrow A_4$). (b) The pipelined/retimed lattice digital filter can achieve a sampling period of 2 time units.

Unfolding

The **unfolding** transformation is similar to loop unrolling. In *J*-unfolding, each node is replaced by *J* nodes and each edge is replaced by *J* edges. The *J*-unfolded data flow graph executes *J* iterations of the original algorithm [Parhi, 1991].

The unfolding transformation can unravel the hidden concurrency in a data flow program. The achievable iteration period for a *J*-unfolded data flow graph is $1/J$ times the critical path length of the unfolded data flow graph. By exploiting interiteration concurrency, unfolding can lead to a lower iteration period in the context of a software programmable multiprocessor implementation.

The unfolding transformation can also be applied in the context of hardware design. If we apply an unfolding transformation on a (word-serial) nonrecursive algorithm, the resulting data flow graph represents a **word-parallel** (or simply parallel) algorithm that processes multiple samples or words in parallel every clock cycle. If we apply 2-unfolding to the 3-tap FIR filter in Fig. 18.1(a), we can obtain the data flow graph of Fig. 18.2.

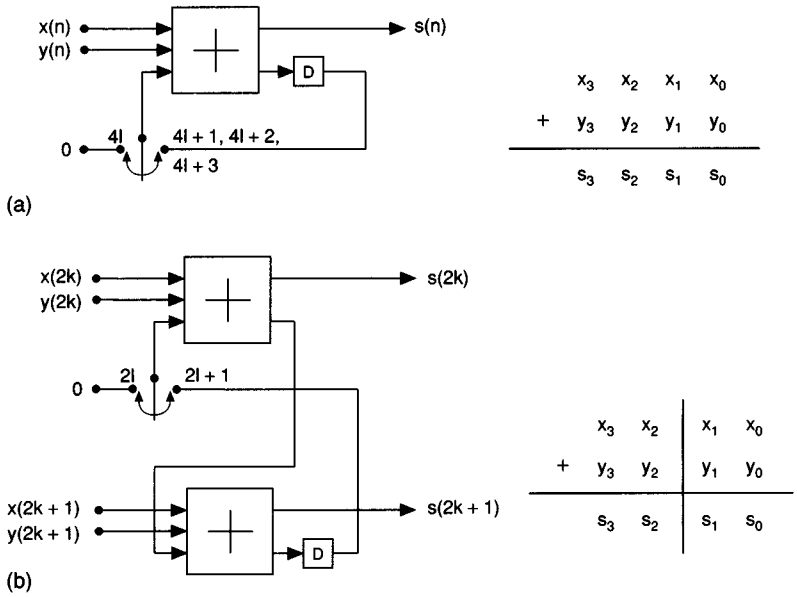


FIGURE 18.4 (a) A least-significant-bit first bit-serial adder for word length of 4; (b) a digit-serial adder with digit size 2 obtained by two-unfolding of the bit-serial adder. The bit position 0 stands for least significant bit.

Because the unfolding algorithm is based on graph theoretic approach, it can also be applied at the bit level. Thus, unfolding of a **bit-serial** data flow program by a factor of J leads to a **digit-serial** program with digit size J. The *digit size* represents the number of bits processed per clock cycle. The digit-serial architecture is clocked at the same rate as the bit-serial (assuming that the clock rate is limited by the communication I/O bound much before reaching the computation bound of the bit-serial program). Because the digit-serial program processes J bits per clock cycle the effective bit rate of the digit-serial architecture is J times higher. A simple example of this unfolding is illustrated in Fig. 18.4, where the bit-serial adder in Fig. 18.4(a) is unfolded by a factor of 2 to obtain the digit-serial adder in Fig. 18.4(b) for digit size 2 for a word length of 4. In obvious ways, the unfolding transformation can be applied to both word level and bit level simultaneously to generate word-parallel digit-serial architectures. Such architectures process multiple words per clock cycle and process a digit of each word (not the entire word).

Folding Transformation

The **folding** transformation is the reverse of the unfolding transformation. While the unfolding transformation is simpler, the folding transformation is more difficult [Parhi et al., 1992].

The folding transformation can be applied to fold a bit-parallel architecture to a digit-serial or bit-serial one or to fold a digit-serial architecture to a bit-serial one. It can also be applied to fold an algorithm data flow graph to a hardware data flow for a specified folding set. The folding set indicates the processor in which and the time partition at which a task is executed. A specified folding set may be infeasible, and this needs to be detected first. The folding transformation performs a preprocessing step to detect feasibility and in the feasible case transforms the algorithm data flow graph to an equivalent pipelined/retimed data flow graph that can be folded. For the special case of regular data flow graphs and for linear space–time mappings, the folding tranformation reduces to **systolic** array design.

In the folded architecture, each edge in the algorithm data flow graph is mapped to a communicating edge in the hardware architecture data flow graph. Consider an edge $U \rightarrow V$ in the algorithm data flow graph with associated number of delays $i(U \rightarrow V)$. Let the tasks U and V be mapped to the hardware units H_U and H_V respectively. Assume that N time partitions are available, i.e., the iteration period is N. A modulo operation determines the time partition. For example, the time unit 18 for $N = 4$ corresponds to time partition 18 modulo

4 or 2. Let the tasks U and V be executed in time partitions u and v, i.e., the lth iterations of tasks U and V are executed in time units $Nl + u$ and $Nl + v$, respectively. The $i(U \rightarrow V)$ delays in the edge $U \rightarrow V$ implies that the result of the lth iteration of U is used for the $(l + i)$th iteration of V. The $(l + i)$th iteration of V is executed in time unit $N(l + i) + v$. Thus the number of storage units needed in the folded edge corresponding to the edge $U \rightarrow V$ is

$$D_F(U \rightarrow V) = N(l + i) + v - Nl - u - P_u = Ni + v - u - P_u$$

where P_u is the level of pipelining of the hardware operator H_U. The $D_F(U \rightarrow V)$ delays should be connected to the edge between H_U and H_V and this signal should be switched to the input of H_V at time partition v. If the $D_F(U \rightarrow V)$'s as calculated here were always nonnegative for all edges $U \rightarrow V$, then the problem would be solved. However, some $D_F()$'s would be negative. The algorithm data flow graph needs to be pipelined and retimed such that all the $D_F()$'s are nonnegative. This can be formulated by simple inequalities using the retiming variables. The retiming formulation can be solved as a path problem, and the retiming variables can be determined if a solution exists. The algorithm data flow graph can be retimed for folding and the calculation of the $D_F()$'s can be repeated. The folded hardware architecture data flow graph can now be completed. The folding technique is illustrated in Fig. 18.5. The algorithm data flow graph of a two-stage pipelined lattice recursive digital filter of Fig. 18.3(a) is folded for the folding set shown in Fig. 18.5. Fig. 18.5(a) shows the pipelined/retimed data flow graph (preprocessed for folding) and Fig. 18.5(b) shows the hardware architecture data flow graph obtained after folding.

As indicated before, a special case of folding can address systolic array design for regular data flow graphs and for linear mappings. The systolic architectures make use of extensive pipelining and local communication and operate in a synchronous manner [Kung, 1988]. The systolic processors can also be made to operate in an asynchronous manner, and such systems are often referred to as wavefront processors. Systolic architectures have been designed for a variety of applications including convolution, matrix solvers, matrix decomposition, and filtering.

Look-Ahead Technique

The **look-ahead** technique is a very powerful technique for pipelining of recursive signal processing algorithms [Parhi and Messerschmitt, 1989]. This technique can transform a sequential recursive algorithm to an equivalent concurrent one, which can then be realized using pipelining or parallel processing or both. This technique has been successfully applied to pipeline many signal processing algorithms, including recursive digital filters (in direct form and lattice form), adaptive lattice digital filters, two-dimensional recursive digital filters, Viterbi decoders, Huffman decoders, and finite state machines. This research demonstrated that the recursive signal processing algorithms can be operated at high speed. This is an important result since modern signal processing applications in radar and image processing and particularly in high-definition and super-high-definition television video signal processing require very high throughput. Traditional algorithms and topologies cannot be used for such high-speed applications because of the inherent speed bound of the algorithm created by the feedback loops. The look-ahead transformation creates additional concurrency in the signal processing algorithms and the speed bound of the transformed algorithms is increased substantially. The look-ahead transformation is not free from its drawbacks. It is accompanied by an increase in the hardware overhead. This difficulty has encouraged us to develop inherently pipelinable topologies for recursive signal processing algorithms. Fortunately, this is possible to achieve in adaptive digital filters using relaxations on the look-ahead or by the use of relaxed look-ahead [Shanbhag and Parhi, 1992].

To begin, consider a time-invariant one-pole recursive digital filter transfer function

$$H(z) = \frac{X(z)}{U(z)} = \frac{1}{1 - az^{-1}}$$

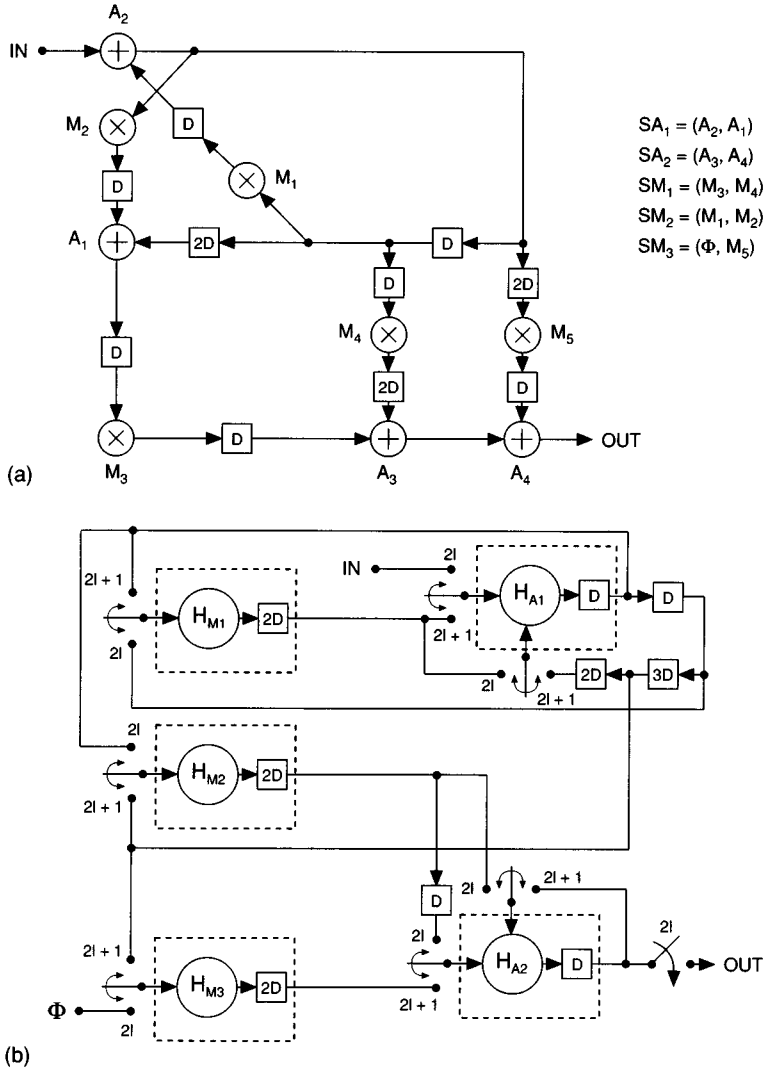


FIGURE 18.5 (a) A pipelined/retimed data flow graph obtained from Fig. 18.3(a) by preprocessing for folding; (b) the folded hardware architecture data flow graph. In our folding notation, the tasks are ordered within a set and the ordering represents the time partition in which the task is executed. For example, $SA_1 = (A_2, A_1)$ implies that A_2 and A_1 are, respectively, executed in even and odd time partitions in the same processor. The notation **Φ** represents a null operation.

described by the difference equation

$$x(n) = ax(n-1) + u(n)$$

and shown in Fig. 18.6(a). The maximum achievable speed in this system is limited by the operating speed of one multiply–add operation. To increase the speed of this system by a factor of 2, we can express $x(n)$ in terms of $x(n-2)$ by substitution of one recursion within the other:

$$x(n) = a[ax(n-2) + u(n-1)] + u(n) = a^2x(n-2) + au(n-1) + u(n)$$

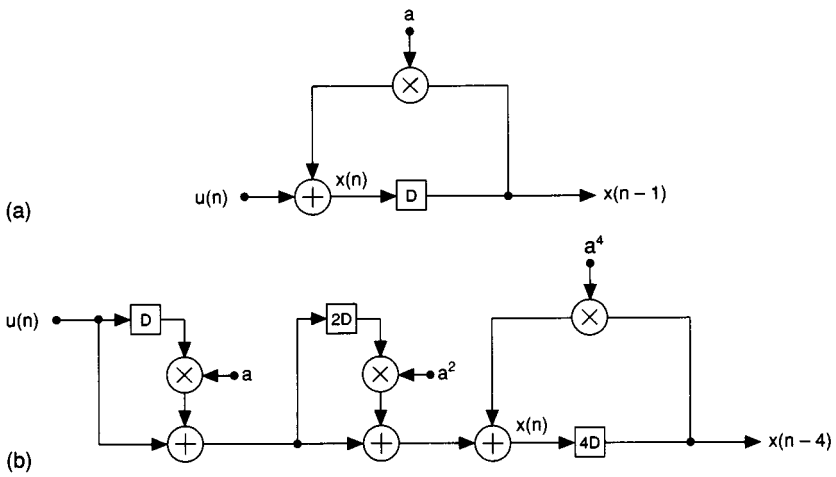


FIGURE 18.6 (a) A first-order recursive digital filter; (b) a four-stage pipelinable equivalent filter obtained by look-ahead computation.

The transfer function of the emulated second-order system is given by

$$H(z) = \frac{1 + az^{-1}}{1 - a^2 z^{-2}}$$

and is obtained by using a pole-zero cancellation at $-a$. In the modified system, $x(n)$ is computed using $x(n-2)$ as opposed to $x(n-1)$; thus we *look ahead*. The modified system has two delays in the multiply–add feedback loop, and these two delays can be distributed to pipeline the multiply–add operation by two stages. Of course, the additional multiply–add operation that represents one zero would also need to be pipelined by two stages to keep up with the sample rate of the system. To increase the speed by four times, we can rewrite the transfer function as:

$$H(z) = \frac{(1 + az^{-1})(1 + a^2 z^{-2})}{(1 - a^4 z^{-4})}$$

This system is shown in Fig. 18.6(b). Arbitrary speed increase is possible. However, for power-of-two speed increase the hardware complexity grows logarithmically with speed-up factor. The same technique can be applied to any higher-order system. For example, a second-order recursive filter with transfer function

$$H(z) = \frac{1}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}}$$

can be modified to

$$H(z) = \frac{1 + 2r\cos\theta z^{-1} + r^2 z^{-2}}{1 - 2r^2\cos 2\theta z^{-2} + r^4 z^{-4}}$$

for a twofold increase in speed. In this example, the output $y(n)$ is computed using $y(n-2)$ and $y(n-4)$; thus, it is referred to as *scattered look-ahead*.

While look-ahead can transform any recursive digital filter transfer function to pipelined form, it leads to a hardware overhead proportional to $N \log_2 M$, where N is the filter order and M is the speed-up factor. Instead of starting with a sequential digital filter transfer function obtained by traditional design approaches and transforming it for pipelining, it is more desirable to use a constrained filter design program that can satisfy the filter spectrum and the pipelining constraint. The pipelining constraint is satisfied by expressing the denominator of the transfer function in scattered look-ahead form. Such filter design programs have now been developed in both time domain and frequency domain. The advantage of the constrained filter design approach is that we can obtain pipelined digital filters with marginal or zero hardware overhead compared with sequential digital filters. The pipelined transfer functions can also be mapped to pipelined lattice digital filters. The reader might note that the data flow graph of Fig. 18.3(a) was obtained by this approach.

The look-ahead pipelining can also be applied for the design of transversal and adaptive lattice digital filters. Although look-ahead transformation can be used to modify the adaptive filter recursions to create concurrency, this requires large hardware overhead. The adaptive filters are based on weight update operations, and the weights are adapted based on the current error. Finally, the error becomes close to zero and the filter coefficients have been adapted. Thus, making relaxations on the error can reduce the hardware overhead substantially without degradation of the convergence behavior of the adaptive filter. Three types of relaxations of look-ahead are possible. These are referred to as *sum relaxation, product relaxation,* and *delay relaxation.* To illustrate these three relaxations, consider the weight update recursion

$$w(n+1) = a(n)w(n) + f(n)$$

where the term $a(n)$ is typically 1 for transversal least mean square (LMS) adaptive filters and of the form $(1 - \varepsilon(n))$ for lattice LMS adaptive digital filters, and $f(n) = \mu e(n)u(n)$ where μ is a constant, $e(n)$ is the error, and $u(n)$ is the input. The use of look-ahead transforms the above recursion to

$$w(n+M) = \prod_{i=0}^{M-1} a(n+M-i-1) \quad w(n)$$

$$+ \left[1a(n+M-1) \prod_{i=0}^{1} a(n+M-i-1) \ldots \prod_{i=0}^{M-2} a(n+M-i-1) \right] \begin{bmatrix} f(n+M-1) \\ f(n+M-2) \\ \cdot \\ \cdot \\ \cdot \\ f(n) \end{bmatrix}$$

In sum relaxation, we only retain the single term dependent on the current input for the last term of the look-ahead recursion. The relaxed recursion after sum relaxation is given by

$$w(n+M) = \prod_{i=0}^{M-1} a(n+M-i-1) \quad w(n) + f(n+M-1)$$

In lattice digital filters, the coefficient $a(n)$ is close to 1 for all n, since it can be expressed as $(1 - \varepsilon(n))$ and $\varepsilon(n)$ is close to zero for all n and is positive. The product relaxation on the above equation leads to

$$w(n+M) = (1 - M\varepsilon(n+M-1))\, w(n) + f(n+M-1)$$

The delay relaxation assumes the signal to be slowly varying or to be constant over D samples and replaces the look-ahead by

$$w(n+M) = (1 - M\varepsilon(n+M-1))\ w(n) + f(n+M-D-1)$$

These three types of relaxations make it possible to implement pipelined transversal and lattice adaptive digital filters with marginal increase in hardware overhead. Relaxations on the weight update operations change the convergence behavior of the adaptive filter, and we are forced to examine carefully the convergence behavior of the relaxed look-ahead adaptive digital filters. It has been shown that the relaxed look-ahead adaptive digital filters do not suffer from degradation in adaptation behavior. Futhermore, when coding, the use of pipelined adaptive filters could lead to a dramatic increase in pixel rate with no degradation in signal-to-noise ratio of the coded image and no increase in hardware overhead [Shanbhag and Parhi, 1992].

The concurrency created by look-ahead and relaxed look-ahead transformations can also be exploited in the form of parallel processing. Furthermore, for a constant speed, concurrent architectures (especially the pipelined architectures) can also lead to low power consumption.

Associativity Transformation

The addition operations in many signal processing algorithms can be interchanged since the add operations satisfy associativity. Thus, it is possible to move the add operations outside the critical loops to increase the maximum achievable speed of the system. As an example of the associative transformation, consider the realization of a second-order recursion $x(n) = 5/8x(n-1) - 3/4x(n-2) + u(n)$. Two possible realizations are shown in Fig. 18.7(a). The realization on the left contains one multiplication and two add operations in the critical inner loop, whereas the realization on the right contains one multiplication and one add operation in the critical inner loop. The realization on the left can be transformed to the realization on the right using the associativity transformation. Figure 18.7(b) shows a bit-serial implementation of this second-order recursion for the realization on the right for a word length of 8. This bit-serial system can be operated in a functionally correct manner for any word length greater than or equal to 5 since the inner loop computation latency is 5 cycles. On the other hand, if associativity were not exploited, then the minimum realizable word length would be 6. Thus, associativity can improve the achievable speed of the system.

Distributivity

Another local transformation that is often useful is distributivity. In this transformation, a computation $(A \times B) + (A \times C)$ may be reorganized as $A \times (B + C)$. Thus, the number of hardware units can be reduced from two multipliers and one adder to one multiplier and one adder.

Arithmetic Processor Architectures

In addition to algorithms and architecture designs, it is also important to address implementation styles and arithmetic processor architectures.

Most DSP systems use fixed-point hardware arithmetic operators. While many number system representations are possible, the two's complement number system is the most popular number system. The other number systems include the residue number system, the redundant or signed-digit number system, and the logarithmic number system. The residue and logarithmic number systems are rarely used or are used in very special cases such as nonrecursive digital filters. Shifting or scaling and division are difficult in the residue number system. Difficulty with addition and the overhead associated with logarithm and antilogarithm converters reduce the attractiveness of the logarithm number system. The use of the redundant number system leads to carry-free operation but is accompanied by the overhead associated with redundant-to-two's complement conversion. Another approach often used is distributed arithmetic. This approach has recently been used in a few video transformation chips.

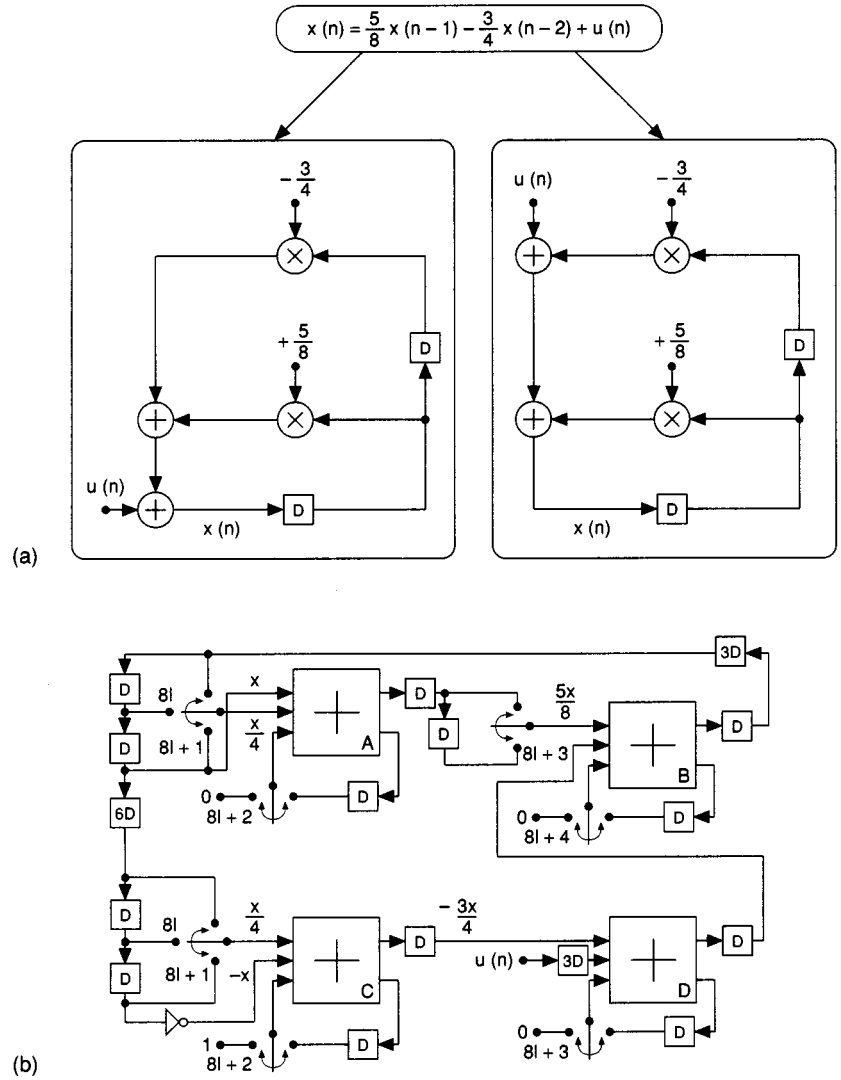


FIGURE 18.7 (a) Two associative realizations of a second-order recursion; (b) an efficient bit-serial realization of the recursion for a word length of 8.

The simplest arithmetic operation is addition. Multiplication can be realized as a series of add-shift operations, and division and square-root can be realized as a series of controlled add–subtract operations. The conventional two's complement adder involves carry ripple operation. This limits the throughput of the adder operation. In DSP, however, the combined multiply–add operation is most common. Carry–save operations have been used to realize pipelined multiply-adders using fewer pipelining latches. In conventional pipelined two's complement multiplier, the multiplication time is approximately two times the bit-level addition time. Recently, a technique has been proposed to reduce the multiplication time from $2W$ bit-level binary adder times to $1.25W$ bit-level binary adder times where W is the word length. This technique is based on the use of hybrid number system representation, where one input operand is in two's complement number representation and the other in redundant number representation [Srinivas and Parhi, 1992]. Using an efficient sign-select redundant-to-two's complement conversion technique, this multiplier can be made to operate faster and, in the pipelined mode, would require fewer pipelining latches and less silicon area.

Computer-Aided Design

With progress in the theory of architectures, the computer-aided design (CAD) systems for DSP application also become more powerful. In early 1980, the first silicon compiler system for signal processing was developed at the University of Edinburgh and was referred to as the FIRST design system. This system only addressed the computer-aided design of bit-serial signal processing systems. Since then more powerful systems have been developed. The Cathedral I system from Katholieke Universiteit Leuven and the BSSC (bit-serial silicon compiler) from GE Research Center in Schenectady, New York, also addressed synthesis of bit-serial circuits. The Cathedral system has now gone through many revisions, and the new versions can systhesize parallel multiprocessor data paths and can perform more powerful scheduling and allocation. The Lager design tool at the University of California at Berkeley was developed to synthesize the DSP algorithms using parametrizable macro building blocks (such as ALU, RAM, ROM). This system has also gone through many revisions. The Hyper system also developed at the University of California at Berkeley and the MARS design system developed at the University of Minnesota at Minneapolis perform higher level transformations and perform scheduling and allocation. These CAD tools are crucial to rapid prototyping of high-performance DSP integrated circuits.

Future VLSI DSP Systems

Future VLSI systems will make use of a combination of many types of architectures such as dedicated and programmable. These systems can be designed successfully with proper understanding of the algorithms, applications, theory of architectures, and with the use of advanced CAD systems.

Defining Terms

Bit serial: Processing of one bit per clock cycle. If word length is W, then one sample or word is processed in W clock cycles. In contrast, all W bits of a word are processed in the same clock cycle in a bit-parallel system.

Digit serial: Processing of more than one but not all bits in one clock cycle. If the digit size is W_1 and the word length is W, then the word is processed in W/W_1 clock cycles. If $W_1 = 1$, then the system is referred to as a bit-serial and if $W_1 = W$, then the system is referred to as a bit-parallel system. In general, the digit size W_1 need not be a divisor of the word length W, since the least and most significant bits of consecutive words can be overlapped and processed in the same clock cycle.

Folding: The technique of mapping many tasks to a single processor.

Look-ahead: The technique of computing a state $x(n)$ usng previous state $x(n - M)$ without requiring the intermediate states $x(n - 1)$ through $x(n - M + 1)$. This is referred to as a M-step look-ahead. In the case of higher-order computations, there are two forms of look-ahead: clustered look-ahead and scattered look-ahead. In clustered look-ahead, $x(n)$ is computed using the clustered states $x(n - M - N + 1)$ through $x(n - M)$ for an Nth order computation. In scattered look-ahead, $x(n)$ is computed using the scattered states $x(n - iM)$ where i varies from 1 to N.

Parallel processing: Processing of multiple tasks independently by different processors. This also increases the throughput.

Pipelining: A technique to increase throughput. A long task is divided into components, and each component is distributed to one processor. A new task can begin even though the former tasks have not been completed. In the pipelined operation, different components of different tasks are executed at the same time by different processors. Pipelining leads to an increase in the system latency, i.e., the time elapsed between the starting of a task and the completion of the task.

Retiming: The technique of moving the delays around the system. Retiming does not alter the latency of the system.

Systolic: Flow of data in a rhythmic fashion from a memory through many processors, returning to the memory just as blood flows

Unfolding: The technique of transforming a program that describes one iteration of an algorithm to another equivalent program that describes multiple iterations of the same algorithm.

Word parallel: Processing of multiple words in the same clock cycle.

Related Topic

95.1 Introduction

References

A.P. Chandrakasan, S. Sheng, and R.W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid State Circuits*, vol. 27(4), pp. 473–484, April 1992.

S.Y. Kung, *VLSI Array Processors*, Englewood Cliffs, N.J.: Prentice-Hall, 1988.

E.A. Lee and D.G. Messerschmitt, "Pipeline interleaved programmable DSP's," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 35(9), pp. 1320–1345, September 1987.

C.E. Leiserson, F. Rose, and J. Saxe, "Optimizing synchronous circuitry by retiming," *Proc. 3rd Caltech Conf. VLSI,* Pasadena, Calif., pp. 87–116, March 1983.

K.K. Parhi, "Algorithm transformation techniques for concurrent processors," *Proc. IEEE*, vol. 77(12), pp. 1879–1895, December 1989.

K.K. Parhi, "Systematic approach for design of digit-serial processing architectures," *IEEE Trans. Circuits Systems*, vol. 38(4), pp. 358–375, April 1991.

K.K. Parhi and D.G. Messerschmitt, "Pipeline interleaving and parallelism in recursive digital filters," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37(7), pp. 1099–1135, July 1989.

K.K. Parhi, C.Y. Wang, and A.P. Brown, "Synthesis of control circuits in folded pipelined DSP architectures," *IEEE J. Solid State Circuits*, vol. 27(1), pp. 29–43, January 1992.

N.R. Shanbhag, and K.K. Parhi, "A pipelined adaptive lattice filter architecture," *Proc. 1992 IEEE Int. Symp. Circuits and Systems*, San Diego, May 1992.

H.R. Srinivas and K.K. Parhi, "High-speed VLSI arithmetic processor architectures using hybrid number representation," *J. VLSI Signal Processing*, vol. 4(2/3), pp. 177–198, 1992.

Further Information

A detailed video tutorial on "Implementation and Synthesis of VLSI Signal Processing Systems" presented by K.K. Parhi and J.M. Rabaey in March 1992 can be purchased from the customer service department of IEEE, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

Special architectures for video communications can be found in the book *VLSI Implementations for Image Communications*, published as the fourth volume of the series *Advances in Image Communications* (edited by Peter Pirsch) by the Elsevier Science Publishing Co. in 1993. The informative article "Research on VLSI for Digital Video Systems in Japan," published by K.K. Parhi in the fourth volume of the *1991 Office of Naval Research Asian Office Scientific Information Bulletin* (pages 93–98), provides examples of video codec designs using special architectures. For video programmable digital signal processor approaches, see I. Tamitani, H. Harasaki, and T. Nishitani, "A Real-Time HDTV Signal Processor: HD-VSP," published in *IEEE Transactions on Circuits and Systems for Video Technology*, March 1991, and T. Fujii, T. Sawabe, N. Ohta, and S. Ono, "Implementation of Super High-Definition Image Processing on HiPIPE," published in *1991 IEEE International Symposium on Circuits and Systems,* held in June 1991 in Singapore (pages 348–351).

The *IEEE Design and Test of Computers* published three special issues related to computer-aided design of special architectures; these issues were published in October 1990 (addressing high-level synthesis), December 1990 (addressing silicon compilations), and June 1991 (addressing rapid prototyping).

Descriptions of various CAD systems can be found in the following references. The description of the FIRST system can be found in the article "A Silicon Compiler for VLSI Signal Processing," by P. Denyer et al. in the *Proceedings of the ESSCIRC* conference held in Brussels in September 1982 (pages 215–218). The Cathedral system has been described in R. Jain et al., "Custom Design of a VLSI PCM-FDM Transmultiplexor from System Specifications to Circuit Layout Using a Computer Aided Design System," published in *IEEE Journal of Solid State Circuits* in February 1986 (pages 73–85). The Lager system has been described in "An Integrated Automatic Layout Generation System for DSP Circuits," by J. Rabaey, S. Pope, and R. Brodersen, published in the July 1985 issue of the *IEEE Transactions on Computer Aided Design* (pages 285–296). The description of the MARS Design System can be found in C.-Y. Wang and K.K. Parhi, "High-Level DSP Synthesis Using MARS System,"

published in *Proceedings of the 1992 IEEE International Symposium on Circuits and Systems* in San Diego, May 1992. A tutorial article on high-level synthesis can be found in "The High-Level Synthesis of Digital Systems," by M.C. McFarland, A. Parker, and R. Composano, published in the February 1990 issue of the *Proceedings of the IEEE* (pages 310–318).

Articles on pipelined multipliers can be found in T.G. Noll et al., "A Pipelined 330 MHZ Multiplier," *IEEE Journal of Solid State Circuits*, June 1986 (pages 411–416) and in M. Hatamian and G. Cash, "A 70-MHz 8-Bit × 8-Bit-Parallel Pipelined Multiplier in 2.5 μm CMOS," *IEEE Journal of Solid State Circuits*, 1986.

Technical articles on special architectures and chips for signal and image processing appear at different places, including proceedings of conferences such as IEEE Workshop on VLSI Signal Processing, IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE International Symposium on Circuits and Systems, IEEE International Solid State Circuits Conference, IEEE Customs Integrated Circuits Conference, IEEE International Conference on Computer Design, ACM/IEEE Design Automation Conference, ACM/IEEE International Conference on Computer Aided Design, International Conference on Application Specific Array Processors, and journals such as *IEEE Transactions on Signal Processing, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems: Part II: Analog and Digital Signal Processing, IEEE Transactions on Computers, IEEE Journal of Solid State Circuits, IEEE Signal Processing Magazine, IEEE Design and Test Magazine,* and *Journal of VLSI Signal Processing.*

18.2 Signal Processing Chips and Applications

Rulph Chassaing and Bill Bitler

Recent advances in very large scale integration (VLSI) have contributed to the current digital signal processors. These processors are just special-purpose fast microprocessors characterized by architectures and instructions suitable for real-time digital signal processing (DSP) applications. The commercial DSP processor, a little more than a decade old, has emerged because of the ever-increasing number of signal processing applications. DSP processors are now being utilized in a number of applications from communications and controls to speech and image processing. They have found their way into talking toys and music synthesizers. A number of texts [such as Chassaing, 1992] and articles [such as Ahmed and Kline, 1991] have been written, discussing the applications that use DSP processors and the recent advances in DSP systems.

DSP Processors

Digital signal processors are currently available from a number of companies, including Texas Instruments, Inc. (Texas), Motorola, Inc. (Arizona), Analog Devices, Inc. (Massachusetts), AT&T (New Jersey), and NEC (California). These processors are categorized as either **fixed-point** or **floating-point processors.** Several companies are now supporting both types of processors. **Special-purpose digital signal processors,** designed for a specific signal processing application such as for fast Fourier transform (FFT), have also emerged. Currently available digital signal processors range from simple, low cost processing units through high performance units such as Texas Instruments' (TI) TMS320C40 (Chassaing and Martin, 1995) and TMS320C80, and Analog Devices[1] ADSP-21060 SHARC (Chassaing and Ayers, 1996).

One of the first-generation digital signal processors is the (N-MOS technology) TMS32010, introduced by Texas Instruments in 1982. This first-generation fixed-point processor is based on the Harvard architecture, with a fast on-chip hardware multiplier/accumulator, and with data and instructions in separate memory spaces, allowing for concurrent accesses. This type of **pipelining feature** enables the processor to execute one instruction while fetching at the same time the next instruction. Other features include 144 (16-bit) words of on-chip data RAM and a 16-bit by 16-bit multiply operation in one instruction cycle time of 200 ns. Since many instructions can be executed in one single cycle, the TMS32010 is capable of executing 5 million instructions per second (MIPS). Major drawbacks of this first-generation processor are its limited **on-chip memory** size and much slower execution time for accessing external memory. Improved versions of this first-generation processor are now available in C-MOS technology, with a faster instruction cycle time of 160 ns.

The second-generation fixed-point processor TMS32020, introduced in 1985 by TI, was quickly followed by an improved C-MOS version TMS320C25 [Chassaing and Horning, 1990] in 1986. Features of the TMS320C25 include 544 (16-bit) words of on-chip data RAM, separate program and data memory spaces (each 64 K words), and an instruction cycle time of 100 ns, enabling the TMS320C25 to execute 10 MIPS. A faster version, TI's fixed-point TMS320C50 processor, is available with an instruction cycle time of 35 ns.

The third-generation TMS320C30 (by TI) supports fixed- as well as floating-point operations [Chassaing, 1992]. Features of this processor include 32-bit by 32-bit floating-point multiply operations in one instruction cycle time of 60 ns. Since a number of instructions, such as load and store, multiply and add, can be performed in parallel (in one cycle time), the TMS320C30 can execute a pair of parallel instructions in 30 ns, allowing for 33.3 MIPS. The Harvard-based architecture of the fixed-point processors was abandoned for one allowing four levels of pipelining with three subsequent instructions being consequently fetched, decoded, and read while the current instruction is being executed. The TMS320C30 has 2 K words of on-chip memory and a total of 16 million words of addressable memory spaces for program, data, and input/output. Specialized instructions are available to make common DSP algorithms such as filtering and spectral analysis execute fast and efficiently. The architecture of the TMS320C30 was designed to take advantage of higher-level languages such as C and ADA. The TMS320C31 and TMS320C32, recent members of the third-generation floating-point processors, are available with a 40 ns instruction cycle time.

DSP starter kits (DSK) are inexpensive development systems available from TI and based on both the fixed-point TMS320C50 and the floating-point TMS320C31 processors. We will discuss both the fixed-point TMS320C25 and the floating-point TMS320C30 digital signal processors, including the development tools available for each of these processors and DSP applications.

Fixed-Point TMS320C25-Based Development System

TMS320C25-based development systems are now available from a number of companies such as Hyperception Inc., Texas, and Atlanta Signal Processors, Inc., Georgia. The Software Development System (SWDS), available from TI includes a board containing the TMS320C25, which plugs into a slot on an IBM compatible PC. Within the SWDS environment, a program can be developed, assembled, and run. Debugging aids supported by the SWDS include single-stepping, setting of breakpoints, and display/modification of registers.

A typical workstation consists of:

1. An IBM compatible PC. Commercially available DSP packages (such as from Hyperception or Atlanta Signal Processors) include a number of utilities and filter design techniques.
2. The SWDS package, which includes an assembler, a linker, a debug monitor, and a **C compiler.**
3. Input/output alternatives such as TI's analog interface board (AIB) or analog interface chip (AIC).

The AIB includes a 12-bit analog-to-digital converter (ADC) and a 12-bit digital-to-analog converter (DAC). A maximum sampling rate of 40 kHz can be obtained. With (input) antialiasing and (output) reconstruction filters mounted on a header on the AIB, different input/output (I/O) filter bandwidths can be achieved. Instructions such as **IN** and **OUT** can be used for input/output accesses. The AIC, which provides an inexpensive I/O alternative, includes 14-bit ADC and DAC, antialiasing/reconstruction filters, all on a single C-MOS chip. Two inputs and one output are available on the AIC. (A TMS320C25/AIC interface diagram and communication routines can be found in Chassaing and Horning, 1990.) The TLC32047 AIC is a recent member of the TLC32040 family of voiceband analog interface circuits, with a maximum sampling rate of 25 kHz.

Implementation of a Finite Impulse Response Filter with the TMS320C25

The convolution equation

$$\begin{aligned} y(n) &= \sum_{k=0}^{N-1} h(k)x(n-k) \qquad (18.1)\\ &= h(0)x(n) + h(1)x(n-1) + \ldots + h(N-2)x(n-(N-2)) \\ &\quad + h(N-1)x(n-(N-1)) \end{aligned}$$

TABLE 18.1 TMS320C25 Memory Organization for Convolution

	Input Samples		
Coefficients	Time n	Time $n+1$	Time $n+2$
PC $\rightarrow h(N-1)$	$x(n)$	$x(n+1)$	$x(n+2)$
$h(N-2)$	$x(n-1)$	$x(n)$	$x(n+1)$
.	.	.	.
.	.	.	.
.	.	.	.
$h(2)$	$x(n-(N-3))$	$x(n-(N-4))$	$x(n-(N-5))$
$h(1)$	$x(n-(N-2))$	$x(n-(N-3))$	$x(n-(N-4))$
$h(0)$	AR1 $\rightarrow x(n-(N-1))$	$x(n-(N-2))$	$x(n-(N-3))$

represents a finite impulse response (FIR) filter with length N. The memory organization for the coefficients $h(k)$ and the input samples $x(n-k)$ is shown in Table 18.1. The coefficients are placed within a specified internal program memory space and the input samples within a specified data memory space. The program counter (PC) initially points at the memory location that contains the last coefficient $h(N-1)$, for example at memory address FF00h (in hex). One of the (8) auxiliary registers points at the memory address of the last or least recent input sample. The most recent sample is represented by $x(n)$. The following program segment implements (18.1):

```
LARP    AR1
RPTK    N-1
MACD    FF00h,*-
APAC
```

The first instruction selects auxiliary register AR1, which will be used for indirect addressing. The second instruction RPTK causes the subsequent MACD instruction to execute N times (repeated $N-1$ times). The MACD instruction has the following functions:

1. Multiplies the coefficient value $h(N-1)$ by the input sample value $x(n-(N-1))$.
2. Accumulates any previous product stored in a special register (TR).
3. Copies the data memory sample value into the location of the next-higher memory. This "data move" is to model the input sample delays associated with the next unit of time $n+1$.

The last instruction APAC accumulates the last multiply operation $h(0)x(n)$.

At time $n+1$, the convolution Eq. (18.1) becomes

$$y(n+1) = h(0)x(n+1) + h(1)x(n) + \ldots + h(N-2)x(n-(N-3)) + h(N-1)x(n-(N-2)) \tag{18.2}$$

The previous program segment can be placed within a loop, with the PC and the auxiliary register AR1 reinitialized (see the memory organization of the samples $x(k)$ associated with time $n+1$ in Table 18.1). Note that the last multiply operation is $h(0)x(.)$, where $x(.)$ represents the newest sample. This process can be continuously repeated for time $n+2$, $n+3$, and so on.

The characteristics of a frequency selective FIR filter are specified by a set of coefficients that can be readily obtained using commercially available filter design packages. These coefficients can be placed within a generic FIR program. Within 5–10 minutes, an FIR filter can be implemented in real time. This includes finding the coefficients; assembling, linking and downloading the FIR program into the SWDS; and observing the desired frequency response displayed on a spectrum analyzer. A different FIR filter can be quickly obtained since the only necessary change in the generic program is to substitute a new set of coefficients.

The approach for modeling the sample delays involves moving the data. A different scheme is used with the floating-point TMS320C30 processor with a circular mode of addressing.

TABLE 18.2 TMS320C30 Memory Organization for Convolution

Coefficients	Time n	Time $n+1$	Time $n+2$
AR0 $\rightarrow h(N-1)$	AR1 $\rightarrow x(n-(N-1))$	$x(n+1)$	$x(n+1)$
$h(N-2)$	$x(n-(N-2))$	AR1 $\rightarrow x(n-(N-2))$	$x(n+2)$
$h(N-3)$	$x(n-(N-3))$	$x(n-(N-3))$	AR1 $\rightarrow x(n-(N-3))$
.	.	.	.
.	.	.	.
.	.	.	.
$h(1)$	$x(n-1)$	$x(n-1)$	$x(n-1)$
$h(0)$	$x(n)$	$x(n)$	$x(n)$

Floating-Point TMS320C30-Based Development System

TMS320C30-based DSP development systems are also currently available from a number of companies. The following are available from Texas Instruments:

1. An evaluation module (EVM). The EVM is a powerful, yet relatively inexpensive 8-bit card that plugs into a slot on an IBM AT compatible. It includes the third-generation TMS320C30, 16 K of user RAM, and an AIC for I/O. A serial port connector available on the EVM can be used to interface the TMS320C30 to other input/output devices (the TMS320C30 has two serial ports). An additional AIC can be interfaced to the TMS320C30 through this serial port connector. A very powerful, yet inexpensive, analog evaluation fixture, available from Burr-Brown (Arizona), can also be readily interfaced to the serial port on the EVM. This complete two-channel analog evaluation fixture includes an 18-bit DSP102 ADC, an 18-bit DSP202 DAC, antialiasing and reconstruction filters. The ADC has a maximum sampling rate of 200 kHz.
2. An XDS1000 emulator—powerful but quite expensive. A module can be readily built as a target system to interface to the XDS1000 [Chassaing, 1992]. This module contains the TMS320C30, 16 K of static RAM. Two connectors are included on this module, for interfacing to either an AIC module or to a second-generation analog interface board (AIB). The AIC was discussed in conjunction with the TMS320C25. The AIB includes Burr-Brown's 16-bit ADC and DAC with a maximum sampling rate of 58 kHz. An AIC is also included on this newer AIB version.

EVM Tools

The EVM package includes an assembler, a linker, a simulator, a C compiler, and a C source debugger. The second-generation TMS320C25 fixed-point processor is supported by C with some degrees of success. The architecture and instruction set of the third-generation TMS320C30 processor facilitate the development of high-level language compilers. An optimizer option is available with the C compiler for the TMS320C30. A C-code program can be readily compiled, assembled, linked, and downloaded into either a simulator or the EVM for real-time processing. A run-time support library of C functions, included with the EVM package, can be used during linking. During simulation, the input data can be retrieved from a file and the output data written into a file. Input and output port addresses can be appropriately specified. Within a real-time processing environment with the EVM, the C source debugger can be used. One can single-step through a C-code program while observing the equivalent step(s) through the assembly code. Both the C code and the corresponding assembly code can be viewed through the EVM windows. One can also monitor at the same time the contents of registers, memory locations, and so on.

Implementation of a Finite Impulse Response Filter with the TMS320C30

Consider again the convolution equation, Eq. (18.1), which represents an FIR filter. Table 18.2 shows the TMS320C30 memory organization used for the coefficients and the input samples. Initially, all the input samples can be set to zero. The newest sample $x(n)$, at time n, can be retrieved from an ADC using the following instructions:

```
        FLOAT     *AR3,R3
        STF       R3, *AR1++%
```

These two instructions cause an input value $x(n)$, retrieved from an input port address specified by auxiliary register AR3, to be loaded into a register R3 (one of eight 40-bit-wide extended precision registers), then stored in a memory location pointed by AR1 (AR1 would be first initialized to point at the "bottom" or higher-memory address of the table for the input samples). AR1 is then postincremented in a circular fashion, designated with the modulo operator %, to point at the oldest sample $x(n - (N - 1))$, as shown in Table 18.2. The size of the circular buffer must first be specified. The following program segment implements (18.1):

```
        RPTS      LENGTH-1
        MPYF      *AR0++%,*AR1++%,R0
||      ADDF      R0,R2,R2
        ADDF      R0,R2
```

The repeat "single" instruction RPTS causes the next (multiply) floating-point instruction MPYF to be executed LENGTH times (repeated LENGTH-1), where LENGTH is the length of the FIR filter. Furthermore, since the first ADDF addition instruction is in parallel (designated by ||) with the MPYF instruction, it is also executed LENGTH times. From Table 18.2, AR0, one of the eight available auxiliary registers, initially points at the memory address (a table address) which contains the coefficient $h(N-1)$, and a second auxiliary register AR1 now points to the address of the oldest input sample $x(n-(N-1))$. The second indirect addressing mode instruction multiplies the content in memory (address pointed by AR0) $h(N-1)$ by the content in memory (address pointed by AR1) $x(n-N-1)$), with the result stored in R0. Concurrently (in parallel), the content of R0 is added to the content of R2, with the result stored in R2. Initially R0 and R2 are set to zero; hence, the resulting value in R2 is *not* the product of the first multiply operation. After the first multiply operation, both AR0 and AR1 are incremented, and $h(N-2)$ is multiplied by $x(n-(N-2))$. Concurrently, the result of the first multiply operation (stored in R0) is accumulated into R2. The second addition instruction, executed only once, accumulates the last product $h(0)x(n)$ (similar to the APAC instruction associated with the fixed-point TMS320C25). The overall result yields an output value $y(n)$ at time n. After the last multiply operation, both AR0 and AR1 are postincremented to point at the "top" or lower-memory address of each circular buffer. The process can then be repeated for time $n+1$ in order to obtain a second output value $y(n+1)$. Note that the newest sample $x(n+1)$ would be retrieved from an ADC using the FLOAT and STF instructions, then placed at the top memory location of the buffer (table) containing the samples, overwriting the initial value $x(n-(N-1))$. AR1 is then incremented to point at the address containing $x(n-(N-2))$, and the previous four instructions can be repeated. The last multiply operation involves $h(0)$ and $x(.)$, where $x(.)$ is the newest sample $x(n+1)$, at time $n+1$. The foregoing procedure would be repeated to produce an output $y(n+2)$, $y(n+3)$, and so on. Each output value would be converted to a fixed-point equivalent value before being sent to a DAC. The frequency response of an FIR filter with 41 coefficients and a center frequency of 2.5 kHz, obtained from a signal analyzer, is displayed in Fig. 18.8.

FIR and IIR Implementation Using C and Assembly Code

A real-time implementation of a 45-coefficient bandpass FIR filter and a sixth-order IIR filter with 345 samples, using C code and TMS320C30 code, is discussed in Chassaing and Bitler [1991]. Tables 18.3 and 18.4 show a comparison of execution times of those two filters. The C language FIR filter, implemented without the modulo operator %, and compiled with a C compiler V4.1, executed two times slower[1] than an equivalent assembly language filter (which has a similar execution time as one implemented with a filter routine in assembly, called by a C program). The C language IIR filter ran 1.3 times slower than the corresponding assembly language IIR filter. These slower execution times may be acceptable for many applications. Where execution speed is crucial,

[1] 1.5 times slower using a newer C compiler V4.4.

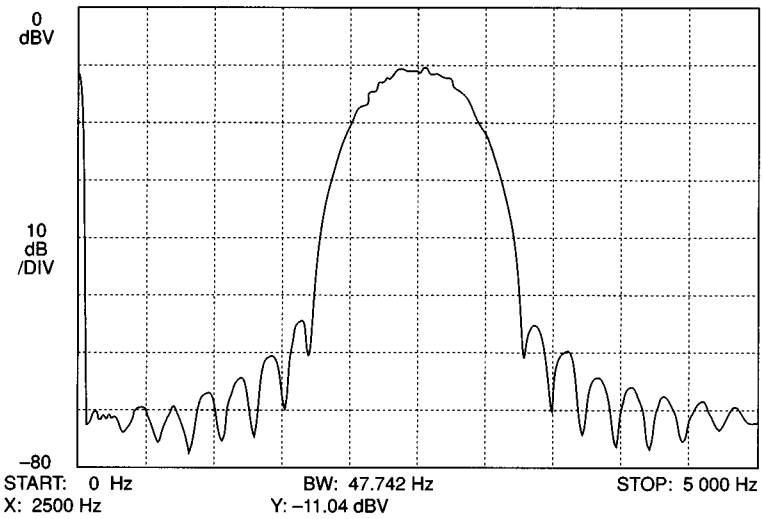


FIGURE 18.8 Frequency response of 41-coefficient FIR filter.

TABLE 18.3 Execution Time and Program Size of FIR Filter

FIR (45 samples)	Execution Time (msec)	Size (words)
C with modulo	4.16	122
C without modulo	0.338	116
C-called assembly	0.1666	74
Assembly	0.1652	27

TABLE 18.4 Execution Time and Program Size of 6th-Order IIR Filter

IIR (345 samples)	Execution Time (msec)	Size (words)
C	1.575	109
Assembly	1.18	29

a time-critical function may be written in assembly and called from a C program. In applications where speed is not absolutely crucial, C provides a better environment because of its portability and maintainability.

Real-Time Applications

A number of applications are discussed in Chassaing and Horning (1990) using TMS320C25 code and in Chassaing (1992) using TMS320C30 and C code. These applications include multirate and adaptive filtering, modulation techniques, and graphic and parametric equalizers. Two applications are briefly discussed here: a ten-band multirate filter and a video line rate analysis.

1. The functional block diagram of the multirate filter is shown in Fig. 18.9. The multirate design provides a significant reduction in processing time and data storage, compared to an equivalent single-rate design. With multirate filtering, we can use a decimation operation in order to obtain a sample rate reduction or an interpolation operation (as shown in Fig. 18.9) in order to obtain a sample rate increase [Crochiere and Rabiner, 1983]. A pseudorandom noise generator implemented in software provides the input noise to the ten octave band filters. Each octave band filter consists of three 1/3-octave filters (each with 41 coefficients), which can be individually controlled. A controlled noise source can be obtained with this design. Since each 1/3-octave band filter can be turned *on* or *off*, the noise spectrum can be shaped accordingly. The interpolation filter is a low-pass FIR filter with a 2:1 data-rate increase, yielding two sample outputs for each input sample. The sample rate of the highest octave-band filter is set at 32,768 samples per second, with each successively lower band processing at half the rate of the next-higher band. The multirate filter (a nine-band version) was implemented with the TMS320C25 [Chassaing et al., 1990]. Figure 18.10 shows the three 1/3-octave band filters of band 10 implemented with the EVM

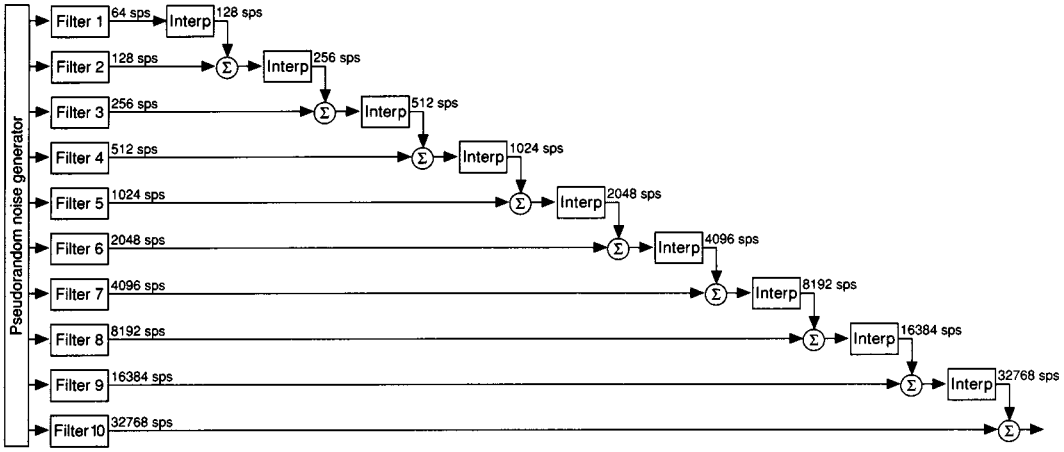


FIGURE 18.9 Multirate filter functional block diagram.

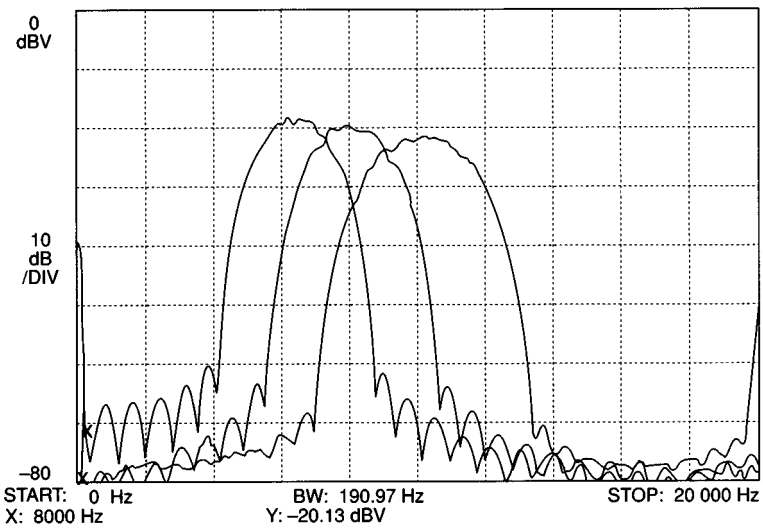


FIGURE 18.10 Frequency responses of the 1/3-octave band ten filters.

in conjunction with the two-channel analog fixture (made by Burr-Brown). The center frequency of the middle 1/3-octave band 10 filter is at approximately 8 kHz since the coefficients were designed for a center frequency of 1/4 the sampling rate (the middle 1/3-octave band 9 filter would be centered at 4 kHz, the middle 1/3-octave band 8 filter at 2 kHz, and so on). Note that the center frequency of the middle 1/3-octave band 1 filter would be at 2 Hz if the highest sampling rate is set at 4 kHz. Observe from Fig. 18.10 that the crossover frequencies occur at the 3-dB points. Since the main processing time of the multirate filter (implemented in assembly code) was measured to be 8.8 ms, the maximum sampling rate was limited to 58 ksps.

2. A video line rate analysis implemented entirely in C code is discussed in Chassaing and Bitler [1992]. A module was built to sample a video line of information. This module included a 9.8-MHz clock, a high sampling rate 8-bit ADC and appropriate support circuitry (comparator, FIFO buffer, etc.). Interactive features allowed for the selection of one (out of 256) horizontal lines of information and the execution of algorithms for digital filtering, averaging, and edge enhancement, with the resulting effects displayed on the PC screen. Figure 18.11 shows the display of a horizontal line (line #125) of information

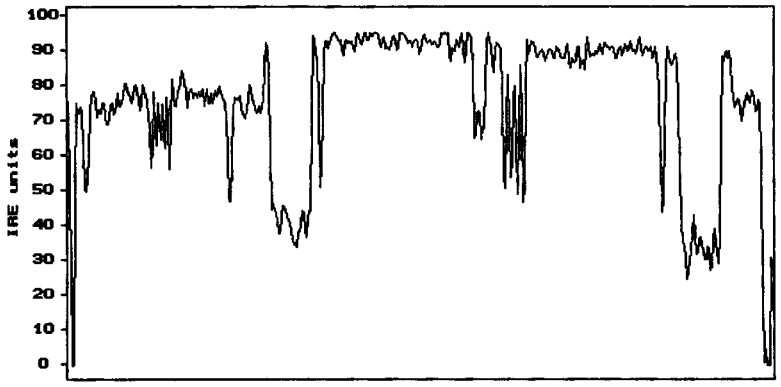


FIGURE 18.11 Display of a horizontal line of video signal.

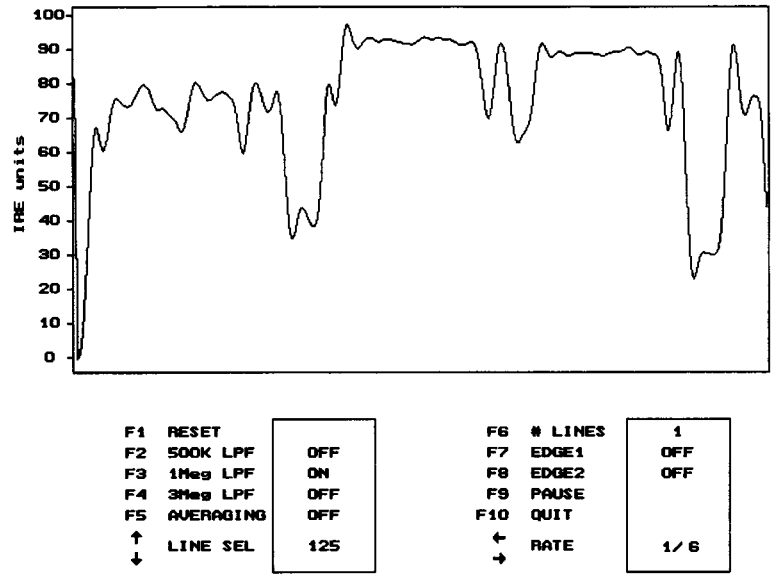


FIGURE 18.12 Video line signal with 1-MHz filtering.

obtained from a test chart with a charge coupled device (CCD) camera. The function key **F3** selects the 1-MHz low-pass filter resulting in the display shown in Fig. 18.12. The 3-MHz filter (with **F4**) would pass more of the higher-frequency components of the signal but with less noise reduction. **F5** implements the noise averaging algorithm. The effect of the edge enhancement algorithm (with **F7**) is displayed in Fig. 18.13.

Conclusions and Future Directions

DSP processors have been used extensively in a number of applications, even in non-DSP applications such as graphics. The fourth-generation floating-point TMS320C40, code compatible with the TMS320C30, features an instruction cycle time of 40 ns and six serial ports. The fifth-generation fixed-point TMS320C50, code compatible with the first two generations of fixed-point processors, features an instruction cycle time of 35 ns and 10 K words (16-bit) of on-chip data and program memory. Currently, both the fixed-point and floating-point processors are being supported by TI.

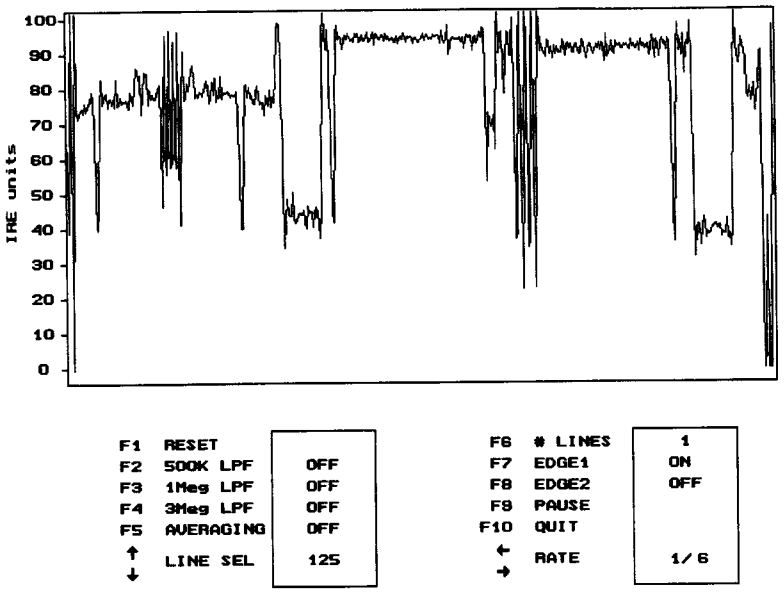


FIGURE 18.13 Video line signal with edge enhancement.

Defining Terms

C compiler: Program that translates C code into assembly code.

Digital signal processor: Special-purpose microprocessor with an architecture suitable for fast execution of signal processing algorithms.

Fixed-point processor: A processor capable of operating on scaled integer and fractional data values.

Floating-point processor: Processor capable of operating on integers as well as on fractional data values without scaling.

On-chip memory: Internal memory available on the digital signal processor.

Pipelining feature: Feature that permits parallel operations of fetching, decoding, reading, and executing.

Special-purpose digital signal processor: Digital signal processor with special feature for handling a specific signal processing application, such as FFT.

Related Topics

14.3 Design and Implementation of Digital Filters • 79.1 IC Logic Family Operation and Characteristics

References

H. M. Ahmed and R. B. Kline, "Recent advances in DSP systems," *IEEE Communications Magazine,* 1991.

R. Chassaing, *Digital Signal Processing with C and the TMS320C30*, New York: Wiley, 1992.

R. Chassaing and R. Ayers, "Digital signal processing with the SHARC," in *Proceedings of the 1996 ASEE Annual Conference,* 1996.

R. Chassaing and B. Bitler, "Real-time digital filters in C," in *Proceedings of the 1991 ASEE Annual Conference,* 1991.

R. Chassaing and B. Bitler, "A video line rate analysis using the TMS320C30 floating-point digital signal processor," in *Proceedings of the 1992 ASEE Annual Conference,* 1992.

R. Chassaing and D. W. Horning, *Digital Signal Processing with the TMS320C25,* New York: Wiley, 1990.

R. Chassaing and P. Martin, "Parallel processing with the TMS320C40," in *Proceedings of the 1995 ASEE Annual Conference,* 1995.

R. Chassaing, W.A. Peterson, and D. W. Horning, "A TMS320C25-based multirate filter," *IEEE Micro,* 1990.

R.E. Crochiere and L.R. Rabiner, *Multirate Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1983.
K. S. Lin (ed.), *Digital Signal Processing Applications with the TMS320 Family. Theory, Algorithms, and Implementations,* vol. 1, Texas Instruments Inc., Texas, 1989.
A. V. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1989.
P. Papamichalis (ed.), *Digital Signal Processing Applications with the TMS320 Family. Theory, Algorithms, and Implementations,* vol. 3, Texas Instruments, Inc., Texas, 1990.

Further Information

Rulph Chassaing teaches hands-on workshops on digital signal processing using C and the TMS320C30, offered at Roger Williams University in Bristol, RI, 02809. He offered a one-week workshop in August 1996, supported by the National Science Foundation (NSF). He will offer two workshops in August 1997, supported by NSF, using the TMS320C30 and the TMS320C31. Workshops on the TMS320 family of digital signal processors are offered by Texas Instruments, Inc. at various locations.

A tutorial "Digital Signal Processing Comes of Age" can be found in the *IEEE Spectrum,* May 1996.

Schroeter, J., Mehta, S.K., Carter, G.C. "Acoustic Signal Processing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

19

Acoustic Signal Processing

Juergen Schroeter
Acoustics Research Dept., AT&T Bell Laboratories

Sanjay K. Mehta
NUWC Detachment

G. Clifford Carter
NUWC Detachment

19.1 Digital Signal Processing in Audio and Electroacoustics
Steerable Microphone Arrays • Digital Hearing Aids • Spatial Processing • Audio Coding • Echo Cancellation • Active Noise and Sound Control

19.2 Underwater Acoustical Signal Processing s
What Is Underwater Acoustical Signal Processing? • Technical Overview • Underwater Propagation • Processing Functions • Advanced Signal Processing • Application

19.1 Digital Signal Processing in Audio and Electroacoustics

Juergen Schroeter

In this section we will focus on advances in algorithms and technologies in digital signal processing (DSP) that have already had or, most likely, will soon have, a major impact on **audio** and **electroacoustics** (A&E). Because A&E embraces a wide range of topics, it is impossible for us to go here into any depth in any one of them. Instead, this section will try to give a compressed overview of the topics the author judges to be most important.

In the following, we will look into steerable microphone arrays, digital hearing aids, spatial processing, audio coding, echo cancellation, and active noise and sound control. We will *not* cover basic techniques in digital recording [Pohlmann, 1989] and computer music [Moore, 1990].

Steerable Microphone Arrays

Steerable microphone arrays have controllable directional characteristics. One important application is in teleconferencing. Here, sound pickup can be highly degraded by reverberation and room noise. One solution to this problem is to utilize highly directional microphones. Instead of pointing such a microphone manually to a desired talker, steerable microphone arrays can be used for reliable automatic tracking of speakers as they move around in a noisy room or auditorium, if combined with a suitable speech detection algorithm.

Figure 19.1 depicts the simplest kind of steerable array using N microphones that are uniformly spaced with distance d along the linear x-axis. It can be shown that the response of this system to a plane wave impinging at an angle θ is:

$$H(j\omega) = \sum_{n=0}^{N-1} a_n e^{-j(\omega/c)nd\cos\theta} \tag{19.1}$$

Here, $j = \sqrt{-1}$, ω is the radian frequency, and c is the speed of sound. Equation (19.1) is a spatial filter with coefficients a_n and the delay operator $z^{-1} = \exp(-jd\omega/c\cos\theta)$. Therefore, we can apply finite impulse response (FIR) filter theory. For example, we could taper the weights a_n to suppress sidelobes of the array. We also have to guard against spatial aliasing, that is, grating lobes that make the directional characteristic of the array

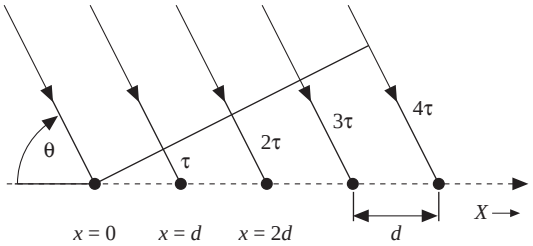


FIGURE 19.1 A linear array of N microphones (here, $N = 5$; $\tau = d/c \cos\theta$).

FIGURE 19.2 Three superimposed linear arrays depicted by large, midsize, and small circles. The largest array covers the low frequencies, the midsize array covers the midrange frequencies, and the smallest covers the high frequencies.

ambiguous. The array is steered to an angle θ_0 by introducing appropriate delays into the N microphone lines. In Eq. (19.1), we can incorporate these delays by letting a_n

$$a_n = e^{-j\omega\tau_0} \; e^{+j(\omega/c)nd\cos\theta_0} \tag{19.2}$$

Here τ_0 is an overall delay equal to or larger than $Nd/c \cos\theta_0$ that ensures causality, while the second term in Eq. (19.2) cancels the corresponding term in Eq. (19.1) at $\theta = \theta_0$. Due to the axial symmetry of the one-dimensional (linear, 1-D) array, the directivity of the array is a figure of revolution around the x-axis. Therefore, in case we want the array to point to a *single* direction in space, we need a 2-D array.

Since most of the energy of typical room noise and the highest level of reverberation in a room is at low frequencies, one would like to use arrays that have their highest directivity (i.e., narrowest beamwidth) at low frequencies. Unfortunately, this need collides with the physics of arrays: the smaller the array relative to the wavelength, the wider the beam. (Again, the corresponding notion in filter theory is that systems with shorter impulse responses have wider bandwidth.) One solution to this problem is to superimpose different-size arrays and filter each output by the appropriate bandpass filter, similar to a crossover network used in two- or three-way loudspeaker designs. Such a superposition of three five-element arrays is shown in Fig. 19.2. Note that we only need nine microphones in this example, instead of $5 \times 3 = 15$.

Another interesting application is the use of an array to mitigate discrete noise sources in a room. For this, we need to attach an FIR filter to each of the microphone signal outputs. For any given frequency, one can show that N microphones can produce $N - 1$ nulls in the directional characteristic of the array. Similarly, attaching an M-point FIR filter to each of the microphones, we can get these zeros at $M - 1$ frequencies. The weights for these filters have to be adapted, usually under the constraint that the transfer function (frequency characteristic) of the array for the *desired* source is optimally flat. In practical tests, systems of this kind work nicely in (almost) anechoic environments. Their performance degrades, however, with increasing reverberation.

More information on microphone arrays can be found in Flanagan et al. [1991]; in particular, they describe how to make arrays adapt to changing talker positions in a room by constantly scanning the room with a moving search beam and by switching the main beam accordingly. Current research issues are, among others, 3-D arrays and how to take advantage of low-order wall reflections.

Digital Hearing Aids

Commonly used hearing aids attempt to compensate for sensorineural (cochlear) hearing loss by delivering an amplifed acoustic signal to the external ear canal. As will be pointed out below, the most important problem is how to find the best aid for a given patient.

Historically, technology has been the limiting factor in hearing aids. Early on, carbon hearing aids provided a limited gain and a narrow, peaky frequency response. Nowadays, hearing aids have a broader bandwidth and a flatter frequency response. Consequently, more people can benefit from the improved technology. With the advent of digital technology, the promise is that even more people would be able to do so. Unfortunately, as will be pointed out below, we have not fulfilled this promise yet.

We distinguish between *analog*, *digitally controlled analog*, and *digital* hearing aids. Analog hearing aids contain only (low-power) pre-amp, filter(s), (optional) automatic gain control (AGC) or compressor, power amp, and output limiter. Digitally controlled aids have certain additional components: one kind adds a digital controller to monitor and adjust the analog components of the aid. Another kind contains switched-capacitor circuits that represent sampled signals in analog form, in effect allowing simple discrete-time processing (e.g., filtering). Aids with switched-capacitor circuits have a lower power consumption compared to digital aids. Digital aids—none are yet commercially available—contain A/D and D/A converters and at least one programmable digital signal processing (DSP) chip, allowing for the use of sophisticated DSP algorithms, (small) microphone arrays, speech enhancement in noise, etc. Experts disagree, however, as to the usefulness of these techniques. To date, the most successful approach seems to be to ensure that all parts of the signal get amplified so that they are clearly audible but not too loud and to "let the brain sort out signal and noise."

Hearing aids pose a tremendous challenge for the DSP engineer, as well as for the audiologist and acoustician. Due to the continuing progress in chip technology, the physical size of a digital aid should no longer be a serious problem in the near future; however, power consumption will still be a problem for quite some time. Besides the obvious necessity of avoiding howling (acoustic feedback), for example, by employing sophisticated models of the electroacoustic transducers, acoustic leaks, and ear canal to control the aid accordingly, there is a much more fundamental problem: since DSP allows complex schemes of splitting, filtering, compressing, and (re-) combining the signal, hearing aid performance is no longer limited by bottlenecks in technology. It is still limited, however, by the lack of basic knowledge about how to map an arbitrary input signal (i.e., speech from a desired speaker) onto the reduced capabilities of the auditory system of the targeted wearer of the aid. Hence, the *selection and fitting of an appropriate aid* becomes the most important issue. This serious problem is illustrated in Fig. 19.3.

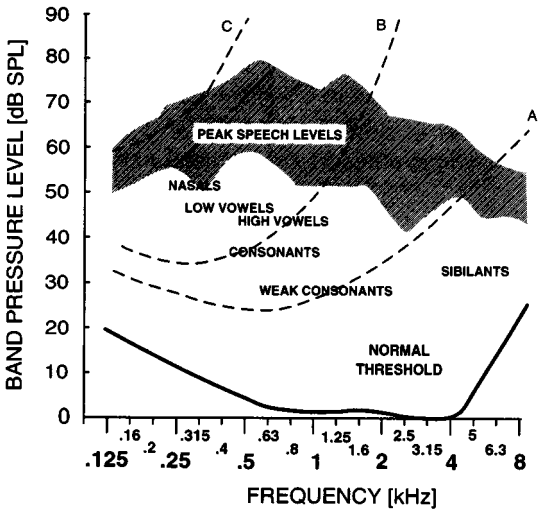


FIGURE 19.3 Peak third-octave band levels of normal to loud speech (hatched) and typical levels/dominant frequencies of speech sounds (identifiers). Both can be compared to the third-octave threshold of normal-hearing people (solid line), thresholds for a mildly hearing-impaired person (A), for a severely hearing-impaired person (B), and for a profoundly hearing-impaired person (C). For example, for person (A), sibilants and some weak consonants in a normal conversation cannot be perceived. (*Source:* H. Levitt, "Speech discrimination ability in the hearing impaired: spectrum considerations," in *The Vanderbilt Hearing-Aid Report: State of the Art-Research Needs*, G.A. Studebaker and F.H. Bess (Eds.), Monographs in Contemporary Audiology, Upper Darby, Pa., 1982, p. 34. With permission.)

It is important to note that for speech presented at a constant level, a linear (no compression) hearing aid can be tuned to do as well as a hearing aid with compression. However, if parameters like signal and background noise levels change dynamically, compression aids, in particular those with two bands or more, should have an advantage.

While a patient usually has no problem telling whether setting A or B is "clearer," adjusting more than just 2–3 (usually interdependent) parameters is very time consuming. For a multiparameter aid, an efficient fitting procedure that maximizes a certain objective is needed. Possible objectives are, for example, **intelligibility maximization** or **loudness restoration**. The latter objective is assumed in the following.

It is known that an impaired ear has a reduced dynamic range. Therefore, the procedure for fitting a patient with a hearing aid could estimate the so-called loudness-growth function (LGF) that relates the sound pressure

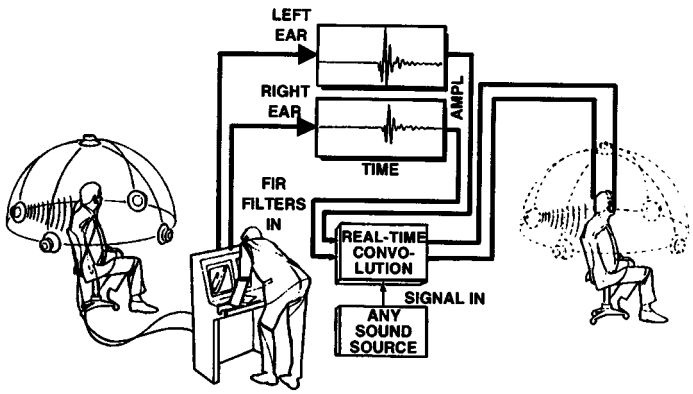


FIGURE 19.4 Measuring and using transfer functions of the external ear for binaural mixing (FIR = finite impulse response). (*Source:* E.M. Wenzel, Localization in virtual acoustic displays, *Presence*, vol. 1, p. 91, 1992. With permission.)

level of a specific (band-limited) sound to its loudness. An efficient way of measuring the LGF is described by Allen et al. [1990]. Once the LGF of an impaired ear is known, a multiband hearing aid can implement the necessary compression for each band [Villchur, 1973]. Note, however, that this assumes that interactions between the bands can be neglected (problem of summation of partial loudnesses). This might not be valid for aids with a large number of bands. Other open questions include the choice of widths and filter shape of the bands, and optimization of dynamic aspects of the compression (e.g., time constants). For aids with just two bands, the crossover frequency is a crucial parameter that is difficult to optimize.

Spatial Processing

In spatial processing, audio signals are modified to give them new spatial attributes, such as, for example, the perception of having been recorded in a specific concert hall. The auditory system—using only the two ears as inputs—is capable of perceiving the direction and distance of a sound source with a high degree of accuracy, by exploiting **binaural** and **monaural** spectral cues. Wave propagation in the ear canal is essentially one-dimensional. Hence, the 3-D spatial information is coded by sound diffraction into spectral information before the sound enters the ear canal. The sound diffraction is caused by the head/torso (on the order of 20-dB and 600-μs **interaural** level difference and delay, respectively) and at the two pinnae (auriculae); see, for example, Shaw [1980]. Binaural techniques like the one discussed below can be used for evaluating room and concert-hall acoustics (optionally in reduced-scale model rooms using a miniature dummy head), for noise assessment (e.g., in cars), and for "Kunstkopfstereophonie" (dummy-head stereophony). In addition, there are techniques for *loudspeaker reproduction* (like "Q-Sound") that try to extend the range in horizontal angle of traditional stereo speakers by using interaural cross cancellation. Largely an open question is how to reproduce spatial information for large audiences, for example, in movie theaters.

Figure 19.4 illustrates the technique for filtering a single-channel source using measured headrelated transfer functions, in effect, creating a virtual sound source in a given direction of the listener's auditory space (assuming plane waves, i.e., infinite source distance). On the left in this figure, the measurement of head-related transfer functions is shown. Focusing on the *left* ear for a moment (subscript l), we need to estimate the so-called free-field transfer function (subscript ff) for given angles of incidence in the horizontal plane (azimuth φ) and vertical plane (elevation δ):

$$H_{\mathrm{ff,l}}(j\omega, \varphi, \delta) = P_{\mathrm{probe,l}}(j\omega, \varphi, \delta)/P_{\mathrm{ref}}(j\omega) \tag{19.3}$$

where $P_{\mathrm{probe,l}}$ is the Fourier transform of the sound pressure measured in the subject's left ear, and P_{ref} is the Fourier transform of the pressure measured at a suitable reference point in the free field without the subject being present (e.g., at the midpoint between the two ears). (Note that P_{ref} is independent of the direction of sound incidence since we assume an anechoic environment.) The middle of Fig. 19.4 depicts the convolution

of any "dry" (e.g., mono, low reverberation) source with the stored $H_{ff,l}(j\omega, \varphi, \delta)$s and corresponding $H_{ff,r}(j\omega, \varphi, \delta)$s. On the right side in the figure, the resulting binaural signals are reproduced via equalized headphones. The equalization ensures that a sound source with a flat spectrum (e.g., white noise) does not suffer any perceivable coloration for any direction (φ, δ).

Implemented in a real-time "binaural mixing console," the above scheme can be used to create "virtual" sound sources. When combined with an appropriate scheme for interpolating head-related transfer functions, moving sound sources can be mimicked realistically. Furthermore, it is possible to superimpose early reflections of a hypothetical recording room, each filterered by the appropriate head-related transfer function. Such inclusion of a room in the simulation makes the spatial reproduction more robust against individual differences between "recording" and "listening" ears, in particular, if the listener's head movements are fed back to the binaural mixing console. (Head movements are useful for disambiguating spatial cues.) Finally, such a system can be used to create "virtual acoustic displays," for example, for pilots and astronauts [Wenzel, 1992]. Other research issues are, for example, the required accuracy of the head-related transfer functions, intersubject variability, and psychoacoustic aspects of room simulations.

Audio Coding

Audio coding is concerned with compressing (reducing the bit rate) of audio signals. The uncompressed digital audio of compact disks (CDs) is recorded at a rate of 705.6 kbit/s for each of the two channels of a stereo signal (i.e., 16 bit/sample, 44.1-kHz sampling rate; 1411.2 kbit/s total). This is too high a bit rate for digital audio broadcasting (DAB) or for transmission via end-to-end digital telephone connections (integrated services digital network, ISDN). Current audio coding algorithms provide at least "better than FM" quality at a combined rate of 128 kbit/s for the two stereo channels (2 ISDN B channels!), "transparent coding" at rates of 96 to 128 kbit/s per mono channel, and "studio quality" at rates between 128 and 196 kbit/s per mono channel. (While a large number of people will be able to detect distortions in the first class of coders, even so-called "golden ears" should not be able to detect any differences between original and coded versions of known "critical" test signals; the highest quality category adds a safety margin for editing, filtering, and/or recoding.)

To compress audio signals by a factor as large as eleven while maintaining a quality exceeding that of a local FM radio station requires sophisticated algorithms for reducing the **irrelevance and redundancy** in a given signal. A large portion (but usually less than 50%) of the bit-rate reduction in an audio coder is due to the first of the two mechanisms. Eliminating irrelevant portions of an input signal is done with the help of psychoacoustic models. It is obvious that a coder can eliminate portions of the input signal that—when played back—will be below the threshold of hearing. More complicated is the case when we have multiple signal components that tend to cover each other, that is, when weaker components cannot be heard due to the presence of stronger components. This effect is called *masking*. To let a coder take advantage of masking effects, we need to use good masking models. Masking can be modeled in the time domain where we distinguish so-called simultaneous masking (masker and maskee occur at the *same* time), forward masking (masker occurs *before* maskee), and backward masking (masker occurs *after* maskee). Simultaneous masking usually is modeled in the frequency domain. This latter case is illustrated in Fig. 19.5.

Audio coders that employ common frequency-domain models of masking start out by splitting and subsampling the input signal into different frequency bands (using filterbanks such as subband filterbanks or time-frequency transforms). Then, the masking threshold (i.e., *predicted* masked threshold) is determined, followed by quantization of the spectral information and (optional) noiseless compression using variable-length coding. The encoding process is completed by multiplexing the spectral information with side information, adding error protection, etc.

The first stage, the filter bank, has the following requirements. First, decomposing and then simply reconstructing the signal should not lead to distortions ("perfect reconstruction filterbank"). This results in the advantage that all distortions are due to the quantization of the spectral data. Since each quantizer works on band-limited data, the distortion (also band-limited due to refiltering) is controllable by using the masking models described above. Second, the bandwidths of the filters should be narrow to provide sufficient coding gain. On the other hand, the length of the impulse responses of the filters should be short enough (time resolution of the coder!) to avoid so-called pre-echoes, that is, backward spreading of distortion components

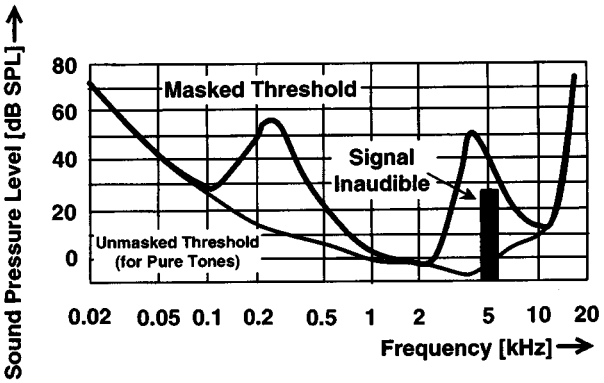


FIGURE 19.5 Masked threshold in the frequency domain for a hypothetical input signal. In the vicinity of high-level spectral components, signal components below the current masked threshold cannot be heard.

that result from sudden onsets (e.g., castanets). These two contradictory requirements, obviously, have to be worked out by a compromise. "**Critical band**" filters have the shortest impulse responses needed for coding of transient signals. On the other hand, the optimum frequency resolution (i.e., the one resulting in the highest coding gain) for a typical signal can be achieved by using, for example, a 2048-point modified discrete cosine transform (MDCT).

In the second stage, the (time-varying) masking threshold as determined by the psychoacoustic model usually controls an iterative analysis-by-synthesis quantization and coding loop. It can incorporate rules for masking of tones by noise and of noise by tones, though little is known in the psychoacoustic literature for more general signals. Quantizer step sizes can be set and bits can be allocated according to the known spectral estimate, by block companding with transmission of the scale factors as side information or iteratively in a variable-length coding loop (Huffman coding). In the latter case, one can low-pass filter the signal if the total required bit rate is too high.

The decoder has to invert the processing steps of the encoder, that is, do the error correction, perform Huffman decoding, and reconstruct the filter signals or the inverse-transformed time-domain signal. Since the decoder is significantly less complex than the encoder, it is usually implemented on a single DSP chip, while the encoder uses several DSP chips.

Current research topics encompass tonality measures and time-frequency representations of signals. More information can be found in Johnston and Brandenburg [1991].

Echo Cancellation

Echo cancellers were first deployed in the U.S. telephone network in 1979. Today, they are virtually ubiquitous in long-distance telephone circuits where they cancel so-called line echoes (i.e., *electrical* echoes) resulting from nonperfect hybrids (the devices that couple local two-wire to long-distance four-wire circuits). In satellite circuits, echoes bouncing back from the far end of a telephone connection with a round-trip delay of about 600 ms are very annoying and disruptive. *Acoustic* echo cancellation—where the echo path is characterized by the transfer function $H(z)$ between a loudspeaker and a microphone in a room (e.g., in a speakerphone)—is crucial for teleconferencing where two or more parties are connected via full-duplex links. Here, echo cancellation can also alleviate acoustic feedback ("howling").

The principle of acoustic echo cancellation is depicted in Fig. 19.6(a). The echo path $H(z)$ is cancelled by modeling $H(z)$ by an adaptive filter and subtracting the filter's output $\hat{y}(t)$ from the microphone signal $y(t)$. The adaptability of the filter is necessary since $H(z)$ changes appreciably with movement of people or objects in the room and because periodic measurements of the room would be impractical. *Acoustic* echo cancellation is more challenging than cancelling line echoes for several reasons. First, room impulse responses $h(t)$ are longer than 200 ms compared to less than 20 ms for line echo cancellers. Second, the echo path of a room $h(t)$ is likely to change constantly (note that even small changes in temperature can cause significant changes of h). Third,

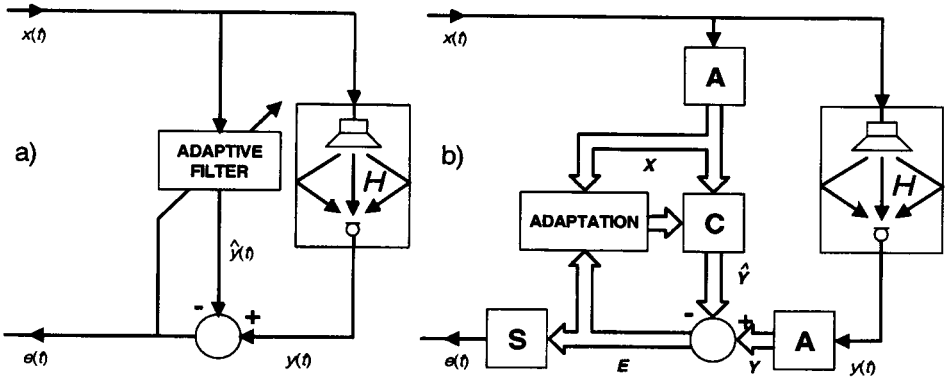


FIGURE 19.6 (a) Principle of using an echo canceller in teleconferencing. (b) Realization of the echo canceller in subbands. (After M. M. Sondhi and W. Kellermann, "Adaptive echo cancellation for speech signals," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds., New York: Marcel Dekker, 1991. By courtesy of Marcel Dekker, Inc.)

teleconferencing eventually will demand larger audio bandwidths (e.g., 7 kHz) compared to standard telephone connections (about 3.2 kHz). Finally, we note that echo cancellation in a stereo setup (*two* microphones and *two* loudspeakers at each end) is an even harder problem on which very little work has been done so far.

It is obvious that the initially unknown echo path $H(z)$ has to be "learned" by the canceller. It is also clear that for adaptation to work there needs to be a nonzero input signal $x(t)$ that excites all the eigenmodes of the system (resonances, or "peaks" of the system magnitude response $|H(j\omega)|$). Another important problem is how to handle double-talk (speakers at both ends are talking simultaneously). In such a case, the canceller could easily get confused by the speech from the near end that acts as an uncorrelated noise in the adaptation. Finally, the convergence rate, that is, how fast the canceller adapts to a change in the echo path, is an important measure to compare different algorithms.

Adaptive filter theory suggests several algorithms for use in echo cancellation. The most popular one is the so-called *least-mean square* (LMS) algorithm that models the echo path by an FIR filter with an impulse response $\hat{h}(t)$. Using vector notation h for the true echo path impulse response, $\hat{\mathrm{h}}$ for its estimate, and x for the excitation time signal, an estimate of the echo is obtained by $\hat{y}(t) = \hat{\mathrm{h}}'\mathrm{x}$, where the prime denotes vector transpose. A reasonable objective for a canceller is to minimize the instantaneous squared error $e^2(t)$, where $e(t) = y(t) - \hat{y}(t)$. The time derivative of $\hat{\mathrm{h}}$ can be set to

$$\frac{\mathrm{d}\hat{\mathbf{h}}}{\mathrm{d}t} = -\mu\nabla_{\hat{\mathbf{h}}}\, e^2(t) = -2\mu e(t)\nabla_{\hat{\mathbf{h}}}\, e(t) = 2\mu e(t)\mathbf{x} \tag{19.4}$$

resulting in the simple update equation $\hat{h}_{k+1} = \hat{h}_k + \alpha e_k x_k$, where α (or μ) control the rate of change. In practice, whenever the far-end signal $x(t)$ is low in power, it is a good idea to freeze the canceller by setting $\alpha = 0$. Sophisticated logic is needed to detect double talk. When it occurs, then also set $\alpha = 0$. It can be shown that the spread of the eigenvalues of the autocorrelation matrix of $x(t)$ determines the convergence rate, where the slowest-converging eigenmode corresponds to the smallest eigenvalue. Since the eigenvalues themselves scale with the power of the predominant spectral components in $x(t)$, setting $\alpha = 2\mu/(\mathrm{x}'\mathrm{x})$ will make the convergence rate independent of the far-end power. This is the *normalized* LMS method. Even then, however, all eigenmodes will converge at the same rate only if $x(t)$ is white noise. Therefore, pre-whitening the far-end signal will help in speeding up convergence.

The LMS method is an iterative approach to echo cancellation. An example of a noniterative, block-oriented approach is the *least-squares* (LS) algorithm. Solving a system of equations to get $\hat{\mathrm{h}}$, however, is computationally more costly. This cost can be reduced considerably by running the LS method on a sample-by-sample basis and by taking advantage of the fact that the new signal vectors are the old vectors with the oldest sample dropped and one new sample added. This is the *recursive least-squares* (RLS) algorithm. It also has the advantage

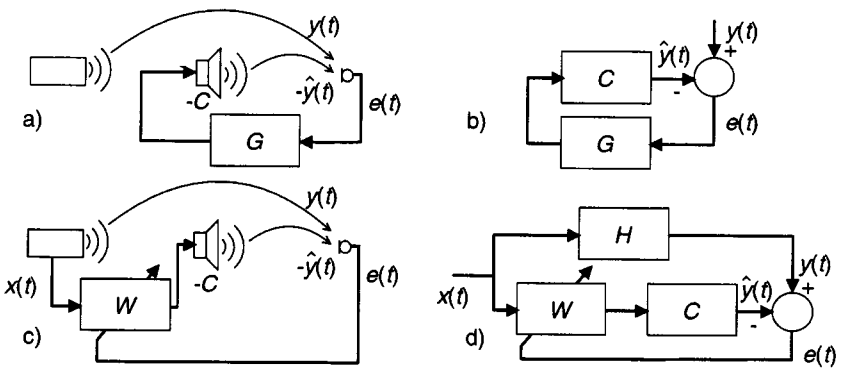


FIGURE 19.7 Two principles of active noise control. Feedback control system (a) and (b); feedforward control system (c) and (d). Physical block diagrams (a) and (c), and equivalent electrical forms (b) and (d). (After P. A. Nelson and S. J. Elliott, *Active Control of Sound*, London: Academic Press, 1992. With permission.)

of normalizing x by multiplying it with the inverse of its autocorrelation matrix. This, in effect, equalizes the adaptation rate of all eigenmodes.

Another interesting approach is outlined in Fig. 19.6(b). As in subband coding (discussed earlier), splitting the signals x and y into subbands with analysis filterbanks **A**, doing the cancellation in bands, and resynthesizing the outgoing ("error") signal e through a synthesis filterbank **S** also reduces the eigenvalue spread of each bandpass signal compared to the eigenvalue spread of the fullband signal. This is true for the eigenvalues that correspond to the "center" (i.e., unattenuated) portions of each band. It turns out, however, that the slowly converging "transition-band" eigenmodes get attenuated significantly by the synthesis filter **S**. The main advantage of the subband approach is the reduction in computational complexity due to the down-sampling of the filterbank signals. The drawback of the subband approach, however, is the introduction of the combined delay of **A** and **S**. Eliminating the analysis filterbank on $y(t)$ and moving the synthesis filterbank into the adaptation branch $\hat{Y}$ will remove this delay with the result that the canceller will not be able to model the earliest portions of the echo-path impulse response $h(t)$. To alleviate this problem, we could add in parallel a fullband echo canceller with a short filter. Further information and an extensive bibliography can be found in Haensler [1992].

Active Noise and Sound Control

Active noise control (ANC) is a way to reduce the sound pressure level of a given noise source through electroacoustic means. ANC and echo cancellation are somewhat related. While even *acoustic* echo cancellation is actually done on electrical signals, ANC could be labeled "wave cancellation," since it involves using one or more secondary acoustic or vibrational sources. Another important difference is the fact that in ANC one usually would like to cancel a given noise in a whole *region* in space, while echo cancellation commonly involves only one microphone picking up the echo signal at a single *point* in space. Finally, the transfer function of the transducer used to generate a cancellation ("secondary source") signal needs to be considered in ANC.

Active sound control (ASC) can be viewed as an offspring of ANC. In ASC, instead of trying to cancel a given sound field, one tries to control specific spatial and temporal characteristics of the sound field. One application is in adaptive sound reproduction systems. Here, ASC aims at solving the large-audience spatial reproduction problem mentioned in the spatial processing section of this chapter.

Two important principles of ANC are depicted in Fig. 19.7. In the upper half [Fig. 19.7(a) and (b)], a feedback loop is formed between the controller $G(s)$ and the transfer function $C(s)$ of the secondary source, and the acoustic path to the error microphone. Control theory suggests that $E/Y = 1/[1 + C(s)G(s)]$, where $E(s)$ and $Y(s)$ are Laplace transforms of $e(t)$ and $y(t)$, respectively. Obviously, if we could make C a real constant and $G \rightarrow \infty$, we would get a "zone of quiet" around the error microphone. Unfortunately, in practice, $C(s)$ will introduce at least a delay, thus causing stability problems for too large a magnitude $|G|$ at high enough frequencies. The system can be kept stable, for example, by including a low-pass filter in G and by positioning the secondary source in close vicinity to the error microphone. A highly successful application of the feedback

control in ANC is in *active* hearing protective devices (HPDs) and high-quality headsets and "motional-feedback" loudspeakers. *Passive* HPDs offer little or no noise attenuation at low frequencies due to inherent physical limitations. Since the volume enclosed by earmuffs is rather small, HPDs can benefit from the increase in low-frequency attenuation brought about by feedback-control ANC. Finally, note that the same circuit can be used for high-quality reproduction of a communications signal $s(t)$ fed into a headset by subtracting $s(t)$ electrically from $e(t)$. The resulting transfer function is $E/S = C(s)G(s)/[1 + C(s)G(s)]$ assuming $Y(s) = 0$. Thus, a high loop gain $|G(s)|$ will ensure both, a high noise attenuation at low frequencies and a faithful bass reproduction of the communications signal.

The principle of the feedforward control method in ANC is outlined in the lower half of Fig. 19.6(c) and (d). The obvious difference to the feedback control method is that the separate reference signal $x(t)$ is used. Here, cancellation is achieved for the filter transfer function $W = H(s)/C(s)$ which is most often implemented by an adaptive filter. The fact that $x(t)$ reaches the ANC system earlier than $e(t)$ allows for a causal filter, needed in broadband systems. However, a potential problem with this method is the possibility of feedback of the secondary source signal $\hat{y}(t)$ into the path of the reference signal $x(t)$. This is obviously the case when $x(t)$ is picked up by a microphone in a duct just upstream of the secondary source C. An elegant solution for ANC in a duct without explicit feedback cancellation is to use a recursive filter W.

Single error signal/single secondary source systems cannot achieve *global* cancellation or sound control in a room. An intuitive argument for this fact is that one needs *at least* as many secondary sources and error microphones as there are orthogonal wave modes in the room. Since the number of wave modes in a room below a given frequency is approximately proportional to the third power of this frequency, it is clear that ANC (and ASC) is practical only at low frequencies. In practice, using small (point-source) transducers, it turns out that one should use more error microphones than secondary sources. Examples of such multidimensional ANC systems are employed for cancelling the lowest few harmonics of the engine noise in an airplane cabin and in a passenger car. In both of these cases, the adaptive filter matrix is controlled by a multiple-error version of the LMS algorithm. Further information can be found in Nelson and Elliott [1992].

Summary and Acknowledgment

In this section, we have touched upon several topics in audio and electroacoustics. The reader may be reminded that the author's choice of these topics was biased by his background in communication acoustics (and by his lack of knowledge in music). Furthermore, ongoing efforts in integrating different communication modalities into systems for teleconferencing [see, e.g., Flanagan et al., 1990] had a profound effect in focusing this contribution. Experts in topics covered in this contribution, like Jont Allen, David Berkley, Gary Elko, Joe Hall, Jim Johnston, Mead Killion, Harry Levitt, Dennis Morgan, and—last, but not least—Mohan Sondhi, are gratefully acknowledged for their patience and help.

Defining Terms

Audio: Science of processing signals that are within the frequency range of hearing, that is, roughly between 20 Hz and 20 kHz. Also name for this kind of signal.

Critical bands: Broadly used to refer to psychoacoustic phenomena of limited frequency resolution in the cochlea. More specifically, the concept of critical bands evolved in experiments on the audibility of a tone in noise of varying bandwidth, centered around the frequency of the tone. Increasing the noise bandwidth beyond a certain critical value has little effect on the audibility of the tone.

Electroacoustics: Science of interfacing between acoustical waves and corresponding electrical signals. This includes the engineering of transducers (e.g., loudspeakers and microphones), but also parts of the psychology of hearing, following the notion that it is not necessary to present to the ear signal components that cannot be perceived.

Intelligibility maximization and loudness restoration: Two different objectives in fitting hearing aids. Maximizing intelligibility involves conducting laborious intelligibility tests. Loudness restoration involves measuring the mapping between a given sound level and its perceived loudness. Here, we assume that recreating the loudness a normal hearing person would perceive is close to maximizing the intelligibility of speech.

Irrelevance and redundancy: In audio coding, irrelevant portions of an audio signal can be removed without perceptual effect. Once removed, however, they cannot be regenerated in the decoder. Contrary to this, redundant portions of a signal that have been removed in the encoder can be regenerated in the decoder. The "lacking" irrelevant parts of an original signal constitute the major cause for a (misleadingly) low signal-to-noise ratio (SNR) of the decoded signal while its subjective quality can still be high.

Monaural/interaural/binaural: *Monaural* attributes of ear input signals (e.g., timbre, loudness) require, in principle, only one ear to be detected. *Interaural* attributes of ear input signals (e.g., localization in the horizontal plane) depend on differences between, or ratios of measures of, the two ear input signals (e.g., delay and level differences). Psychoacoustic effects (e.g., cocktail-party effect) that depend on the fact that we have two ears are termed *binaural*.

Related Topics

15.2 Speech Enhancement and Noise Reduction • 73.2 Noise

References

J.B. Allen, J.L. Hall, and P.S. Jeng, "Loudness growth in 1/2-octave bands (LGOB) — A procedure for the assessment of loudness," *J. Acoust. Soc. Am.*, vol. 88, no. 2, pp. 745–753, 1990.

J.L. Flanagan, D.A. Berkley, and K.L. Shipley, "Integrated information modalities for human/machine communication: HuMaNet, an experimental system for conferencing," *J. of Visual Communication and Image Representation*, vol. 1, no. 2, pp. 113–126, 1990.

J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, 1991.

E. Haensler, "The hands-free telephone problem—An annotated bibliography," *Signal Processing*, vol. 27, pp. 259–271, 1992.

J.D. Johnston and K. Brandenburg, "Wideband coding—perceptual considerations for speech and music," in *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi, Eds., New York: Marcel Dekker, 1991.

F.R. Moore, *Elements of Computer Music*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

P. A. Nelson and S.J. Elliott, *Active Control of Sound*, London: Academic Press, 1992.

K. C. Pohlmann, *Principles of Digital Audio*, 2nd ed., Carmel, Ind.: SAMS/Macmillan Computer Publishing, 1989.

E. A. G. Shaw, "The acoustics of the external ear," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Studebaker and I. Hochberg, Eds., Baltimore, Md.: University Park Press, 1980.

E. Villchur, "Signal processing to improve speech intelligibility in perceptive deafness," *J. Acoust. Soc. Am.*, vol. 53, no. 6, pp. 1646–1657, 1973.

E.M. Wenzel, "Localization in virtual acoustic displays," *Presence*, vol. 1, pp. 80–107, 1992.

Further Information

A highly informative article that is complementary to this contribution is the one by P. J. Bloom, "High-quality digital audio in the entertainment industry: An overview of achievements and challenges," *IEEE-ASSP Magazine*, Oct. 1985. An excellent introduction to the fundamentals of audio, including music synthesis and digital recording, is contained in the 1992 book *Music Speech Audio*, by W. J. Strong and G. R. Plitnik, available from Soundprint, 2250 North 800 East, Provo, UT 84604 (ISBN 0-9611938-2-4). *Oversampling Delta-Sigma Data Converters* is a 1992 collection of papers edited by J. C. Candy and G. C. Temes. It is available from IEEE Press (IEEE order number PC0274-1). Specific issues of the *Journal of Rehabilitation Research and Development* (ISSN 007-506X), published by the Veterans Administration, are a good source of information on hearing aids, in particular the Fall 1987 issue. *Spatial Hearing* is the title of a 1982 book by J. Blauert, available from MIT Press (ISBN 0-262-02190-0). Anyone interested in *Psychoacoustics* should look into the 1990 book of this title by E. Zwicker and H. Fastl, available from Springer-Verlag (ISBN 0-387-52600-5).

The Institute of Electrical and Electronics Engineers (IEEE) *Transactions on Speech and Audio Processing* is keeping up-to-date on algorithms in audio. Every two to three years, a workshop on applications of signal processing to audio and electroacoustics covers the latest advances in areas introduced in this article. IEEE can be reached at 445 Hoes Lane, Piscataway, NJ 08855-1331, ph. (908) 981-0060. *The Journal of the Audio Engineering Society* (AES) is another useful source of information on audio. The AES can be reached at 60 East 42nd St., Suite 2520, New York, NY 10165-0075, ph. (212) 661-8528. *The Journal of the Acoustical Society of America* (ASA) contains information on physical, psychological, and physiological acoustics, as well as on acoustic signal processing, among other things. ASA's "Auditory Demonstrations" CD contains examples of signals demonstrating hearing-related phenomena ranging from "critical bands" over "pitch" to "binaural beats." ASA can be reached at 500 Sunnyside Blvd., Woodbury, NY 11797-2999, ph. (516) 576-2360.

19.2 Underwater Acoustical Signal Processing

Sanjay K. Mehta and G. Clifford Carter

What Is Underwater Acoustical Signal Processing?

The use of acoustical signals that have propagated through water to detect, classify, and localize underwater objects is referred to as underwater acoustical signal processing.

Why Exploit Sound for Underwater Applications?

It has been found that acoustic energy propagates better under water than other types of energy. For example, both light and radio waves (used for satellite or above-ground communications) are attenuated to a far greater degree under water than are sound waves. For this reason sound waves have generally been used to extract information about underwater objects. A typical underwater acoustical signal processing scenario is shown in Fig. 19.8.

Technical Overview

In underwater acoustics, a number of units are used: distances of nautical miles (1852 m), yards (0.9144 m) and kiloyards; speeds of knots (nautical mile/h); depths of fathoms (6 ft or 1.8288 m); and bearing of degrees (0.01745 rad). However, in the past two decades there has been a conscious effort to be totally metric, i.e., to use MKS or Standard International units.

Underwater acoustic signals to be processed for detection, classification, and localization can be characterized from a statistical point of view. When time averages of each waveform are the same as the ensemble average of waveforms, the signals are ergodic. When the statistics do not change with time, the signals are said to be stationary. The spatial equivalent to stationary is homogeneous. For many introductory problems, only stationary signals and homogeneous noise are considered; more complex problems involve nonstationary, inhomogeneous environments.

Acoustic waveforms of interest have a probability density function (PDF); for example, the PDF may be Gaussian or in the case of clicking, sharp noise spikes, or crackling ice noise, the PDF may be non-Gaussian. In addition to being characterized by a PDF, signals can be characterized in the frequency domain by their power spectral density functions, which are Fourier transforms of the autocorrelation functions. White signals, which are uncorrelated from sample to sample, have a delta function autocorrelation and equivalently a flat (constant) power spectral density. Ocean signals in general are much more colorful and not limited to being stationary.

Passive sonar signals are primarily modeled as random signals. Their first-order PDFs are typically Gaussian; one exception is a stable sinusoidal signal that is non-Gaussian and has a power spectral density function that is a Dirac delta function in the frequency domain. However, in the ocean environment, an arbitrarily narrow frequency width is never observed, and signals have some finite narrow bandwidth. Indeed, the full spectrum of most underwater signals is quite "colorful."

Received active sonar signals can be viewed as consisting of the results of a deterministic component (known transmit waveform) convolved with the medium and reflector transfer functions and a random (noise) component. Moreover, the **Doppler** imparted (frequency shift) to the reflected signal makes the total system effect nonlinear, thereby complicating analysis and processing of these signals.

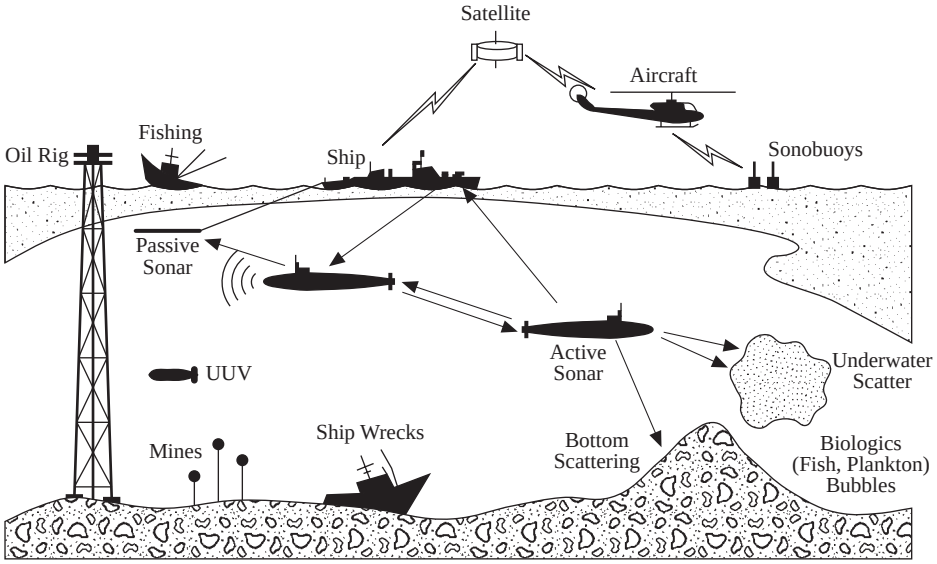


FIGURE 19.8 Active and passive underwater acoustical signal processing.

SONAR

SONAR, "SOund NAvigation and Ranging," the acronym adopted in the 1940s, similar to the popular RADAR, "RAdio Detection And Ranging," involves the use of sound to explore the ocean and underwater objects

- *Passive sonar* uses sound radiated from the underwater object itself. The duration of the radiated sound may be short or long in time and narrow or broad in frequency. Only one-way transmission through the ocean, from the acoustic source to a receiving sensor, is involved in this case.
- *Active sonar* involves echo-ranging where an acoustical signal is transmitted from a source, and reflected echoes are received back from the object. Here one is concerned with two-way transmissions from a transmitter to an object and back to a receiving sensor. There are three types of active sonar systems:

  1. *Monostatic:* In this most common form, the source and receiver are either identical or distinct but located on the same platform (e.g., a surface ship).
  2. *Bistatic:* In this form, the transmitter and receiver are on different platforms.
  3. *Multistatic:* Here, a single (or more) source or transmitter and multiple receivers, which can be located on different receiving platforms or ships, are used.

The performance of sonar systems can be assessed by the passive and active *sonar equations*. The major parameters in the sonar equation, measured in **decibels (dB)**, are as follows:

L_S = source level
L_N = noise level
N_{DI} = directivity index
N_{TS} = echo level or target strength
N_{RD} = recognition differential

Here, L_S is the target-radiated signal strength (for passive) or transmitted signal strength (for active), and L_N is the total background noise level. N_{DI}, or DI, is the directivity index, which is a measure of the capability of a receiving array to discriminate against unwanted noise. N_{TS} is the received echo level or target strength. Underwater objects with large values of N_{TS} are more easily detectable with active sonar than are those with small values of N_{TS}. In general, N_{TS} varies as a function of object size, aspect angle, (i.e., the direction at which impinging acoustic energy reaches the underwater object), and reflection angle (the direction at which the impinging acoustic energy is reflected off the underwater object). N_{RD} is the recognition differential of the processing system.

TABLE 19.1 Expressions for Sound Speed in Meters per Second

Expression	Limits
$c = 1492.9 + 3(T-10) - 6 \times 10^{-3}(T-10)^2$	$-2 \le T \le 24.5°$
$\quad 4 \times 10^{-2}(T-18)^2 + 1.2(S-35)$	$30 \le S \le 42$
$\quad - 10^{-2}(T-18)(S-35) + D/61$	$0 \le D \le 1.000$
$c = 1449.2 + 4.6T - 5.5 \times 10^{-2}T^2$	$0 \le T \le 35°$
$\quad + 2.9 \times 10^{-4}T^3 + (1.34 - 10^{-2}T)(S-35)$	$0 \le S \le 45$
$\quad + 1.6 \times 10^{-2}D$	$0 \le D \le 1{,}000$
$c = 1448.96 + 4.591T - 5.304 \times 10^{-2}T^2$	$0 \le T \le 30°$
$\quad + 2.374 \times 10^{-4}T^3 + 1.340(S-35)$	$30 \le S \le 40$
$\quad + 1.630 \times 10^{-2}D + 1.675 \times 10^{-7}D^2$	$0 \le D \le 8{,}000$
$\quad - 1.025 \times 10^{-2}T(S-35) - 7.139 \times 10^{-13}TD^3$	

D = depth, in meters. S = salinity, in parts per thousand. T = temperature, in degrees Celsius.

The **figure of merit** (FOM), a basic performance measure involving parameters of the sonar system, ocean, and target, is computed for active and passive sonar systems (in dBs) as follows:
For passive sonar,

$$\text{FOM}_P = L_S - (L_N - N_{DI}) - N_{RD} \tag{19.5}$$

For active sonar,

$$\text{FOM}_A = (L_S + N_{TS}) - (L_N - N_{DI}) - N_{RD} \tag{19.6}$$

Sonar systems, for a given set of parameters of the sonar equations, are designed so that the FOM exceeds the acoustic propagation loss. The amount above the FOM is called the *signal excess*. When two sonar systems are compared, the one with the largest signal excess is said to hold the *acoustic advantage*. However, it should be noted that the set of parameters in the preceding FOM equations is not unique. Depending on the design or parameter measurability conditions, different parameters can be combined or expanded in terms of quantities such as frequency dependency of the sonar system in particular ocean conditions, speed and bearing of the receiving or transmitting platforms, reverberation loss, and so forth. Furthermore, due to multipaths, differences in sonar system equipment and operation, and the constantly changing nature of the ocean medium, the FOM parameters fluctuate with time. Thus, the FOM is not an absolute measure of performance but rather an expected value of performance over time in a stochastic sense [for details, see Urick, 1983].

Underwater Propagation

Speed/Velocity of Sound

Sound speed, c, in the ocean, in general lies between 1450–1540 m/s and varies as a function of several physical parameters, such as temperature, salinity, and pressure (depth). Variations in sound speed can significantly affect the propagation (range or quality) of sound in the ocean. Table 19.1 gives approximate expressions for sound speed as a function of these physical parameters.

Sound Velocity Profiles

Sound rays that are normal (perpendicular) to the acoustic wavefront can be traced from the source to the receiver by a process called ray tracing.[1] In general, the acoustic ray paths are not straight, but bend in a manner analogous to optical rays focused by a lens. In underwater acoustics, the ray paths are determined by the **sound velocity profile** (SVP) or *sound speed profile* (SSP), that is, the speed of sound in water as a function of water

[1]Ray tracing models are used for high-frequency signals and in deep water. Generally, if the depth-to-wavelength ratio is 100 or more, ray tracing models are accurate. Below that, corrections must be made to the ray trace models. In shallow water or low frequencies, i.e., when depth-to-wavelength is about 30 or less, "mode theory" models are used.

depth. The sound speed not only varies with depth but also varies in different regions of the ocean and with time as well. In deep water, the SVP fluctuates the most in the upper ocean due to variations of temperature and weather. Just below the sea surface is the *surface layer* where the sound speed is greatly affected by temperature and wind action. Below this layer lies the *seasonal thermocline* where the temperature and speed decrease with depth, and the variations are seasonal. In the next layer, the *main thermocline*, the temperature and speed decrease with depth and surface conditions or seasons have little effect. Finally, there is the *deep isothermal layer* where the temperature is nearly constant at 39°F, and the sound velocity increases almost linearly with depth. A typical deep water sound velocity profile as a function of depth is shown in Fig. 19.9.

If the sound speed is a minimum at a certain depth below the surface, then this depth is called axis of the underwater sound channel.[2] The sound velocity increases both above and below this axis. When the sound wave travels through a medium with a sound speed gradient, the direction of travel of sound wave is bent towards the area of lower sound speed.

Although the definition of shallow water can be signal dependent, in terms of depth-to-wavelength ratio, water depth of less than 1000 meters is generally referred to as shallow water. In shallow water the SVP is irregular and difficult to predict because of large surface temperature and salinity variations, wind effects, and multiple reflections of sound from the ocean bottom.

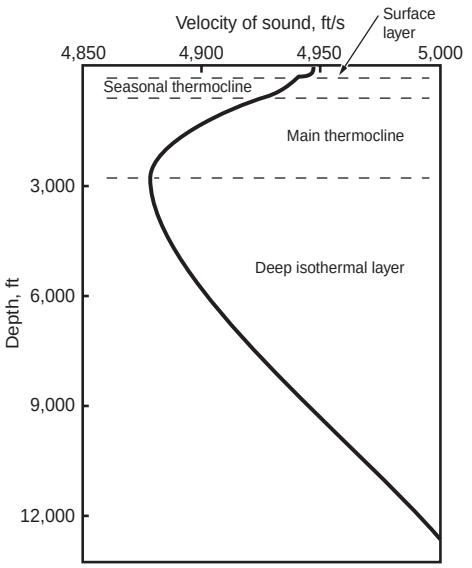


FIGURE 19.9 A typical sound velocity profile (SVP).

Propagation Modes

In general, there are three dominant propagation paths that depend on the distance or range between the acoustic source and the receiver (Fig. 19.10).

- *Direct Path:* Sound energy travels in (nominal) straight line path between the source and receiver, usually present at short ranges.
- *Bottom Bounce Path:* Sound energy is reflected from the ocean bottom (present at intermediate ranges).
- *Convergence Zone (CZ) Path:* Sound energy converges at longer ranges where multiple acoustic ray paths add or recombine coherently to reinforce the presence of acoustic energy from the radiating/reflecting source.

Figure 19.11 shows the propagation loss as a function of range for different frequencies of the signal. Note the recombination of energy at the convergence zones.

Multipaths

The ocean contains multiple acoustic paths that split the acoustic energy. When the receiving system can resolve these multiple paths (or multipaths), then they should be recombined by optimal signal processing to fully exploit the available acoustic energy for detection [Chan, 1989]. It is also theoretically possible to exploit the geometrical properties of multipaths present in the bottom bounce path by investigation of the apparent aperture created by the different path arrivals to localize the energy source. In the case of first-order bottom bounce transmission, i.e., only one bottom interaction, there are four paths (from source to receiver):

1. A bottom bounce ray path (B).
2. A surface interaction followed by a bottom interaction (SB).
3. A bottom bounce followed by a surface interaction (BS).
4. A path that first hits the surface, then the bottom, and finally the surface (SBS).

Typical first-order bottom bounce ocean propagation paths are depicted in Fig. 19.12.

[2]Often called the SOFAR (Sound Fixing and Ranging) channel.

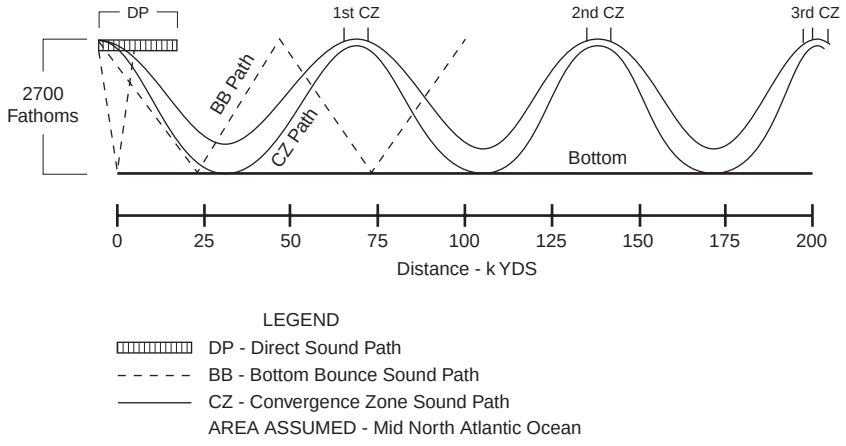


FIGURE 19.10 Typical sound paths between source and receiver. (*Source:* A.W. Cox, *Sonar and Underwater Sound*, Lexington, Mass., Lexington Books, D.C. Health and Company, 1974, p. 25. With permission.)

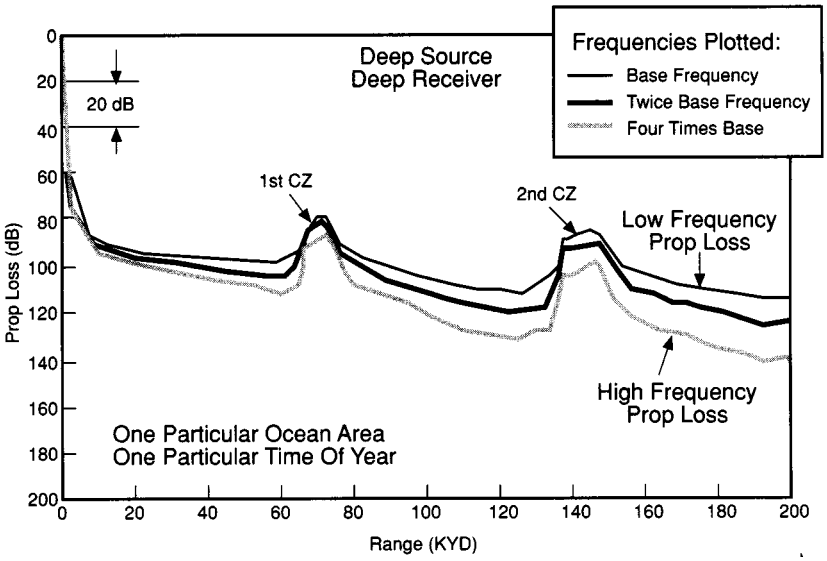


FIGURE 19.11 Propagation loss as a function of range.

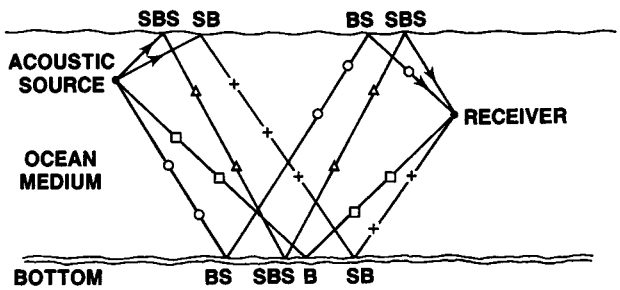


FIGURE 19.12 Multipaths for a first-order bottom bounce propagation model.

Performance Limitations

In a typical reception of a signal wavefront, noise and interference can degrade the performance of a sonar system and limit the system's ability to detect signals in the underwater environment. The noise or interference could be sounds from a school of fish, shipping (surface or subsurface) noise, active transmission interference (e.g., jammers), or interference when multiple receivers or sonar systems are in operation simultaneously. Also, the ambient noise may have unusual vertical or horizontal directivity and in some environments, such as the Arctic, the noise due to ice motion may produce unfamiliar interference. Unwanted backscatters, similar to the headlights of a car driving in fog, can cause a signal-induced noise that degrades processing gain without proper processing. Some other performance-limiting factors are the loss of signal level and acoustic coherence due to boundary interaction as a function of grazing angle; the radiated pattern (signal level) of the object and its spatial coherence; the presence of surface, bottom, and volume **reverberation** (in active sonar); signal spreading owing to the modulating effect of surface motion; biologic noise as a function of time (both time of day and time of year); and statistics of the noise in the medium. (Does the noise arrive in the same or at different ray path angles as the signal?)

Hydrophone Sensors and Output

Hydrophone sensors are underwater microphones capable of operating in water and under hydrostatic pressure. These sensors receive radiated and reflected acoustic energy that arrives through the multiple paths of the ocean medium from a variety of sources and reflectors. As with a microphone, hydrophones convert acoustic pressure to electrical voltages or to optical signals.

A block diagram model of a stationary acoustic source, $s(t)$, input to M unique hydrophone receivers is shown in Fig. 19.13. Multipaths from the source to each receiver can be characterized by the source to (each individual) receiver impulse response. The inverse Fourier transforms of these impulse responses are the transfer functions shown in the block diagram as $A_j(f)$, where the subscript, $j = 1,\ldots, M$, denotes the appropriate source-to-receiver transfer function. For widely spaced receivers, there will be a different transfer function from the source to each receiver. Also, for multiple sources and widely spaced receivers, there will be a different transfer function from each source to each receiver. The receiver outputs from a single source are modeled as being corrupted by additive noise, $n_j(t)$, as shown in Fig. 19.13.

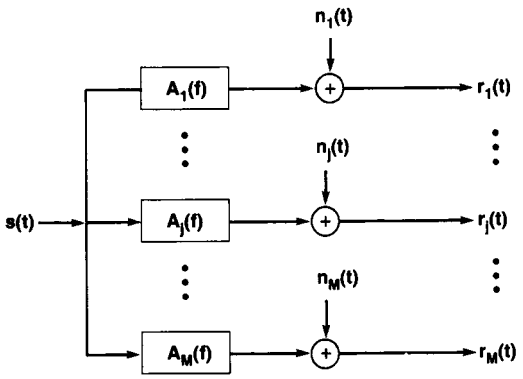


FIGURE 19.13 Hydrophone receiver model: source signal $s(t)$ through medium filter $A_j(t)$, corrupted by additive noise received at one of M hydrophones.

Processing Functions

Beamforming

Beamforming is a process in which outputs from the hydrophone sensors of an array are coherently combined by delaying and summing the outputs to provide enhanced detection and estimation. In underwater applications, one is trying to detect a directional (single direction) signal in the presence of normalized background noise that is ideally isotropic (nondirectional). By arranging the hydrophone (array) sensors in different physical geometries and electronically steering them in a particular direction, one can increase the signal-to-noise ratio (**SNR**) in a given direction by rejecting or canceling the noise in other directions. There are many different kinds of arrays (e.g., equally spaced line, continuous line, circular, cylindrical, spherical, or random sonobuoy arrays). The beam pattern specifies the response of these arrays to the variation in direction. In the simplest case, the increase in SNR due to the beamformer, called the *array gain* (in dB), is given by

$$\mathrm{AG} = 10 \log \frac{\mathrm{SNR}_{\text{array (output)}}}{\mathrm{SNR}_{\text{single sensor (input)}}} \tag{19.7}$$

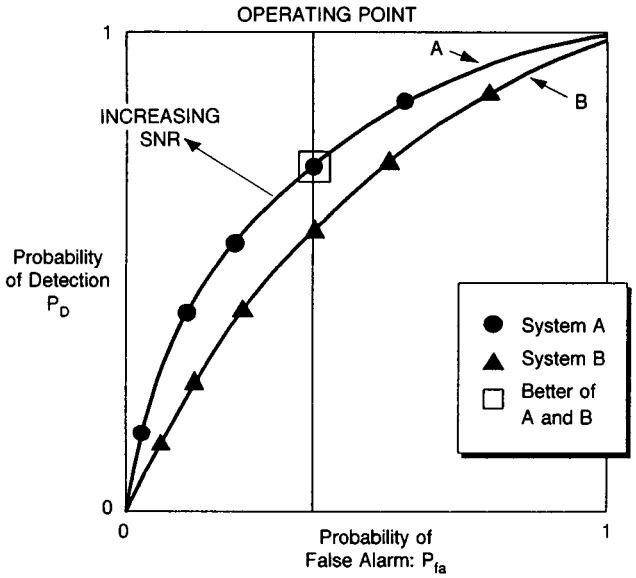


FIGURE 19.14 Typical ROC curves. Note points (0,0) and (1,1) are on all ROC curves; upper curve represents higher P_D for fixed P_{fa} and hence better performance by having higher SNR or processing time.

Detection

Detection of signals in the presence of noise, using classical Bayes or Neyman-Pearson decision criteria, is based on hypothesis testing. In the simplest binary hypothesis case, the detection problem is posed as two hypotheses:

- H_0: Signal is not present (referred to as the null hypothesis).
- H_1: Signal is present.

For a received wavefront, H_0 relates to the noise-only case and H_1 to the signal-plus-noise case. Complex hypotheses (M-hypotheses) can also be formed if detecting a signal among a variety of sources is required.

Probability is a measure, between zero and unity, of how likely an event is to occur. For a received wavefront the likelihood ratio, Λ, is the ratio of P_{H1} (probability that hypothesis H_1 is true) to P_{H0} (probability that hypothesis H_0 is true). A decision (detection) is made by comparing the likelihood, or logarithm of the likelihood ratio called the log-likelihood ratio, to a predetermined threshold η. That is, if $\Lambda = P_{H1}/P_{H0} > \eta$, a decision is made that the signal is present.

Probability of detection, P_D, measures the likelihood of detecting an event or object when the event does occur. *Probability of false alarm*, P_{fa}, is a measure of the likelihood of saying something happened when the event did NOT occur. **Receiver operating characteristics (ROC)** curves plot P_D versus P_{fa} for a particular (sonar signal) processing system. A single plot of P_D versus P_{fa} for one system must fix the SNR and processing time. The threshold η is varied to sweep out the ROC curve. The curve is often plotted on either log-log scale or "probability" scale. In comparing a variety of processing systems one would like to select the system (or develop a new one) that maximizes the P_D for every given P_{fa}. Processing systems must operate on their ROC curves, but most processing systems allow the operator to select where on the ROC curve the system is operated by adjusting a threshold; low thresholds ensure a high probability of detection at the expense of high false alarm rate. A sketch of two monotonically increasing ROC curves is given in Fig. 19.14. By proper adjustment of the decision threshold, one can trade off detection performance for false alarm performance. Since the points (0,0) and (1,1) are on all ROC curves, one can always guarantee 100% probability of detection with an arbitrarily low threshold (albeit at the expense of 100% probability of false alarm) or 0% probability of false alarm with an arbitrarily high threshold (albeit at the expense of 0% probability of detection). The (log) *likelihood detector* is a detector that achieves the maximum probability of detection for fixed probability of false alarm; it is shown in Fig. 19.15 for detecting Gaussian signals reflected or radiated from the stationary objects modeled in Fig. 19.13. For moving objects more complicated time compression or Doppler compensation processing is

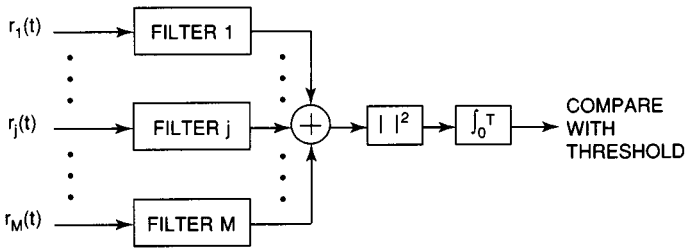


FIGURE 19.15 Log likelihood detector structure for uncorrelated Gaussian noise in the received signal $r_j(t)$, $j = 1,...,M$.

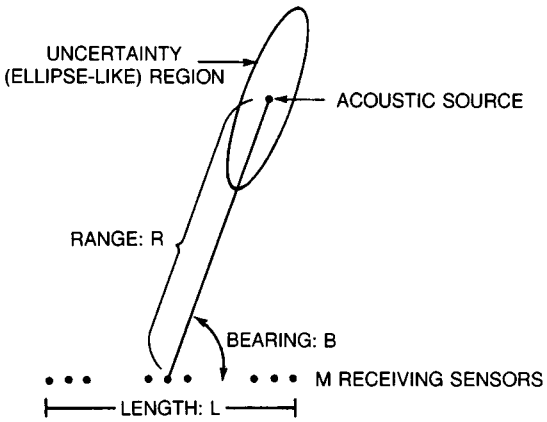


FIGURE 19.16 Array geometry used to estimate source position. (*Source:* G.C. Carter, "Coherence and time delay estimation," *Proceedings IEEE*, vol. 75, no. 2, p. 251, © 1987 IEEE. With permission.)

required. For spiky non-Gaussian noise, other signal processing is required; indeed, clipping prior to filtering improves detection performance, by "eliminating" strong noise "pulses".

In active sonar, the filters are matched to the known transmitted waveforms. If the object (acoustic reflector) has motion, it will induce Doppler on the reflected signal, and the receiver will be complicated by the addition of a bank of Doppler compensators. Returns from a moving object are shifted in frequency by $\Delta f = (2v/c)f$, where v is the relative velocity (range rate) between the source and object, c is the speed of sound in water, and f is the operating frequency of the source transmitter.

In passive sonar, at low SNR, the optimal filters in Fig. 19.15 (so-called Eckart filters) are functions of $G_{ss}^{1/2}(f)/G_{nn}(f)$, where f is frequency in hertz, $G_{ss}(f)$ is the signal power spectrum, and $G_{nn}(f)$ is the noise power spectrum [see page 484 Carter (1993)].

Estimation/Localization

The second function of underwater acoustic signal processing estimates the parameters that localize the position of the detected object. The source position is estimated in range, bearing, and depth, typically from the underlying parameter of **time delay** associated with the acoustic wavefront. The statistical uncertainty of the positional estimates is important. Knowledge of the first order probability density function or its first- and second-order moments, the mean (expected value) and the variance, are vital to understanding the expected performance of the processing system. In the passive case, the ability to estimate range is extremely limited by the geometry of the measurements; indeed, the variance of passive range estimates can be extremely large, especially when the true range to the acoustic source is long when compared with the aperture length of the receiving array. Figure 19.16 depicts direct path passive ranging uncertainty from a collinear array with sensors clustered so as to minimize the bearing and uncertainty region. Beyond the direct path, multipath signals can be processed to estimate source depth covertly. Range estimation accuracy is not difficult with the active sonar, but active sonar is not covert, which for some applications can be important.

Classification

The third function of sonar signal processing is classification. This function determines the type of object that has radiated or reflected acoustic energy. For example, was the sonar signal return from a school of fish or a reflection from the ocean bottom? The action one takes is highly dependent upon this important function. The amount of radiated or reflected signal power relative to the background noise (that is, SNR) necessary to achieve good classification may be higher than for detection. Also, the type of signal processing required for classification may be different than the type of processing for detection. Processing methods that are developed on the basis of detection might not have the requisite SNR to adequately perform the classification function. Classifiers are, in general, divided into feature (or clue) extractors followed by a classifier decision box. A key to successful classification is feature extraction. Performance of classifiers is plotted as in ROC detection curves as probability of deciding on class A, given A was actually present, or $P(A/A)$, versus the probability of deciding on class B, given that A was present, i.e., $P(B/A)$, for two different classes of objects, A and B. Of course, for the same class of objects, one could also plot $P(B/B)$ versus $P(A/B)$.

Motion Analysis or Tracking

The fourth function of underwater acoustic signal processing is to perform contact (or target) motion analysis (TMA), that is, to estimate parameters of bearing and speed. Generally, nonlinear filtering methods, including Kalman-Bucy filters, are applied; typically these methods rely on a state space model for the motion of the contact. For example, the underlying model of motion could assume a straight-line course and constant speed of the contact of interest. When the acoustic source of interest behaves like the model, then results consistent with the basic theory can be expected. It is also possible to incorporate motion compensation into the signal processing detection function. For example, in the active sonar case, proper waveform selection and processing can reduce the degradation of detector performance caused by uncompensated Doppler. Moreover, joint detection and estimation can provide clues to the TMA and classification processes. For example, if the processor simultaneously estimates depth in the process of performing detection, then a submerged object would not be classified as a surface object. Also, joint detection and estimation using Doppler for detection can directly improve contact motion estimates.

Normalization

Another important signal processing function for the detection of weak signals in the presence of unknown and (temporal and spatial) varying noise is normalization. The statistics of noise or reverberation for oceans typically varies in time, frequency, and/or bearing from measurement to measurement and location to location. To detect a weak signal in a broadband, nonstationary, and inhomogeneous background, it is usually desirable to make the noise background statistics as uniform as possible for the variations in time, frequency, and/or bearing. The noise background estimates are first obtained from a window of resolution cells (which usually surrounds the test data cell). These estimates are then used to normalize the test cell, thus reducing the effects of the background noise on detection. Window length and distance from the test cell are two of the parameters that can be adjusted to obtain accurate estimates of the different types of stationary or nonstationary noise.

Advanced Signal Processing

Adaptive Beamforming

Beamforming was discussed in an earlier section. The cancellation of noise through beamforming can also be done adaptively, which can improve the array gain further. Some of the various adaptive beamforming techniques are [Knight et al., 1981], Dicanne, sidelobe cancellers, maximum entropy array processing, and maximum-likelihood (ML) array processing.

Coherence Processing

Coherence is a normalized (to lie between zero and unity) cross-spectral density function that is a measure of the similarity of received signals and noise between any sensors of the array. The complex coherence function between two wide-sense-stationary processes x and y is defined by

$$\gamma_{xy}(f) = \frac{G_{xy}(f)}{\sqrt{G_{xx}(f)G_{yy}(f)}} \tag{19.8}$$

where, as before, f is the frequency in hertz and G is the power spectrum function. Array gain depends on the coherence of the signal and noise between the sensors of the array. To increase the array gain, it is necessary to have good coherence among the sensors for the signal, but poor coherence (incoherent) for the noise. Coherence of the signal between sensors improves with decreasing separation between the sensors, frequency of the received waveform, total bandwidth, and integration time. Loss of coherence of the signal could be due to ocean motion, object motion, multipaths, reverberation, or scattering. The coherence function has many uses, including measurement of SNR or array gain, system identification, and determination of time delays [Carter, 1987].

Acoustic Data Fusion

Acoustic data fusion is a technique that combines information from multiple receivers or receiving platforms about a common object or channel. Instead of each receiver making a decision, relevant information from the different receivers is sent to a common control unit where the acoustic data is combined and processed (hence the name *data fusion*). After fusion, a decision can be relayed or "fed" back to each of the receivers. If data transmission is a concern, due to time constraints, cost, or security, other techniques can be used in which each receiver makes a decision and transmits only the decision. The control unit makes a global decision based on the decisions of all the receivers and relays this global decision back to the receivers. This is called "distributed detection." The receivers can then be asked to re-evaluate their individual decisions based on the new global decision. This process could continue until all the receivers are in agreement or could be terminated whenever an acceptable level of consensus is attained.

An advantage of data fusion is that the receivers can be located at different ranges (e.g., on two different ships), in different mediums (shallow or deep water, or even at the surface), and at different bearings from the object, thus giving comprehensive information about the object or the underwater acoustic channel.

Application

Since World War II, in addition to military applications, there has been an expansion in commercial and industrial underwater acoustics applications. Table 19.2 lists the military and nonmilitary functions of sonar along with some of the current applications.

Defining Terms

Decibels (dB): Logarithmic scale of representing the ratio of two quantities given as $10 \log_{10}(P_1/P_0)$ for power level ratios and $20 \log_{10}(V_1/V_0)$ for comparing acoustic pressure or voltage ratios. A standard reference pressure or intensity level in SI units is equal to 1 micropascal (1 pascal = 1 newton per square meter = 10 dyne per square centimeter).

Doppler shift: Shift in frequency of transmitted waveform due to the relative motion between the source and object.

Figure of merit/sonar equation: Performance evaluation measure for the various target and equipment parameters of a sonar system. It is a subset of the broader sonar performance given by the *sonar equations*, which includes reverberation effects.

Hydrophone: Receiving sensors that convert sound energy into electrical or optical energy (analogous to underwater microphones).

Receiver operating characteristics (ROC) curves: Plots of the probability of detection (likelihood of detecting the object when the object is present) versus the probability of false alarm (likelihood of detecting the object when the object is not present) for a particular processing system.

TABLE 19.2 Underwater Acoustics Applications

Function	Description
Military	
Detection	Deciding if a target is present or not.
Classification	Deciding if a detected target does or does not belong to a specific class.
Localization	Measuring at least one of the instantaneous positions and velocity components of a target (either relative or absolute), such as range, bearing, range rate, or bearing rate.
Navigation	Determining, controlling, and/or steering a course through a medium (includes avoidance of obstacles and the boundaries of the medium).
Communications	Instead of a wire link, transmitting and receiving acoustic power and information.
Control	Using a sound-activated release mechanism.
Position marking	Transmitting a sound signal continuously (beacons) or transmitting only when suitably interrogated (transponders).
Depth sounding	Sending short pulses downward and timing the bottom return.
Acoustic-speedometers	Using pairs of transducers pointing obliquely downwards to obtain speed over the bottom from the Doppler shift of the bottom return.
Commercial Applications:	
Industrial	**Oceanological**
Fish finders/fish herding	Subbottom geological mapping
Oil and mineral explorations	Ocean topography
River flow meter	Bathyvelocimeter
Acoustic holography	Emergency telephone
Viscosimeter	Seismic simulation and measurement
Acoustic ship docking system	Biological signal and noise measurement
Ultrasonic grinding/drilling	Sonar calibration

Reverberation/clutter: Inhomogeneities, such as dust, sea organisms, schools of fish, sea mounds on the bottom of the sea, form mass density discontinuities in the ocean medium. When an acoustic wave strikes these inhomogeneities, some of the acoustic energy is reflected and reradiated. The sum total of all such reradiations is called reverberation. Reverberation is present only in active sonar, and in the case where the object echoes are completely masked by reverberation, the sonar system is said to be "reverberation limited."

SONAR: Acronym for "SOund NAvigation and Ranging," adopted in the 1940s, involves the use of sound to explore the ocean and underwater objects.

Sound velocity profile (SVP): Description of the speed of sound in water as a function of water depth.

SNR: The signal-to-noise (power) ratios, usually measured in decibels (dB).

Time delay: The time (delay) difference in seconds from when an acoustic wavefront impinges on one hydrophone or receiver until it strikes another.

Related Topic

16.1 Spectral Analysis

References

L. Brekhovskikh and Yu. Lysanov, *Fundamentals of Ocean Acoustics*, New York.: Springer-Verlag, 1982.
W.S. Burdic, *Underwater Acoustic System Analysis*, Englewood Cliffs, N.J.: Prentice-Hall, 1984.
G.C. Carter, Coherence and time delay estimation, Piscataway, NJ: IEEE Press, 1993.
A.W. Cox, *Sonar and Underwater Sound*, Lexington, Mass.: Lexington Books, D.C. Health and Company, 1974.

W.C. Knight, R.G. Pridham, and S.M. Kay, "Digital signal processing for sonar," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1451–1506, Nov. 1981.
R.O. Nielsen, *Sonar Signal Processing*, Boston: Artech House, 1991.
A.V. Oppenheim, Ed., *Applications of Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1980.
R.J. Urick, *Principles of Underwater Sound*, New York.: McGraw-Hill, 1983.
H.L. Van Trees, *Detection, Estimation, and Modulation Theory*, New York: John Wiley & Sons, 1968.
L.J. Ziomek, *Underwater Acoustics, A Linear Systems Theory Approach*, New York: Academic Press, 1985.

Further Information

Journal of Acoustical Society of America (JASA), *IEEE Transactions on Signal Processing* (formerly the *IEEE Transactions on Acoustics, Speech and Signal Processing*), and *IEEE Journal of Oceanic Engineering* are professional journals providing current information on underwater acoustical signal processing.

The annual meetings of the International Conference on Acoustics, Speech and Signal Processing, sponsored by the IEEE, and the biannual meetings of the Acoustical Society of America are a good source for current trends and technologies.

A detailed tutorial on *Digital Signal Processing for Sonar* by W.C. Knight et al. is an informative and detailed tutorial on the subject [Knight et al., 1981].

Principe, J.C. "Artificial Neural Networks"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

20

Artificial Neural Networks

20.1 Definitions and Scope
Introduction • Definitions and Style of Computation • ANN Types and Applications
20.2 Multilayer Perceptrons
Function of Each PE • How to Train MLPs • Applying Back-Propagation in Practice • *A Posteriori* Probabilities
20.3 Radial Basis Function Networks
20.4 Time Lagged Networks
Memory Structures • Training-Focused TLN Architectures
20.5 Hebbian Learning and Principal Component Analysis Networks
Hebbian Learning • Principal Component Analysis • Associative Memories
20.6 Competitive Learning and Kohonen Networks

Jose C. Principe
University of Florida

20.1 Definitions and Scope

Introduction

Artificial neural networks (ANN) are among the newest signal-processing technologies in the engineer's toolbox. The field is highly interdisciplinary, but our approach will restrict the view to the engineering perspective. In engineering, neural networks serve two important functions: as pattern classifiers and as nonlinear adaptive filters. We will provide a brief overview of the theory, learning rules, and applications of the most important neural network models.

Definitions and Style of Computation

An ANN is an adaptive, most often nonlinear system that learns to perform a function (an input/output map) from data. Adaptive means that the system parameters are changed during operation, normally called the *training phase*. After the training phase the ANN parameters are fixed and the system is deployed to solve the problem at hand (the *testing phase*). The ANN is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the *learning rule*. The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The nonlinear nature of the neural network processing elements (PEs) provides the system with lots of flexibility to achieve practically any desired input/output map, i.e., some ANNs are *universal mappers*.

There is a style in neural computation that is worth describing (Fig. 20.1). An input is presented to the network and a corresponding desired or target response set at the output (when this is the case the training is called *supervised*). An error is composed from the difference between the desired response and the system

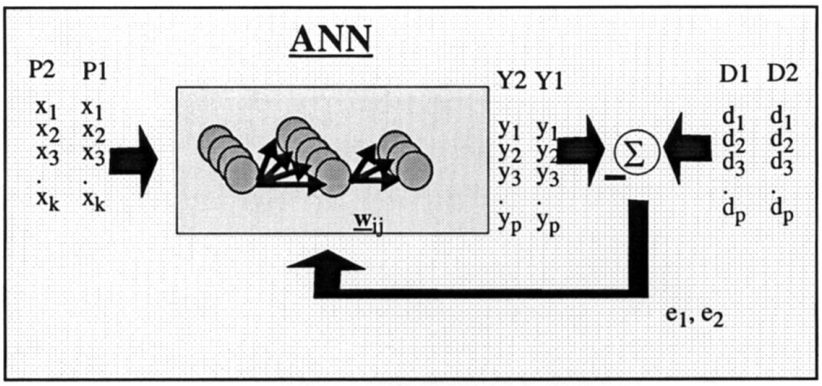


FIGURE 20.1 The style of neural computation.

output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable. It is clear from this description that the performance hinges heavily on the data. If one does not have data that cover a significant portion of the operating conditions or if they are noisy, then neural network technology is probably not the right solution. On the other hand, if there is plenty of data and the problem is poorly understood to derive an approximate model, then neural network technology is a good choice.

This operating procedure should be contrasted with the traditional engineering design, made of exhaustive subsystem specifications and intercommunication protocols. In ANNs, the designer chooses the network topology, the performance function, the learning rule, and the criterion to stop the training phase, but the system automatically adjusts the parameters. So, it is difficult to bring *a priori* information into the design, and when the system does not work properly it is also hard to incrementally refine the solution. But ANN-based solutions are extremely efficient in terms of development time and resources, and in many difficult problems ANNs provide performance that is difficult to match with other technologies. Denker 10 years ago said that "ANNs are the second best way to implement a solution" motivated by the simplicity of their design and because of their universality, only shadowed by the traditional design obtained by studying the physics of the problem. At present, ANNs are emerging as the technology of choice for many applications, such as pattern recognition, prediction, system identification, and control.

ANN Types and Applications

It is always risky to establish a taxonomy of a technology, but our motivation is one of providing a quick overview of the application areas and the most popular topologies and learning paradigms.

Application	Topology	Supervised Learning	Unsupervised Learning
Association	Hopfield [Zurada, 1992; Haykin, 1994]	—	Hebbian [Zurada, 1992; Haykin, 1994; Kung, 1993]
	Multilayer perceptron [Zurada, 1992; Haykin, 1994; Bishop, 1995]	Back-propagation [Zurada, 1992; Haykin, 1994; Bishop, 1995]	—
	Linear associative mem. [Zurada, 1992; Haykin, 1994]	—	Hebbian
Pattern recognition	Multilayer perceptron [Zurada, 1992; Haykin, 1994; Bishop, 1995]	Back-propagation	—
	Radial basis functions [Zurada, 1992; Bishop, 1995]	Least mean square	*k*-means [Bishop, 1995]
Feature extraction	Competitive [Zurada, 1992; Haykin, 1994]	—	Competitive
	Kohonen [Zurada, 1992; Haykin, 1994]	—	Kohonen
	Multilayer perceptron [Kung, 1993]	Back-propagation	—
	Principal comp. anal. [Zurada, 1992; Kung, 1993]	—	Oja's [Zurada, 1992; Kung, 1993]
Prediction, system ID	Time-lagged networks [Zurada, 1992; Kung, 1993; de Vries and Principe, 1992]	Back-propagation through time [Zurada, 1992]	—
	Fully recurrent nets [Zurada, 1992]		

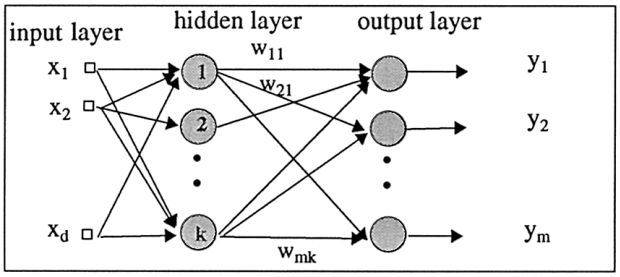


FIGURE 20.2 MLP with one hidden layer (*d*-*k*-*m*).

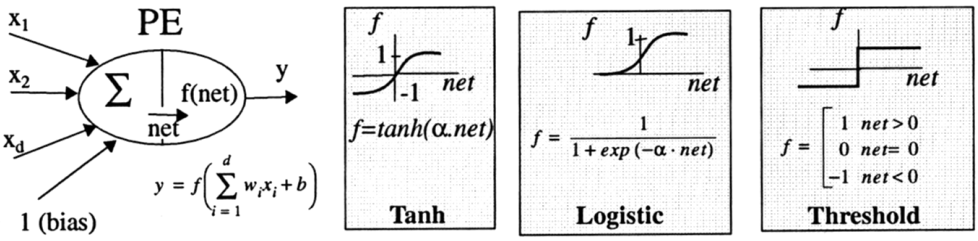


FIGURE 20.3 A PE and the most common nonlinearities.

It is clear that *multilayer perceptrons* (MLPs), the *back-propagation algorithm* and its extensions — time-lagged networks (TLN) and back-propagation through time (BPTT), respectively — hold a prominent position in ANN technology. It is therefore only natural to spend most of our overview presenting the theory and tools of back-propagation learning. It is also important to notice that *Hebbian learning* (and its extension, the Oja rule) is also a very useful (and biologically plausible) learning mechanism. It is an *unsupervised learning* method since there is no need to specify the desired or target response to the ANN.

20.2 Multilayer Perceptrons

Multilayer perceptrons are a layered arrangement of nonlinear PEs as shown in Fig. 20.2. The layer that receives the input is called the *input layer*, and the layer that produces the output is the *output layer*. The layers that do not have direct access to the external world are called *hidden layers*. A layered network with just the input and output layers is called the *perceptron*. Each connection between PEs is weighted by a scalar, w_i, called a *weight*, which is adapted during learning.

The PEs in the MLP are composed of an adder followed by a smooth saturating nonlinearity of the sigmoid type (Fig. 20.3). The most common saturating nonlinearities are the logistic function and the hyperbolic tangent. The threshold is used in other nets. The importance of the MLP is that it is a universal mapper (implements arbitrary input/output maps) when the topology has at least two hidden layers and sufficient number of PEs [Haykin, 1994]. Even MLPs with a single hidden layer are able to approximate continuous input/output maps. This means that rarely we will need to choose topologies with more than two hidden layers. But these are existence proofs, so the issue that we must solve as engineers is to choose how many layers and how many PEs in each layer are required to produce good results.

Many problems in engineering can be thought of in terms of a transformation of an input space, containing the input, to an output space where the desired response exists. For instance, dividing data into classes can be thought of as transforming the input into 0 and 1 responses that will code the classes [Bishop, 1995]. Likewise, identification of an unknown system can also be framed as a mapping (function approximation) from the input to the system output [Kung, 1993]. The MLP is highly recommended for these applications.

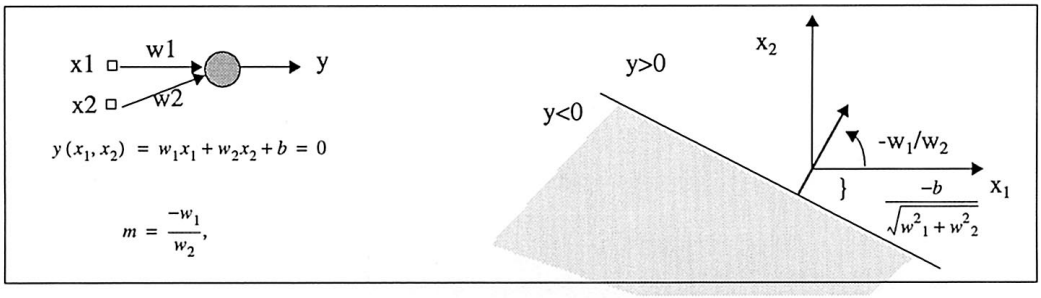


FIGURE 20.4 A two-input PE and its separation surface.

Function of Each PE

Let us study briefly the function of a single PE with two inputs [Zurada, 1992]. If the nonlinearity is the threshold nonlinearity we can immediately see that the output is simply 1 and –1. The surface that divides these subspaces is called a *separation surface,* and in this case it is a line of equation

$$y(w_1, w_2) = w_1 x_1 + w_2 x_2 + b = 0 \tag{20.1}$$

i.e., the PE weights and the bias control the orientation and position of the separation line, respectively (Fig. 20.4). In many dimensions the separation surface becomes an hyperplane of dimension one less than the dimensionality of the input space. So, each PE creates a dichotomy in the input space. For smooth nonlinearities the separation surface is not crisp; it becomes fuzzy but the same principles apply. In this case, the size of the weights controls the width of the fuzzy boundary (larger weights shrink the fuzzy boundary).

The perceptron input/output map is built from a juxtaposition of linear separation surfaces, so the perceptron gives zero classification error only for *linearly separable classes* (i.e., classes that can be exactly classified by hyperplanes).

When one adds one layer to the perceptron creating a one hidden layer MLP, the type of separation surfaces changes drastically. It can be shown that this learning machine is able to create "bumps" in the input space, i.e., an area of high response surrounded by low responses [Zurada, 1992]. The function of each PE is always the same, no matter if the PE is part of a perceptron or an MLP. However, notice that the output layer in the MLP works with the result of hidden layer activations, creating an embedding of functions and producing more complex separation surfaces. The one-hidden-layer MLP is able to produce *nonlinear separation surfaces.*

If one adds an extra layer (i.e., two hidden layers), the learning machine now can combine at will bumps, which can be interpreted as a *universal mapper,* since there is evidence that any function can be approximated by localized bumps. One important aspect to remember is that changing a single weight in the MLP can drastically change the location of the separation surfaces; i.e., the MLP achieves the input/output map through the interplay of all its weights.

How to Train MLPs

One fundamental issue is how to adapt the weights w_i of the MLP to achieve a given input/output map. The core ideas have been around for many years in optimization, and they are extensions of well-known engineering principles, such as the *least mean square (LMS) algorithm* of adaptive filtering [Haykin, 1994]. Let us review the theory here. Assume that we have a linear PE (f(net) = net) and that one wants to adapt the weights as to minimize the square difference between the desired signal and the PE response (Fig. 20.5).

This problem has an analytical solution known as the *least squares* [Haykin, 1994]. The optimal weights are obtained as the product of the inverse of the input autocorrelation function (R^{-1}) and the cross-correlation vector (**P**) between the input and the desired response. The analytical solution is equivalent to a search for the minimum of the quadratic performance surface $J(w_i)$ using gradient descent, where the weights at each iteration k are adjusted by

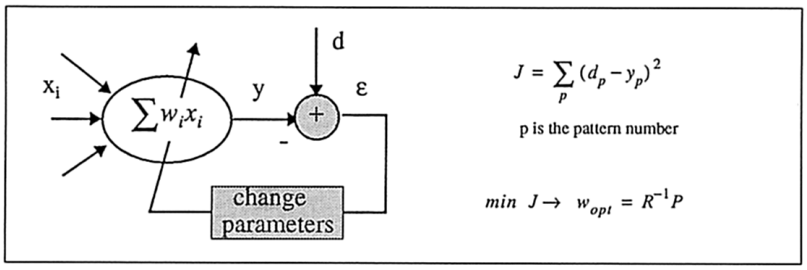


FIGURE 20.5 Computing analytically optimal weights for the linear PE.

$$w_i(k+1) = w_i(k) - \eta \nabla J_i(k) \qquad \nabla J_i = \frac{\partial J}{\partial w_i} \tag{20.2}$$

where η is a small constant called the *step size*, and $\nabla J(k)$ is the gradient of the performance surface at iteration k. Bernard Widrow in the late 1960s proposed a very efficient estimate to compute the gradient at each iteration

$$\nabla J_i(k) = \frac{\partial}{\partial w_i} J(k) \qquad \sim \frac{1}{2} \frac{\partial}{\partial w_i} \left(\varepsilon^2(k)\right) = -\varepsilon(k)\, x_i(k) \tag{20.3}$$

which when substituted into Eq. (20.2) produces the so-called *LMS algorithm*. He showed that the LMS converged to the analytic solution provided the step size η is small enough. Since it is a steepest descent procedure, the largest step size is limited by the inverse of the largest eigenvalue of the input autocorrelation matrix. The larger the step size (below this limit), the faster is the convergence, but the final values will "rattle" around the optimal value in a basin that has a radius proportional to the step size. Hence, there is a fundamental trade-off between speed of convergence and accuracy in the final weight values. One great appeal of the LMS algorithm is that it is very efficient (just one multiplication per weight) and requires only local quantities to be computed.

The LMS algorithm can be framed as a computation of partial derivatives of the cost with respect to the unknowns, i.e., the weight values. In fact, with the chainrule one writes

$$\frac{\partial J}{\partial w_i} = \frac{\partial J}{\partial y} \frac{\partial y}{\partial w_i} = \frac{\partial}{\partial y} \left(\sum (d - y)^2\right) \frac{\partial}{\partial w_i} \left(\sum w_i x_i\right) = -\varepsilon x_i \tag{20.4}$$

we obtain the LMS algorithm for the linear PE. What happens if the PE is nonlinear? If the nonlinearity is differentiable (smooth), we still can apply the same method, because of the chain rule, which prescribes that (Fig. 20.6)

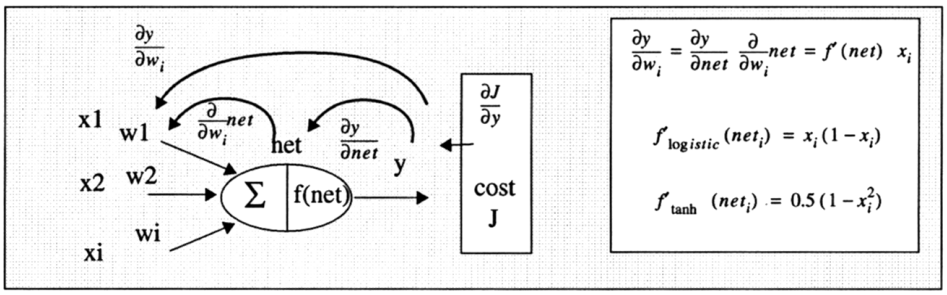


FIGURE 20.6 How to extend LMS to nonlinear PEs with the chain rule.

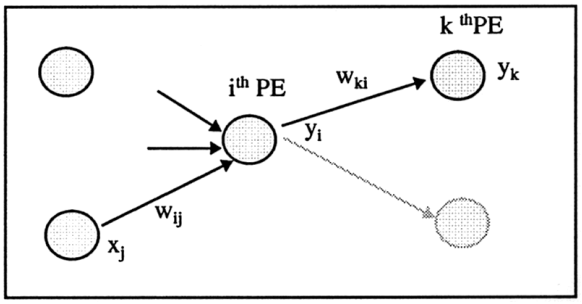


FIGURE 20.7 How to adapt the weights connected to *i*th PE.

$$\frac{\partial J}{\partial w_i} = \frac{\partial J}{\partial y}\frac{\partial y}{\partial \text{net}}\frac{\partial}{\partial w_i} net = -(d - y) f'(\text{net})\, x_i = -\varepsilon f'(\text{net})\, x_i \tag{20.5}$$

where $f'(\text{net})$ is the derivative of the nonlinearity computed at the operating point. Equation (20.5) is known as the *delta rule*, and it will train the perceptron [Haykin, 1994]. Note that throughout the derivation we skipped the pattern index p for simplicity, but this rule is applied for each input pattern. However, the delta rule cannot train MLPs since it requires the knowledge of the error signal at each PE.

The principle of the ordered derivatives can be extended to multilayer networks, provided we organize the computations in flows of activation and error propagation. The principle is very easy to understand, but a little complex to formulate in equation form [Haykin, 1994].

Suppose that we want to adapt the weights connected to a hidden layer PE, the *i*th PE (Fig. 20.7). One can decompose the computation of the partial derivative of the cost with respect to the weight w_{ij} as

$$\frac{\partial J}{\partial w_{ij}} = \underbrace{\frac{\partial J}{\partial y_i}}_{1}\underbrace{\frac{\partial y_i}{\partial \text{net}_i}\frac{\partial}{\partial w_{ij}}\text{net}_i}_{2} \tag{20.6}$$

i.e., the partial derivative with respect to the weight is the product of the partial derivative with respect to the PE state — part 1 in Eq. (20.6) — times the partial derivative of the local activation to the weights — part 2 in Eq. (20.6). This last quantity is exactly the same as for the nonlinear PE ($f'(\text{net}_i)x_j$), so the big issue is the computation of $\frac{\partial J}{\partial y}$. For an output PE, $\frac{\partial J}{\partial y}$ becomes the injected error ε in Eq. (20.4). For the hidden *i*th PE $\frac{\partial J}{\partial y}$ is evaluated by summing all the errors that reach the PE from the top layer through the topology when the injected errors ε_k are clamped at the top layer, or in an equation

$$\frac{\partial J}{\partial y_i} = \left(\sum_k \frac{\partial J}{\partial y_k}\frac{\partial y_k}{\partial \text{net}_k}\frac{\partial}{\partial y_i}\text{net}_k\right) = \sum_k \varepsilon_k f'(\text{net}_k) w_{ki} \tag{20.7}$$

Substituting back in Eq. (20.6) we finally get

$$\frac{\partial J}{\partial w_{ij}} = \underbrace{-x_j f'(\text{net}_i)}_{1}\underbrace{\left(\sum_k \varepsilon_k\, f'(\text{net}_k)\, w_{ki}\right)}_{2} \tag{20.8}$$

This equation embodies the *back-propagation training algorithm* [Haykin, 1994; Bishop, 1995]. It can be rewritten as the product of a local activation (part 1) and a local error (part 2), exactly as the LMS and the delta rules. But now the local error is a composition of errors that flow through the topology, which becomes equivalent to the existence of a desired response at the PE.

There is an intrinsic flow in the implementation of the back-propagation algorithm: first, inputs are applied to the net and activations computed everywhere to yield the output activation. Second, the external errors are computed by subtracting the net output from the desired response. Third, these external errors are utilized in Eq. (20.8) to compute the local errors for the layer immediately preceding the output layer, and the computations chained up to the input layer. Once all the local errors are available, Eq. (20.2) can be used to update every weight. These three steps are then repeated for other training patterns until the error is acceptable.

Step three is equivalent to injecting the external errors in the *dual topology* and back-propagating them up to the input layer [Haykin, 1994]. The dual topology is obtained from the original one by reversing data flow and substituting summing junctions by splitting nodes and vice versa. The error at each PE of the dual topology is then multiplied by the activation of the original network to compute the weight updates. So, effectively the dual topology is being used to compute the local errors which makes the procedure highly efficient. This is the reason back-propagation trains a network of N weights with a number of multiplications proportional to N, $(O(N))$, instead of $(O(N^2))$ for previous methods of computing partial derivatives known in control theory. Using the dual topology to implement back-propagation is the best and most general method to program the algorithm in a digital computer.

Applying Back-Propagation in Practice

Now that we know an algorithm to train MLPs, let us see what are the practical issues to apply it. We will address the following aspects: size of training set vs. weights, search procedures, how to stop training, and how to set the topology for maximum generalization.

Size of Training Set

The size of the training set is very important for good performance. Remember that the ANN gets its information from the training set. If the training data do not cover the full range of operating conditions, the system may perform badly when deployed. Under no circumstances should the training set be less than the number of weights in the ANN. A good size of the training data is ten times the number of weights in the network, with the lower limit being set around three times the number of weights (these values should be taken as an indication, subject to experimentation for each case) [Haykin, 1994].

Search Procedures

Searching along the direction of the gradient is fine if the performance surface is quadratic. However, in ANNs rarely is this the case, because of the use of nonlinear PEs and topologies with several layers. So, gradient descent can be caught in local minima, which makes the search very slow in regions of small curvature. One efficient way to speed up the search in regions of small curvature and, at the same time, to stabilize it in narrow valleys is to include a momentum term in the weight adaptation

$$w_{ij}(n+1) = w_{ij}(n) + \eta\delta(n)x_j(n) + \alpha\left(w_{ij}(n) - w_{ij}(n-1)\right) \tag{20.9}$$

The value of momentum α should be set experimentally between 0.5 and 0.9. There are many more modifications to the conventional gradient search, such as adaptive step sizes, annealed noise, conjugate gradients, and second-order methods (using information contained in the Hessian matrix), but the simplicity and power of momentum learning is hard to beat [Haykin, 1994; Bishop, 1995].

How to Stop Training

The stop criterion is a fundamental aspect of training. The simple ideas of capping the number of iterations or of letting the system train until a predetermined error value are not recommended. The reason is that we want the ANN to perform well in the test set data; i.e., we would like the system to perform well in data it

never saw before (good *generalization*) [Bishop, 1995]. The error in the training set tends to decrease with iteration when the ANN has enough degrees of freedom to represent the input/output map. However, the system may be remembering the training patterns (*overfitting*) instead of finding the underlying mapping rule. This is called *overtraining*. To avoid overtraining the performance in a *validation set*, i.e., a set of input data that the system never saw before, must be checked regularly during training (i.e., once every 50 passes over the training set). The training should be stopped when the performance in the validation set starts to increase, despite the fact that the performance in the training set continues to decrease. This method is called *cross validation*. The validation set should be 10% of the training set, and distinct from it.

Size of the Topology

The size of the topology should also be carefully selected. If the number of layers or the size of each layer is too small, the network does not have enough degrees of freedom to classify the data or to approximate the function, and the performance suffers.

On the other hand, if the size of the network is too large, performance may also suffer. This is the phenomenon of *overfitting* that we mentioned above. But one alternative way to control it is to reduce the size of the network. There are basically two procedures to set the size of the network: either one starts small and adds new PEs or one starts with a large network and prunes PEs [Haykin, 1994]. One quick way to prune the network is to impose a penalty term in the performance function — a *regularizing term* — such as limiting the slope of the input/output map [Bishop, 1995]. A regularization term that can be implemented locally is

$$w_{ij}(n+1) = w_{ij}(n)\left(1 - \frac{\lambda}{\left(1 + w_{ij}(n)\right)^2}\right) + \eta\delta_i(n)\,x_j(n) \tag{20.10}$$

where λ is the *weight decay* parameter and δ the local error. Weight decay tends to drive unimportant weights to zero.

A Posteriori Probabilities

We will finish the discussion of the MLP by noting that this topology when trained with the mean square error is able to estimate directly at its outputs *a posteriori* probabilities, i.e., the probability that a given input pattern belongs to a given class [Bishop, 1995]. This property is very useful because the MLP outputs can be interpreted as probabilities and operated as numbers. In order to guarantee this property, one has to make sure that each class is attributed to one output PE, that the topology is sufficiently large to represent the mapping, that the training has converged to the absolute minimum, and that the outputs are normalized between 0 and 1. The first requirements are met by good design, while the last can be easily enforced if the *softmax activation* is used as the output PE [Bishop, 1995],

$$y = \frac{\exp(\text{net})}{\sum_j \exp(\text{net}_j)} \tag{20.11}$$

20.3 Radial Basis Function Networks

The radial basis function (RBF) network constitutes another way of implementing arbitrary input/output mappings. The most significant difference between the MLP and RBF lies in the PE nonlinearity. While the PE in the MLP responds to the full input space, the PE in the RBF is local, normally a Gaussian kernel in the input space. Hence, it only responds to inputs that are close to its center; i.e., it has basically a *local response*.

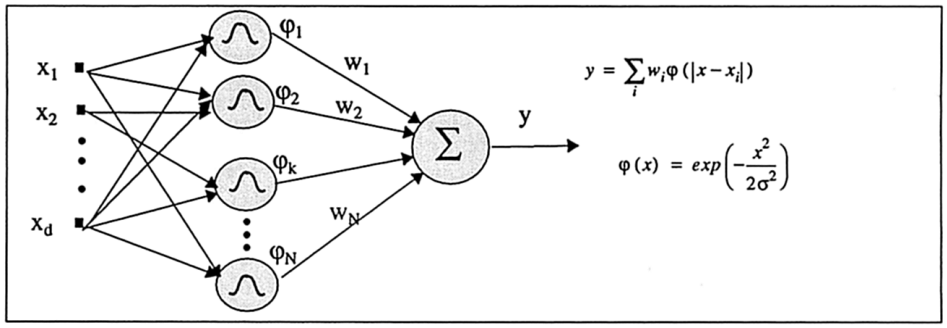


FIGURE 20.8 Radial Basis Function (RBF) network.

The RBF network is also a layered net with the hidden layer built from Gaussian kernels and a linear (or nonlinear) output layer (Fig. 20.8). Training of the RBF network is done normally in two stages [Haykin, 1994]: first, the centers x_i are adaptively placed in the input space using competitive learning or k means clustering [Bishop, 1995], which are unsupervised procedures. Competitive learning is explained later in the chapter. The variances of each Gaussian are chosen as a percentage (30 to 50%) to the distance to the nearest center. The goal is to cover adequately the input data distribution. Once the RBF is located, the second layer weights w_i are trained using the LMS procedure.

RBF networks are easy to work with, they train very fast, and they have shown good properties both for function approximation as classification. The problem is that they require lots of Gaussian kernels in high-dimensional spaces.

20.4 Time-Lagged Networks

The MLP is the most common neural network topology, but it can only handle instantaneous information, since the system has no memory and it is feedforward. In engineering, the processing of signals that exist in time requires systems with memory, i.e., linear filters. Another alternative to implement memory is to use feedback, which gives rise to *recurrent networks*. Fully recurrent networks are difficult to train and to stabilize, so it is preferable to develop topologies based on MLPs but where explicit subsystems to store the past information are included. These subsystems are called *short-term memory structures* [de Vries and Principe, 1992]. The combination of an MLP with short-term memory structures is called a *time-lagged network* (*TLN*). The memory structures can be eventually recurrent, but the feedback is local, so stability is still easy to guarantee. Here, we will cover just one TLN topology, called *focused*, where the memory is at the input layer. The most general TLN have memory added anywhere in the network, but they require other more-involved training strategies (BPTT [Haykin, 1994]). The interested reader is referred to de Vries and Principe [1992] for further details.

The function of a short-term memory in the focused TLN is to represent the past of the input signal, while the nonlinear PEs provide the mapping as in the MLP (Fig. 20.9).

Memory Structures

The simplest memory structure is built from a *tap delay line* (Fig. 20.10). The *memory by delays* is a single-input, multiple-output system that has no free parameters except its size K. The tap delay memory is the memory utilized in the *time-delay neural network* (TDNN) which has been utilized successfully in speech recognition and system identification [Kung, 1993].

A different mechanism for linear memory is the *feedback* (Fig. 20.11). Feedback allows the system to remember past events because of the exponential decay of the response. This memory has limited resolution because of the low pass required for long memories. But notice that unlike the memory by delay, memory by feedback provides the learning system with a free parameter μ that controls the length of the memory. Memory by feedback has been used in Elman and Jordan networks [Haykin, 1994].

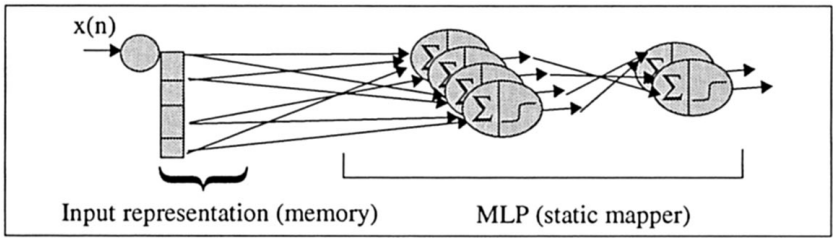


FIGURE 20.9 A focused TLN.

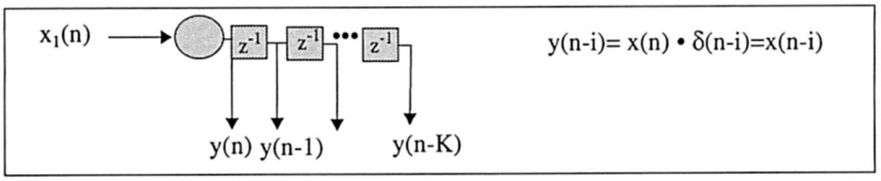


FIGURE 20.10 Tap delay line memory.

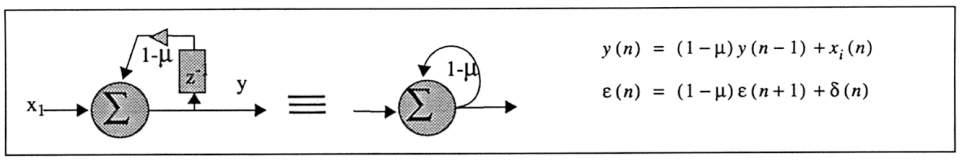


FIGURE 20.11 Memory by feedback (context PE).

It is possible to combine the advantages of memory by feedback with the ones of the memory by delays in linear systems called *dispersive delay lines*. The most studied of these memories is a cascade of low-pass functions called the *gamma memory* [de Vries and Principe, 1992]. The gamma memory has a free parameter μ that controls and decouples memory depth from resolution of the memory. *Memory depth D* is defined as the first moment of the impulse response from the input to the last tap *K*, while *memory resolution R* is the number of taps per unit time. For the gamma memory $D = K/\mu$, and $R = \mu$; i.e., changing μ modifies the memory depth and resolution inversely. This recursive parameter μ can be adapted with the output MSE as the other network parameters; i.e., the ANN is able to choose the best memory depth to minimize the output error, which is unlike the tap delay memory.

Training-Focused TLN Architectures

The appeal of the focused architecture is that the MLP weights can be still adapted with back-propagation. However, the input/output mapping produced by these networks is static. The input memory layer is bringing in past input information to establish the value of the mapping.

As we know in engineering, the size of the memory is fundamental to identify, for instance, an unknown plant or to perform prediction with a small error. But note now that with the focused TLN the models for system identification become nonlinear (i.e., nonlinear moving average — NMA).

When the tap delay implements the short-term memory, straight back-propagation can be utilized since the only adaptive parameters are the MLP weights. When the gamma memory is utilized (or the context PE), the recursive parameter is adapted in a total adaptive framework (or the parameter is preset by some external consideration). The equations to adapt the context PE and the gamma memory are shown in Figs. 20.11 and 20.12, respectively. For the context PE $\delta(n)$ refers to the total error that is back-propagated from the MLP and that reaches the dual context PE.

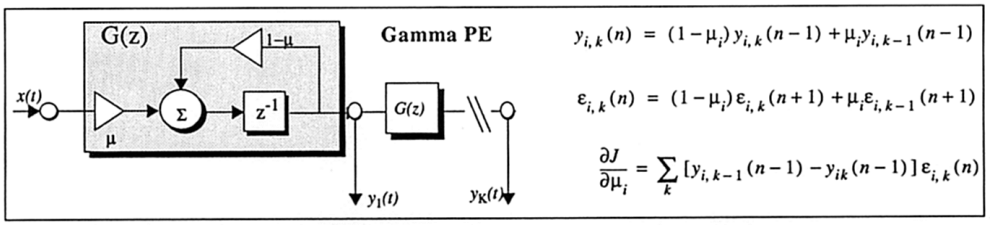


FIGURE 20.12 Gamma memory (dispersive delay line).

20.5 Hebbian Learning and Principal Component Analysis Networks

Hebbian Learning

Hebbian learning is an unsupervised learning rule that *captures similarity* between an input and an output through *correlation*. To adapt a weight w_i using Hebbian learning we adjust the weights according to $\Delta w_i = \eta x_i y$ or in an equation [Haykin, 1994]

$$w_i(n+1) = w_i(n) + \eta\, x_i(n) y(n) \tag{20.12}$$

where η is the step size, x_i is the ith input and y is the PE output.

The output of the single PE is an inner product between the input and the weight vector (formula in Fig. 20.13). It measures the similarity between the two vectors — i.e., if the input is close to the weight vector the output y is large; otherwise it is small. The weights are computed by an outer product of the input X and output Y, i.e., $W = XY^T$, where T means transpose. The problem of Hebbian learning is that it is unstable; i.e., the weights will keep on growing with the number of iterations [Haykin, 1994].

Oja proposed to stabilize the Hebbian rule by normalizing the new weight by its size, which gives the rule [Haykin, 1994]:

$$w_i(n+1) = w_i(n) + \eta\, y(n)\left[x_i(n) - y(n)\, w_i(n)\right] \tag{20.13}$$

The weights now converge to finite values. They still define in the input space the direction where the data cluster has its largest projection, which corresponds to the eigenvector with the largest eigenvalue of the input correlation matrix [Kung, 1993]. The output of the PE provides the largest eigenvalue of the input correlation matrix.

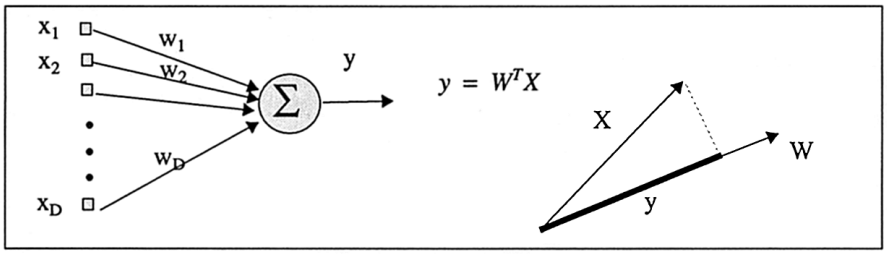


FIGURE 20.13 Hebbian PE.

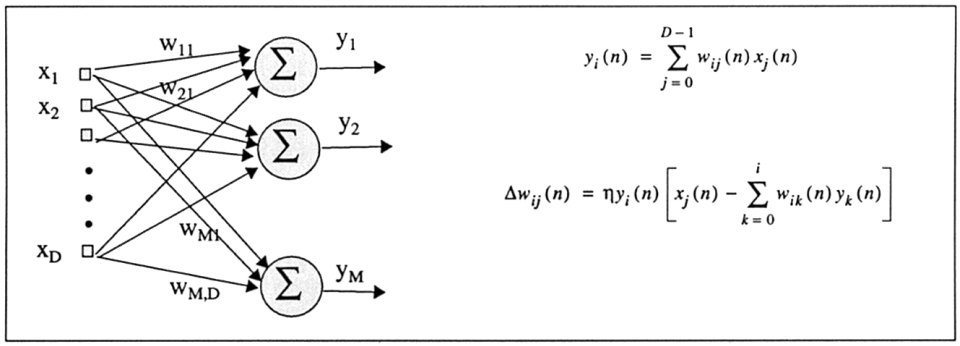


FIGURE 20.14 PCA network.

Principal Component Analysis

Principal component analysis (PCA) is a well-known technique in signal processing that is used to project a signal into a signal-specific basis. The importance of PCA analysis is that it provides *the best linear projection* to a subspace in terms of preserving the signal energy [Haykin, 1994]. Normally, PCA is computed analytically through a singular value decomposition. PCA networks offer an alternative to this computation by providing an iterative implementation that may be preferred for real-time operation in embedded systems.

The PCA network is a one-layer network with linear-processing elements (Fig. 20.14). One can extend Oja's rule for many-output PEs (less or equal to the number of input PEs), according to the formula shown in Fig. 20.14 which is called the Sanger's rule [Haykin, 1994]. The weight matrix rows (that contain the weights connected to the output PEs in descending order) are the eigenvectors of the input correlation matrix. If we set the number of output PEs equal to $M < D$, we will be projecting the input data onto the M largest principal components. Their outputs will be proportional to the M largest eigenvalues. Note that we are performing an eigendecomposition through an iterative procedure.

Associative Memories

Hebbian learning is also the rule to create *associative memories* [Zurada, 1992]. The most-utilized associative memory implements *heteroassociation*, where the system is able to associate an input X to a designated output Y which can be of a different dimension (Fig. 20.15). So, in heteroassociation the signal Y works as the desired response.

We can train such a memory using Hebbian learning or LMS, but the LMS provides a more efficient encoding of information. Associative memories differ from conventional computer memories in several respects. First, they are content addressable, and the information is distributed throughout the network, so they are robust to noise in the input. With nonlinear PEs or recurrent connections (as in the famous Hopfield network) [Haykin, 1994] they display the important property of *pattern completion;* i.e., when the input is distorted or only partially available, the recall can still be perfect.

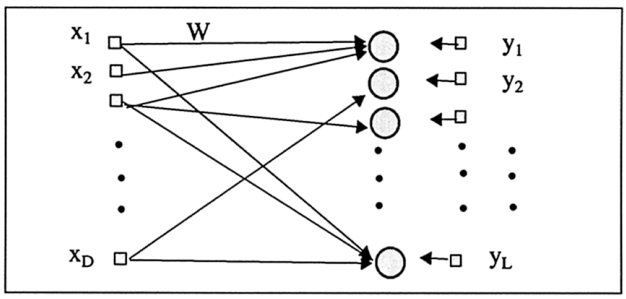


FIGURE 20.15 Associative memory (heteroassociation).

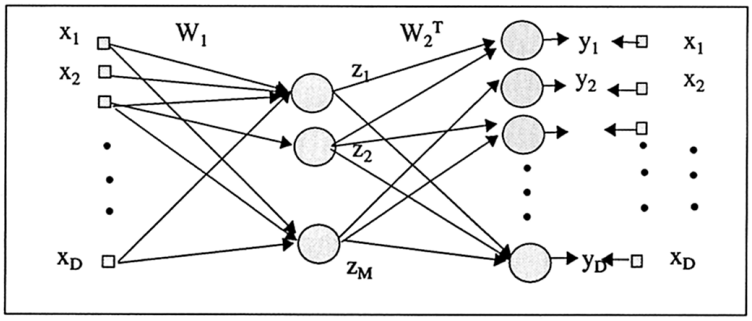


FIGURE 20.16 Autoassociator.

A special case of associative memories is called the *autoassociator* (Fig. 20.16), where the training output of size D is equal to the input signal (also a size D) [Kung, 1993]. Note that the hidden layer has fewer PEs ($M \ll D$) than the input (bottleneck layer). $W_1 = W_2^T$ is enforced. The function of this network is one of *encoding or data reduction*. The training of this network (W_2 matrix) is done with LMS. It can be shown that this network also implements PCA with M components, even when the hidden layer is built from nonlinear PEs.

20.6 Competitive Learning and Kohonen Networks

Competition is a very efficient way to divide the computing resources of a network. Instead of having each output PE more or less sensitive to the full input space, as in the associative memories, in a competitive network each PE specializes into a piece of the input space and represents it [Haykin, 1994]. Competitive networks are linear, single-layer nets (Fig. 20.17). Their functionality is directly related to the competitive learning rule, which belongs to the unsupervised category. First, only the PE that has the largest output gets its weights updated. The weights of the winning PE are updated according to the formula in Fig. 20.17 in such a way that they approach the present input. The step size exactly controls how much is this adjustment (see Fig. 20.17).

Notice that there is an intrinsic nonlinearity in the learning rule: only the PE that has the largest output (the winner) has its weights updated. All the other weights remain unchanged. This is the mechanism that allows the competitive net PEs to specialize.

Competitive networks are used for clustering; i.e., an M output PE net will seek M clusters in the input space. The weights of each PE will correspond to the centers of mass of one of the M clusters of input samples. When a given pattern is shown to the trained net, only one of the outputs will be active and can be used to *label* the sample as belonging to one of the clusters. No more information about the input data is preserved.

Competitive learning is one of the fundamental components of the Kohonen self-organizing feature map (SOFM) network, which is also a single-layer network with linear PEs [Haykin, 1994]. Kohonen learning creates annealed competition in the output space, by adapting not only the winner PE weights but also their spatial

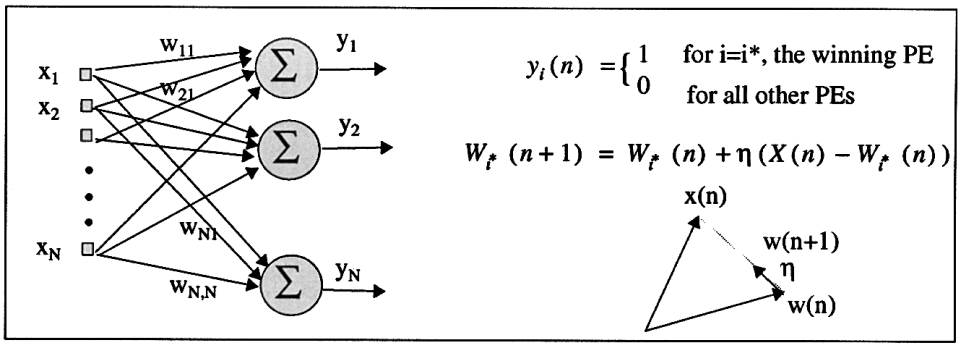


FIGURE 20.17 Competitive neural network.

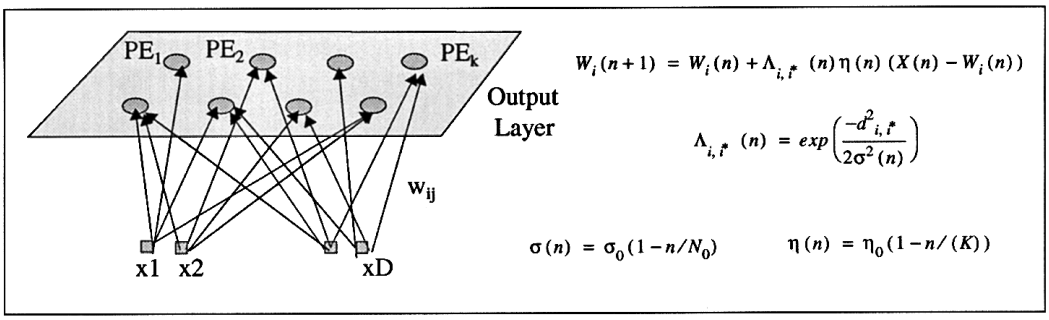


FIGURE 20.18 Kohonen SOFM.

neighbors using a Gaussian neighborhood function Λ. The output PEs are arranged in linear or two-dimensional neighborhoods (Fig. 20.18)

Kohonen SOFM networks produce a mapping between the continuous input space to the discrete output space preserving topological properties of the input space (i.e., local neighbors in the input space are mapped to neighbors in the output space). During training, both the spatial neighborhoods and the learning constant are decreased slowly by starting with a large neighborhood σ_0, and decreasing it (N_0 controls the scheduling). The initial step size η_0 also needs to be scheduled (by K).

The Kohonen SOFM network is useful to project the input to a subspace as an alternative to PCA networks. The topological properties of the output space provide more information about the input than straight clustering.

References

C. M. Bishop, *Neural Networks for Pattern Recognition,* New York: Oxford University Press, 1995.

de Vries and J. C. Principe, "The gamma model — a new neural model for temporal processing," *Neural Networks,* Vol. 5, pp. 565–576, 1992.

S. Haykin, *Neural Networks: A Comprehensive Foundation,* New York: Macmillan, 1994.

S. Y. Kung, *Digital Neural Networks,* Englewood Cliffs, N.J.: Prentice-Hall, 1993.

J. M. Zurada, *Artificial Neural Systems,* West Publishing, 1992.

Further Information

The literature in this field is voluminous. We decided to limit the references to text books for an engineering audience, with different levels of sophistication. Zurada is the most accessible text, Haykin the most comprehensive. Kung provides interesting applications of both PCA networks and nonlinear signal processing and system identification. Bishop concentrates on the design of pattern classifiers.

Interested readers are directed to the following journals for more information: *IEEE Transactions on Signal Processing, IEEE Tranactions on Neural Networks, Neural Networks, Neural Computation,* and *Proceedings of the Neural Information Processing System Conference* (NIPS).

Etter, D.M. “Computing Environments for Digital Signal Processing”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

21

Computing Environments for Digital Signal Processing

21.1 MATLAB Environment
21.2 Example 1: Signal Analysis
21.3 Example 2: Filter Design and Analysis
21.4 Example 3: Multirate Signal Processing

Delores M. Etter
University of Colorado

Computing environments provided by many software tools and packages allow users to design, simulate, and implement digital signal processing (DSP) techniques with speed, accuracy, and confidence. With access to libraries of high-performance algorithms and to advanced visualization capabilities, we can design and analyze systems using the equations and notations that we use to think about signal processing problems; we do not have to translate the equations and techniques into a different notation and syntax. The graphics interface provides an integral part of this design environment, and is accessible from any point within our algorithms. Within this type of computing environment, we are more productive. But, even more important, we develop better solutions because we have so many more tools for analyzing solutions, for experimenting with **"what if" questions**, and for developing extensive simulations to test our solutions. To illustrate the power of these environments, we present a brief description of MATLAB, one of the most popular technical computing environments in both industry and academia, and then present three examples that use MATLAB.

21.1 MATLAB Environment

MATLAB is an integrated technical environment designed to provide accelerated DSP design capabilities. In addition to the basic software package that contains powerful functions for numeric computations, advanced graphics and visualization capabilities, a high-level programming language, and tools for designing **graphical user interfaces (GUI)**, MATLAB also provides a number of application-specific **toolboxes** that contain specialized libraries of functions. The discussion and examples that follow in this article use capabilities from the Signal Processing Toolbox. Other toolboxes that are applicable to solving signal processing problems include the following: Control Systems, Frequency Domain System Identification, Fuzzy Logic, Higher-Order Spectral Analysis, Image Processing, LMI (Linear Matrix Inequality) Control, Model Predictive Control, μ-Analysis and Synthesis, Neural Networks, Optimization, Partial Differential Equations, QFT (Quantitation Feedback Theory) Control, Robust Control, Signal Processing, Splines, Statistics, Symbolic Math, System Identification, and Wavelets.

An interactive environment for modeling, analyzing, and simulating a wide variety of dynamic systems is also provided by MATLAB through SIMULINK—a graphical user interface designed to construct block diagram

models using **"drag-and-drop" operations**. Simulations of the block diagrams can be used to test a number of "what-if" questions. Special purpose block libraries are available for DSP algorithm development, and include a DSP Blockset, a Fixed-Point Blockset, and a Nonlinear Control Design Blockset.

In order to bridge the gap between interactive prototyping and embedded systems, MATLAB has developed a compiler to generate optimized C code from MATLAB code. Automatic C code generation eliminates manual coding and algorithm recoding, thus providing a hierarchical framework for designing, simulating, and prototyping DSP solutions.

21.2 Example 1: Signal Analysis

One of the most common DSP applications is the analysis of signals that have been collected from experiments or from a physical environment. These signals are typically stored in data files, and often need preprocessing steps applied to them before we are able to extract the desired information. Preprocessing can include removing means or linear trends, filtering noise, removing anomalies, and interpolating for missing data. Once the data is ready to analyze, we are usually interested in statistical information (mean, median, variance, autocorrelation, etc.) along with an estimate of the distribution of the values (uniform, Gaussian, etc.). The frequency content of a signal is also important to determine; if the signal is non-stationary, the frequency content needs to be determined using short time windows.

To illustrate the use of MATLAB in computing some of the steps mentioned above, we use a speech signal collected at 8 kHz. After loading the signal from a data file, we will remove any linear trend that might have been introduced in the collection process (this also removes any constant term). Figure 21.1 contains a plot of the signal which clearly shows the time-varying nature of the signal. Figure 21.2 contains a histogram of the distribution of the values, showing that the values are closer to a Laplacian or Gamma distribution than to a uniform or Gaussian distribution. Figure 21.3 contains a spectogram which displays the frequency content of the signal computed using short overlapping time windows. The MATLAB code that generated these plots is shown in Fig. 21.4. This code illustrates some of the important characteristics of high-level computational tools. The fundamental data structure is a matrix, and all operations and functions are designed to work with matrices. Hence, loops are rarely necessary, and thus the code is generally much shorter, more readable, and more self-documenting.

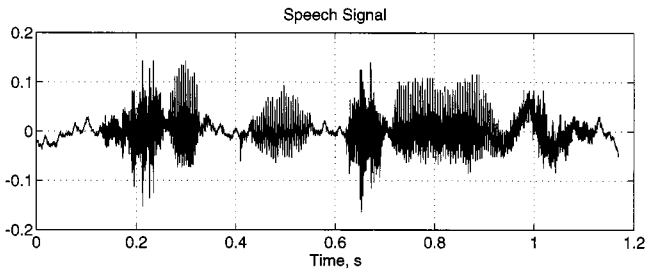


FIGURE 21.1

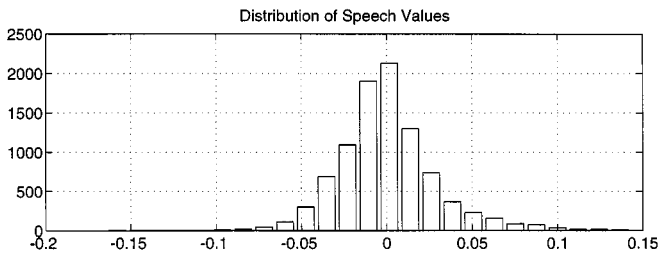


FIGURE 21.2

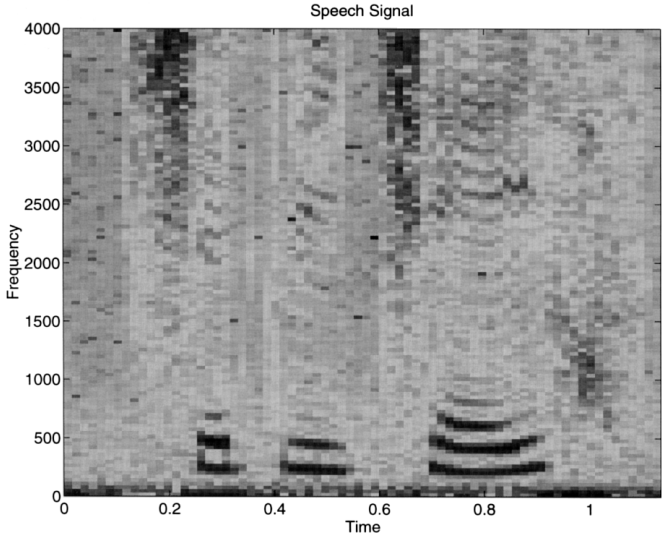


FIGURE 21.3

```
%  These MATLAB statements read and process a speech file.
%
clear,clf               % clear memory and figure window
load speech.dat;        % load speech data file
T = 1/8000;             % specify sampling time
s = detrend(speech);    % remove mean and linear trend from speech
N = length(s);          % determine the number of speech points
t = (0:N-1)*T;          % specify time signal in seconds
%
% plot speech signal
%
subplot(2,1,1),plot(t,s),title('Speech Signal'),
   xlabel('Time, s'),grid,pause
%
% determine and plot histogram of speech signal using 25 bins
%
clf
subplot(2,1,1),hist(s,25),title('Distribution of Speech Values'),grid,pause
%
  plot a spectogram of the speech signal using windows of 256 pts
%
clf
specgram(s,256,8000),title('Speech Signal'),pause
```

FIGURE 21.4

21.3 Example 2: Filter Design and Analysis

MATLAB gives us a number of different options for designing both IIR and FIR digital filters. We can design classical IIR filters (Butterworth, Chebyshev type I, Chebyshev type II, and elliptic) that are lowpass, highpass, bandpass, or bandstop filters. We can also use other techniques, such as the Yule-Walker technique, to design IIR filters with arbitrary passbands. Several techniques allow us to design FIR filters using windowed least squares techniques. The Parks-McClellan algorithm uses the Remez exchange algorithm to design filters with

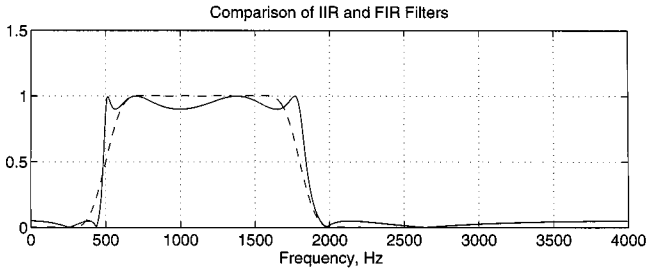


FIGURE 21.5

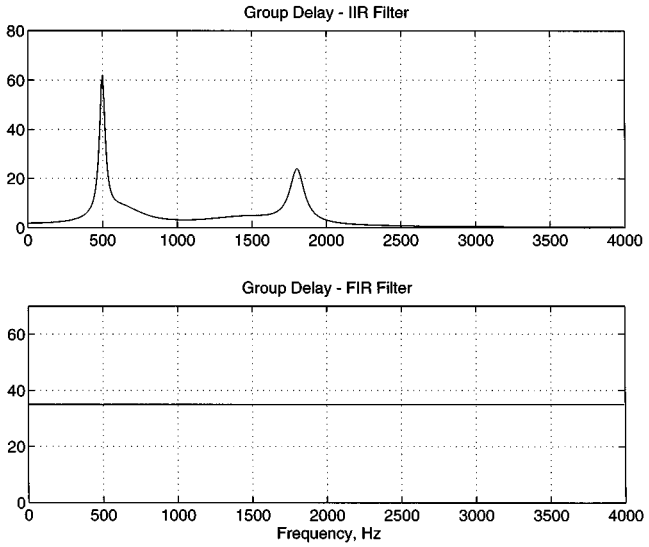


FIGURE 21.6

an optimal fit to an arbitrary desired response. Once a filter is designed, it can be easily translated to other forms, including transfer functions, impulse responses, and poles/zeros.

Assume that we are going to analyze the dial tones from a telephone network that uses dual-tone multifrequency (DTMF) signaling. In this system, pairs of tones are used to signal each character on the telephone keypad. For example, the digit 1 is represented by tones at 697 Hz and 1209 Hz. All of the tones are between 697 Hz and 1633 Hz. Thus, before analyzing the signal to determine the two tones that it contains, we might want to filter out all signals outside of the band that contains all possible tones in order to increase the signal-to-noise ratio. In this example, we design a bandpass filter with a passband between 500 Hz and 1800 Hz. Designs are compared using an elliptic IIR filter of order 8 and a causal FIR filter of order 70. Figure 21.5 contains magnitude plots of the two filters, and clearly shows the characteristics of the filters. The elliptic filter has sharp transitions with ripple in the passband and in the stopband, while the FIR filter (which also uses a Hamming window) is flat in the passband and the stopband, but has wider transition bands. Figure 21.6 contains the group delays for the two filters. The FIR filter has a linear phase response, and thus the group delay is a fixed value of 35 samples; the IIR filter has a nonlinear phase, but has a relatively constant delay in the passband. Figure 21.7 contains the corresponding impulse responses, illustrating the finite impulse response of the FIR filter and the infinite impulse response of the IIR filter. Figure 21.8 contains the pole/zero plot for the IIR solution. The code for performing the designs and generating all the plots is shown in Fig. 21.9.

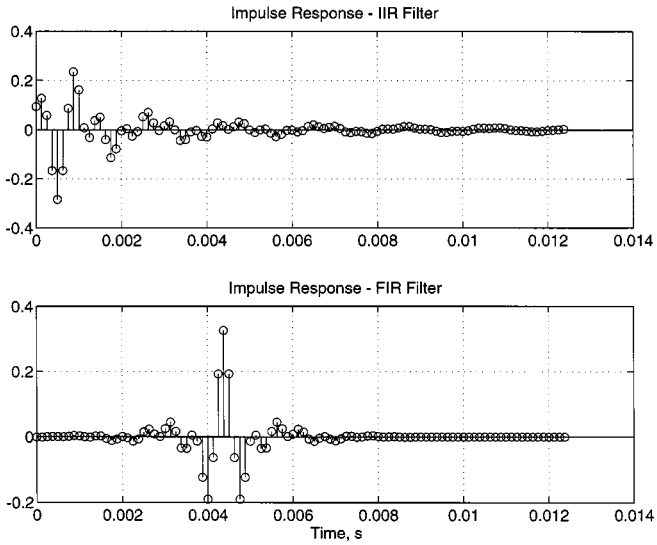


FIGURE 21.7

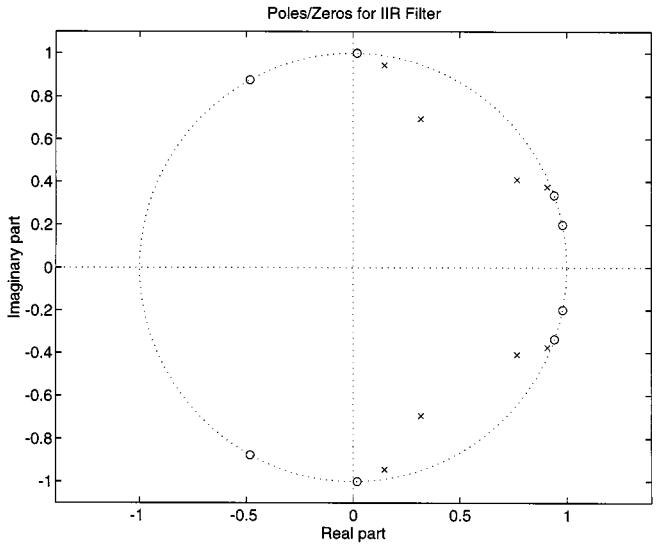


FIGURE 21.8

21.4 Example 3: Multirate Signal Processing

Given a signal that has been collected or computed using a process that eliminates or minimizes aliasing from components above the Nyquist frequency (one-half the sampling frequency), we have a great deal of flexibility in modifying the sampling rate. For example, if the frequency content of the signal is much lower than the Nyquist frequency, then the sampling rate can be reduced without losing any of the signal content. This "decimation" process allows us to compress the signal into a form that requires less memory requirements. An "interpolation" process can be used to interpolate new data points between points of the decimated signal in such a way that the frequency content of the new signal is essentially the same as the original signal. The decimation process requires a reduction of data points by an integer factor, *M,* such as a factor of 3. The

```
%  These MATLAB statements design and analyze IIR and FIR filters.
%
clear,clf                              % clear memory and figure window
Fs = 8000;                             % specify sampling frequency
band = [500/4000 1800/4000];           % specify passband in normalized freq.
Rp = -20*log10(.9);                    % compute passband ripple
Rs = -20*log10(.05);                   % compute stopband ripple
[B1,A1] = ellip(4,Rp,Rs,band);         % design elliptic passband filter
B2 = fir1(70,band);                    % design causal FIR filter
[H1,f] = freqz(B1,A1,512,Fs);          % compute frequency content of filters
[H2,f] = freqz(B2,1,512,Fs);
mag_H1 = abs(H1); mag_H2 = abs(H2); % compute magnitude of filters
[gd1,f] = grpdelay(B1,A1,512,Fs);      % compute group delay of filters
[gd2,f] = grpdelay(B2,1,512,Fs);
%
% plot filter magnitudes
%
subplot(2,1,1),plot(f,abs(H1),f,abs(H2),'--'),xlabel('Frequency, Hz'),
   title('Comparison of IIR and FIR Filters'),grid,pause
%
% plot group delays
%
clf
subplot(2,1,1),plot(f,gd1),title('Group Delay - IIR Filter'),grid,
subplot(2,1,2),plot(f,gd2),title('Group Delay - FIR Filter'),
  xlabel('Frequency, Hz'),grid,pause
%
% compute and plot impulse responses
%
clf
[h1,t] = impz(B1,A1,100,Fs);
[h2,t] = impz(B2,1,100,Fs);
subplot(2,1,1),stem(t,h1),title('Impulse Response - IIR Filter'),grid,
subplot(2,1,2),stem(t,h2),title('Impulse Response - FIR Filter'),grid,
   xlabel('Time, s'),pause
%
% determine and plot poles and zeros of IIR filter
%
clf
zplane(B1,A1),title('Poles/Zeros for IIR Filter'),pause
```

FIGURE 21.9

interpolation process requires that an integral number of points, L-1, be interpolated between existing points, such as interpolation of 5 new points between existing pairs of points. The decimation process increases a sampling interval by M, and the interpolation process decreases a sampling interval by a factor of L. MATLAB contains functions for decimation and interpolation, as well as a function for a resampling of a signal using a non-integer factor of P/Q where P and Q are integers.

Consider a signal that is one sinusoid modulated by another sinusoid. The signal has been sampled at a frequency chosen to provide efficient storage of the data. However, when plotting the data for further analysis, we want to interpolate by a factor of 8 so that the signal looks smoother. Therefore, we use the MATLAB interpolation function. Figure 21.10 contains plots of the original and interpolated time signals. Figure 21.11 contains frequency plots to confirm that the interpolation did not significantly affect the frequency content. Figure 21.12 contains the MATLAB code for this process.

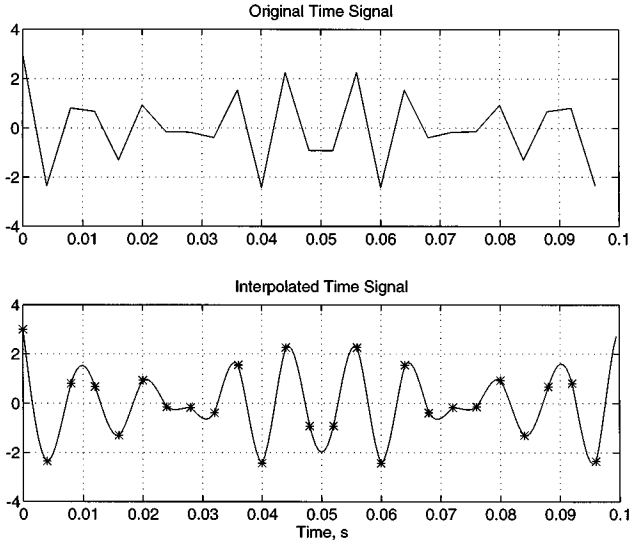


FIGURE 21.10

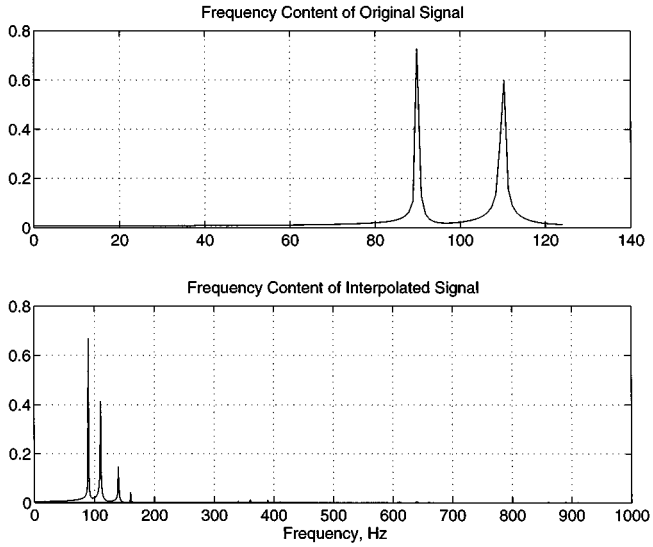


FIGURE 21.11

Defining Terms

Drag and drop operation: Graphical operation for building diagrams by selecting, copying, and moving icons using a mouse or track ball.

Graphical user interface (GUI): Interface using pull-down menus, push buttons, sliders, and other point-and-click icons.

Toolbox: Library of specialized functions.

"What if" question: Question that allows a user to determine the effect of parameter changes in a problem solution.

```
%  These MATLAB statements interpolate a signal by a factor of 8.
%
clear,clf                              % clear memory and figure window
N1 = 256; N2 = 2048;                   % specify numbers of points
T1 = 0.004; T2 = 0.0005;               % specify time intervals
k1 = 0:255; k2 = 0:2047;               % specify time index
t1 = k1*T1; t2 = k2*T2;                % specify time signals
x1 = 3*cos(20*pi*t1).*cos(200*pi*t1);  % generate original signal
x2 = interp(x1,8);                     % interpolate by a factor of 8
%
% plot original and interpolated signals
%
subplot(2,1,1),plot(t1(1:25),x1(1:25)),grid,
   title('Original Time Signal'),
subplot(2,1,2),plot(t2(1:200),x2(1:200),t1(1:25),x1(1:25),'*'),grid,
   title('Interpolated Time Signal'),xlabel('Time, s'),pause
%
% compute and plot frequency content
%
X1 = fft(x1); X2 = fft(x2);
f1 = k1/(N1*T1); f2 = k2/(N2*T2);
subplot(2,1,1),plot(f1(1:128),abs(X1(1:128))/N1),grid,
   title('Frequency Content of Original Signal'),
subplot(2,1,2),plot(f2(1:1024),abs(X2(1:1024))/N2),grid,
   title('Frequency Content of Interpolated Signal'),
   xlabel('Frequency, Hz'),pause
```

FIGURE 21.12

Related Topics

14.3 Design and Implementation of Digital Filters • 14.4 Signal Restoration • 15.1 Coding, Transmission, and Storage

References

Buck, Daniel, and Singer, *Computer Explorations in Signals and Systems Using MATLAB,* Englewood Cliffs, N.J.: Prentice-Hall, 1997.

Burris, McClellan, and Oppenheim, *Computer-Based Exercises for Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1994.

Etter, *Engineering Problem Solving with MATLAB,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1997.

Etter, *Introduction to MATLAB for Engineers and Scientists,* Englewood Cliffs, N.J.: Prentice-Hall, 1996.

Garcia, *Numerical Methods for Physics,* Englewood Cliffs, N.J.: Prentice-Hall, 1994.

Hanselman and Kuo, *MATLAB Tools for Control System Analysis and Design,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1995.

Jang, Sun, and Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence,* Englewood Cliffs, N.J.: Prentice-Hall, 1997.

Kamen and Heck, *Fundamentals of Signals and Systems Using MATLAB,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1997.

Marcus, *Matrices and MATLAB: A Tutorial,* Englewood Cliffs, N.J.: Prentice-Hall, 1993.

Polking, *Ordinary Differential Equations Using MATLAB,* Englewood Cliffs, N.J.: Prentice-Hall, 1995.

Van Loan, *Introduction to Scientific Computing: A Matrix Vector Approach Using MATLAB,* Englewood Cliffs, N.J.: Prentice-Hall, 1997.

Further Information

For further information on MATLAB, here are e-mail addresses, WWW sites, and other resources locations:

E-mail addresses:

news-notes@mathworks.com (MATLAB *News & Notes* editor)
support@mathworks.com (technical support for all products)
info@mathworks.com (general information)

Web sites:

http://www.mathworks.com (the MathWorks home page)
http://education.mathworks.com (educational products and services)

Other resources:

ftp.mathworks.com (FTP server)
comp.soft-sys.matlab (usenet newsgroup)

Steadman, J.W. "Section III – Electronics"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

The Cheetah disc drive is produced by Seagate Technology, Scotts Valley, California, and has been dubbed the industry's fastest disc drive. The Cheetah is the world's first-announced drive to utilize 10,000-rpm technology. The increased rotational remarkably increases data transfer rates to 15 Mbytes/sec which is 40% greater than that of 7,200-rpm drives. The 10,000-rpm rotational rate also significantly reduces the seek time.

Seagate's pioneering of the 10,000-rpm technology enables OEMs, VARs, and system integrators to take advantage of performance levels that were previously unattainable. Seagate has developed and manufactured some of the industry's highest-performance disc drives which not only enable users to achieve higher levels of system performance, but will also introduce exciting new electronic applications. (Photo courtesy of Seagate Technology.)

III

Electronics

22 Semiconductors *G.S. Gildenblat, B. Gelmont, M. Milkovic, A. Elshabini-Riad, F.W. Stephenson, I.A. Bhutta, D.C. Look*
Physical Properties • Diodes • Electrical Equivalent Circuit Models and Device Simulators for Semiconductor Devices • Electrical Characterization of Semiconductors

23 Semiconductor Manufacturing *H.G. Parks, W. Needham, S. Rajaram, C. Rafferty*
Processes • Testing • Electrical Characterization of Interconnections • Process Modeling and Simulation

24 Transistors *S. Soclof, J. Watson, J.R. Brews*
Junction Field-Effect Transistors • Bipolar Transistors • The Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET)

25 Integrated Circuits *J.E. Brewer, M.R. Zargham, S. Tragoudas, S. Tewksbury*
Integrated Circuit Technology • Layout, Placement, and Routing • Application-Specific Integrated Circuits

26 Surface Mount Technology *G.R. Blackwell*
Definition and Considerations • SMT Design, Assembly, and Test Overview • Surface Mount Device (SMD) Definitions • Substrate Design Guidelines • Thermal Design Considerations • Adhesives • Solder Paste and Joint Formation • Parts Inspection and Placement • Reflow Soldering • Cleaning • Prototype Systems

27 Operational Amplifiers *E.J. Kennedy, J.V. Wait*
Ideal and Practical Models • Applications

28 Amplifiers *G.L. Carpenter, J. Choma, Jr.*
Large Signal Analysis • Small Signal Analysis

29 Active Filters *R.E. Massara, J.W. Steadman, B.M. Wilamowski, J.A. Svoboda*
Synthesis of Low-Pass Forms • Realization • Generalized Impedance Converters and Simulated Impedances

30 Power Electronics *K. Rajashekara, A.K.S. Bhat, B.K. Bose*
Power Semiconductor Devices • Power Conversion • Power Supplies • Converter Control of Machines

31 Optoelectronics *J. Hecht, L.S. Watkins, R.A. Becker*
Lasers • Sources and Detectors • Circuits

32 D/A and A/D Converters *S.A.R. Garrod*
D/A and A/D Circuits

33 Thermal Management of Electronics *A. Bar-Cohen*
Heat Transfer Fundamentals • Chip Module Thermal Resistance

34 Digital and Analog Electronic Design Automation *A. Dewey*
Design Entry • Synthesis • Verification • Physical Design • Test

John W. Steadman
University of Wyoming

THE TRULY INCREDIBLE CHANGES in the technology associated with electronics over the past three decades have certainly been the driving force for most of the growth in the field of electrical engineering. Recall that 30 years ago the transistor was a novel device and that the majority of electronic systems still used vacuum tubes. Then look at the section headings in the following chapters and appreciate the range of ways that electronics has impacted electrical engineering. Amplifiers, integrated circuits, filters, power electronics, and optoelectronics are examples of how electronics transformed the practice of electrical engineering in such diverse fields as power generation and distribution, communications, signal processing, and computers.

The various contributors to this section have done an outstanding job of providing concise and practical coverage of this immense field. By necessity, the content ranges from rather theoretical considerations, such as physical principles of semiconductors, to quite practical issues such as printed circuit board technology and circuits for active filter realizations. There are areas of overlap with other chapters in the *Handbook*, such as those covering electrical effects and devices, biomedical electronics, digital devices, and computers. The contributors to this section, however, have maintained a focus on providing practical and useful information directly related to electronics as needed by a practicing electrical engineer.

The author(s) of each chapter was given the task of providing broad coverage of the field while being restricted to only a few pages of text. As a result, the information content is quite high and tends to treat the main principles or most useful topics in each area without giving the details or extensions of the subject. This practice, followed throughout the *Handbook*, is what makes it a valuable new work in electrical engineering. In most cases the information here will be complete enough. When this is not the case, the references will point the way to whatever added information is necessary.

Nomenclature

Symbol	Quantity	Unit
A	area	m^2
A_i	current gain	
A_v	terminal voltage gain	
α_i	ionization coefficient	
B	bandwidth	Hz
C	velocity of light in vacuum	2.998×10^8 m/s
C	specific heat	W/kg K
C_c	coupling capacitor	
C_E	emitter bypass capacitor	
C_j	junction capacitance	F
E	energy	J
ϵ_o	permittivity constant	8.85×10^{-12} F/m
f	focal length	m
F	luminous flux	lumen
F	radiational factor	
ϕ	pn-junction contact potential	V
g_m	transconductance	S
h	Planck's constant	6.626×10^{-34} J·s
h	heat transfer coefficient	
h_{FE}	common-emitter direct current gain	
h_{re}	small-signal current gain	
η	quantum efficiency	
i_b	incremental base current	A
I	illuminance	lumen/cm
I_B	direct base current	A
I_D	diode forward current	A
I_E	direct emitter current	A
I_s	reverse saturation current	A
J	current density	A/m^2
k	Boltzmann constant	1.38×10^{-23} J/K
k	wavenumber	rad/m
k	wave vector	
k	attenuation	
k	thermal conductivity	W/m K
λ	carrier mean free path	m
λ	wavelength	m
μ	magnetic permeability	H/m
μ	viscosity	kg/ms
μ_n	electron mobility	
n	electron density	electrons/cm^3
n	refractive index	
ν	light frequency	Hz

Symbol	Quantity	Unit
p	hole density	holes/cm^3
Pr	Prandtl number	
ψ_{bk}	Bloch wave function	
q	electronic charge	1.6×10^{-19} C
q	heat flow	W
R_B	base resistor	
Re	Reynolds number	
R_g	generator internal resistance	Ω
R_G	total resistance	Ω
σ	conductivity	S
σ	Stefan-Boltzmann constant	5.67×10^{-8} W/m^2 K^4
T	absolute temperature	K
τ	momentum relaxation time	s
θ	volumetric flow rate	m^3/s
v	electron velocity	m/s
V_{BE}	direct base-emitter voltage	V
V_{CC}	direct voltage supply	V
V_T	thermal voltage	mV
V_Z	Zener voltage	V
W	power	W
Z_o	characteristic impedance	Ω

Gildenblat, G.S., Gelmont, B., Milkovic, M., Elshabini-Riad, A., Stephenson, F.W., Bhutta, I.A., Look, D.C. “Semiconductors”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

22

Semiconductors

Gennady Sh. Gildenblat
The Pennsylvania State University

Boris Gelmont
University of Virginia

Miram Milkovic
Analog Technology Consultants

Aicha Elshabini-Riad
Virginia Polytechnic Institute and State University

F.W. Stephenson
Virginia Polytechnic Institute and State University

Imran A. Bhutta
RFPP

David C. Look
Wright State University

22.1 Physical Properties
Energy Bands • Electrons and Holes • Transport Properties • Hall Effect • Electrical Breakdown • Optical Properties and Recombination Processes • Nanostructure Engineering • Disordered Semiconductors

22.2 Diodes
pn-Junction Diode • *pn*-Junction with Applied Voltage • Forward-Biased Diode • I_D-V_D Characteristic • DC and Large-Signal Model • High Forward Current Effects • Large-Signal Piecewise Linear Model • Small-Signal Incremental Model • Large-Signal Switching Behavior of a *pn*-Diode • Diode Reverse Breakdown • Zener and Avalanche Diodes • Varactor Diodes • Tunnel Diodes • Photodiodes and Solar Cells • Schottky Barrier Diode

22.3 Electrical Equivalent Circuit Models and Device Simulators for Semiconductor Devices
Overview of Equivalent Circuit Models • Overview of Semiconductor Device Simulators

22.4 Electrical Characterization of Semiconductors
Theory • Determination of Resistivity and Hall Coefficient • Data Analysis • Sources of Error

22.1 Physical Properties

Gennady Sh. Gildenblat and Boris Gelmont

Electronic applications of semiconductors are based on our ability to vary their properties on a very small scale. In conventional semiconductor devices, one can easily alter charge carrier concentrations, fields, and current densities over distances of 0.1–10 μm. Even smaller characteristic lengths of 10–100 nm are feasible in materials with an engineered band structure. This section reviews the essential physics underlying modern semiconductor technology.

Energy Bands

In crystalline semiconductors atoms are arranged in periodic arrays known as crystalline lattices. The lattice structure of silicon is shown in Fig. 22.1. Germanium and diamond have the same structure but with different interatomic distances. As a consequence of this periodic arrangement, the allowed energy levels of electrons are grouped into **energy bands**, as shown in Fig. 22.2. The probability that an electron will occupy an allowed quantum state with energy E is

$$f = [1 + \exp(E - F)/k_B T]^{-1} \tag{22.1}$$

Here $k_B = 1/11{,}606$ eV/K denotes the Boltzmann constant, T is the absolute temperature, and F is a parameter known as the Fermi level. If the energy $E > F + 3k_B T$, then $f(E) < 0.05$ and these states are mostly empty. Similarly, the states with $E < F - 3k_B T$ are mostly occupied by electrons. In a typical metal [Fig. 22.2(a)], the

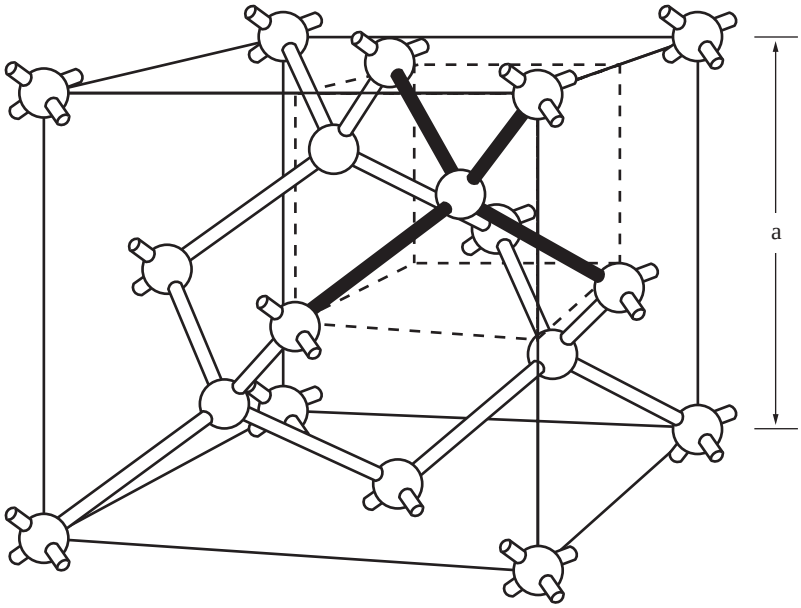


FIGURE 22.1 Crystalline lattice of silicon, $a = 5.43$ Å at 300°C.

energy level $E = F$ is allowed, and only one energy band is partially filled. (In metals like aluminum, the partially filled band in Fig. 22.2(a) may actually represent a combination of several overlapping bands.) The remaining energy bands are either completely filled or totally empty. Obviously, the empty energy bands do not contribute to the charge transfer. It is a fundamental result of solid-state physics that energy bands that are completely filled also do not contribute. What happens is that in the filled bands the average velocity of electrons is equal to zero. In semiconductors (and insulators) the Fermi level falls within a forbidden **energy gap** so that two of the energy bands are partially filled by electrons and may give rise to electron current. The upper partially filled band is called the **conduction band** while the lower is known as the **valence band**. The number of electrons in the conduction band of a semiconductor is relatively small and can be easily changed by adding impurities. In metals, the number of free carriers is large and is not sensitive to doping.

A more detailed description of energy bands in a crystalline semiconductor is based on the Bloch theorem, which states that an electron wave function has the form (Bloch wave)

$$\Psi_{bk} = u_{bk}(r)\ \exp(ikr) \tag{22.2}$$

where **r** is the radius vector of electron, the modulating function $u_{bk}(\mathbf{r})$ has the periodicity of the lattice, and the quantum state is characterized by wave vector **k** and the band number b. Physically, (22.2) means that an electron wave propagates through a periodic lattice without attenuation. For each energy band one can consider the dispersion law $E = E_b(\mathbf{k})$. Since (see Fig. 22.2b) in the conduction band only the states with energies close to the bottom, E_c, are occupied, it suffices to consider the $E(\mathbf{k})$ dependence near E_c. The simplified band diagrams of Si and GaAs are shown in Fig. 22.3.

Electrons and Holes

The concentration of electrons in the valence band can be controlled by introducing impurity atoms. For example, the substitutional doping of Si with As results in a local energy level with an energy about $\Delta W_d \approx 45$ meV below the conduction band edge, E_c [Fig. 22.2(b)]. At room temperature this impurity center is readily ionized, and (in the absence of other impurities) the concentration of electrons is close to the concentration of As atoms. Impurities of this type are known as donors.

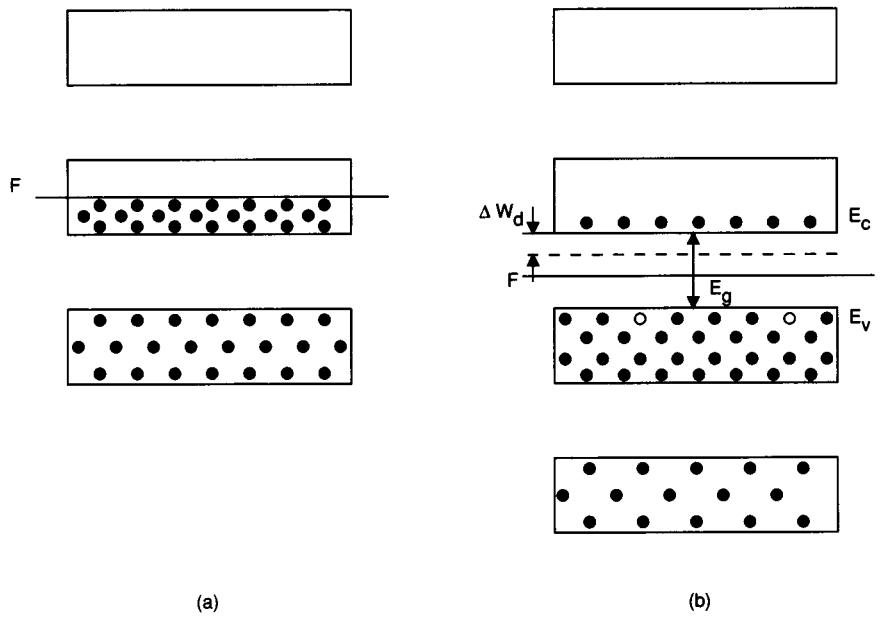


FIGURE 22.2 Band diagrams of metal (a) and semiconductor (b); ●, electron; ○, missing electron (hole).

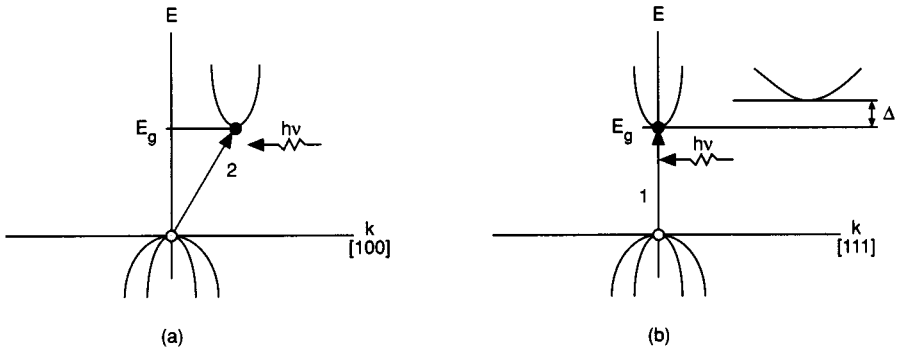


FIGURE 22.3 Simplified $E(\mathrm{k})$ dependence for Si (a) and GaAs (b). At room temperature $E_g(\mathrm{Si}) = 1.12$ eV, $E_g(\mathrm{GaAs}) = 1.43$ eV, and $\Delta = 0.31$ eV; (1) and (2) indicate direct and indirect band-to-band transitions.

While considering the contribution $\mathbf{j}_p$ of the predominantly filled valence band to the current density, it is convenient to concentrate on the few missing electrons. This is achieved as follows: let $v(k)$ be the velocity of electron described by the wave function (20.2). Then

$$\mathbf{j}_p = -q\sum_{\substack{\text{filled}\\ \text{states}}} \mathbf{v}(\mathbf{k}) = -q\left[\sum_{\text{all states}} \mathbf{v}(\mathbf{k}) - \sum_{\substack{\text{empty}\\ \text{states}}} \mathbf{v}(\mathbf{k})\right] = q\sum_{\substack{\text{empty}\\ \text{states}}} \mathbf{v}(\mathbf{k}) \tag{22.3}$$

Here we have noted again that a completely filled band does not contribute to the current density. The picture emerging from (22.3) is that of particles (known as **holes**) with the charge $+q$ and velocities corresponding to those of missing electrons. The concentration of holes in the valence band is controlled by adding acceptor-type impurities (such as boron in silicon), which form local energy levels close to the top of the valence band. At room temperature these energy levels are occupied by electrons that come from the valence band and leave

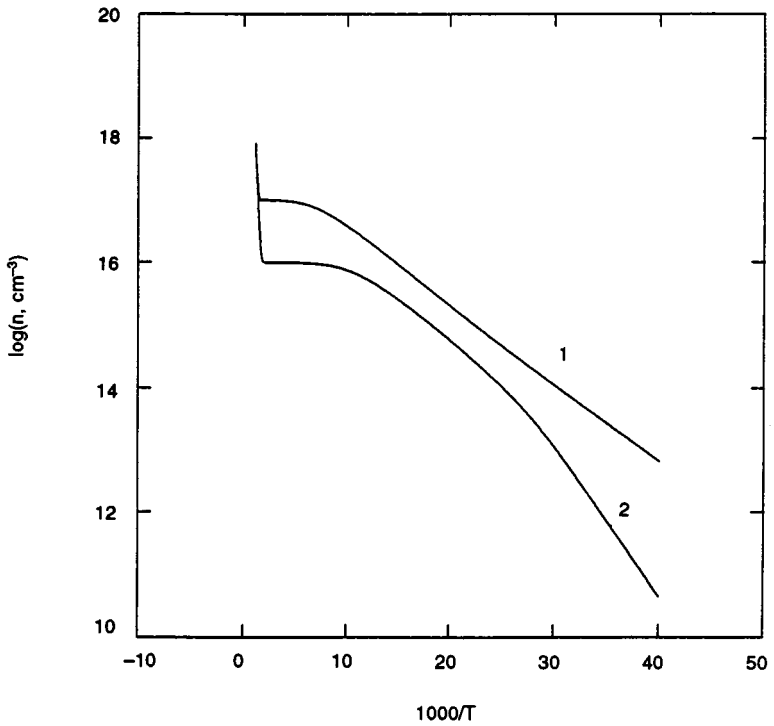


FIGURE 22.4 The inverse temperature dependence of electron concentration in Si; 1: $N_d = 10^{17}$ cm^{-3}, $N_a = 0$; 2: $N_d = 10^{16}$ cm^{-3}, $N_a = 10^{14}$ cm^{-3}.

the holes behind. Assuming that the Fermi level is removed from both E_c and E_v by at least $3k_BT$ (a nondegenerate semiconductor), the concentrations of electrons and holes are given by

$$n = N_c \exp[(F - E_c)/k_BT] \tag{22.4}$$

and

$$p = N_v \exp[(E_v - F)/k_BT] \tag{22.5}$$

where $N_c = 2\ (2m_n^*\pi k_BT)^{3/2}/h^3$ and $N_v = 2(2m_p^*\pi k_BT)^{3/2}/h^3$ are the effective densities of states in the conduction and valence bands, respectively, h is Plank constant, and the effective masses m_n^* and m_p^* depend on the details of the band structure [Pierret, 1987].

In a nondegenerate semiconductor, $np = N_cN_v \exp(-E_g/k_BT) \triangleq n_i^2$ is independent of the doping level. The neutrality condition can be used to show that in an n-type ($n > p$) semiconductor at or below room temperature

$$n(n + N_a)(N_d - N_a - n)^{-1} = (N_c/2)\ \exp(-\Delta W_d/k_BT) \tag{22.6}$$

where N_d and N_a denote the concentrations of **donors** and **acceptors**, respectively.

Corresponding temperature dependence is shown for silicon in Fig. 22.4. Around room temperature $n = N_d - N_a$, while at low temperatures n is an exponential function of temperature with the activation energy $\Delta W_d/2$ for $n > N_a$ and ΔW_d for $n < N_a$. The reduction of n compared with the net impurity concentration $N_d - N_a$ is known as a freeze-out effect. This effect does not take place in the heavily doped semiconductors.

For temperatures $T > T_i = (E_g/2k_B)/\ln[\sqrt{N_cN_v}/(N_d - N_a)]$ the electron concentration $n \approx n_i >> N_d - N_a$ is no longer dependent on the doping level (Fig. 22.4). In this so-called intrinsic regime electrons come directly from the valence band. A loss of technological control over n and p makes this regime unattractive for electronic

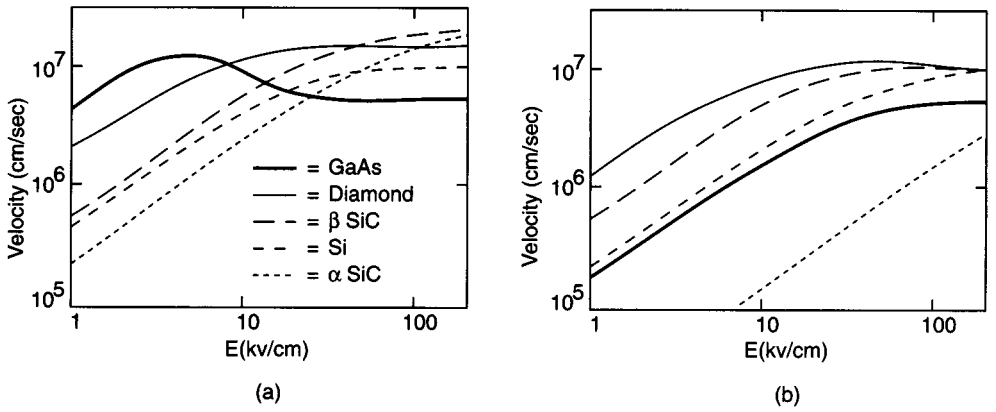


FIGURE 22.5 Electron (a) and hole (b) drift velocity versus electric field dependence for several semiconductors at $N_d = 10^{17}$ cm^{-3}. (*Source:* R.J. Trew, J.-B. Yan, and L.M. Mack, *Proc. IEEE,* vol. 79, no. 5, p. 602, May 1991. © 1991 IEEE.)

applications. Since $T_i \propto E_g$ the transition to the intrinsic region can be delayed by using widegap semiconductors. Both silicon carbide (several types of SiC with different lattice structures are available with E_g = 2.2–2.86 eV) and diamond (E_g = 5.5 eV) have been used to fabricate diodes and transistors operating in the 300–700°C temperature range.

Transport Properties

In a semiconductor the motion of an electron is affected by frequent collisions with **phonons** (quanta of lattice vibrations), impurities, and crystal imperfections. In weak uniform electric fields, $\mathscr{E}$, the carrier drift velocity, $\mathbf{v}_d$, is determined by the balance of the electric and collision forces:

$$m_n^* \mathbf{v}_d / \tau = -q\mathscr{E} \tag{22.7}$$

where τ is the momentum relaxation time. Consequently $\mathbf{v}_d = -\mu_n \mathscr{E}$, where $\mu_n = q\tau/m_n^*$ is the electron mobility. For an *n*-type semiconductor with uniform electron density, *n*, the current density $\mathbf{j}_n = -qn\mathbf{v}_d$ and we obtain Ohm's law $\mathbf{j}_n = \sigma\mathscr{E}$ with the conductivity $\sigma = qn\mu_n$. The momentum relaxation time can be approximately expressed as

$$1/\tau = 1/\tau_{ii} + 1/\tau_{ni} + 1/\tau_{ac} + 1/\tau_{npo} + 1/\tau_{po} + 1/\tau_{pe} + \ldots \tag{22.8}$$

where τ_{ii}, τ_{ni}, τ_{ac}, τ_{npo}, τ_{po}, τ_{pe} are the relaxation times due to ionized impurity, neutral impurity, acoustic phonon, nonpolar optical, polar optical, and piezoelectric scattering, respectively.

In the presence of concentration gradients, electron current density is given by the drift-diffusion equation

$$\mathbf{j}_n = qn\mu_n\mathscr{E} + qD_n\nabla n \tag{22.9}$$

where the diffusion coefficient D_n is related to mobility by the Einstein relation $D_n = (k_B T/q)\mu_n$.

A similar equation can be written for holes and the total current density is $\mathbf{j} = \mathbf{j}_n + \mathbf{j}_p$. The right-hand side of (22.9) may contain additional terms corresponding to temperature gradient and compositional nonuniformity of the material [Wolfe et al., 1989].

In sufficiently strong electric fields the drift velocity is no longer proportional to the electric field. Typical velocity–field dependencies for several semiconductors are shown in Fig. 22.5. In GaAs $v_d(\mathscr{E})$ dependence is not monotonic, which results in negative differential conductivity. Physically, this effect is related to the transfer of electrons from the conduction band to a secondary valley (see Fig. 22.3).

The limiting value v_s of the drift velocity in a strong electric field is known as the saturation velocity and is usually within the 10^7–$3{\cdot}10^7$ cm/s range. As semiconductor device dimensions are scaled down to the submicrometer range, v_s becomes an important parameter that determines the upper limits of device performance.

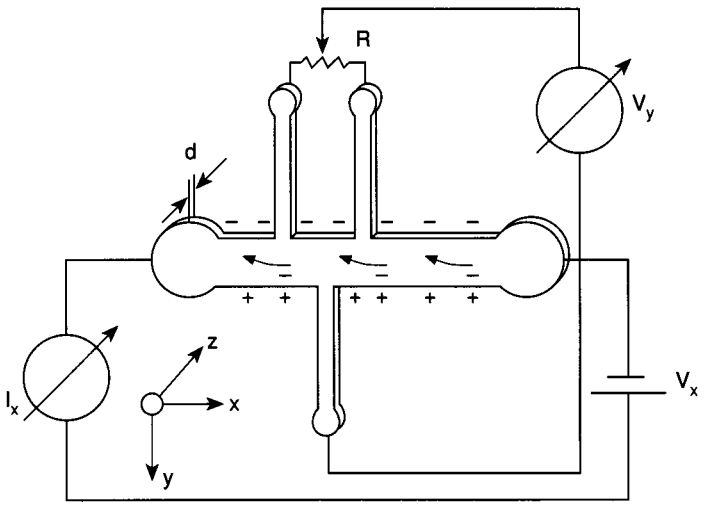


FIGURE 22.6 Experimental setup for Hall effect measurements in a long two-dimensional sample. The Hall angle is determined by a setting of the rheostat that renders $j_y = 0$. Magnetic field $B = B_z$. (*Source:* K.W. Böer, *Surveys of Semiconductor Physics,* New York: Chapman & Hall, 1990, p. 760. With permission.)

The curves shown in Fig. 22.5 were obtained for uniform semiconductors under steady-state conditions. Strictly speaking, this is not the case with actual semiconductor devices, where velocity can "overshoot" the value shown in Fig. 22.5. This effect is important for Si devices shorter than 0.1μm (0.25 μm for GaAs devices) [Shur, 1990; Ferry, 1991]. In such extreme cases the drift-diffusion equation (22.9) is no longer adequate, and the analysis is based on the Boltzmann transport equation

$$\frac{\partial f}{\partial t} + \mathbf{v}\nabla f + q\mathscr{E}\ \nabla_{\mathrm{p}} f = \left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}} \tag{22.10}$$

Here f denotes the distribution function (number of electrons per unit volume of the phase space, i.e., $f = dn/d^3rd^3p$), $\mathbf{v}$ is electron velocity, $\mathbf{p}$ is momentum, and $(\partial f/\partial t)_{\mathrm{coll}}$ is the "collision integral" describing the change of f caused by collision processes described earlier. For the purpose of semiconductor modeling, Eq. (22.10) can be solved directly using various numerical techniques, including the method of moments (hydrodynamic modeling) or Monte Carlo approach. The drift-diffusion equation (22.9) follows from (22.10) as a special case. For even shorter devices quantum effects become important and device modeling may involve quantum transport theory [Ferry, 1991].

Hall Effect

In a uniform magnetic field electrons move along circular orbits in a plane normal to the magnetic field $\mathbf{B}$ with the angular (cyclotron) frequency $\omega_c = qB/m_n^\star$. For a uniform semiconductor the current density satisfies the equation

$$\mathbf{j} = \sigma(\mathscr{E} + R_H[\mathbf{jB}]) \tag{22.11}$$

In the usual weak-field limit $\omega_c\tau << 1$ the Hall coefficient $R_H = -r/nq$ and the Hall factor r depend on the dominating scattering mode. It varies between $3\pi/8 \approx 1.18$ (acoustic phonon scattering) and $315\pi/518 \approx 1.93$ (ionized impurity scattering).

The Hall coefficient can be measured as $R_H = V_y d/I_x B$ using the test structure shown in Fig. 22.6. In this expression V_y is the Hall voltage corresponding to $I_y = 0$ and d denotes the film thickness.

Combining the results of the Hall and conductivity measurements one can extract the carrier concentration type (the signs of V_y are opposite for n-type and p-type semiconductors) and Hall mobility $\mu_H = r\mu$:

$$\mu_H = -R_H\sigma, \quad n = -r/qR_H \tag{22.12}$$

Measurements of this type are routinely used to extract concentration and mobility in doped semiconductors. The weak-field Hall effect is also used for the purpose of magnetic field measurements.

In strong magnetic fields $\omega_c\tau >> 1$ and on the average an electron completes several circular orbits without a collision. Instead of the conventional $E_b(\mathbf{k})$ dependence, the allowed electron energy levels in the magnetic field are given by ($\hbar = h/2\pi$; $s = 0, 1, 2, \ldots$)

$$E_s = \hbar\omega_c\,(s + 1/2) + \hbar^2 k_z^2 / 2m_n^* \tag{22.13}$$

The first term in Eq. (22.13) describes the so-called Landau levels, while the second corresponds to the kinetic energy of motion along the magnetic field $B = B_z$. In a pseudo-two-dimensional system like the channel of a field-effect transistor the second term in Eq. (22.13) does not appear, since the motion of electrons occurs in the plane perpendicular to the magnetic field.[1] In such a structure the electron density of states (number of allowed quantum states per unit energy interval) is peaked at the Landau level. Since $\omega_c \propto B$, the positions of these peaks relative to the Fermi level are controlled by the magnetic field.

The most striking consequence of this phenomenon is the quantum Hall effect, which manifests itself as a stepwise change of the Hall resistance $\rho_{xy} = V_y/I_x$ as a function of magnetic field (see Fig. 22.7). At low temperature (required to establish the condition $\tau << \omega_c^{-1}$) it can be shown [von Klitzing, 1986] that

$$\rho_{xy} = h/sq^2 \tag{22.14}$$

where s is the number of the highest occupied Landau level. Accordingly, when the increased magnetic field pushes the sth Landau level above the Fermi level, ρ_{xy} changes from h/sq^2 to $h/(s–1)q^2$. This stepwise change of ρ_{xy} is seen in Fig. 22.7. Localized states produced by crystal defects determine the shape of the $\rho_{xy}(B)$ dependence between the plateaus given by Eq. (22.14). They are also responsible for the disappearance of $\rho_{xx} = V_x/I_x$ between the transition points (see Fig. 22.7). The quantized Hall resistance ρ_{xy} is expressed in terms of fundamental constants and can be used as a resistance standard that permits one to measure an electrical resistance with better accuracy than any wire resistor standard. In an ultraquantum magnetic field, i.e., when only the lowest Landau level is occupied, plateaus of the Hall resistance are also observed at fractional s (the fractional quantum Hall effect). These plateaus are related to the Coulomb interaction of electrons.

Electrical Breakdown

In sufficiently strong electric fields a measurable fraction of electrons (or holes) acquires sufficient energy to break the valence bond. Such an event (called impact ionization) results in the creation of an electron–hole pair by the energetic electron. Both the primary and secondary electrons as well as the hole are accelerated by the electric field and may participate in further acts of impact ionization. Usually, the impact ionization is balanced by recombination processes. If the applied voltage is high enough, however, the process of electron multiplication leads to avalanche breakdown. The threshold energy E_{th} (the minimum electron energy required to produce an electron–hole pair) is determined by energy and momentum conservation laws. The latter usually results in $E_{\text{th}} > E_g$, as shown in Table 22.1.

The field dependence of the impact ionization is usually described by the impact ionization coefficient α_i, defined as the average number of electron–hole pairs created by a charge carrier per unit distance traveled. A simple analytical expression for α_i [Okuto and Crowell, 1972] can be written as

$$\alpha_i = (\lambda/x)\exp\left(a - \sqrt{a^2 + x^2}\right) \tag{22.15}$$

[1]To simplify the matter we do not discuss surface subbands, which is justified as long as only the lowest of them is occupied.

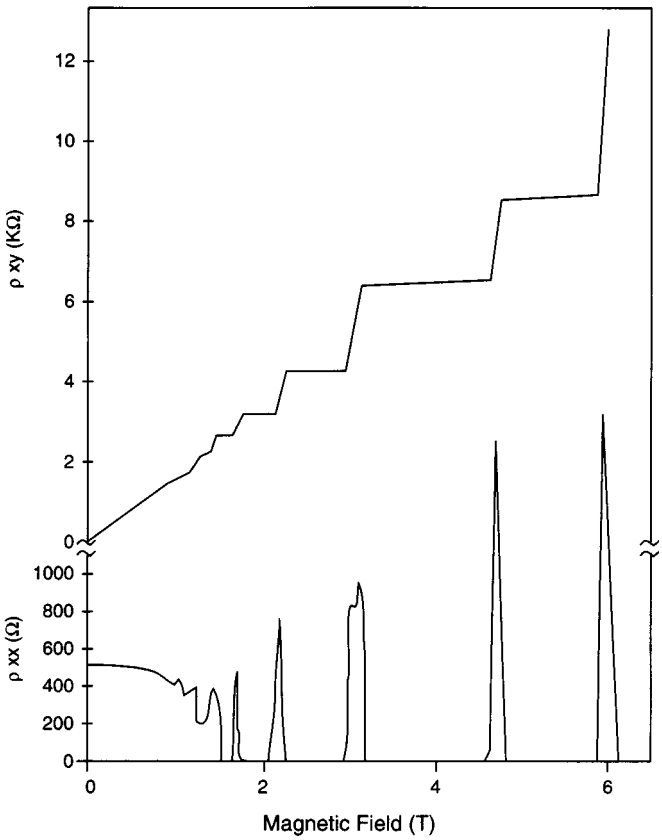


FIGURE 22.7 Experimental curves for the Hall resistance $\rho_{xy} = \mathscr{E}_y/j_x$ and the resistivity $\rho_{xx} = \mathscr{E}_x/j_x$ of a heterostructure as a function of the magnetic field at a fixed carrier density. (*Source:* K. von Klitzing, *Rev. Modern Phys.*, vol. 58, no. 3, p. 525, 1986. With permission.)

TABLE 22.1 Impact Ionization Threshold Energy (eV)

Semiconductor	Si	Ge	GaAs	GaP	InSb
Energy gap, E_g	1.1	0.7	1.4	2.3	0.2
E_{th}, electron-initiated	1.18	0.76	1.7	2.6	0.2
E_{th}, hole-initiated	1.71	0.88	1.4	2.3	0.2

where $x = q\mathscr{E}\lambda/E_{th}$, $a = 0.217\ (E_{th}/E_{opt})^{1.14}$, λ is the carrier mean free path, and E_{opt} is the optical phonon energy ($E_{opt} = 0.063$ eV for Si at 300°C).

An alternative breakdown mechanism is tunneling breakdown, which occurs in highly doped semiconductors when electrons may tunnel from occupied states in the valence band into the empty states of the conduction band.

Optical Properties and Recombination Processes

If the energy of an incident **photon** $\hbar\omega > E_g$, then the energy conservation law permits a direct band-to-band transition, as indicated in Fig. 22.2(b). Because the photon's momentum is negligible compared to that of an electron or hole, the electron's momentum $\hbar\mathbf{k}$ does not change in a direct transition. Consequently, direct transitions are possible only in direct-gap semiconductors where the conduction band minimum and the valence band maximum occur at the same $\mathbf{k}$. The same is true for the reverse transition, where the electron is transferred

from the conduction to the valence band and a photon is emitted. Direct-gap semiconductors (e.g., GaAs) are widely used in optoelectronics.

In indirect-band materials [e.g., Si, see Fig. 22.3(a)], a band-to-band transition requires a change of momentum that cannot be accomplished by absorption or emission of a photon. Indirect band-to-band transitions require the emission or absorption of a phonon and are much less probable than direct transitions.

For $\hbar\omega < E_g$ [i.e., for $\lambda > \lambda_c = 1.24\ \mu m/E_g\ (eV)$ – cutoff wavelength] band-to-band transitions do not occur, but light can be absorbed by a variety of the so-called subgap processes. These processes include the absorption by free carriers, formation of excitons (bound electron–hole pairs whose formation requires less energy than the creation of a free electron and a free hole), transitions involving localized states (e.g., from an acceptor state to the conduction band), and phonon absorption. Both band-to-band and subgap processes may be responsible for the increase of the free charge carriers concentration. The resulting reduction of the resistivity of illuminated semiconductors is called *photoconductivity* and is used in photodetectors.

In a strong magnetic field ($\omega_c\tau >> 1$) the absorption of microwave radiation is peaked at $\omega = \omega_c$. At this frequency the photon energy is equal to the distance between two Landau levels, i.e., $\hbar\omega = E_{S+1} - E_S$ with reference to Eq. (22.13). This effect, known as cyclotron resonance, is used to measure the effective masses of charge carriers in semiconductors [in a simplest case of isotropic $E(\mathbf{k})$ dependence, $m_n^\star = qB/\omega_c$].

In indirect-gap materials like silicon, the generation and annihilation (or recombination) of electron–hole pairs is often a two-step process. First, an electron (or a hole) is trapped in a localized state (called a recombination center) with the energy near the center of the energy gap. In a second step, the electron (or hole) is transferred to the valence (conduction) band. The net rate of recombination per unit volume per unit time is given by the Shockley–Read–Hall theory as

$$R = \frac{np - n_i^2}{\tau_n(p + p_1) + \tau_p(n + n_1)} \tag{22.16}$$

where τ_n, τ_p, p_1, and n_1 are parameters depending on the concentration and the physical nature of recombination centers and temperature. Note that the sign of R indicates the tendency of a semiconductor toward equilibrium (where $np = n_i^2$, and $R = 0$). For example, in the depleted region $np < n_i^2$ and $R < 0$, so that charge carriers are generated.

Shockley–Read–Hall recombination is the dominating recombination mechanism in moderately doped silicon. Other recombination mechanisms (e.g., Auger) become important in heavily doped semiconductors [Wolfe et al., 1989; Shur, 1990; Ferry, 1991].

The recombination processes are fundamental for semiconductor device theory, where they are usually modeled using the continuity equation

$$\frac{\partial n}{\partial t} = div\,\frac{\mathbf{j}_n}{q} - R \tag{22.17}$$

Nanostructure Engineering

Epitaxial growth techniques, especially molecular beam epitaxy and metal-organic chemical vapor deposition, allow monolayer control in the chemical composition process. Both single thin layers and superlattices can be obtained by such methods. The electronic properties of these structures are of interest for potential device applications. In a single quantum well, electrons are bound in the confining well potential. For example, in a rectangular quantum well of width b and infinite walls, the allowed energy levels are

$$E_s(\mathbf{k}) = \pi^2 s^2 \hbar^2/(2m_n^\star b^2) + \hbar^2 k^2/(2m_n^\star), \quad s = 1, 2, 3, \ldots \tag{22.18}$$

where $\mathbf{k}$ is the electron wave vector parallel to the plane of the semiconductor layer. The charge carriers in quantum wells exhibit confined particle behavior. Since $E_s \propto b^{-2}$, well structures can be grown with distance

between energy levels equal to a desired photon energy. Furthermore, the photoluminescence intensity is enhanced because of carrier confinement. These properties are advantageous in fabrication of lasers and photodetectors.

If a quantum well is placed between two thin barriers, the tunneling probability is greatly enhanced when the energy level in the quantum well coincides with the Fermi energy (resonant tunneling). The distance between this "resonant" energy level and the Fermi level is controlled by the applied voltage. Consequently, the current peaks at the voltage corresponding to the resonant tunneling condition. The resulting negative differential resistance effect has been used to fabricate microwave generators operating at both room and cryogenic temperatures.

Two kinds of superlattices are possible: compositional and doping. Compositional superlattices are made of alternating layers of semiconductors with different energy gaps. Doping superlattices consist of alternating *n*- and *p*-type layers of the same semiconductor. The potential is modulated by electric fields arising from the charged dopants. Compositional superlattices can be grown as lattice matched or as strained layers. The latter are used for modification of the band structure, which depends on the lattice constant to produce desirable properties.

In superlattices energy levels of individual quantum wells are split into minibands as a result of electron tunneling through the wide-bandgap layers. This occurs if the electron mean free path is larger than the superlattice period. In such structures the electron motion perpendicular to the layer is quantized. In a one-dimensional tight binding approximation the miniband can be described as

$$E(k) = E_o[1 - \cos(ka)] \tag{22.19}$$

where a is the superlattice period and E_o is the half-width of the energy band. The electron group velocity

$$v = \hbar^{-1}\partial E(k)/\partial k = (E_o a/\hbar)\ \sin(ka) \tag{22.20}$$

is a decreasing function of k (and hence of energy) for $k > \pi/2a$. The higher energy states with $k > \pi/2a$ may become occupied if the electrons are heated by the external field. As a result, a negative differential resistance can be achieved at high electric fields. The weak-field mobility in a superlattice may exceed that of the bulk material because of the separation of dopants if only barriers are doped. In such modulated structures, the increased spatial separation between electrons and holes is also responsible for a strong increase in recombination lifetimes.

Disordered Semiconductors

Both amorphous and heavily doped semiconductors are finding increasing applications in semiconductor technology. The electronic processes in these materials have specific features arising from the lack of long-range order.

Amorphous semiconductors do not have a crystalline lattice, and their properties are determined by the arrangement of the nearest neighboring atoms. Even so, experimental data show that the forbidden energy band concept can be applied to characterize their electrical properties. However, the disordered nature of these materials results in a large number of localized quantum states with energies within the energy gap. The localized states in the upper and lower half of the gap behave like acceptors and donors, respectively. As an example, consider the density of states in hydrogenated amorphous silicon (a-Si) shown in Fig. 22.8. The distribution of the localized states is not symmetrical with respect to the middle of the energy gap. In particular, the undoped hydrogenated amorphous silicon is an *n*-type semiconductor.

Usually amorphous semiconductors are not sensitive to the presence of impurity atoms, which saturate all their chemical bonds in the flexible network of the host atoms. (Compare this with a situation in crystalline silicon where an arsenic impurity can form only four chemical bonds with the host lattice, leaving the fifth responsible for the formation of the donor state.) Consequently, the doping of amorphous semiconductors is difficult to accomplish. However, in hydrogenated a-Si (which can be prepared by the glow discharge decomposition of silane), the density of the localized states is considerably reduced and the conductivity of this material can be controlled by doping. As in crystalline semiconductors, the charge carrier concentration in hydrogenated

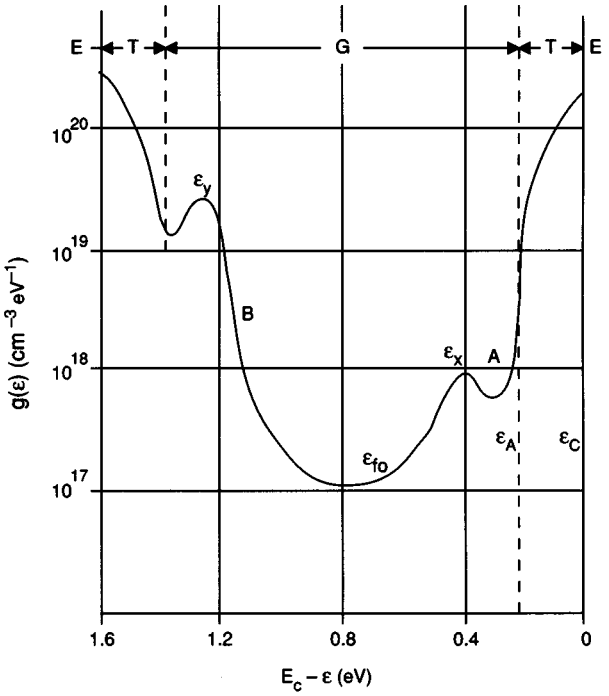


FIGURE 22.8 Experimentally determined density of states for a-Si. *A* and *B* are acceptor-like and donor-like states, respectively. The arrow marks the position of the Fermi level ε_{fo} in undoped hydrogenated a-Si. The energy spectrum is divided into extended states *E*, band-tail states *T*, and gap states *G*. (*Source:* M.H. Brodsky, Ed., *Amorphous Semiconductors,* 2nd ed., Berlin: Springer-Verlag, 1985. With permission.)

a-Si can also be affected by light and strong field effects. The a-Si is used in applications that require deposition of thin-film semiconductors over large areas [xerography, solar cells, thin-film transistors (TFT) for liquid-crystal displays]. The a-Si device performance degrades with time under electric stress (TFTs) or under illumination (Staebler–Wronski effect) because of the creation of new localized states.

An impurity band in crystalline semiconductors is another example of a disordered system. Indeed, the impurity atoms are randomly distributed within the host lattice. For lightly doped semiconductors at room temperature, the random potential associated with charged impurities can usually be ignored. As the doping level increases, however, a single energy level of a donor or an acceptor is transformed into an energy band with a width determined by impurity concentrations. Unless the degree of compensation is unusually high, this reduces the activation energy compared to lightly doped semiconductors. The activation energy is further reduced by the overlap of the wave functions associated with the individual donor or acceptor states.

For sufficiently heavy doping, i.e., for $N_d > N_{dc} = (0.2/a_B)^3$, the ionization energy is reduced to zero, and the transition to metal-type conductivity (the Anderson–Mott transition) takes place. In this expression the effective electron Bohr radius $a_B = \hbar/\sqrt{2m_n^* E_i}$, where E_i is the ionization energy of the donor state. For silicon, $N_{dc} \approx 3.8 \cdot 10^{18}$ cm^{-3}. This effect explains the absence of freeze-out in heavily doped semiconductors.

Defining Terms

Conduction/valence band: The upper/lower of the two partially filled bands in a semiconductor.

Donors/acceptors: Impurities that can be used to increase the concentration of electrons/holes in a semiconductor.

Energy band: Continuous interval of energy levels that are allowed in the periodic potential field of the crystalline lattice.

Energy gap: The width of the energy interval between the top of the valence band and the bottom of the conduction band.

Hole: Fictitious positive charge representing the motion of electrons in the valence band of a semiconductor; the number of holes equals the number of unoccupied quantum states in the valence band.
Phonon: Quantum of lattice vibration.
Photon: Quantum of electromagnetic radiation.

Related Topic

52.1 Introduction

References

D.K. Ferry, *Semiconductors,* New York: Macmillan, 1991.
Y. Okuto and C.R. Crowell, *Phys. Rev.*, vol. B6, p. 3076, 1972.
R.F. Pierret, *Advanced Semiconductor Fundamentals,* Reading, Mass.: Addison-Wesley, 1987.
M. Shur, *Physics of Semiconductor Devices,* Englewood Cliffs, N.J.: Prentice-Hall, 1990.
K. von Klitzing, *Rev. Modern Phys.*, vol. 58, p. 519, 1986.
C.M. Wolfe, N. Holonyak, and G.E. Stilman, *Physical Properties of Semiconductors,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

Further Information

Engineering aspects of semiconductor physics are often discussed in the *IEEE Transactions on Electron Devices*, *Journal of Applied Physics*, and *Solid-State Electronics*.

22.2 Diodes

Miran Milkovic

Diodes are the most widely used devices in low- and high-speed electronic circuits and in rectifiers and power supplies. Other applications are in voltage regulators, detectors, and demodulators. Rectifier diodes are capable of conducting several hundred amperes in the forward direction and less than 1 μA in the reverse direction. Zener diodes are ordinary diodes operated in the Zener or avalanche region and are used as voltage regulators. Varactor diodes are ordinary diodes used in reverse biasing as voltage-dependent capacitors. Tunnel diodes and quantum well devices have a negative differential resistance and are capable of operating in the upper gigahertz region. Photodiodes are ordinary diodes operated in the reverse direction. They are sensitive to light and are used as light sensors. Solar cells are diodes which convert light energy into electrical energy. Schottky diodes, also known as metal-semiconductor diodes, are extremely fast because they are **majority carrier** devices.

pn-Junction Diode

A *pn*-diode is a semiconductor device having a *p*-region, a *n*-region, and a junction between the regions. Modern planar semiconductor *pn*-junction diodes are fabricated by **diffusion** or implantation of impurities into a semiconductor. An *n*-type semiconductor has a relatively large density of free electrons to conduct electric current, and the *p*-type semiconductor has a relatively large concentration of "free" holes to conduct electric current. The *pn*-junction is formed during the fabrication process. There is a large concentration of holes in the *p*-semiconductor and a large concentration of electrons in the *n*-semiconductor. Because of their large concentration gradients, holes and electrons start to diffuse across the junction. As holes move across the junction, negative immobile charges (**acceptors**) are uncovered on the *p* side, and positive immobile charges (**donors**) are uncovered on the *n* side due to the movement of electrons across the junction. When sufficient numbers of the immobile charges on both sides of the junction are uncovered, a potential energy barrier voltage V_0 is created by the uncovered acceptors and donors. This **barrier voltage** prevents further diffusion of holes and electrons across the junction. The charge distribution of acceptors and donors establishes an opposing

electric field, E, which at equilibrium prevents a further diffusion of carriers across the junction. This equilibrium can be regarded as the flow of two equal and opposite currents across the junction, such that the net current across the junction is equal to zero. Thus, one component represents the diffusion of carriers across the junction and the other component represents the **drift** of carriers across the junction due to the electric field E in the junction. The barrier voltage V_0 is, according to the **Boltzmann relation**, [Grove, 1967; Foustad, 1994]

$$V_0 = V_T \ln[p_p/p_n] \tag{22.21}$$

In this equation, p_p is the concentration of holes in the p-material and p_n is the concentration of holes in the n-material. V_T is the thermal voltage. V_T = 26 mV at room temperature (300 K). With

$$p_p \approx N_A \quad \text{and} \quad p_n \approx \frac{n_i^2}{N_D}$$

where n_i is the intrinsic concentration, the barrier voltage V_0 becomes approximately [Sze, 1985; Fonstad, 1994]

$$V_0 = V_T \ln[N_A N_D / n_i^2] \tag{22.22}$$

Here N_A denotes the concentration of immobile acceptors on the p side of the junction and N_D is the concentration of immobile donors on the n side of the junction. A depletion layer of immobile acceptors and donors causes an electric field E across the junction. For silicon, V_0 is at room temperature T = 300°K, typically V_0 = 0.67 V for an abrupt junction with $N_A = 10^{17}$ at/cm^3 and $N_D = 10^{15}$ at/cm^3. The depletion layer width is typically about 4 μm, and the electric field E is about 60 kV/cm. Note the magnitude of the electric field across the junction.

pn-Junction with Applied Voltage

If the externally applied voltage V_D to the diode is opposite to the barrier voltage V_0, then p_p in the Boltzmann relation in Eq. (22.21) is altered to

$$p_p = p_n \exp(V_0 - V_D)/V_T \tag{22.23}$$

This implies that the effective barrier voltage is reduced and the diffusion of carriers across the junction, is increased. Accordingly the concentration of diffusing holes into the n material is at $x = 0$,

$$p_n(x = 0) = p_n \exp V_D/V_T \tag{22.24}$$

and accordingly the concentration of electrons

$$n_n(x = 0) = n_n \exp V_D/V_T \tag{22.25}$$

Most modern planar diodes are unsymmetrical. Figure 22.9 shows a pn-diode with the n region W_n much shorter than the diffusion length L_{pn} of holes in the n-semiconductor region. This results in a linear **concentration gradient** of injected diffusing holes in the n region given by

$$dp/dx = -(p_n \exp V_D/V_T - p_n)/W_n \tag{22.26}$$

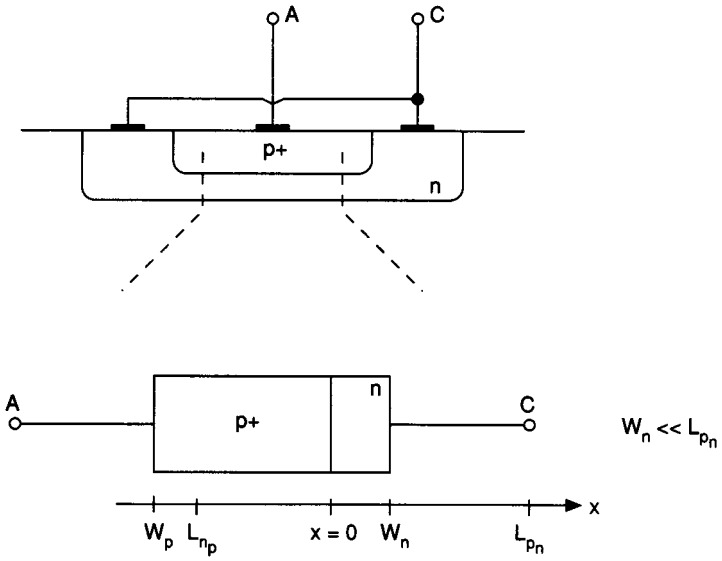


FIGURE 22.9 Planar diodes are fabricated in planar technology. Most modern diodes are unsymmetrical; thus $W_n << L_{pn}$. The *p*-type region is more highly doped than the *n* region.

The diffusion gradient is negative since the concentration of positive holes decreases with distance due to the hole–electron recombinations. The equation for the hole diffusion current is

$$I_p = -qA_jD_p\ dp/dx \tag{22.27}$$

where A_j is the junction area, D_p is the **diffusion constant** for holes, and q is the elementary charge.

By combining of above equations we obtain

$$I_p = (qA_jD_pp_n/W_n)\ (\exp\ V_D/V_T - 1) \tag{22.28}$$

In the *p*-semiconductor we assume that $L_{np} << W_p$; then

$$dn/dx = n_p\ \exp(V_D/V_T - 1) \tag{22.29}$$

By substituting this into the electron diffusion equation,

$$I_n = qA_jD_n\ dn/dx \tag{22.30}$$

we obtain

$$I_n = (qA_jD_nn_p)/L_{np}(\exp\ V_D/V_T - 1) \tag{22.31}$$

Thus, the total junction diffusion current is

$$I_D = I_p + I_n = \{qA_jD_pp_n/W_n + qA_jD_nn_p/L_{np}\}\ (\exp\ V_D/V_T - 1) \tag{22.32}$$

Since the recombination of the injected carriers establishes a diffusion gradient, this in turn yields a flow of current proportional to the slope. For $|-V_D| >> V_T$, i.e., $V_D = -0.1\ V$,

$$I_S = (qA_iD_pp_n/W_n + qA_i\ D_n\ n_p/L_{np}) \tag{22.33}$$

Here I_S denotes the **reverse saturation current**. In practical junctions, the p region is usually much more heavily doped than the n region; thus $n_p << p_n$. Also, since $W_n << L_{np}$ in Eq. (22.33), we obtain

$$I_S = qA_jD_p p_n / W_n = qA_j D_p \, n_i^2 / W_n N_D \tag{22.34}$$

The reverse saturation current in short diodes is mainly determined by the diffusion constant D_p and the width W_n of the n region, by intrinsic concentration n_i, by the doping concentration N_D in the n region, and by the diode area A_j. (In reality, I_S is also slightly dependent on the reverse voltage [Phillips, 1962].)

If V_D is made positive, the exponential term in Eq. (22.32) rapidly becomes larger than one; thus

$$I_D = I_S \exp V_D / V_T \tag{22.35}$$

where I_D is the diode forward current and I_S is the reverse saturation current.

Another mechanism predominates the reverse current I_S in silicon. Due to the recombination centers in the depletion region, a generation-recombination hole–electron current I_G is generated in the depletion region [Phillips, 1962; Sze, 1985].

$$I_G = KqA_jeX_d \tag{22.36a}$$

Here e is the generation rate unit volume, A_j is the junction area, q is the elementary charge, X_d is the depletion layer thickness, and K is a dimensional constant. I_G is proportional to the thickness X_d of the depletion layer and to the junction area A_j. Since X_d increases with the square root of the reverse voltage, I_G increases accordingly, yielding a slight slope in the reverse I-V characteristic. The forward I-V characteristic of the practical diode is only slightly affected (slope $m = 2$) at very small forward currents ($I_D = 1$ nA to 1 μA). In practical diodes $n \approx 1$ at small to medium currents ($I_D = 1$ μA to 10 mA). At large currents ($I_D > 10$ mA), $m = 1$ to 2 due to the high current effects [Phillips, 1962] and due to the series bulk resistance of the diode.

The reverse current I_R in silicon is voltage dependent. The predominant effect is the voltage dependence of the generation-recombination current I_G and to a smaller extent the voltage dependence of I_S.

The total reverse current of the diode is thus equal to

$$I_R = I_G + I_S \tag{22.36b}$$

Forward-Biased Diode

For most practical applications

$$I_D = I_S \exp V_D / mV_T \tag{22.37}$$

where I_S is the reverse saturation current (about 10^{-14} A for a small-signal diode); $V_T = kT/q$ is the thermal voltage equal to 26 mV at room temperature; k = Boltzmann's constant, $1.38 \cdot 10^{-23}$ J/K; T is the absolute temperature in kelvin; q is the elementary charge $1.602 \cdot 10^{-19}$ C; m is the **ideality factor**, $m = 1$ for medium currents, $m = 2$ for very small and very large currents; I_S is part of the total reverse current I_R of the diode $I_R = I_S + I_G$; and I_S is the reverse saturation current and I_G is the generation-recombination current, also called diode leakage current because I_G is not a part of the carrier diffusion process in the diode. I_D is exponentially related to V_D in Fig. 22.10.

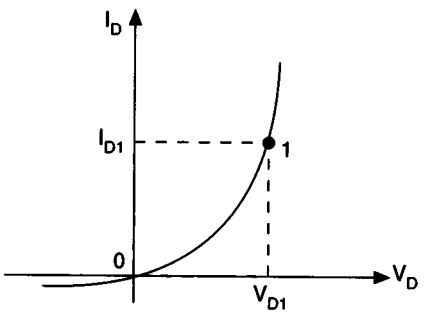


FIGURE 22.10 I_D versus V_D of a diode.

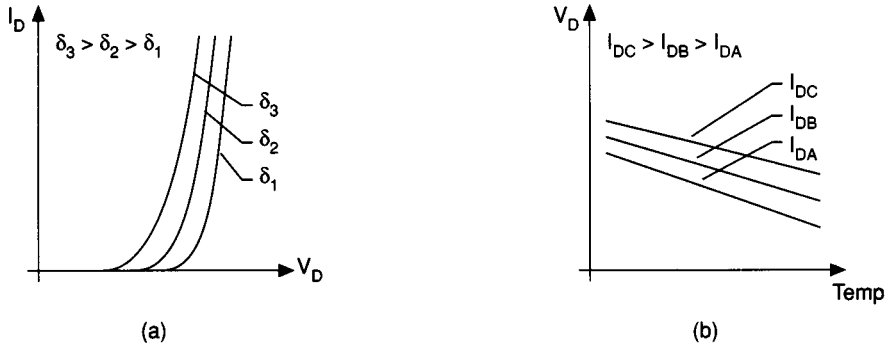


FIGURE 22.11 (a) I_D versus V_D of a diode at three different temperatures $\delta_3 > \delta_2 > \delta_1$. (b) $V_D = f(\text{Temp})$, $I_{DC} > I_{DB} > I_{DA}$.

Temperature Dependence of V_D

Equation (22.37) solved for V_D yields

$$V_D = mV_T \ln(I_D/I_S) \tag{22.38}$$

at constant current I_D, the diode voltage V_D is temperature dependent because V_T and I_S are temperature dependent. Assume $m = 1$. The reverse saturation current I_S from Eq. (22.34) is

$$I_S = qA_j n_i^2 D_p / W_n N_D = B_1 n_i^2 D_p = B_2 n_i^2 \mu_p$$

where $D_p = V_T \mu_p$. With $\mu_p = B_3 T^{-n}$ and for n_i^2

$$n_i^2 = B_4 T^{\gamma} \exp(-V_{G0}/V_T) \tag{22.39}$$

where $\gamma = 4 - n$, and V_{G0} is the extrapolated **bandgap energy** [Gray and Meyer, 1993]. With Eq. (22.39) into Eq. (22.38), the derivative dV_D/dT for $I_D = \text{const}$ yields

$$dV_D/dT = (V_D - V_{G0})/T - \gamma k/q \tag{22.40}$$

At room temperature ($T = 300$ K), and $V_D = 0.65$ V, $V_{G0} = 1.2$ V, $\gamma = 3$, $V_T = 26$ mV, and $k/q = 86$ μV/degree, one gets $dV_D/dT \approx -2.1$ mV/degree.The **temperature coefficient** TC of V_D is thus

$$TC = dV_D/V_D\ dT = 1/T - V_{G0}/V_D T - \gamma k/qV_D \tag{22.41}$$

For the above case $TC \approx -0.32\%$/degree. In practical applications it is more convenient to use the expression

$$V_D(\delta_2) = V_D(\delta_1) - TC(\delta_2 - \delta_1) \tag{22.42}$$

where δ_1 and δ_2 are temperatures in degrees Celsius. For $TC = -0.32\%$/degree and $V_D = 0.65$ V at $\delta_1 = 27°C$, $V_D = 0.618$ V at $\delta_2 = 37°C$. Both dV_D/dT and TC are I_D dependent. At higher I_D, both dV_D/dT and TC are smaller than at a lower I_D, as shown in Fig. 22.11.

I_D-V_D Characteristic

From the I_D-V_D characteristic of the diode one can find for $m = 1$

$$I_{D1} = I_S \exp(V_{D1}/V_T) \quad \text{and} \quad I_{D2} = I_S \exp(V_{D2}/V_T) \tag{22.43}$$

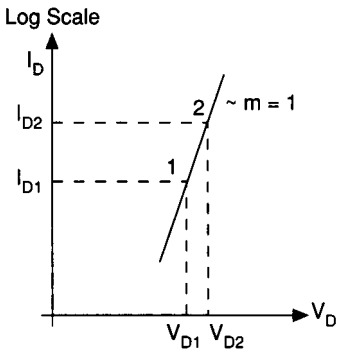


FIGURE 22.12 I_D versus V_D of a diode on a semi-logarithmic plot.

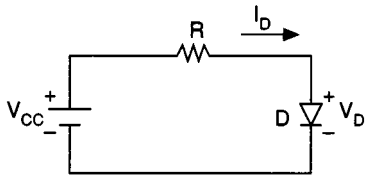


FIGURE 22.13 Diode-resistor biasing circuit.

Thus, the ratio of currents is

$$I_{D2}/I_{D1} = \exp\,(V_{D2} - V_{D1})/V_T \tag{22.44}$$

or the difference voltage

$$V_{D2} - V_{D1} = V_T \ln(I_{D2}/I_{D1}) \tag{22.45}$$

in terms of base 10 logarithm

$$V_{D2} - V_{D1} = V_T\, 2.3\, \log(I_{D2}/I_{D1}) \tag{22.46}$$

For $(I_{D2}/I_{D1}) = 10$ (one decade), $V_{D2} - V_{D1} = \sim 60$ mV, or $V_{D2} - V_{D1} = 17.4$ mV for $(I_{D2}/I_{D1}) = 2$. In a typical example, $m = 1$, $V_D = 0.67$ V at $I_D = 100$ μA. At $I_D = 200$ μA, $V_D = 0.67$ V + 17.4 mV = 0.687 V.

DC and Large-Signal Model

The diode equation in Eq. (22.37) is widely utilized in diode circuit design. I_S and m can sometimes be found from the data book or they can be determined from measured I_D and V_D. From two measurements of I_D and V_D, for example, $I_D = 0.2$ mA at $V_D = 0.670$ V and $I_D = 10$ mA at $V_D = 0.772$ V, one can find $m = 1.012$ and $I_S = 1.78 \cdot 10^{-15}$ A for the particular diode. A practical application of the large-signal diode model is shown in Fig. 22.13. Here, the current I_D through the series resistor R and a diode D is to be found,

$$I_D = (V_{CC} - V_D)/R \tag{22.47}$$

The equation is implicit and cannot be solved for I_D since V_D is a function of I_D. Here, V_D and I_D are determined by using iteration. By assuming $V_D = V_{D0} = 0.6$ V (cut-in voltage), the first iteration yields

$$I_D(1) = (5\text{ V} - 0.6\text{ V})/1\text{ k}\Omega = 4.4\text{ mA}$$

Next, the first iteration voltage $V_D(1)$ is calculated (by using m and I_S above and $I_{D1} = 4.4$ mA), thus

$$V_D(1) = mV_T[\ln\, I_D(1)/I_S] = 1.012 \times 26\text{ mV}\ln(4.4\text{ mA}/1.78 \cdot 10^{-15}\text{ A})$$

$$= 0.751\text{ V}$$

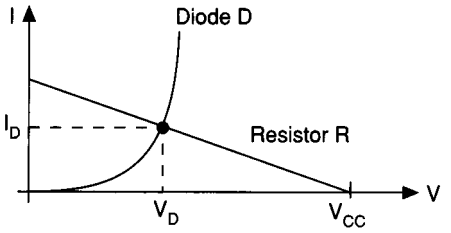


FIGURE 22.14 Graphical analysis of a diode-resistor circuit.

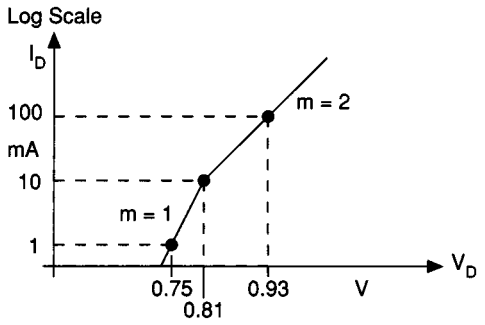
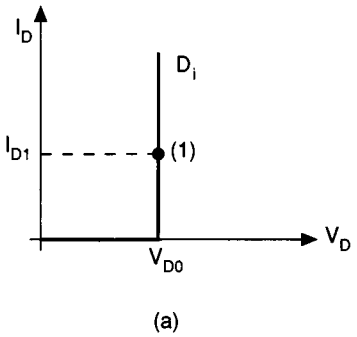


FIGURE 22.15 I_D versus V_D of a diode at low and high forward currents.



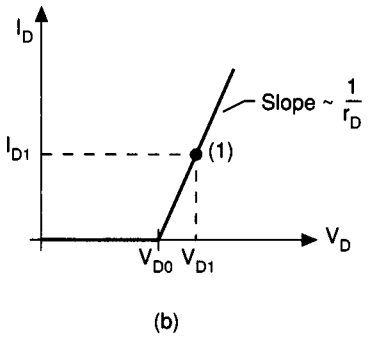


FIGURE 22.16 (a) Simplified piecewise linear model of a diode; (b) improved piecewise linear model of a diode. The diode cut-in voltage V_{D0} is defined as the voltage V_D at a very small current I_D typically at about 1 nA. For silicon diodes this voltage is typically $V_{D0} = 0.6$ V.

From the second iteration $I_D(2) = [V_{CC} - V_D(1)]/R = 4.25$ mA and thus $V_D(2) = 0.75$ V. The third iteration yields $I_D(3) = 4.25$ mA, and $V_D(3) = 0.75$ V. These are the actual values of I_D and V_D for the above example, since the second and the third iterations are almost equal.

Graphical analysis (in Fig. 22.14) is another way to analyze the circuit in Fig. 22.13. Here the load line R is drawn with the diode I-V characteristic, where $V_{CC} = V_D + I_D R$. This type of analysis is illustrative but not well suited for a numerical analysis.

High Forward Current Effects

In the *pn*-junction diode analysis it was assumed that the density of injected carriers from the *p* region into the *n* region is small compared to the density of majority carriers in that region. Thus, all of the forward voltage V_D appears across the junction. Therefore, the injected carriers move only because of the diffusion. At high forward currents this is not the case anymore. When the voltage drop across the bulk resistance becomes comparable with the voltage across the junction, the effective applied voltage is reduced [Phillips, 1962]. Due to the electric field created by the voltage drop in the bulk (neutral) regions, the current is not only a diffusion current anymore. The drift current due to the voltage drop across the bulk region opposes the diffusion current. The net effect is that, first, the current becomes proportional to twice the diffusion constant, second, the high-level current becomes independent of resistivity, and, third, the magnitude of the exponent is reduced by a factor of two in Eq. (22.37). The effect of high forward current on the I-V characteristic is shown in Fig. 22.15. In all practical designs, $m \approx 2$ at $I_D \geq 20$ mA in small-signal silicon diodes.

Large-Signal Piecewise Linear Model

Piecewise linear model of a diode is a very useful tool for quick circuit design containing diodes. Here, the diode is represented by asymptotes and not by the exponential *I-V* curve. The simplest piecewise linear model is shown in Fig. 22.16(a). Here D_i is an ideal diode with $V_D = 0$ at $I_D \geq 0$, in series with V_{D0}, where V_{D0} is the diode cut-in or threshold voltage. The current in the diode will start to flow at $V_D \geq V_{D0}$.

An improved model is shown in Fig. 22.16(b), where V_{D0} is again the diode voltage at a very small current I_{D0}, r_D is the extrapolated diode resistance, and I_{D1} is the diode current in operating point 1. Thus, the diode voltage is

$$V_{D1} = V_{D0} + I_{D1}\, r_D \tag{22.48}$$

where V_{D1} is the diode voltage at I_{D1}. V_{D0} for silicon is about 0.60 V. r_D is estimated from the fact that V_D in a real diode is changing per decade of current by m 2.3 V_T. Thus, V_D changes about 60 mV for a decade change of current I_D at $m = 1$. Thus in a 0.1 to 10 mA current change, V_D changes about 120 mV, which corresponds to $anr_D \approx 120$ mV/10 mA = 12 Ω.

The foregoing method is an approximation; however, it is quite practical for first-hand calculations. To compare this with the above iterative approach let us assume $m = 1$, $V_{D0} = 0.60$ V, $r_D = 12$ Ω, $V_{CC} = 5$ V, $R =$ 1 kΩ. The current $I_{D1} = [V_{CC} - V_{D0}]/(R + r_D) = 4.34$ mA compared with $I_{D1} = 4.25$ mA in the iterative approach.

Small-Signal Incremental Model

In the small-signal **incremental model**, the diode is represented by linear elements. In small-signal (incremental) analysis, the diode voltage signals are assumed to be about $V_T/2$ or less, thus much smaller than the dc voltage V_D across the diode. In the forward-biased diode, three elements are of practical interest: **incremental resistance** (or small-signal or differential resistance) r_d, the **diffusion capacitance** C_d, and the **junction capacitance** C_j.

Incremental Resistance, r_d

For small signals the diode represents a small-signal resistance (often called incremental or differential resistance) r_d in the operating point (I_D, V_D) where

$$r_d = dV_D/dI_D = mV_T/I_S \exp(V_D/mV_T) = mV_T/I_D \tag{22.49}$$

In Fig. 22.17, r_d is shown as the tangent in the dc operating point (V_D, I_D). Note that r_d is independent of the geometry of the device and inversely proportional to the diode dc current. Thus for $I_D = 1$ mA, $m = 1$ and $V_T = 26$ mV, the incremental resistance is $r_d = 26$ Ω.

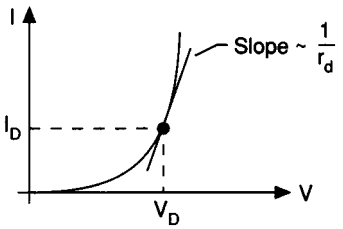


FIGURE 22.17 Small-signal incremental resistance r_d of a diode.

Diffusion Capacitance, C_d

C_d is associated with the injection of holes and electrons in the forward-biased diode. In steady state, holes and electrons are injected across the junction. Hole and electron currents flow due to the diffusion gradients on both sides of the junction in Fig. 22.18. In a short diode, holes are traveling a distance $W_n << L_{p_n}$. For injected holes, and since $w_n << L_{p_n}$

$$I_p = dq_p/dt = dq_p\, v/dx \tag{22.50}$$

where v is the average carrier velocity, D_p is the diffusion constant for holes and W_n is the travel distance of holes. By integrating of Eq. (22.50) one gets

$$I_p \int_0^{W_n} dx = v \int_0^{Q_p} dq_p$$

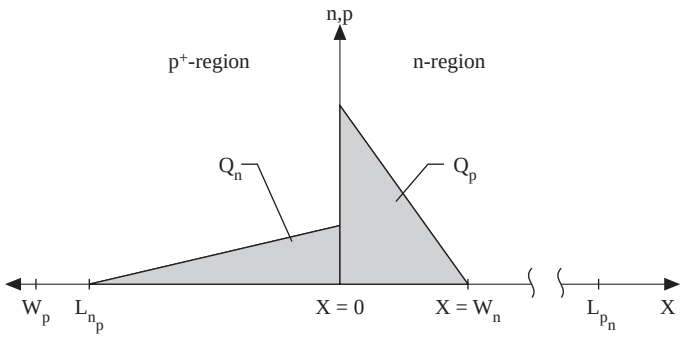


FIGURE 22.18 Minority carrier charge injection in a diode.

and the charge Q_p of holes becomes

$$Q_p = I_p W_n / v = I_p \tau_p. \tag{22.51}$$

$\tau_p = W_n / v$ is the transit time holes travel the distance W_n. Similarly, for electron charge Q_n, since $W_p >> L_{np}$

$$Q_n = I_n L_{np} / v = I_n \tau_n. \tag{22.52}$$

Thus the total diffusion charge Q_d is

$$Q_d = Q_p + Q_n, \tag{22.53}$$

and the total transit time is

$$\tau_F = \tau_p + \tau_n, \tag{22.54}$$

and with $I_p + I_n = I_D = I_S \exp V_D / mV_T$ and Eqs. (22.51), (22.52), and (22.54) one gets

$$Q_d = \tau_F I_S \exp V_D / mV_T = \tau_F I_D. \tag{22.55}$$

The total diffusion capacitance is

$$C_d = C_p + C_n = dQ_d / dV_D = Q_d / mV_T \tag{22.56}$$

and from Eqs. (22.55) and (22.56)

$$C_d = I_D \tau_F / mV_T. \tag{22.57}$$

C_d is thus directly proportional to I_D and to the carrier transit time τ_F. For an unsymmetrical diode with $W_n << L_{pn}$ and $N_A >> N_D$ [Gray and Meyer, 1984]

$$\tau_F \approx W_n^2 / 2D_p \tag{22.58}$$

τ_F is usually given in data books or it can be measured.

For $W_n = 6\ \mu$ and $D_p = 14\ \text{cm}^2/\text{s}$, $\tau_F \approx 13$ ns, $I_D = 1$ mA, $V_T = 26$ mV, and $m = 1$, the diffusion capacitance is $C_d = 500$ pF.

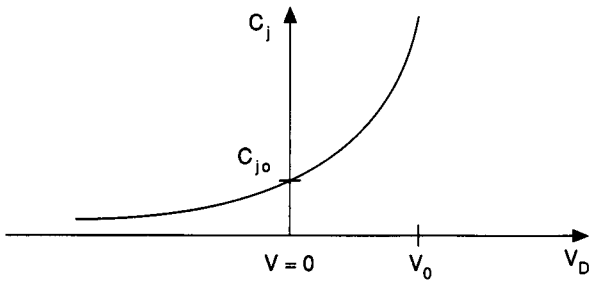


FIGURE 22.19 Depletion capacitance C_j of a diode versus diode voltage V_R.

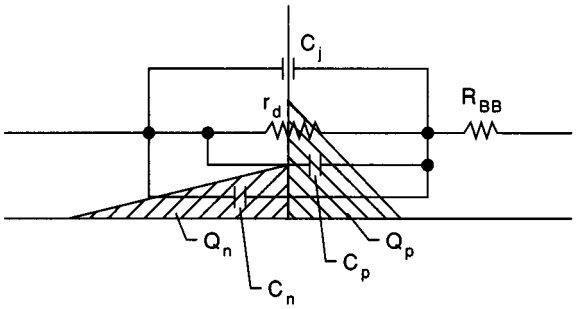


FIGURE 22.20 Simplified small-signal model of a diode.

Depletion Capacitance, C_j

The depletion region is always present in a *pn*-diode. Because of the immobile ions in the depletion region, the junction acts as a voltage-dependent plate capacitor C_j [Gray and Meyer, 1993; Horenstein, 1990]

$$C_j = C_{j0} \Big/ \sqrt{V_0 - V_D} \tag{22.59}$$

V_D is the diode voltage (positive value for forward biasing, negative value for reverse biasing), and C_{j0} is the zero bias depletion capacitance; A_j is the junction diode area:

$$C_{j0} = KA_j \tag{22.60}$$

K is a proportionality constant dependent on diode doping, and A_j is the diode area. C_j is voltage dependent. As V_D increases, C_j increases in a forward-biased diode in Fig. 22.19. For $V_0 = 0.7$ V and $V_D = -10$ V and $C_{j0} =$ 3 pF, the diode depletion capacitance is $C_j = 0.75$ pF. In Fig. 22.20 the small-signal model of the diode is shown. The total small-signal time constant τ_d is thus (by neglecting the bulk series diode resistance R_{BB})

$$\tau_d = r_d(C_d + C_j) = r_dC_d + r_dC_j = \tau_F + r_dC_j \tag{22.61}$$

τ_d is thus current dependent. At small I_D the r_dC_j product is predominant. For high-speed operation r_dC_j must be kept much smaller than τ_F. This is achieved by a large operating current I_D. The diode behaves to a first approximation as a frequency-dependent element. In the reverse operation, the diode behaves as a high ohmic resistor $R_p \approx V_R/I_G$ in parallel with the capacitor C_j. In forward small-signal operation, the diode behaves as a resistor r_d in parallel with the capacitors C_j and C_d (R_p is neglected). Thus, the diode is in a first approximation, a low-pass network.

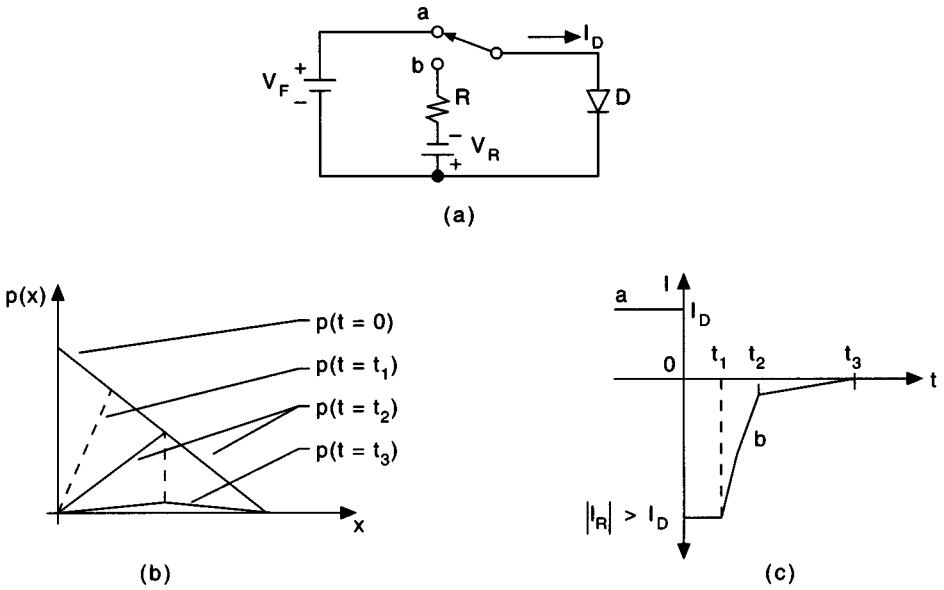


FIGURE 22.21 (a) Diode is switched from forward into reverse direction; (b) concentration of holes in the *n* region; (c) diode turns off in three time intervals.

Large-Signal Switching Behavior of a *pn*-Diode

When a forward-biased diode is switched from the forward into the reverse direction, the stored charge Q_d of **minority carriers** must first be removed. The charge of minority carriers in the forward-biased unsymmetrical diode is from Eqs. (22.55) and (22.58)

$$Q_d = I_D\tau_F = I_D W_n^2/2D_p \tag{22.62}$$

where $W_n << L_{pn}$ is assumed. τ_F is minimized by making W_n very small. Very low-lifetime τ_F is required for high-speed diodes. **Carrier lifetime** τ_F is reduced by adding a large concentration of recombination centers into the junction. This is common practice in the fabrication of high-speed computer diodes [Phillips, 1962]. The charge Q_d is stored mainly in the *n* region in the form of a concentration gradient of holes in Fig. 22.21(a). The diode is turned off by moving the switch from position (a) into position (b) [Fig. 22.21(a)]. The removal of carriers is done in three time intervals. During the time interval t_1, also called the recovery phase, a constant reverse current $|I_R| = V_R/R$ flows in the diode. During the time interval $t_2 - t_1$ the charge in the diode is reduced by about 1/2 of the original charge. During the third interval $t_3 - t_2$, the residual charge is removed.

If during the interval t_1, $|I_R| >> I_D$, then Q_d is reduced only by flow of reverse diffusion current; no holes arrive at the metal contact [Gugenbuehl et al., 1962], and

$$t_1 \approx \tau_F\,(I_D/|I_R|)^2 \tag{22.63}$$

During time interval $t_2 - t_1$, when $|I_R| = I_D$, in Fig. 22.21(b),

$$t_2 - t_1 \approx \tau_F I_D/|I_R| \tag{22.64}$$

The residual charge is removed during the time $t_3 - t_2 \approx 0.5\,\tau_F$.

Diode Reverse Breakdown

Avalanche breakdown occurs in a reverse-biased plane junction when the critical electric field E_{crt} at the junction within the depletion region reaches about $3 \cdot 10^5$ V/cm for junction doping densities of about 10^{15} to 10^{16} at/cm^3 [Gray and Meyer, 1984]. At this electric field E_{crt}, the minority carriers traveling (as reverse current) in the depletion region acquire sufficient energy to create new hole–electron pairs in collision with atoms. These energetic pairs are able to create new pairs, etc. This process is called the avalanche process and leads to a sudden increase of the reverse current I_R in a diode. The current is then limited only by the external circuitry. The avalanche current is not destructive as long as the local junction temperature does not create local hot spots, i.e., melting of material at the junction. Figure 22.22 shows a typical *I-V* characteristic for a junction diode in the avalanche breakdown. The effect of breakdown is seen by the large increase of the reverse current I_R when V_R reaches $-BV$. Here BV is the actual breakdown voltage. It was found that $I_{RA} = M\, I_R$, where I_{RA} is the avalanche reverse current at BV, M is the multiplication factor, and I_R is the reverse current not in the breakdown region. M is defined as

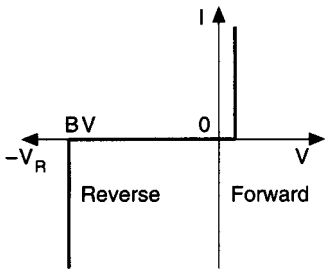


FIGURE 22.22 Reverse breakdown voltage of a diode at $-V_R = BV$.

$$M = 1/[1 - V_R/BV]^n \tag{22.65}$$

where $n = 3$ to 6. As $V_R = BV$, $M \rightarrow \infty$ and $I_{RA} \rightarrow \infty$. The above BV is valid for a strictly plane junction without any curvature. However, in a real planar diode as shown in Fig. 22.9, the *p*-diffusion has a curvature with a finite radius x_j. If the diode is doped unsymmetrically, thus $\sigma_p >> \sigma_n$, then the depletion area penetrates mostly into the *n* region. Because of the finite radius, the breakdown occurs at the radius x_j, rather than in a plane junction [Grove, 1967]. The breakdown voltage is significantly reduced due to the curvature. In very shallow planar diodes, the avalanche breakdown voltage BV can be much smaller than 10 V.

Zener and Avalanche Diodes

Zener diodes (ZD) and avalanche diodes are *pn*-diodes specially built to operate in reverse breakdown. They operate in the reverse direction; however, their operating mechanism is different. In a Zener diode the hole–electron pairs are generated by the electric field by direct transition of carriers from valence band into the conductance band. In an avalanche diode, the hole–electron pairs are generated by impact ionization due to high-energy holes and electrons.

Avalanche and Zener diodes are extensively used as voltage regulators and as overvoltage protection devices. T_C of Zener diodes is negative at $V_Z \leq 3.5$ to 4.5 V and is equal to zero at about $V_Z \approx 5$ V. T_C of a Zener diode operating above 5 V is in general positive. Above 10 V the *pn*-diodes operate as avalanche diodes with a strong positive temperature coefficient. The T_C of a Zener diode is more predictable than that of the avalanche diode. Temperature-compensated Zener diodes utilize the positive T_C of a 7-V Zener diode, which is compensated with a series-connected forward-biased diode with a negative T_C. The disadvantage of Zener diodes is a relatively large electronic noise.

Varactor Diodes

The varactor diode is an ordinary *pn*-diode that uses the voltage-dependent variable capacitance of the diode. The varactor diode is widely used as a voltage-dependent capacitor in electronically tuned radio receivers and in TV.

Tunnel Diodes

The tunnel diode is an ordinary *pn*-junction diode with very heavy doped *n* and *p* regions. Because the junction is very thin, a tunnel effect takes place. An electron can tunnel through the thin depletion layer from the

conduction band of the *n* region directly into the valence band of the *p* region. Tunnel diodes create a negative differential resistance in the forward direction, due to the tunnel effect. Tunnel diodes are used as mixers, oscillators, amplifiers, and detectors. They operate at very high frequencies in the gigahertz bands.

Photodiodes and Solar Cells

Photodiodes are ordinary *pn*-diodes that generate hole–electron pairs when exposed to light. A photocurrent flows across the junction, if the diode is reverse biased. Silicon *pn*-junctions are used to sense light at near-infrared and visible spectra around 0.9 μm. Other materials are used for different spectra.

Solar cells utilize the *pn*-junction to convert light energy into electrical energy. Hole–electron pairs are generated in the semiconductor material by light photons. The carriers are separated by the high electric field in the depletion region across the *pn*-junction.The electric field forces the holes into the *p* region and the electrons into the *n* region. This displacement of mobile charges creates a voltage difference between the two semiconductor regions. Electric power is generated in an external load connected between the terminals to the *p* and *n* regions. The conversion efficiency is relatively low, around 10 to 12%. With the use of new materials, an efficiency of about 30% has been reported. Efficiency up to 45% was achieved by using monochromatic light.

Schottky Barrier Diode

The Schottky barrier diode is a metal-semiconductor diode. Majority carriers carry the electric current. No minority carrier injection takes place. When the diode is forward biased, carriers are injected into the metal, where they reside as majority carriers at an energy level that is higher than the Fermi level in metals. The *I-V* characteristic is similar to conventional diodes. The barrier voltage is small, about 0.2 V for silicon. Since no minority carrier charge exists, the Schottky barrier diodes are very fast. They are used in high-speed electronic circuitry.

Defining Terms

Acceptor: Ionized, negative-charged immobile dopant atom (ion) in a *p*-type semiconductor after the release of a hole.

Avalanche breakdown: In the reverse-biased diode, hole–electron pairs are generated in the depletion region by ionization, thus by the lattice collision with energetic electrons and holes.

Bandgap energy: Energy difference between the conduction band and the valence band in a semiconductor.

Barrier voltage: A voltage which develops across the junction due to uncovered immobile ions on both sides of the junction. Ions are uncovered due to the diffusion of mobile carriers across the junction.

Boltzmann relation: Relates the density of particles in one region to that in an adjacent region, with the potential energy between both regions.

Carrier lifetime: Time an injected minority carrier travels before its recombination with a majority carrier.

Concentration gradient: Difference in carrier concentration.

Diffusion: Movement of free carriers in a semiconductor caused by the difference in carrier densities (concentration gradient). Also movement of dopands during fabrication of diffused diodes.

Diffusion capacitance: Change in charge of injected carriers corresponding to change in forward bias voltage in a diode.

Diffusion constant: Product of the thermal voltage and the mobility in a semiconductor.

Donor: Ionized, positive-charged immobile dopant atom (ion) in an *n*-type semiconductor after the release of an electron.

Drift: Movement of free carriers in a semiconductor due to the electric field.

Ideality factor: The factor determining the deviation from the ideal diode characteristic $m = 1$. At small and large currents $m \approx 2$.

Incremental model: Small-signal differential (incremental) semiconductor diode equivalent RC circuit of a diode, biased in a dc operating point.

Incremental resistance: Small-signal differential (incremental) resistance of a diode, biased in a dc operating point.

Junction capacitance: Change in charge of immobile ions in the depletion region of a diode corresponding to a change in reverse bias voltage on a diode.

Majority carriers: Holes are in majority in a *p*-type semiconductor; electrons are in majority in an *n*-type semiconductor.

Minority carriers: Electrons in a *p*-type semiconductor are in minority; holes are in majority. Similarly, holes are in minority in an *n*-type semiconductor and electrons are in majority.

Reverse breakdown: At the reverse breakdown voltage the diode can conduct a large current in the reverse direction.

Reverse generation-recombination current: Part of the reverse current in a diode caused by the generation of hole–electron pairs in the depletion region. This current is voltage dependent because the depletion region width is voltage dependent.

Reverse saturation current: Part of the reverse current in a diode which is caused by diffusion of minority carriers from the neutral regions to the depletion region. This current is almost independent of the reverse voltage.

Temperature coefficient: Relative variation $\Delta X/X$ of a value X over a temperature range, divided by the difference in temperature ΔT.

Zener breakdown: In the reverse-biased diode, hole–electron pairs are generated by a large electric field in the depletion region.

Related Topic

5.1 Diodes and Rectifiers

References

C.G. Fonstad, *Microelectronic Devices and Circuits,* New York: McGraw-Hill, 1994.
P.R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits,* New York: John Wiley & Sons, 1993.
A.S. Grove, *Physics and Technology of Semiconductor Devices*, New York: John Wiley & Sons, 1967.
W. Gugenbuehl, M.J.O. Strutt, and W. Wunderlin, *Semiconductor Elements,* Basel: Birkhauser Verlag, 1962.
M.N. Horenstein, *Microelectronic Circuits and Devices,* Englewood Cliffs, N.J.: Prentice-Hall, 1990.
A.B. Phillips, *Transistor Engineering,* New York: McGraw-Hill, 1962.
S.M. Sze, *Semiconductor Devices, Physics, and Technology,* New York: John Wiley & Sons, 1985.

Further Information

A good classical introduction to diodes is found in P. E. Gray and C. L. Searle, *Electronic Principles,* New York: Wiley, 1969. Other sources include S. Soclof, *Applications of Analog Integrated Circuits,* Englewood Cliffs, N.J.: Prentice-Hall, 1985 and E. J. Angelo, Jr., *Electronics: BJT's, FET's and Microcircuits,* New York: McGraw-Hill, 1969.

22.3 Electrical Equivalent Circuit Models and Device Simulators for Semiconductor Devices

Aicha Elshabini-Riad, F. W. Stephenson, and Imran A. Bhutta

In the past 15 years, the electronics industry has seen a tremendous surge in the development of new semiconductor materials, novel devices, and circuits. For the designer to bring these circuits or devices to the market in a timely fashion, he or she must have design tools capable of predicting the device behavior in a variety of circuit configurations and environmental conditions. Equivalent circuit models and semiconductor device simulators represent such design tools.

Overview of Equivalent Circuit Models

Circuit analysis is an important tool in circuit design. It saves considerable time, at the circuit design stage, by providing the designer with a tool for predicting the circuit behavior without actually processing the circuit.

An electronic circuit usually contains active devices, in addition to passive components. While the current and voltage behavior of passive devices is defined by simple relationships, the equivalent relationships in active devices are quite complicated in nature. Therefore, in order to analyze an active circuit, the devices are replaced by equivalent circuit models that give the same output characteristics as the active device itself. These models are made up of passive elements, voltage sources, and current sources. Equivalent circuit models provide the designer with reasonably accurate values for frequencies below 1 GHz for bipolar junction transistors (BJTs), and their use is quite popular in circuit analysis software. Some field-effect transistor (FET) models are accurate up to 10 GHz. As the analysis frequency increases, however, so does the model complexity. Since the equivalent circuit models are based on some fundamental equations describing the device behavior, they can also be used to predict the characteristics of the device itself.

When performing circuit analysis, two important factors that must be taken into account are the speed and accuracy of computation. Sometimes, the computation speed can be considerably improved by simplifying the equivalent circuit model, without significant loss in computation accuracy. For this reason, there are a number of equivalent circuit models, depending on the device application and related conditions. Equivalent circuit models have been developed for diodes, BJTs, and FETs. In this overview, the equivalent circuit models for BJT and FET devices are presented.

Most of the equivalent circuits for BJTs are based on the Ebers–Moll model [1954] or the Gummel–Poon model [1970]. The original Ebers–Moll model was a large signal, nonlinear dc model for BJTs. Since then, a number of improvements have been incorporated to make the model more accurate for various applications. In addition, an accurate model has been introduced by Gummel and Poon.

There are three main types of equivalent circuit models, depending on the device signal strength. On this basis, the models can be classified as follows:

1. Large-signal equivalent circuit model
2. Small-signal equivalent circuit model
3. DC equivalent circuit model

Use of the large-signal or small-signal model depends on the magnitude of the driving source. In applications where the driving currents or the driving voltages have large amplitudes, large-signal models are used. In circuits where the signal does not deviate much from the dc biasing point, small-signal models are more suitable. For dc conditions and very-low-frequency applications, dc equivalent circuit models are used. For dc and very-low-frequency analysis, the circuit element values can be assumed to be lumped, whereas in high-frequency analysis, incremental element values give much more precise results.

Large-Signal Equivalent Circuit Model

Depending on the frequency of operation, large-signal equivalent circuit models can be further classified as (1) high-frequency large-signal equivalent circuit model and (2) low-frequency large-signal equivalent circuit model.

High-Frequency Large-Signal Equivalent Circuit Model of a BJT. In this context, high-frequency denotes frequencies above 10 kHz. In the equivalent circuit model, the transistor is assumed to be composed of two back-to-back diodes. Two current-dependent current sources are added to model the current flowing through the reverse-biased base-collector junction and the forward-biased base-emitter junction. Two junction capacitances, C_{jE} and C_{jC}, model the fixed charges in the emitter-base space charge region and base-collector space charge region, respectively. Two diffusion capacitances, C_{DE} and C_{DC}, model the corresponding charge associated with mobile carriers, while the base resistance, r_b, represents the voltage drop in the base region. All the above circuit elements are very strong functions of operating frequency, signal strength, and bias voltage.

The high-frequency large-signal equivalent circuit model of an *npn* BJT is shown in Fig. 22.23, where the capacitances C_{jE}, C_{jC}, C_{DE}, C_{DC} are defined as follows:

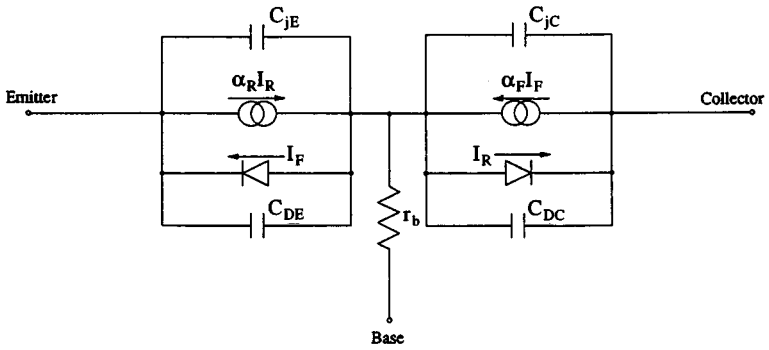


FIGURE 22.23 High-frequency large-signal equivalent circuit model of an *npn* BJT.

$$C_{jE}(V_{B'E'}) = \frac{C_{jEO}}{\left(1 - \dfrac{v_{B'E'}}{\phi_E}\right)^{m_E}} \tag{22.66}$$

$$C_{jC}(V_{B'C'}) = \frac{C_{jCO}}{\left(1 - \dfrac{v_{B'C'}}{\phi_C}\right)^{m_C}} \tag{22.67}$$

$$C_{DE} = \frac{\tau_F I_{CC}}{V_{B'E'}} \tag{22.68}$$

and

$$C_{DC} = \frac{\tau_R I_{EC}}{V_{B'E'}} \tag{22.69}$$

In these equations, $V_{B'E'}$ is the internal base-emitter voltage, C_{jEO} is the base-emitter junction capacitance at $V_{B'E'} = 0$, ϕ_E is the base-emitter barrier potential, and m_E is the base-emitter capacitance gradient factor. Similarly, $V_{B'C'}$ is the internal base-collector voltage, C_{jCO} is the base-collector junction capacitance at $V_{B'C'} = 0$, ϕ_C is the base-collector barrier potential, and m_C is the base-collector capacitance gradient factor. I_{CC} and I_{EC} denote the collector and emitter reference currents, respectively, while τ_F is the total forward transit time, and τ_R is the total reverse transit time. α_R and α_F are the large-signal reverse and forward current gains of a common base transistor, respectively.

This circuit can be made linear by replacing the forward-biased base-emitter diode with a low-value resistor, r_π, while the reverse-biased base-collector diode is replaced with a high-value resistor, r_μ. The junction and diffusion capacitors are lumped together to form C_π and C_μ, while the two current sources are lumped into one source ($g_{mF}V_F - g_{mR}V_R$), where g_{mF} and g_{mR} are the transistor forward and reverse transconductances, respectively. V_F and V_R are the voltages across the forward- and reverse-biased diodes, represented by r_π and r_μ, respectively. r_π is typically about 3 kΩ, while r_μ is more than a few megohms, and C_π is about 120 pF. The linear circuit representation is illustrated in Fig. 22.24.

The Gummel–Poon representation is very similar to the high-frequency large-signal linear circuit model of Fig. 22.24. However, the terms describing the elements are different and a little more involved.

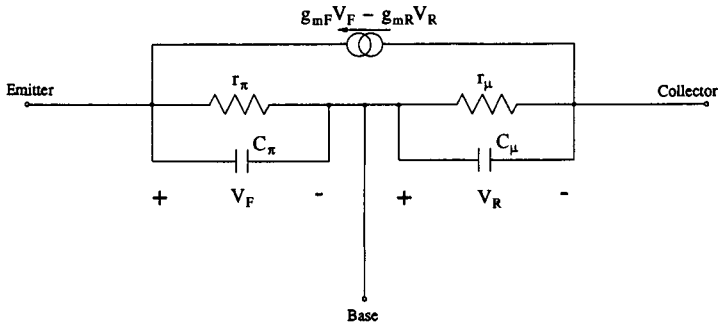


FIGURE 22.24 High-frequency large-signal equivalent circuit model (linear) of an *npn* BJT.

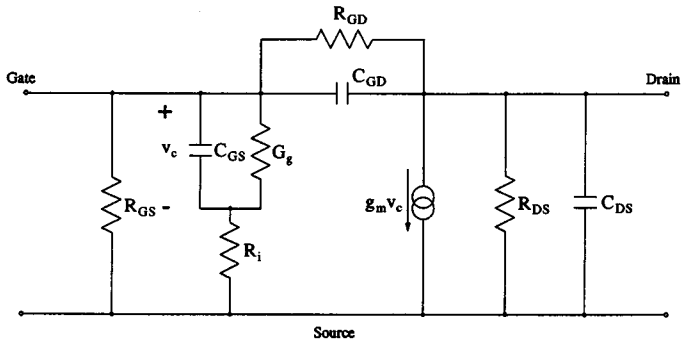


FIGURE 22.25 High-frequency large-signal equivalent circuit model of a FET.

High-Frequency Large-Signal Equivalent Circuit Model of a FET. In the high-frequency large-signal equivalent circuit model of a FET, the fixed charge stored between the gate and the source and between the gate and the drain is modeled by the gate-to-source and the gate-to-drain capacitances, C_{GS} and C_{GD}, respectively. The **mobile charges** between the drain and the source are modeled by the drain-to-source capacitance, C_{DS}. The voltage drop through the active channel is modeled by the drain-to-source resistance, R_{DS}. The current through the channel is modeled by a voltage-controlled current source. For large signals, the gate is sometimes driven into the forward region, and thus the conductance through the gate is modeled by the gate conductance, G_g. The conductance from the gate to the drain and from the gate to the source is modeled by the gate-to-drain and gate-to-source resistances, R_{GD} and R_{GS}, respectively. A variable resistor, R_i, is added to model the gate charging time such that the time constant given by R_iC_{GS} holds the following relationship

$$R_iC_{GS} = \text{constant} \tag{22.70}$$

For MOSFETs, typical element values are: C_{GS} and C_{GD} are in the range of 1–10 pF, C_{DS} is in the range of 0.1–1 pF, R_{DS} is in the range of 1–50 kΩ, R_{GD} is more than 10^{14} Ω, R_{GS} is more than 10^{10} Ω, and g_m is in the range of 0.1–20 mA/V.

Figure 22.25 illustrates the high-frequency large-signal equivalent model of a FET.

Low-Frequency Large-Signal Equivalent Circuit Model of a BJT. In this case, low frequency denotes frequencies below 10 kHz. The low-frequency large-signal equivalent circuit model of a BJT is based on its dc characteristics. Whereas at high frequencies one has to take incremental values to obtain accurate analysis, at low frequencies, the average of these incremental values yields the same level of accuracy in the analysis. Therefore, in low-frequency analysis, the circuit elements of the high-frequency model are replaced by their average values. The low-frequency large-signal equivalent circuit model is shown in Fig. 22.26.

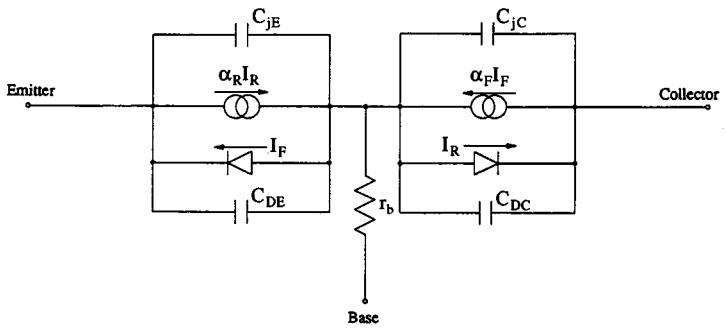


FIGURE 22.26 Low-frequency large-signal equivalent circuit model of an *npn* BJT.

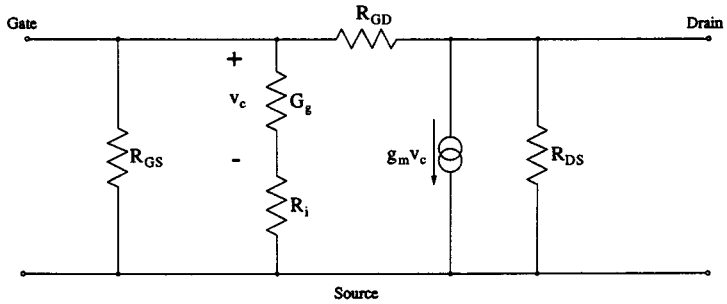


FIGURE 22.27 Low-frequency large-signal equivalent circuit model of a FET.

Low-Frequency Large-Signal Equivalent Circuit Model of a FET. Because of their high reactance values, the gate-to-source, gate-to-drain, and drain-to-source capacitances can be assumed to be open circuits at low frequencies. Therefore, the low-frequency large-signal model is similar to the high-frequency large-signal model, except that it has no capacitances. The resulting circuit describing low-frequency operation is shown in Fig. 22.27.

Small-Signal Equivalent Circuit Model

In a small-signal equivalent circuit model, the signal variations around the dc-bias operating point are very small. Just as for the large-signal model, there are two types of small-signal models, depending upon the operating frequency: (1) the high-frequency small-signal equivalent circuit model and (2) the low-frequency small-signal equivalent circuit model.

High-Frequency Small-Signal Equivalent Circuit Model of a BJT. The high-frequency small-signal equivalent circuit model of a BJT is quite similar to its high-frequency large-signal equivalent circuit model. In the small-signal model, however, in addition to the base resistance r_b, the emitter and collector resistances, r_e and r_c, respectively, are added to the circuit. The emitter resistance is usually very small because of high emitter doping used to obtain better emitter injection efficiency. Therefore, whereas at large signal strengths the effect of r_e is overshadowed by the base resistance, at small signal strengths this emitter resistance cannot be neglected. The collector resistance becomes important in the linear region, where the collector-emitter voltage is low. The high-frequency small-signal equivalent circuit model is shown in Fig. 22.28.

High-Frequency Small-Signal Equivalent Circuit Model of a FET. For small-signal operations, the signal strength is not large enough to forward bias the gate-to-semiconductor diode; hence, no current will flow from the gate to either the drain or the source. Therefore, the gate-to-source and gate-to-drain series resistances, R_{GS} and R_{GD}, can be neglected. Also, since there will be no current flow from the gate to the channel, the gate conductance, G_g, can also be neglected. Figure 22.29 illustrates the high-frequency small-signal equivalent circuit model of a FET.

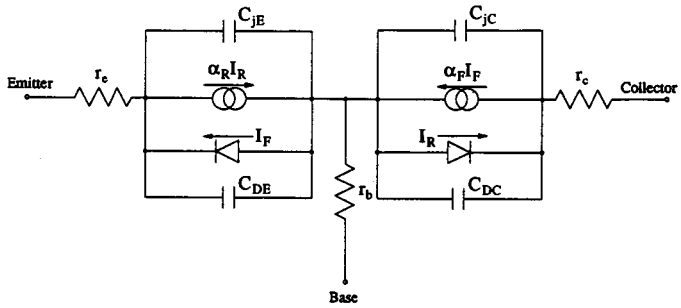


FIGURE 22.28 High-frequency small-signal equivalent circuit model of an *npn* BJT.

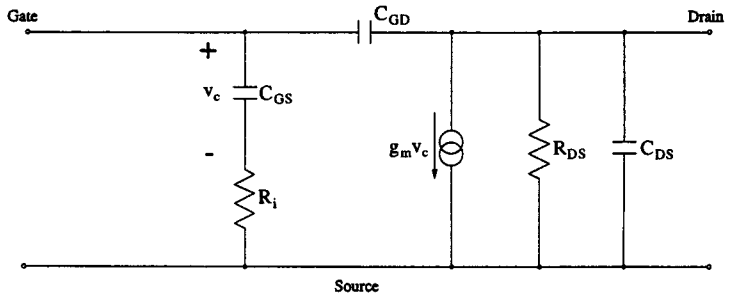


FIGURE 22.29 High-frequency small-signal equivalent circuit model of a FET.

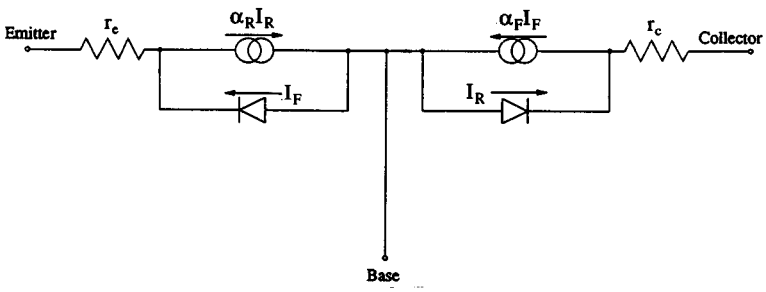


FIGURE 22.30 Low-frequency small-signal equivalent circuit model of an *npn* BJT.

Low-Frequency Small-Signal Equivalent Circuit Model of a BJT. As in the low-frequency large-signal model, the junction capacitances, C_{jC} and C_{jE}, and the diffusion capacitances, C_{DE} and C_{DC}, can be neglected. Furthermore, the base resistance, r_b, can also be neglected, because the voltage drop across the base is not significant and the variations in the base width caused by changes in the collector-base voltage are also very small. The low-frequency small-signal equivalent circuit model is shown in Fig. 22.30.

Low-Frequency Small-Signal Equivalent Circuit Model of a FET. Because the reactances associated with all the capacitances are very high, one can neglect the capacitances for low-frequency analysis. The gate conductance as well as the gate-to-source and gate-to-drain resistances can also be neglected in small-signal operation. The resulting low-frequency equivalent circuit model of a FET is shown in Fig. 22.31.

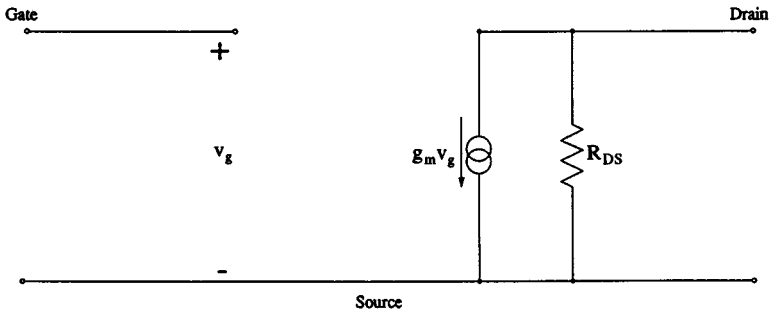


FIGURE 22.31 Low-frequency small-signal equivalent circuit model of a FET.

DC Equivalent Circuit Model

DC Equivalent Circuit Model of a BJT. The dc equivalent circuit model of a BJT is based on the original Ebers–Moll model. Such models are used when the transistor is operated at dc or in applications where the operating frequency is below 1 kHz.

There are two versions of the dc equivalent circuit model—the *injection version* and the *transport version*. The difference between the two versions lies in the choice of the reference current. In the *injection version*, the reference currents are I_F and I_R, the forward- and reverse-biased diode currents, respectively. In the *transport version*, the reference currents are the collector transport current, I_{CC}, and the emitter transport current, I_{CE}. These currents are of the form:

$$I_F = I_{ES}\left[\exp\left(\frac{qV_{BE}}{kT}\right) - 1\right] \tag{22.71}$$

$$I_R = I_{CS}\left[\exp\left(\frac{qV_{BC}}{kT}\right) - 1\right] \tag{22.72}$$

$$I_{CC} = I_S\left[\exp\left(\frac{qV_{BE}}{kT}\right) - 1\right] \tag{22.73}$$

and

$$I_{EC} = I_S\left[\exp\left(\frac{qV_{BC}}{kT}\right) - 1\right] \tag{22.74}$$

In these equations, I_{ES} and I_{CS} are the base-emitter saturation current and the base-collector saturation current, respectively. I_S denotes the saturation current.

In most computer simulations, the *transport version* is usually preferred because of the following conditions:

1. I_{CC} and I_{EC} are ideal over many decades.
2. I_S can specify both reference currents at any given voltage.

The dc equivalent circuit model of a BJT is shown in Fig. 22.32.

DC Equivalent Circuit Model of a FET. In the dc equivalent circuit model of a FET, the gate is considered isolated because the gate-semiconductor interface is formed as a reverse-biased diode and therefore is open circuited. All capacitances are also assumed to represent open circuits. R_{GS}, R_{GD}, and R_{DS} are neglected because

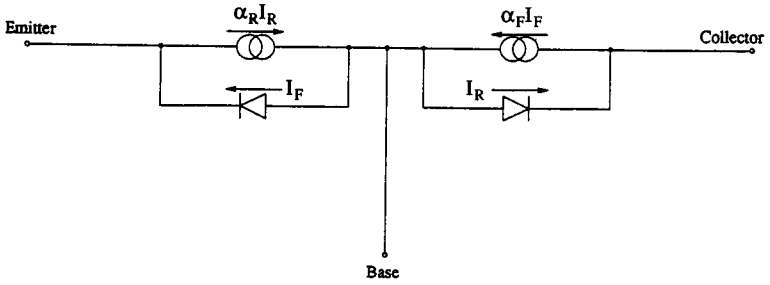


FIGURE 22.32 DC equivalent circuit model (injection version) of an *npn* BJT.

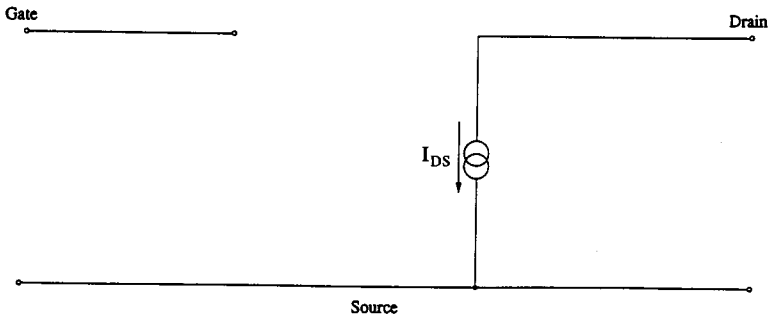


FIGURE 22.33 DC equivalent circuit model of a FET.

there is no conductance through the gate and, because this is a dc analysis, there are no charging effects associated with the gate. The dc equivalent circuit of a FET is illustrated in Fig. 22.33.

Commercially Available Packages

A number of circuit analysis software packages are commercially available, one of the most widely used being SPICE. In this package, the BJT models are a combination of the Gummel–Poon and the modified Ebers–Moll models. Figure 22.34 shows a common emitter transistor circuit and a SPICE input file containing the transistor model. Some other available packages are SLIC, SINC, SITCAP, and Saber.

Equivalent circuit models are basically used to replace the semiconductor device in an electronic circuit. These models are developed from an understanding of the device's current and voltage behavior for novel devices where internal device operation is not well understood. For such situations, the designer has another tool available, the semiconductor device simulator.

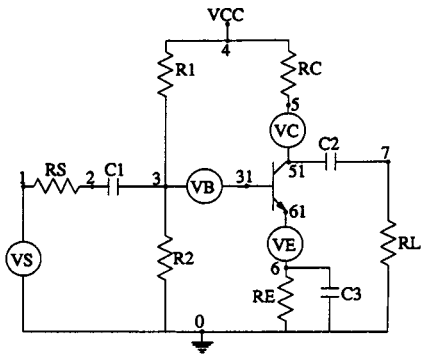
Overview of Semiconductor Device Simulators

Device simulators are based on the physics of semiconductor devices. The input to the simulator takes the form of information about the device under consideration such as material type, device, dimensions, doping concentrations, and operating conditions. Based on this information, the device simulator computes the electric field inside the device and thus predicts carrier concentrations in the different regions of the device. Device simulators can also predict transient behavior, including quantities such as current–voltage characteristics and frequency bandwidth. The three basic approaches to device simulation are (1) the classical approach, (2) the semiclassical approach, and (3) the quantum mechanical approach.

Device Simulators Based on the Classical Approach

The *classical approach* is based on the solution of Poisson's equation and the current continuity equations. The current consists of the drift and the diffusion current components.

```
VS 1 0 DC 0.0 AC 75e-3 sin(0 75m 10k); sin(offset, peak amp, freq, delay, damping, phase)
VCC 4 0 12.0
*
*Resistor elements
*
Rs 1 2 1 Kohm
R1 4 3 225K
R2 3 0 47K
RC 4 5 5.1K
RE 6 0 1K
RL 7 0 2K
*
*Capacitor elements
*
C1 2 3 3.3UFd
C2 51 7 3.3UF
C3 6 0 47UF
*
*Voltage Sources for Current measurements
*
VB 3 31 0.0
VC 5 51 0.0
VE 61 6 0.0
*
*Transistor (Collector-Base-Emitter)
*
Q1 51 31 61 MQNOM
.model MQNOM NPN (BF = 130 CJE = 25pF CJC = 8pF)
*
.AC dec 25 10 10MEG; Freq Variation from 10 Hz- 10 MHz
.tran 1u 200u 100u; Step size, duration, start point
.Probe
.End
```



Transistor model

FIGURE 22.34 Common emitter transistor circuit and SPICE circuit file.

Assumptions. The equations for the classical approach can be obtained by making the following approximations to the Boltzmann transport equation:

1. Carrier temperature is the same throughout the device and is assumed to be equal to the lattice temperature.
2. Quasi steady-state conditions exist.
3. Carrier **mean free path** must be smaller than the distance over which the **quasi-Fermi level** is changing by kT/q.
4. The impurity concentration is constant or varies very slowly along the mean free path of the carrier.
5. The energy band is parabolic.
6. The influence of the boundary conditions is negligible.

For general purposes, even with these assumptions and limitations, the models based on the classical approach give fairly accurate results. The model assumes that the driving force for the carriers is the quasi-Fermi potential gradient, which is also dependent upon the electric field value. Therefore, in some simulators, the quasi-Fermi level distributions are computed and the carrier distribution is estimated from this information.

Equations to Be Solved. With the assumption of a quasi-steady-state condition, the operating wavelength is much larger than the device dimensions. Hence, Maxwell's equations can be reduced to the more familiar Poisson's equation:

$$\nabla^2 \psi = -\frac{\rho}{\epsilon} \tag{22.75}$$

and, for a nonhomogeneous medium,

$$\nabla \cdot \epsilon \, (\nabla \psi) = -\rho \tag{22.76}$$

where ψ denotes the potential of the region under simulation, ϵ denotes the permittivity, and ρ denotes the charge enclosed by this region.

Also from Maxwell's equations, one can determine the current continuity equations for a homogeneous medium as:

$$\nabla \cdot J_n - q\left(\frac{\partial n}{\partial t}\right) = +qU \tag{22.77}$$

where

$$J_n = q\mu_n E + qD_n \nabla \cdot n \tag{22.78}$$

and

$$\nabla \cdot J_p + q\left(\frac{\partial p}{\partial t}\right) = -qU \tag{22.79}$$

where

$$J_p = q\mu_p pE - qD_p \nabla \cdot p \tag{22.80}$$

For nonhomogeneous media, the electric field term in the current expressions is modified to account for the nonuniform **density of states** and the bandgap variation [Lundstrom and Schuelke, 1983].

In the classical approach, the objective is to calculate the potential and the carrier distribution inside the device. Poisson's equation is solved to yield the potential distribution inside the device from which the electric field can be approximated. The electric field distribution is then used in the current continuity equations to obtain the carrier distribution and the current densities. The diffusion coefficients and carrier mobilities are usually field as well as spatially dependent.

The generation-recombination term U is usually specified by the Shockley–Read–Hall relationship [Yoshi et al., 1982]:

$$Rn = \frac{p\,n - n_{ie}^2}{\tau_p(n + n_t) + \tau_n(p + p_t)} \tag{22.81}$$

where p and n are the hole and electron concentrations, respectively, n_{ie} is the effective intrinsic carrier density, τ_p and τ_n are the hole and electron lifetimes, and p_t and n_t are the hole and electron trap densities, respectively.

The electron and hole mobilities are usually specified by the Scharfetter–Gummel empirical formula, as

$$\mu = \mu_0\left[1 + \frac{N}{(N/a) + b} + \frac{(E/c)^2}{(E/c) + d} + (E/e)^2\right]^{-1/2} \tag{22.82}$$

where N is the total ionized impurity concentration, E is the electric field, and a, b, c, d, and e are defined constants [Scharfetter and Gummel, 1969] that have different values for electrons and holes.

Boundary Conditions. Boundary conditions have a large effect on the final solution, and their specific choice is a very important issue. For ohmic contacts, infinite recombination velocities and space charge neutrality conditions are assumed. Therefore, for a p-type material, the ohmic boundary conditions take the form

$$\psi = V_{\text{appl}} + \frac{kT}{q} \ln\left(\frac{n_{ie}}{p}\right) \tag{22.83}$$

$$p = \left[\left(\frac{N_D^+ - N_A^-}{2}\right)^2 + n_{ie}^2\right]^{1/2} - \left(\frac{N_D^+ - N_A^-}{2}\right) \tag{22.84}$$

and

$$n = \frac{n_{ie}^2}{p} \tag{22.85}$$

where V_{appl} is the applied voltage, k is Boltzmann's constant, and N_D^+ and N_A^- are the donor and acceptor ionized impurity concentrations, respectively.

For **Schottky contacts**, the boundary conditions take the form

$$\psi = V_{\text{appl}} + \frac{E_G}{2} - \phi_B \tag{22.86}$$

and

$$n = n_{ie} \exp\left(\frac{(E_G/2) - \phi_B}{kT/q}\right) \tag{22.87}$$

where E_G is the semiconductor bandgap and ϕ_B is the barrier potential. For other boundaries with no current flow across them, the boundary conditions are of the form

$$\frac{\partial \psi}{\partial n} = \frac{\partial \varphi_n}{\partial n} = \frac{\partial \varphi_p}{\partial p} = 0 \tag{22.88}$$

where φ_n and φ_p are the electron and hole quasi-Fermi levels, respectively.

For field-effect devices, the potential under the gate may be obtained either by setting the gradient of the potential near the semiconductor-oxide interface equal to the gradient of potential inside the oxide [Kasai et al., 1982], or by solving Laplace's equation in the oxide layer, or by assuming a Dirichlet boundary condition at the oxide-gate interface and determining the potential at the semiconductor-oxide interface as:

$$\epsilon_{\text{Si}} \left.\frac{\partial \psi}{\partial y}\right|_{\text{Si}} = \epsilon_{\text{Ox}} \frac{\psi_G - \psi_S^\star(x,z)}{T(z)} \tag{22.89}$$

where ϵ_{Si} and ϵ_{Ox} are the permittivities of silicon and the oxide, respectively, ψ_G is the potential at the top of the gate, $\psi_S^*(x,z)$ is the potential of the gate near the interface, and T(z) is the thickness of the gate metal.

Solution Methods. Two of the most popular methods of solving the above equations are finite difference method (FDM) and finite element method (FEM).

In FDM, the region under simulation is divided into rectangular or triangular areas for two-dimensional cases or into cubic or tetrahedron volumes in three-dimensional cases. Each corner or vertex is considered as a node. The differential equations are modified using finite difference approximations, and a set of equations is constructed in matrix form. The finite difference equations are solved iteratively at only these nodes. The most commonly used solvers are Gauss–Seidel/Jacobi (G-S/J) techniques or Newton's technique (NT) [Banks et al., 1983]. FDM has the disadvantage of requiring more nodes than the FEM for the same structure. A new variation of FDM, namely the finite boxes scheme [Franz et al., 1983], however, overcomes this problem by enabling local area refinement. The advantage of FDM is that its computational memory requirement is less than that required for FEM because of the band structure of the matrix.

In FEM, the region under simulation is divided into triangular and quadrilateral regions in two dimensions or into tetrahedra in three dimensions. The regions are placed to have the maximum number of vertices in areas where there is expected to be a large variation of composition or a large variation in the solution. The equations in FEM are modified by multiplying them with some shape function and integrating over the simulated region. In triangular meshes, the shape function is dependent on the area of the triangle and the spatial location of the node. The value of the spatial function is between 0 and 1. The solution at one node is the sum of all the solutions, resulting from the nearby nodes, multiplied by their respective shape functions. The number of nodes required to simulate a region is less than that in FDM; however, the memory requirement is greater.

Device Simulators Based on the Semiclassical Approach

The *semiclassical* approach is based upon the Boltzmann transport equation (BTE) [Engl, 1986] which can be written as:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + v \cdot \nabla_r \pm \frac{q}{(h/2\pi)} E \cdot \nabla_k f = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \tag{22.90}$$

where f represents the carrier distribution in the volume under consideration at any time t, v is the group velocity, E is the electric field, and q and h are the electronic charge and Planck's constant, respectively.

BTE is a simplified form of the Liouville–Von Neumann equation for the density matrix. In this approach, the free flight between two consecutive collisions of the carrier is considered to be under the influence of the electric field, whereas different scattering mechanisms determine how and when the carrier will undergo a collision.

Assumptions. The assumptions for the semiclassical model can be summarized as follows:

1. Carrier-to-carrier interactions are considered to be very weak.
2. Particles cannot gain energy from the electric field during collision.
3. Scattering probability is independent of the electric field.
4. Magnetic field effects are neglected.
5. No electron-to-electron interaction occurs in the collision term.
6. Electric field varies very slowly, i.e., electric field is considered constant for a wave packet describing the particle's motion.
7. The electron and hole gas is not degenerate.
8. Band theory and effective-mass theorems apply to the semiconductor.

Equations to Be Solved. As a starting point, Poisson's equation is solved to obtain the electric field inside the device. Using the Monte Carlo technique (MCT), the BTE is solved to obtain the carrier distribution function, f. In the MCT, the path of one or more carriers, under the influence of external forces, is followed, and from

this information the carrier distribution function is determined. BTE can also be solved by the momentum and energy balance equations.

The carrier distribution function gives the carrier concentrations in the different regions of the device and can also be used to obtain the electron and hole currents, using the following expressions:

$$J_n = -q\int_k v f(r,k,t)d^3k \tag{22.91}$$

and

$$J_p = +q\int_k v f(r,k,t)d^3k \tag{22.92}$$

Device Simulators Based on the Quantum Mechanical Approach

The *quantum mechanical approach* is based on the solution of the Schrodinger wave equation (SWE), which, in its time-independent form, can be represented as

$$\frac{(h/2\pi)^2}{2m}\nabla^2\varphi_n + (E_n + qV)\ \varphi_n = 0 \tag{22.93}$$

where φ_n is the wave function corresponding to the subband n whose minimum energy is E_n, V is the potential of the region, m is the particle mass, and h and q are Planck's constant and the electronic charge, respectively.

Equations to Be Solved. In this approach, the potential distribution inside the device is calculated using Poisson's equation. This potential distribution is then used in the SWE to yield the electron wave vector, which in turn is used to calculate the carrier distribution, using the following expression:

$$n = \sum_n N_n \left|\varphi_n\right|^2 \tag{22.94}$$

where n is the electron concentration and N_n is the concentration of the subband n.

This carrier concentration is again used in Poisson's equation, and new values of φ_n, E_n, and n are calculated. This process is repeated until a self-consistent solution is obtained. The final wave vector is invoked to determine the scattering matrix, after which MCT is used to yield the carrier distribution and current densities.

Commercially Available Device Simulation Packages

The classical approach is the most commonly used procedure since it is the easiest to implement and, in most cases, the fastest technique. Simulators based on the classical approach are available in two-dimensional forms like FEDAS, HESPER, PISCES-II, PISCES-2B, MINIMOS, and BAMBI or three-dimensional forms like TRANAL, SIERRA, FIELDAY, DAVINCI, and CADDETH.

Large-dimension devices, where the carriers travel far from the boundaries, can be simulated based on a one-dimensional approach. Most currently used devices, however, do not fit into this category, and therefore one has to resort to either two- or three-dimensional simulators.

FEDAS (Field Effect Device Analysis System) is a two-dimensional device simulator that simulates MOSFETs, JFETs, and MESFETs by considering only those carriers that form the channel. The Poisson equation is solved everywhere except in the oxide region. Instead of carrying the potential calculation within the oxide region, the potential at the semiconductor-oxide interface is calculated by assuming a mixed boundary condition. FEDAS uses FDM to solve the set of linear equations. A three-dimensional variation of FEDAS is available for the simulation of small geometry MOSFETs.

HESPER (HEterostructure device Simulation Program to Estimate the performance Rigorously) is a two-dimensional device simulator that can be used to simulate heterostructure photodiodes, HBTs, and HEMTs. The simulation starts with the solution of Poisson's equation in which the electron and hole concentrations are described as functions of the composition (composition dependent). The recombination rate is given by the Shockley–Read–Hall relationship. Lifetimes of both types of carriers are assumed to be equal in this model.

PISCES-2B is a two-dimensional device simulator for simulation of diodes, BJTs, MOSFETs, JFETs, and MESFETs. Besides steady-state analysis, transient and ac small-signal analysis can also be performed.

Conclusion

The decision to use an equivalent circuit model or a device simulator depends upon the designer and the required accuracy of prediction. To save computational time, one should use as simple a model as accuracy will allow. At this time, however, the trend is toward developing quantum mechanical models that are more accurate, and with faster computers available, the computational time for these simulators has been considerably reduced.

Defining Terms

Density of states: The total number of charged carrier states per unit volume.
Fermi levels: The energy level at which there is a 50% probability of finding a charged carrier.
Mean free path: The distance traveled by the charged carrier between two collisions.
Mobile charge: The charge due to the free electrons and holes.
Quasi-Fermi levels: Energy levels that specify the carrier concentration inside a semiconductor under non-equilibrium conditions.
Schottky contact: A metal-to-semiconductor contact where, in order to align the Fermi levels on both sides of the junction, the energy band forms a barrier in the majority carrier path.

Related Topics

2.3 Controlled Sources • 35.1 Maxwell Equations

References

R. E. Banks, D. J. Rose, and W. Fitchner, "Numerical methods for semiconductor device simulation," *IEEE Trans. Electron Devices,* vol. ED-30, no. 9, pp. 1031–1041, 1983.

J. J. Ebers and J. L. Moll, "Large signal behavior of junction transistors," *Proc. IRE*, vol. 42, pp. 1761–1772, Dec. 1954.

W. L. Engl, *Process and Device Modeling*, Amsterdam: North-Holland, 1986.

A. F. Franz, G. A. Franz, S. Selberherr, C. Ringhofer, and P. Markowich, "Finite boxes—A generalization of the finite-difference method suitable for semiconductor device simulation," *IEEE Trans. Electron Devices*, vol. ED-30, no. 9, pp. 1070–1082, 1983.

H. K. Gummel and H. C. Poon, "An integral charge control model of bipolar transistors," *Bell Syst. Tech. J.*, vol. 49, pp. 827–852, May-June 1970.

R. Kasai, K. Yokoyama, A. Yoshii, and T. Sudo, "Threshold-voltage analysis of short- and narrow-channel MOSFETs by three-dimensional computer simulation," *IEEE Trans. Electron Devices,* vol. ED-21, no. 5, pp. 870–876, 1982.

M. S. Lundstrom and R. J. Schuelke, "Numerical analysis of heterostructure semiconductor devices," *IEEE Trans. Electron Devices*, vol. ED-30, no. 9, pp. 1151–1159, 1983.

D. L. Scharfetter and H. K. Gummel, "Large-signal analysis of a silicon read diode oscillator," *IEEE Trans. Electron Devices*, vol. ED-16, no. 1, pp. 64–77, 1969.

A. Yoshii, H. Kitazawa, M. Tomzawa, S. Horiguchi, and T. Sudo, "A three dimensional analysis of semiconductor devices," *IEEE Trans. Electron Devices*, vol. ED-29, no. 2, pp. 184–189, 1982.

Further Information

Further information about semiconductor device simulation and equivalent circuit modeling, as well as about the different software packages available, can be found in the following articles and books:

C. M. Snowden, *Semiconductor Device Modeling*, London: Peter Peregrinus Ltd., 1988.
C. M. Snowden, *Introduction to Semiconductor Device Modeling*, Teaneck, N.J.: World Scientific, 1986.
W. L. Engl, *Process and Device Modeling*, Amsterdam: North-Holland, 1986.
J.-H. Chern, J. T. Maeda, L. A. Arledge, Jr., and P. Yang, "SIERRA: A 3-D device simulator for reliability modeling," *IEEE Trans. Computer-Aided Design*, vol. CAD-8, no. 5, pp. 516–527, 1989.
T. Toyabe, H. Masuda, Y. Aoki, H. Shukuri, and T. Hagiwara, "Three-dimensional device simulator CADDETH with highly convergent matrix solution algorithms," *IEEE Trans. Electron Devices*, vol. ED-32, no. 10, pp. 2038–2044, 1985.

PISCES-2B and DAVINCI are softwares developed by TMA Inc., Palo Alto, California 94301.

Hewlett-Packard's first product, the model 200A audio oscillator (preproduction version). William Hewlett and David Packard built an audio oscillator in 1938, from which the famous firm grew. Courtesy of Hewlett-Packard Company.)

22.4 Electrical Characterization of Semiconductors

David C. Look

The huge electronics and computer industries exist primarily because of the unique electrical properties of semiconductor materials, such as Si and GaAs. These materials usually contain impurities and defects in their crystal lattices; such entities can act as **donors** and **acceptors**, and can strongly influence the electrical and optical properties of the charge carriers. Thus, it is extremely important to be able to measure the concentration and **mobility** of these carriers, and the concentrations and energies of the donors and acceptors. All of these quantities can, in principle, be determined by measurement and analysis of the temperature-dependent resistivity and Hall effect. On the simplest level, Hall-effect measurements require only a current source, a voltmeter, and a modest-sized magnet. However, the addition of temperature-control equipment and computer analysis produce a much more powerful instrument that can accurately measure concentrations over a range 10^4 to 10^{20} cm^{-3}. Many commercial instruments are available for such measurements; this chapter section reveals how to make full use of the versatility of the technique.

Theory

A phenomenological equation of motion for electrons of charge –e moving with velocity **v** in the presence of electric field **E** and magnetic field **B** is

$$m^*\dot{\mathbf{v}} = -e\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right) - m^* \frac{\mathbf{v} - \mathbf{v}_{eq}}{\tau} \tag{22.95}$$

where m^* is the **effective mass**, $\mathbf{v}_{eq}$ is the velocity at equilibrium (steady state), and τ is the velocity (or momentum) **relaxation time** (i.e., the time in which oscillatory phase information is lost through collisions). Consider a rectangular sample, as shown in Fig. 22.35(a), with an external electric field $\mathbf{E}_{ex} = E_x\mathbf{x}$ and magnetic field $\mathbf{B} = B_z\mathbf{z}$. (Dimensions x and y are parallel to "ℓ" and "w," respectively, and z is perpendicular to both.) Then, if no current is allowed to flow in the y direction (i.e., $v_y = 0$), the steady-state condition $\dot{\mathbf{v}} = 0$ requires that $E_y = -v_xB_z$, and E_y is known as the Hall field. For electron concentration n, the current density is $j_x = nev_x$; thus, $E_y = -j_xB_z/en \equiv -j_xB_zR_H$, where $R_H = -1/en$, the **Hall coefficient**. Thus, simple measurements of the quantities E_y, j_x, and B_z yield a very important quantity: n.

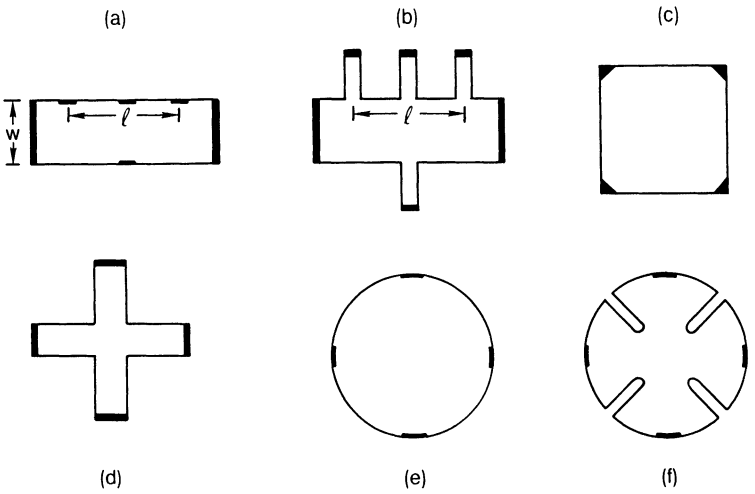


FIGURE 22.35 Various patterns commonly used for resistivity and Hall-effect measurements.

The above analysis assumes that all electrons are moving with the same velocity v (constant τ), which is not true in a semiconductor. A more detailed analysis, allowing for the energy $\mathcal{E}$ dependence of the electrons, gives

$$j_x = \frac{ne^2\langle\tau\rangle}{m^*}E_x \equiv -ne\mu_c E_x \tag{22.96}$$

$$R_H = \frac{E_y}{j_x B} = -\frac{1}{ne}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} = -\frac{r}{en} \tag{22.97}$$

where

$$\langle\tau^n(\mathcal{E})\rangle = \frac{\int_0^\infty \tau^n(\mathcal{E})\mathcal{E}^{3/2}\frac{\partial f_0}{\partial\mathcal{E}}d\mathcal{E}}{\int_0^\infty \mathcal{E}^{3/2}\frac{\partial f_0}{\partial\mathcal{E}}d\mathcal{E}} \rightarrow \frac{\int_0^\infty \tau^n(\mathcal{E})\mathcal{E}^{3/2}e^{-\mathcal{E}/kT}d\mathcal{E}}{\int_0^\infty \mathcal{E}^{3/2}e^{-\mathcal{E}/kT}d\mathcal{E}} \tag{22.98}$$

This formulation is called the **relaxation-time approximation (RTA)** to the Boltzmann transport equation (BTE). Here, f_0 is the Fermi-Dirac **distribution function** and the second equality in Eq. (22.98) holds for non-degenerate electrons (i.e., those describable by Boltzmann statistics). The quantity $\mu_c = e\langle\tau\rangle/m^*$ is known as the conductivity **mobility,** since the quantity $ne\mu$ is just the conductivity σ. The **Hall mobility** is defined as $\mu_H = R_H\sigma = r\mu_c$, and the Hall concentration as $n_H = n/r = -1/eR_H$. Thus, a combined Hall effect and conductivity measurement gives n_H and μ_H, although one would prefer to know n, not n_H; fortunately, however, r is usually within 20% of unity, and is almost never as large as 2. In any case, r can often be calculated or measured so that an accurate value of n can usually be determined. It should also be mentioned that one way to evaluate the expressions in Eq. (22.98) is to define a new variable, $u = \mathcal{E}/kT$, and set u = 10 as the upper limit in the integrals.

The relaxation time, $\tau(\mathcal{E})$, depends on how the electrons interact with the **lattice vibrations,** as well as with extrinsic elements such as charged impurities and defects. For example, acoustical-mode lattice vibrations scatter electrons through the deformation potential (τ_{ac}) and piezoelectric potential (τ_{pe}); optical-mode vibrations through the polar potential (τ_{po}); ionized impurities and defects through the screened coulomb potential

(τ_{ii}); and charged **dislocations,** also through the coulomb potential (τ_{dis}). The strengths of these various scattering mechanisms depend on certain lattice parameters, such as dielectric constants and deformation potentials, and extrinsic factors, such as **donor**, **acceptor,** and dislocation concentrations, N_D, N_A, and N_{dis}, respectively [Rode, 1975; Wiley, 1975; Nag, 1980; Look, 1989; Look, 1998]. The total momentum scattering rate, or inverse **relaxation time,** is

$$\tau^{-1}(\mathcal{E}) = \tau_{ac}^{-1}(\mathcal{E}) + \tau_{pe}^{-1}(\mathcal{E}) + \tau_{po}^{-1}(\mathcal{E}) + \tau_{ii}^{-1}(\mathcal{E}) + \tau_{dis}^{-1}(\mathcal{E}) \tag{22.99}$$

and this expression is then used to determine $\langle\tau^n(\mathcal{E})\rangle$ via Eq. (22.98), and hence, $\mu_H = e\langle\tau^2\rangle/m^*\langle\tau\rangle$. Formulae for τ_{ac}, τ_{pe}, τ_{po}, τ_{ii}, and τ_{dis}, can be found in the literature, but are given below for completeness. For ionized impurity (or defect) scattering, in a non-degenerate, ***n*-type** material:

$$\tau_{ii}(\mathcal{E}) = \frac{2^{9/2}\pi\varepsilon_0^{\,2}(m^*)^{1/2}\mathcal{E}^{3/2}}{e^4(2N_A + n)\left[\ln(1+y) - y/(1+y)\right]} \tag{22.100}$$

where $y = 8\varepsilon_0 m^* kT\mathcal{E}/\hbar^2 e^2 n$. Here, ε_0 is the low-frequency (static) dielectric constant, k is Boltzmann's constant, and $\hbar$ is Planck's constant divided by 2π. [If the sample is ***p*-type**, let $(2N_A+n) \rightarrow (2N_D+p)$]. For acoustic-mode deformation-potential scattering:

$$\tau_{ac}(\mathcal{E}) = \frac{\pi\hbar^4\rho_d s^2 \mathcal{E}^{-1/2}}{2^{1/2} E_1^{\,2}(m^*)^{3/2} kT} \tag{22.101}$$

where ρ_d is the density, s is the speed of sound, and E_1 is the deformation potential. For acoustic-mode piezoelectric-potential scattering:

$$\tau_{p_e}(\mathcal{E}) = \frac{2^{3/2}\pi\hbar^2\varepsilon_0\mathcal{E}^{1/2}}{e^2P^2(m^*)^{1/2}kT} \tag{22.102}$$

where P is the piezoelectric coupling coefficient [$P = (h_{pz}^2/\rho s^2\varepsilon_0)^{1/2}$]. Finally, for polar optic-mode scattering, only a rough approximation can be given because the scattering is inelastic:

$$\tau_{po}(\mathcal{E}) = \frac{2^{3/2}\pi\hbar^2\left(e^{T_{po}/T} - 1\right)\left[0.762\mathcal{E}^{1/2} + 0.824(kT_{po})^{1/2} - 0.235(kT_{po})^{-1/2}\mathcal{E}\right]}{e^2 kT_{po}(m^*)^{1/2}\left(\varepsilon_\infty^{-1} - \varepsilon_0^{-1}\right)} \tag{22.103}$$

where T_{po} is the Debye temperature and ε_∞ is the high-frequency dielectric constant. This formula for $\tau_{po}(\mathcal{E})$ has the following property: if *only* p-o scattering existed, then an accurate BTE calculation of μ_H vs. T [Rode, 1975] would give results almost identical to those obtained by the RTA analysis described above, i.e., by setting $\mu_H = e\langle\tau^2\rangle/m^*\langle\tau\rangle$. However, when other scattering mechanisms are also important, then the RTA solution may not be as reliable. Fortunately, at low temperatures (e.g., below about 150K in GaN), p-o scattering weakens, and the RTA approach is quite accurate. This fact is important because we usually are interested in obtaining a good value of the acceptor concentration N_A from the μ_H vs. T fit, and N_A appears directly only in the ii-scattering formula Eq. (22.100), which is usually dominant at low temperatures.

Dislocation scattering in semiconductor materials is often ignored because it becomes significant only for dislocation densities $N_{dis} > 10^8$ cm^{-2} (note that this is an arreal, not volume, density). Such high densities are rare in most semiconductor devices, such as those fabricated from Si or GaAs, but are indeed quite common in devices based on GaN or other materials that involve mismatched substrates. In GaN grown on Al_2O_3 (sapphire), vertical threading dislocations, typically of concentration 10^{10} cm^{-2} or higher, emanate from the interface up to the surface, and horizontally moving electrons or holes experience a scattering characterized by

$$\tau_{dis}\left(\mathcal{E}\right) = \frac{\hbar^3 \varepsilon_0^{\,2} c^2}{N_{dis} m^* e^4} \frac{\left(1 + 8m^* \lambda^2 \mathcal{E}\right)^{3/2}}{\lambda^4} \tag{22.104}$$

where $\lambda = (\varepsilon_0 kT/e^2 n)^{1/2}$. For high-quality GaN/$Al_2O_3$, $N_{dis} \approx 10^8$ cm^{-2}; in the case of a sample discussed later in this chapter section, this value of N_{dis} drops the 300-K Hall mobility only a minor amount, from 915 to 885 cm^2/V s. However, if this same sample contained the usual concentration of dislocations found in GaN (about 10^{10} cm^{-2}), the mobility would drop to less than 100 cm^2/V s, a typical value found in many other samples.

Before going on, it should be mentioned that a very rough approximation of μ_H, which avoids the integrations of Eq. (22.98), can be obtained by setting $\mathcal{E} \approx kT$ and $\mu \approx e\tau/m^*$ in Eq. (22.99). The latter step (i.e., $\mu^{-1} = \mu_1^{-1} + \mu_2^{-1} + \mu_3^{-1} + \ldots$) is known as **Matthiessen's Rule.** However, with present-day computing power, even that available on PCs, it is not much more difficult to use the RTA analysis.

The fitting of μ_H vs. T data, described above, should be carried out in conjunction with the fitting of n vs. T, which is derived from the **charge-balance equation (CBE)**:

$$n + N_A = \frac{N_D}{1 + n/\phi_D} \tag{22.105}$$

where $\phi_D = (g_0/g_1) N_C' \exp(\alpha_D/k) T^{3/2} \exp(-E_{D0}/kT)$. Here, g_0/g_1 is a degeneracy factor, $N_C' = 2(2\pi m_n^* k)^{3/2}/h^3$, where h is Planck's constant, E_D is the **donor** energy, and E_{D0} and α_D are defined by $E_D = E_{D0} - \alpha_D T$. The above equation describes the simplest type of charge balance, in which the donor (called a single donor) has only one charge-state transition within a few kT of the Fermi energy. An example of such a donor is Si on a Ga site in GaN, for which $g_0 = 1$, and $g_1 = 2$. If there are multiple single donors, then equivalent terms are added on the right-hand side of Eq. (22.105); if there are double or triple donors, or more than one **acceptor,** proper variations of Eq. (22.105) can be found in the literature [Look, 1989].

For a ***p*-type** sample, the nearly equivalent equation is used:

$$p + N_D = \frac{N_A}{1 + p/\phi_A} \tag{22.106}$$

where $\phi_A = (g_1/g_0) N_V' \exp(\alpha_A/k) T^{3/2} \exp(-E_{A0}/kT)$, $N_V' = 2(2\pi m_p^* k)^{3/2}/h^3$, and $E_A = E_{A0} - \alpha_A T$.

Hall samples do not have to be rectangular, and other common shapes are given in Fig. 22.35(c)–(f); in fact, arbitrarily shaped specimens are discussed in the next section. However, the above analysis does assume that n and μ are homogeneous throughout the sample. If n and μ vary with depth (z) only, then the measured quantities are

$$\sigma_{sq} = \int_0^d \sigma(z)\,dz = e\int_0^d n(z)\mu(z)\,dz \tag{22.107}$$

$$R_{Hsq}\sigma_{sq}^{2} = \int_{0}^{d} n(z)\mu^{2}(z)dz \tag{22.108}$$

where d is the sample thickness and where the subscript "sq" denotes a sheet (arreal) quantity (cm^{-2}) rather than a volume quantity (cm^{-3}). If some of the carriers are holes, rather than electrons, then the sign of e for those carriers must be reversed. The general convention is that R_H is negative for electrons and positive for holes. In some cases, the hole and electron contributions to $R_{Hsq}\sigma_{sq}^{2}$ exactly balance at a given temperature, and this quantity vanishes.

Determination of Resistivity and Hall Coefficient

Consider the Hall-bar structure of Fig. 22.35(a) and suppose that current I is flowing along the long direction. Then, if V_c and V_H are the voltages measured along dimensions ℓ and w, respectively, and d is the thickness, one obtains $E_x = V_c/\ell$, $E_y = V_H/w$, $j_x = I/wd$, and

$$\sigma = \rho^{-1} = \frac{j_x}{E_x} = \frac{I\ell}{V_c wd} \tag{22.109}$$

$$R_H = \frac{E_y}{j_x B} = \frac{V_H d}{IB} \tag{22.110}$$

$$\mu_H = R\sigma = \frac{V_H \ell}{V_c wB} \tag{22.111}$$

$$n_H = (eR)^{-1} \tag{22.112}$$

In MKS units, I is in amps (A), V in volts (V), B in Tesla (T), and ℓ, w, and d in meters (m). By realizing that 1 T = 1 V–s m^{-2}, 1 A = 1 coulomb (C)/s, and 1 ohm (Ω) = 1 VA^{-1} then σ is in units of $\Omega^{-1}m^{-1}$, R_H in m^3C^{-1}, μ_H in $m^2V^{-1}s^{-1}$, and n_H in m^{-3}. However, it is more common to denote σ in $\Omega^{-1}cm^{-1}$, R_H in $cm^3\ C^{-1}$, μ_H in $cm^2\ V^{-1}s^{-1}$, and n_H in cm^{-3}, with obvious conversion factors (1 m = 10^2 cm). Because B is often quoted in Gauss (G), it is useful to note that 1 T = 10^4 G.

Clearly, the simple relationships given above will not hold for the nonrectangular shapes shown in Fig. 22.35(c)–(f), several of which are very popular. Fortunately, van der Pauw [1958] has solved the potential problem for a thin layer of arbitrary shape. One of the convenient features of the van der Pauw formulation is that no dimension need be measured for the calculation of sheet resistance or sheet carrier concentration, although a thickness must of course be known for volume resistivity and concentration. Basically, the validity of the van der Pauw method requires that the sample be flat, homogeneous, and isotropic, a singly connected domain (no holes), and have line electrodes on the periphery, projecting to point contacts on the surface, or else have true point contacts on the surface. The last requirement is the most difficult to satisfy, so that much work has gone into determining the effects of finite contact size.

Consider the arbitrarily shaped sample shown in Fig. 22.36(a). Here, a current I flows between contacts 1 and 2, and a voltage V_c is measured between contacts 3 and 4. Let $R_{ij,kl} \equiv V_{kl}/I_{ij}$, where the current enters contact i and leaves contact j, and $V_{kl} = V_k - V_l$. (These definitions, as well as the contact numbering, correspond to ASTM Standard F76.) The resistivity, ρ, with B = 0, is then calculated as follows:

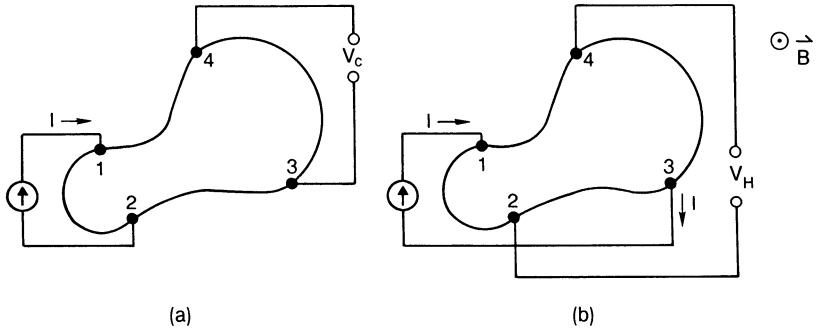


FIGURE 22.36 An arbitrary shape for van der Pauw measurements: (a) resistivity; (b) Hall effect.

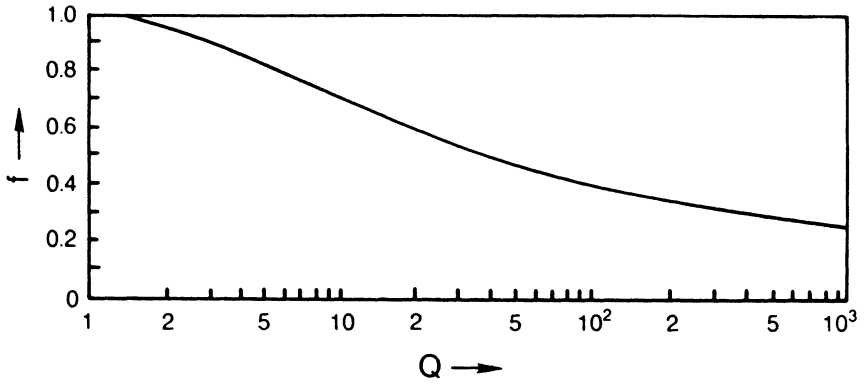


FIGURE 22.37 The resistivity-ratio function used to correct the van der Pauw results for asymmetric sample shape.

$$\rho = \frac{\pi d}{\ln(2)}\left[\frac{R_{21,34} + R_{32,41}}{2}\right] f \tag{22.113}$$

where f is determined from a transcendental equation:

$$\frac{Q-1}{Q+1} = \frac{f}{\ln(2)} \operatorname{arccosh}\left\{\frac{1}{2}\exp\left(\frac{\ln(2)}{f}\right)\right\} \tag{22.114}$$

Here, $Q = R_{21,34}/R_{32,41}$ if this ratio is greater than unity; otherwise, $Q = R_{32,41}/R_{21,34}$. A curve of f vs. Q, accurate to about 2%, is presented in Fig. 22.37 [van der Pauw, 1958]. Also useful is a somewhat simpler analytical procedure for determining f, due to Wasscher and reprinted in Weider [1979]. First, calculate α from

$$Q = \frac{\ln(1/2 - \alpha)}{\ln(1/2 + \alpha)} \tag{22.115}$$

and then calculate f from

$$f = \frac{\ln\left(1/4\right)}{\ln\left(1/2+\alpha\right)+\ln\left(1/2-\alpha\right)} \tag{22.116}$$

It is of course required that $-1/2 < \alpha < 1/2$, but this range of α covers $Q = 0$ to ∞. For example, a ratio $Q = 4.8$ gives a value $\alpha \approx 0.25$, and then $f \approx 0.83$. Thus, the ratio must be fairly large before ρ is appreciably reduced.

It is useful to further average ρ by including the remaining two contact permutations, and also reversing current for all four permutations. Then ρ becomes

$$\rho = \left[\pi d/\ln(2)\right]\left[\left(R_{21,34} - R_{12,34} + R_{32,41} - R_{23,41}\right)f_A + \left(R_{43,12} - R_{34,12} + R_{14,23} - R_{41,23}\right)f_B\right]/8 \tag{22.117}$$

where f_A and f_B are determined from Q_A and Q_B, respectively, by applying either Eq. (22.114) or Eq. (22.115). Here,

$$Q_A = \frac{R_{21,34} - R_{12,34}}{R_{32,41} - R_{23,41}} \tag{22.118}$$

$$Q_B = \frac{R_{43,12} - R_{34,12}}{R_{14,23} - R_{41,23}} \tag{22.119}$$

The **Hall mobility** is determined using the configuration of Fig. 22.36(b), in which the current and voltage contacts are crossed. The **Hall coefficient** becomes

$$R_H = \frac{d}{B}\left[\frac{R_{31,42} + R_{42,13}}{2}\right] \tag{22.120}$$

In general, to minimize magnetoresistive and other effects, it is useful to average over current and magnetic field polarities. Then,

$$\begin{aligned} R_H = \left(d/B\right)\Big[&R_{31,42}(+B) - R_{13,42}(+B) + R_{42,13}(+B) - R_{24,13}(+B) + R_{13,42}(-B) \\ &- R_{31,42}(-B) + R_{24,13}(-B) - R_{42,13}(-B)\Big]/8 \end{aligned} \tag{22.121}$$

Data Analysis

The primary quantities determined from Hall effect and conductivity measurements are the Hall carrier concentration (n_H or p_H) and **mobility** (μ_H). As already discussed, $n_H = -1/eR_H$, where R_H is given by Eq. (22.120) (for a van der Pauw configuration), and $\mu_H = R_H\sigma = R_H/\rho$, where ρ is given by Eq. (22.117). Although simple 300-K values of ρ, n_H, and μ_H are quite important and widely used, it is in temperature-dependent Hall (TDH) measurements that the real power of the Hall technique is demonstrated, because then the **donor** and **acceptor** concentrations and energies can be determined. The methodology is illustrated with a GaN example.

The GaN sample discussed here was a square (6 mm × 6 mm) layer grown on sapphire to a thickness of $d = 20$ μm. Small indium dots were soldered on the corners to provide ohmic contacts, and the Hall measurements were carried out in an apparatus similar to that illustrated in Fig. 22.38. Temperature control was achieved using a He exchange-gas dewar. The temperature dependencies of n_H and μ_H are shown in Figs. 22.39 and 22.40,

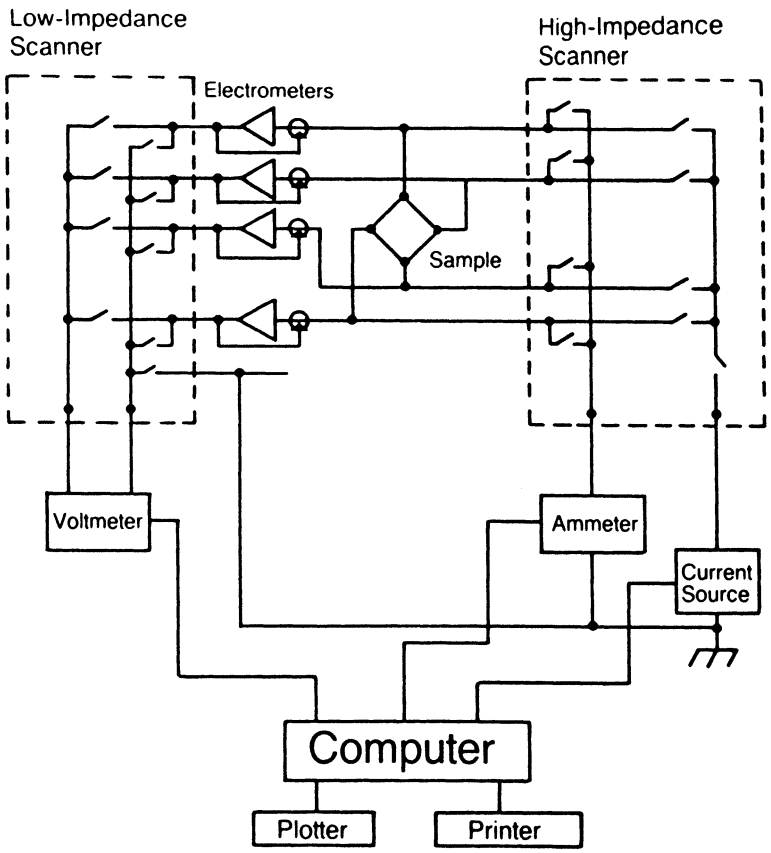


FIGURE 22.38 A schematic diagram of an automated, high-impedance Hall effect apparatus. All components are commercially available.

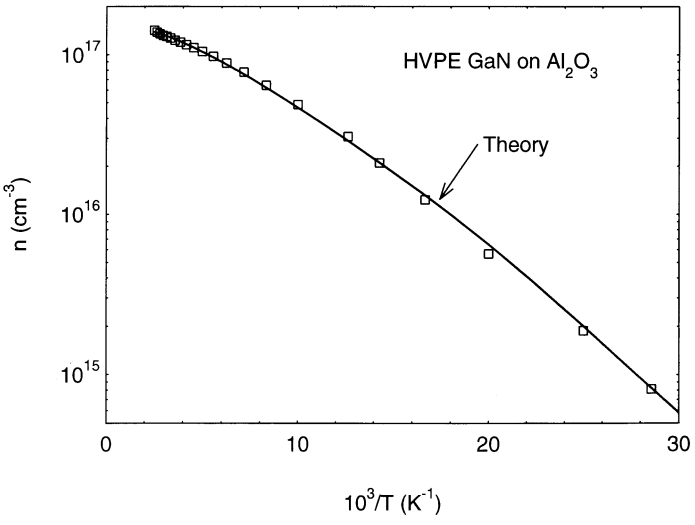


FIGURE 22.39 Hall concentration data (squares) and fit (solid line) vs. inverse temperature.

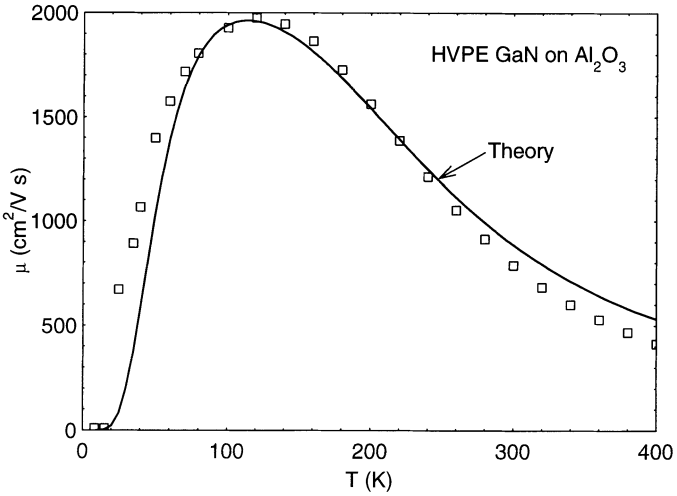


FIGURE 22.40 Hall mobility data (squares) and fit (solid line) vs. temperature.

respectively. The data in these figures have been corrected for a very thin, strongly ***n*-type** layer between the sapphire substrate and GaN layer, as discussed by Look and Molnar [1997].

The solid lines are fits of n_H and μ_H, carried out using MATHCAD software on a personal computer. In many cases, it is sufficient to simply assume $n = n_H$ (i.e., $r = 1$) in Eq. (22.105), but a more accurate answer can be obtained by using the following steps: (1) let $n = n_H = 1/eR_H$ at each T; (2) use Eq. (22.99), Eq. (22.98), and the expression $\mu_H = e\langle\tau^2\rangle/m^*\langle\tau\rangle$ to fit μ_H vs. T and get a value for N_A; (3) calculate $r = \langle\tau^2\rangle/\langle\tau\rangle^2$ at each T; (4) calculate a new $n = rn_H$ at each T; and (5) fit n vs. T with Eq. (22.105) to get values of N_D and E_D. Further iterations can be carried out if desired, but usually add little accuracy. The following parameters were taken from the literature: $P = 0.104$, $\varepsilon_0 = 10.4(8.8542 \times 10^{-12})$ F m^{-1}; $\varepsilon_\infty = 5.47(8.8542 \times 10^{-12})$ F m^{-1}; $T_{po} = 1044$ K; $m^* = 0.22(9.1095 \times 10^{-31})$ kG; $\rho_d = 6.10 \times 10^3$ kg m^{-3}; $s = 6.59 \times 10^3$ m s^{-1}; $g_0 = 1$; $g_1 = 2$; $\alpha_D = 0$; and $N_C' = 4.98 \times 10^{20}$ m^{-3}. The best value for E_1 was found to be 14 eV $= 2.24 \times 10^{-18}$ joules, although 9.2 eV is given by one literature source. The *fitted* parameters are: $N_D = 1.8 \times 10^{17}$ cm^{-3}, $N_A = 2 \times 10^{16}$ cm^{-3}, and $E_D = 18$ meV.

Sources of Error

Contact Size and Placement Effects

Much has been written about this subject over the past few decades [Look, 1989]. Indeed, it is possible to calculate errors due to contact size and placement for any of the structures shown in Fig. 22.35. For (a), (c), and (e), great care is necessary, while for (b), (d), and (f), large or misplaced contacts are not nearly as much of a problem. In general, a good rule of thumb is to keep contact size, and distance from the periphery, each below 10% of the smallest sample-edge dimension. For Hall-bar structures (a) and (b), in which the contacts cover the ends, the ratio $\ell/w > 3$ should be maintained.

Thermomagnetic Errors

Temperature gradients can set up spurious emfs that can modify the measured Hall voltage. Most of these effects, as well as misalignment of the Hall contacts in structure (b), can be averaged out by taking measurements at positive and negative values of both current and magnetic field, and then applying Eq. (22.117) and Eq. (22.121).

Conductive Substrates

If a thin film is grown on a conductive substrate, the substrate conductance may overwhelm the film conductance. If so, and if μ_{sub} and n_{sub} are known, then Eq. (22.107) and Eq. (22.108) can be reduced to a two-layer problem and used to extract μ_{bulk} and n_{bulk}. If the substrate and film are of different types (e.g., a ***p*-type** film

on an ***n*-type** substrate), then a current barrier (p/n junction) will be set up, and the measurement can possibly be made with no correction. However, in this case, the contacts must not overlap both layers.

Depletion Effects in Thin Films

Surface states as well as film/substrate interface states can deplete a thin film of a significant fraction of its charge carriers. Suppose these states lead to surface and interface potentials of ϕ_s and ϕ_i, respectively. Then, regions of width w_s and w_i will be depleted of their free carriers, where

$$w_{s(i)} = \left[\frac{2\varepsilon_0\phi_{s(i)}}{2\left(N_D - N_A\right)}\right]^{1/2} \tag{22.122}$$

It is assumed that $\phi_{s(i)} >> kT/e$, and that $e\phi_{s(i)} >> \mathcal{E}_C - \mathcal{E}_F$. The ***electrical* thickness** of the film will then be given by $d_{elec} = d - w_s - w_i$. Typical values of ϕ_s and ϕ_i are 1 V, so that if $N_D - N_A = 10^{17}$ cm^{-3}, then $w_s + w_i \approx 2000$ Å = 0.2 μm in GaN. Thus, if $d \approx 0.5$ μm, 40% of the electrons will be lost to surface and interface states, and $d_{elec} \approx$ 0.3 μm.

Inhomogeneity

A sample that is inhomogeneous in depth can be analyzed according to Eq. (22.107) and Eq. (22.108), as mentioned above. However, if a sample is *laterally* inhomogeneous, it is nearly always impossible to carry out an accurate analysis. One indication of such inhomogeneity is a resistivity ratio $Q >> 1$ (Fig. 22.37) in a *symmetric* sample, which would be expected to have $Q = 1$. The reader should be warned to *never* attempt an f-correction (Fig. 22.37) in such a case, because the f-correction is valid only for sample-shape asymmetry, not inhomogeneity.

Non-ohmic Contacts

In general, high contact resistances are not a severe problem as long as enough current can be passed to get measurable values of V_c and V_H. The reason is that the voltage measurement contacts carry very little current. However, in some cases, the contacts may set up a p/n junction and significantly distort the current flow. This situation falls under the "inhomogeneity" category, discussed above. Usually, contacts this bad show variations with current magnitude and polarity; thus, for the most reliable Hall measurements, it is a good idea to make sure the values are invariant with respect to the magnitudes and polarities of both current and magnetic field.

Defining Terms

Acceptor: An impurity or lattice defect that can "accept" one or more electrons from donors or the valence band; in the latter case, free holes are left to conduct current in the valence band.

Charge-balance equation (CBE): A mathematical relationship expressing the equality between positive and negative charges in a sample as a function of temperature.

Dislocation: A one-dimensional line defect in a solid, which often extends through the entire lattice. An *edge* dislocation is essentially an extra half-lattice plane inserted into the lattice.

Distribution function: A mathematical relationship describing the distribution of the electrons, as a function of temperature, among all the possible energy states in the lattice, including those arising from the conduction band, valence band, donors, and acceptors.

Donor: An impurity or lattice defect that can "donate" one or more electrons to acceptors or to the conduction band; in the latter case, free electrons are available to conduct current.

Effective mass: The *apparent* mass of an electron or hole with respect to acceleration in an electric field.

Electrical thickness: The "thickness" of a layer in which the current actually flows. In a thin sample, this dimension may be much less than the physical thickness of the sample because some of the charge carriers may be immobilized at surface and interface states.

Hall coefficient: The ratio between the Hall electric field E_y (a field that develops perpendicular to the plane formed by the current and magnetic field directions), and the current density j_x multiplied by the magnetic field strength B_z; i.e., $R_H = E_y/j_xB_z$. The Hall coefficient is closely related to the carrier concentration.

Hall mobility: The Hall coefficient multiplied by the conductivity. This mobility is often nearly equal to the conductivity mobility.

Lattice vibrations: The collective motions of atoms (often called phonons) in a crystal lattice. The phonons can interact with the charge carriers and reduce mobility.

Matthiessen's Rule: The approximation that the inverse of the total mobility is equal to the inverses of the individual components of the mobility; that is, $\mu^{-1} = \mu_1^{-1} + \mu_2^{-1} + \mu_3^{-1} + \ldots$, where μ_i^{-1} denotes the mobility that would result if only scattering mechanism i were present.

Mobility: The ease with which charge carriers move in a crystal lattice.

***n*-type:** The designation of a sample that has a conductivity primarily controlled by electrons.

***p*-type:** The designation of a sample that has a conductivity primarily controlled by holes.

Relaxation time: The time required to nullify a disturbance in the equilibrium energy or momentum distribution of the electrons and holes.

Relaxation time approximation (RTA): A relatively simple analytical solution of the Boltzmann transport equation that is valid for elastic (energy-conserving) scattering processes.

References

Look, D. C., *Electrical Characterization of GaAs Materials and Devices.* Wiley, New York, 1989, Chap. 1.

Look, D. C., Dislocation scattering in GaN, *Phys. Rev. Lett.,* 82, 1237, 1999.

Look, D. C. and Molnar, R. J., Degenerate layer at GaN/sapphire interface: influence on Hall-effect measurements, *Appl. Phys. Lett.,* 70, 3377, 1997.

Nag, B. R., *Electron Transport in Compound Semiconductors,* Springer-Verlag, Berlin, 1980.

Rode, D. L., Low-field electron transport, in *Semiconductors and Semimetals,* Willardson, R. K. and Beer, A. C., Eds., Academic, New York, 1975, Chap. 1.

van der Pauw, L. J., A method of measuring specific resistivity and Hall effect of discs of arbitrary shape, *Philips Res. Repts.,* 13, 1, 1958.

Wieder, H. H., *Laboratory Notes on Electrical and Galvanomagnetic Measurements,* Elsevier, Amsterdam, 1979.

Wiley, J. D., Mobility of holes in III-V compounds, in *Semiconductors and Semimetals,* Willardson, R. K. and Beer, A. C., Eds., Academic, New York, 1975, Chap. 2.

Further Information

Good general references on semiconductor characterization, including techniques other than electrical, are the following: Runyan, W. R., *Semiconductor Measurements and Instrumentation,* McGraw-Hill, New York, 1975; Schroder, D. K., *Semiconductor Material and Device Characterization,* Wiley, New York, 1990; and Orton, J. W. and Blood, P., *The Electrical Characterization of Semiconductors: Measurement of Minority Carrier Properties,* Academic, New York, 1990.

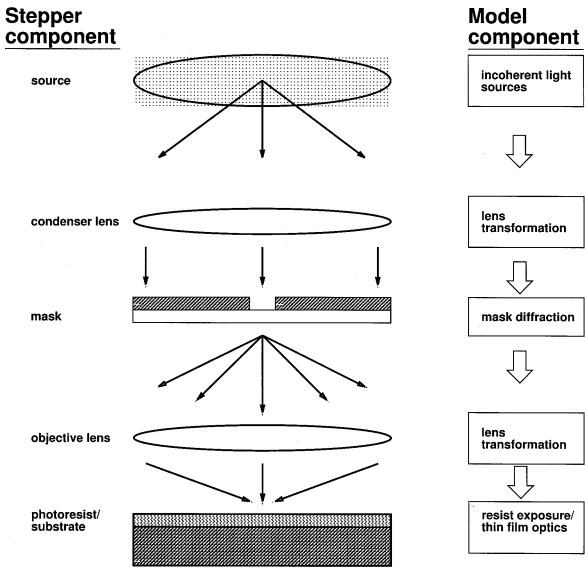


FIGURE 23.57 Optical image formation: each of the components of an imaging system affects the wavefront; between components, light follows free space propagation rules [Leon, 1998].

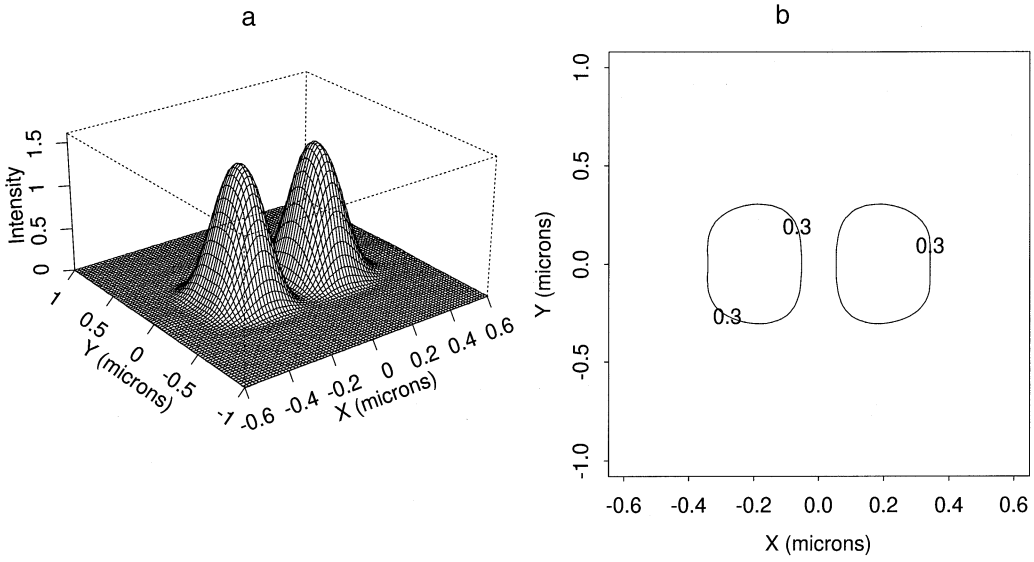


FIGURE 23.58 Aerial image: the light intensity at the resist surface due to imaging through two rectangles of size 0.28 × 0.6 μm separated by 0.12 μm [Watson, 1999] at an illumination wavelength of 248 nm. The half-wavelength feature spacing is well resolved.

Resist development is simulated based on a model introduced by Dill [1975]. The development process is treated as a surface-etching phenomenon, with the etch rate depending only on the local concentration of inhibitor. The evolution of the surface with development time can therefore be treated by the same evolution algorithms used in etching and deposition simulation.

As optical lithography is pushed to its limits in imaging ever smaller features using illumination wavelengths that are not easily reduced, simulation is playing an increasingly important role in the development of resolution-enhancement techniques such as **optical-proximity-correction (OPC)** features, **phase shift masks**, and **off axis** illumination.

Summary and Future Trends

Computer simulation of semiconductor processing has become a widely accepted technique to reduce the high cost and long turnaround time of fabrication trials. **Physically based models** of the classic process steps have been established and widely applied, while new processes such as chemical-mechanical polishing are begetting a new generation of models to simulate them. The increasing speed of computers and the improving understanding of fabrication processes, compared to the increasing cost of experiments in a multibillion dollar fabrication line, will continue to drive the development and refinement of accurate process simulation tools.

Defining Terms

Aerial image: The output of an optical simulator.

Device simulator: A computer simulation program that predicts the relation between current and voltage of an electron device based on its geometrical structure and its dopant atom distribution.

Empirical models: Models based primarily on fitting measured data without an analysis of the underlying phenomena.

Evolution simulator: A computer simulation tool for predicting the change in surface shape under the influence of surface motion rates.

Lithography simulator: A computer simulation tool for predicting the shape of resist features after exposure and development.

Monte Carlo models: Many physical systems can be modeled by following the trajectories of a representative number of individual particles under the applied forces. A random choice is made whenever statistically equally likely outcomes of events are possible.

Optical proximity correction: The modification of mask features in order to counteract undesired diffraction effects around small geometry features.

Off-axis illumination: The use of a non-point illumination source to improve lithographic resolution.

Optical simulator: A computer simulation tool for predicting light intensity at the surface of resist after passing through a projection lithography system.

Oxidation-enhanced diffusion (OED): The diffusion of dopants in the bulk of a wafer is enhanced when oxidation occurs at its surface.

Phase shift masks: The use of partially transmitting features on a mask to improve lithographic resolution.

Physically based models: Models based on fundamental physical and chemical principles.

Process simulator: A computer simulation program that predicts the outcome of the integrated circuit fabrication steps in terms of the geometrical structure and dopant distribution of the wafer.

Stokes flow: The flow of a liquid when body forces and inertial terms are negligible in comparison to viscous resistance.

Transient-enhanced diffusion (TED): The diffusion of dopants in the bulk of a wafer is very much enhanced following ion implantation.

Topography simulator: A computer simulation tool for predicting the net effect of a number of etching and deposition steps on the wafer topography.

References

[Biersack, 1986] J.P. Biersack and L.G. Haggmark, A Monte Carlo computer program for the transport of energetic ions in amorphous targets, *Nucl. Inst. and Meth.*, B13, 100 (1986).

[Cale, 1992] T.S. Cale, G.B. Raupp, and T.H. Gandy, Ballistic transport-reaction prediction of film conformality in tetraethoxysilane O_2 plasma enhanced deposition of silicon dioxide, *J. Vacuum Sci. Technol.*, A10(4), 1128, (1992).

[Chin, 1982] D. Chin, S.Y. Oh, S.M. Hu, R.W. Dutton, and J.L. Moll, Stress in local oxidation, *IEDM Technical Digest*, 228 (1982).

[Deal, 1965] B.E. Deal and A.S. Grove, General relationship for the thermal oxidation of silicon, *J. Appl. Phys.*, 36(12), 3370 (1965).

[Dill, 1975] F.H. Dill, A.R. Neureuther, J. A. Tuttle, and E.J. Walley, Modeling projection printing of positive photoresists, *IEEE Trans. Electron Dev.*, 22, 445 (1975).
[Fahey, 1989] P.M. Fahey, P.B. Griffin, and J.D. Plummer, Point defects and dopant diffusion in silicon, *Rev. Modern Phys.*, 6(12), 289 (1989).
[Fair, 1981] R.B. Fair, Concentration profiles of diffused dopants in silicon, *Impurity Doping*, F.F.Y. Wang (Ed.), North-Holland (1981).
[Gilmer, 1998] G.H. Gilmer, H. Huang, and T. Diaz de la Rubia, Thin film deposition, in *Computational Material Science*, T. Diaz de al Rubia (Ed.), Elsevier, in press.
[Hamaguchi, 1993] S. Hamaguchi, M. Dalvie, R.T. Farouki, and S. Sethuraman, A shock-tracking algorithm for surface evolution under reactive-ion etching, *J. Appl. Phys.*, 74(8), 5172 (1993).
[Hobler, 1986] G. Hobler, E. Langer, and S. Selberherr, Two-dimensional modeling of ion implantation, in *Second Int. Conf. Simulation of Semiconductor Devices and Process*, K. Board and R. Owen, Eds., Pineridge Press, Swansea (1986).
[Kao, 1985] D.-B. Kao, J.P. McVittie, W.D. Nix, and K.C. Saraswat, Two-dimensional silicon oxidation experiments and theory, *IEDM Technical Digest*, 388 (1985).
[Lau, 1990] F. Lau, Modeling of polysilicon diffusion sources, *IEDM Technical Digest*, 737 (1990).
[Leon, 1998] F. Leon, Short course on next generation TCAD: models and methods, *International Electron Device Meeting*, Dec. 13, San Francisco (1998).
[Lim, 1993] D. Lim, S. Yang, S. Morris, and A.F. Tasch, An accurate and computationally efficient model of boron implantation through screen oxide layers into (100) single-crystal silicon, *IEDM*, 291 (1993).
[Massoud, 1985] H.Z. Massoud, J.D. Plummer, and E.A. Irene, Thermal oxidation of silicon in dry oxygen: growth-rate enhancement in the thin regime I. Experimental results, *J. Electrochem. Soc.*, 132, 2685 (1985).
[Oldham, 1979] W.G. Oldham, A.R. Neureuther, C.K. Snug, J.L. Reynolds, and S.N. Nandgaonkar, A general simulator for VLSI lithography and etching processes. Part I. Application to projection lithography, *IEEE Trans. Elect. Dev.*, 26, 712 (1979).
[Oldham, 1980] W.G. Oldham, A.R. Neureuther, C.K. Snug, J.L. Reynolds, and S.N. Nandgaonkar, A general simulator for VLSI lithography and etching processes. Part II. Application to deposition and etching, *IEEE Trans. Elect. Dev.*, 27, 1455 (1980).
[O'Sullivan, 1999] Peter O'Sullivan, private communication (1999).
[O'Toole, 1979] M.M. O'Toole and A.R. Neureuther, Developments in semiconductor microlithography IV, *SPIE*, 174, 22 (1979).
[Rafferty, 1989] C.S. Rafferty, Unpublished.
[Rafferty, 1990] C.S. Rafferty, Two-dimensional modeling of viscous flow in thermal SiO_2, *Extended Abstracts of the Electrochemical Society Spring Meeting*, May 6–11, 423 (1990).
[Rafferty, 1993] C.S. Rafferty, H.-H. Vuong, S.A. Eshraghi, M.D. Giles, M.R. Pinto, and S.J. Hillenius, Explanation of reverse short channel effect by defect gradients, *IEDM Technical Digest*, 311 (1993).
[Robinson, 1974] M.T. Robinson, Computer simulation of atomic displacement cascaded in solids in binary collision approximation, *Phys. Rev.*, B9(12), 5008, (1974).
[Singh, 1992] V. Singh, E.S.G. Shaqfeh, and J.P. McVittie, *J. Vac. Sci. Technol.*, B10(3), 1091 (1992).
[Smy, 1998] T. Smy, R.V. Joshi, N. Tait, S.K. Dew, and M.J. Brett, Deposition and simulation of refractory barriers into high aspect ratio re-entrant features using directional sputtering, *IEDM Technical Digest*, 311 (1998).
[Toh, 1988] K.K.H. Toh, Two-dimensional images with effects of lens aberrations in optical lithography, Memorandum UCB/ERL M88/30, University of California, Berkeley, May 20 (1988).
[Watson, 1999] Patrick Watson, private communication (1999).

For Further Information

Several classic textbooks now exist with good information on numerical methods and process simulation. Among them are:

Physics and Technology of Semiconductor Devices, A.S. Grove, Wiley (1967)
VLSI Technology, edited by S.M. Sze, McGraw Hill (1988 2nd ed.)

Silicon Processing for the VLSI Era, S. Wolf, R.N. Tauber, Vols. 1 & 2, Lattice Press (1986, 1990)
The Finite Element Method, O.C. Zienkiewicz, McGraw-Hill (1977)
Matrix Iterative Analysis, R.S. Varga, Prentice-Hall (1962)

The proceedings of the annual conference SISPAD (Simulation of Semiconductor Processes and Devices), the annual International Electron Device Meeting (IEDM), and the bi-annual meetings of the Electrochemical Society and Materials Research Society, are among the main outlets of process simulation work.

Soclof, S., Watson, J., Brews, J.R. “Transistors”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

24

Transistors

24.1 Junction Field-Effect Transistors
JFET Biasing • Transfer Characteristics • JFET Output Resistance • Source Follower • Frequency and Time-Domain Response • Voltage-Variable Resistor

24.2 Bipolar Transistors
Biasing the Bipolar Transistor • Small-Signal Operation • A Small-Signal Equivalent Circuit • Low-Frequency Performance • The Emitter-Follower or Common-Collector (CC) Circuit • The Common-Emitter Bypass Capacitor C_E • High-Frequency Response • Complete Response • Design Comments • Integrated Circuits • The Degenerate Common-Emitter Stage • The Difference Amplifier • The Current Mirror • The Difference Stage with Current Mirror Biasing • The Current Mirror as a Load

24.3 The Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET)
Current-Voltage Characteristics • Important Device Parameters • Limitations upon Miniaturization

Sidney Soclof
California State University, Los Angeles

Joseph Watson
University of Wales, Swansea

John R. Brews
The University of Arizona

24.1 Junction Field-Effect Transistors

Sidney Soclof

A junction field-effect transistor, or JFET, is a type of transistor in which the current flow through the device between the drain and source electrodes is controlled by the voltage applied to the gate electrode. A simple physical model of the JFET is shown in Fig. 24.1. In this JFET an *n*-type conducting channel exists between drain and source. The gate is a p^+ region that surrounds the *n*-type channel. The gate-to-channel *pn* junction is normally kept reverse-biased. As the reverse bias voltage between gate and channel increases, the depletion region width increases, as shown in Fig. 24.2. The depletion region extends mostly into the *n*-type channel because of the heavy doping on the p^+ side. The depletion region is depleted of mobile charge carriers and thus cannot contribute to the conduction of current between drain and source. Thus as the gate voltage increases, the cross-sectional areas of the *n*-type channel available for current flow decreases. This reduces the current flow between drain and source. As the gate voltage increases, the channel gets further constricted, and the current flow gets smaller. Finally when the depletion regions meet in the middle of the channel, as shown in Fig. 24.3, the channel is pinched off in its entirety between source and drain. At this point the current flow between drain and source is reduced to essentially zero. This voltage is called the **pinch-off voltage,** V_P. The pinch-off voltage is also represented by V_{GS} (off) as being the gate-to-source voltage that turns the drain-to-source current I_{DS} off. We have been considering here an *n*-channel JFET. The complementary device is the *p*-channel JFET that has an n^+ gate region surrounding a *p*-type channel. The operation of a *p*-channel JFET is the same as for an *n*-channel device, except the algebraic signs of all dc voltages and currents are reversed.

We have been considering the case for V_{DS} small compared to the pinch-off voltage such that the channel is essentially uniform from drain to source, as shown in Fig. 24.4(a). Now let's see what happens as V_{DS} increases. As an example let's assume an *n*-channel JFET with a pinch-off voltage of $V_P = -4$ V. We will see what happens

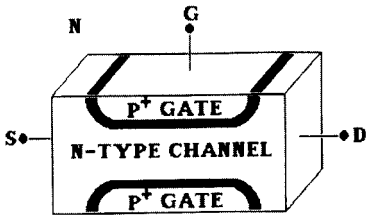


FIGURE 24.1

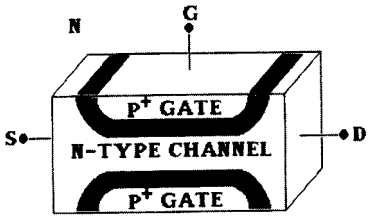


FIGURE 24.2

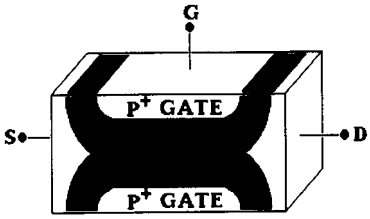


FIGURE 24.3

for the case of $V_{GS} = 0$ as V_{DS} increases. In Fig. 24.4(a) the situation is shown for the case of $V_{DS} = 0$ in which the JFET is fully "on" and there is a uniform channel from source to drain. This is at point A on the I_{DS} vs. V_{DS} curve of Fig. 24.5. The drain-to-source conductance is at its maximum value of g_{ds} (on), and the drain-to-source resistance is correspondingly at its minimum value of r_{ds} (on). Now let's consider the case of $V_{DS} = +1$ V, as shown in Fig. 24.4(b). The gate-to-channel bias voltage at the source end is still $V_{GS} = 0$. The gate-to-channel bias voltage at the drain end is $V_{GD} = V_{GS} - V_{DS} = -1$ V, so the depletion region will be wider at the drain end of the channel than at the source end. The channel will thus be narrower at the drain end than at the source end, and this will result in a decrease in the channel conductance g_{ds} and, correspondingly, an increase in the channel resistance r_{ds}. So the slope of the I_{DS} vs. V_{DS} curve that corresponds to the channel conductance will be smaller at $V_{DS} = 1$ V than it was at $V_{DS} = 0$, as shown at point B on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

In Fig. 24.4(c) the situation for $V_{DS} = +2$ V is shown. The gate-to-channel bias voltage at the source end is still $V_{GS} = 0$, but the gate-to-channel bias voltage at the drain end is now $V_{GD} = V_{GS} - V_{DS} = -2$ V, so the depletion region will now be substantially wider at the drain end of the channel than at the source end. This leads to a further constriction of the channel at the drain end, and this will again result in a decrease in the channel conductance g_{ds} and, correspondingly, an increase in the channel resistance r_{ds}. So the slope of the I_{DS} vs. V_{DS} curve will be smaller at $V_{DS} = 2$ V than it was at $V_{DS} = 1$ V, as shown at point C on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

In Fig. 24.4(d) the situation for $V_{DS} = +3$ V is shown, and this corresponds to point D on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

When $V_{DS} = +4$ V, the gate-to-channel bias voltage will be $V_{GD} = V_{GS} - V_{DS} = 0 - 4$ V $= -4$ V $= V_P$. As a result the channel is now pinched off at the drain end but is still wide open at the source end since $V_{GS} = 0$, as shown in Fig. 24.4(e). It is very important to note that the channel is pinched off just for a very short distance at the drain end so that the drain-to-source current I_{DS} can still continue to flow. This is not at all the same situation as for the case of $V_{GS} = V_P$, where the channel is pinched off in its entirety, all the way from source to drain. When this happens, it is like having a big block of insulator the entire distance between source and drain, and I_{DS} is reduced to essentially zero. The situation for $V_{DS} = +4$ V $= -V_P$ is shown at point E on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

For $V_{DS} > +4$ V, the current essentially saturates and doesn't increase much with further increases in V_{DS}. As V_{DS} increases above +4 V, the pinched-off region at the drain end of the channel gets wider, which increases r_{ds}. This increase in r_{ds} essentially counterbalances the increase in V_{DS} such that I_{DS} does not increase much. This region of the I_{DS} vs. V_{DS} curve in which the channel is pinched off at the drain end is called the **active region** and is also known as the *saturated region.* It is called the active region because when the JFET is to be used as an amplifier, it should be biased and operated in this region. The saturated value of drain current up in the active region for the case of $V_{GS} = 0$ is called the **drain saturation current,** I_{DSS} (the third subscript S

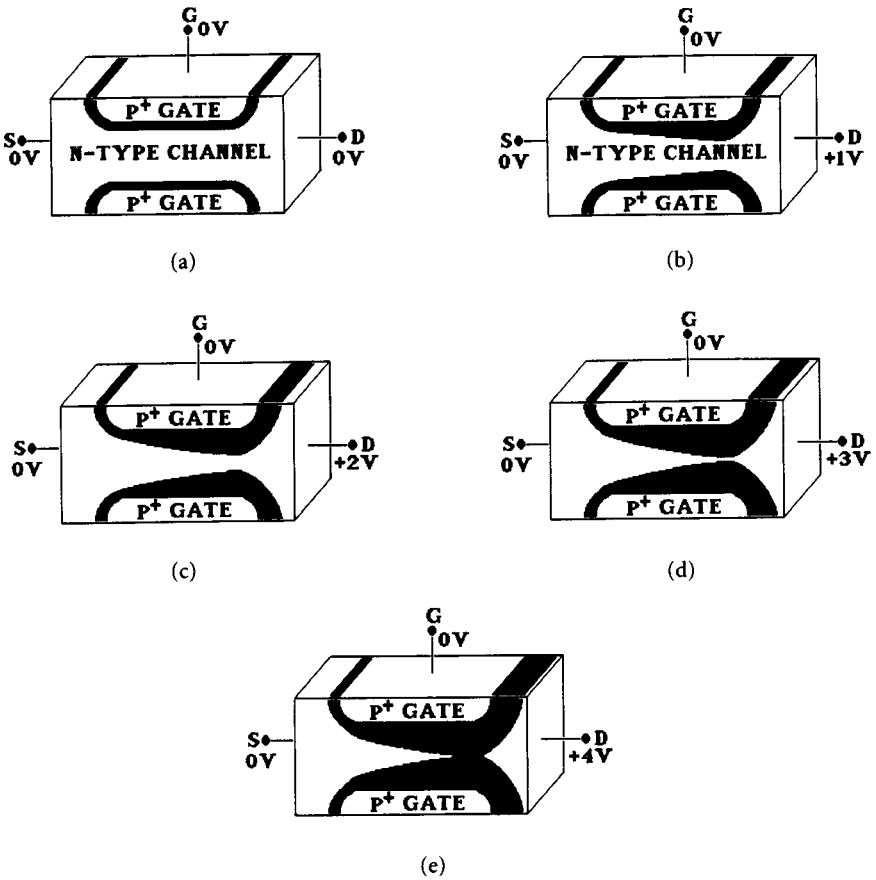


FIGURE 24.4

refers to I_{DS} under the condition of the gate *shorted* to the source). Since there is not really a true saturation of current in the active region, I_{DSS} is usually specified at some value of V_{DS}. For most JFETs, the values of I_{DSS} fall in the range of 1 to 30 mA.

The region below the active region where $V_{DS} < +4\text{ V} = -V_P$ has several names. It is called the **nonsaturated region**, the **triode region**, and the **ohmic region**. The term *triode region* apparently originates from the similarity of the shape of the curves to that of the vacuum tube triode. The term *ohmic region* is due to the variation of I_{DS} with V_{DS} as in Ohm's law, although this variation is nonlinear except for the region of V_{DS} that is small compared to the pinch-off voltage where I_{DS} will have an approximately linear variation with V_{DS}.

The upper limit of the active region is marked by the onset of the breakdown of the gate-to-channel *pn* junction. This will occur at the drain end at a voltage designated as BV_{DG}, or BV_{DS}, since $V_{GS} = 0$. This breakdown voltage is generally in the 30- to 150-V range for most JFETs.

So far we have looked at the I_{DS} vs. V_{DS} curve only for the case of $V_{GS} = 0$. In Fig. 24.6 a family of curves of I_{DS} vs. V_{DS} for various constant values of V_{GS} is presented. This is called the *drain characteristics*, also known as the *output characteristics*, since the output side of the JFET is usually the drain side. In the active region where I_{DS} is relatively independent of V_{DS}, a simple approximate equation relating I_{DS} to V_{GS} is the square-law *transfer equation* as given by $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$. When $V_{GS} = 0$, $I_{DS} = I_{DSS}$ as expected, and as $V_{GS} \rightarrow V_P$, $I_{DS} \rightarrow 0$. The lower boundary of the active region is controlled by the condition that the channel be pinched off at the drain end. To meet this condition

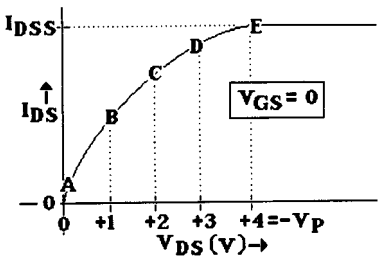


FIGURE 24.5

the basic requirement is that the gate-to-channel bias voltage at the drain end of the channel, V_{GD}, be greater than the pinch-off voltage V_P. For the example under consideration with $V_P = -4$ V, this means that $V_{GD} = V_{GS} - V_{DS}$ must be more negative than –4 V. Therefore, $V_{DS} - V_{GS} \geq +4$ V. Thus, for $V_{GS} = 0$, the active region will begin at $V_{DS} = +4$ V. When $V_{GS} = -1$ V, the active region will begin at $V_{DS} = +3$ V, for now $V_{GD} = -4$ V. When $V_{GS} = -2$ V, the active region begins at $V_{DS} = +2$ V, and when $V_{GS} = -3$ V, the active region begins at $V_{DS} = +1$ V. The dotted line in Fig. 24.6 marks the boundary between the nonsaturated and active regions.

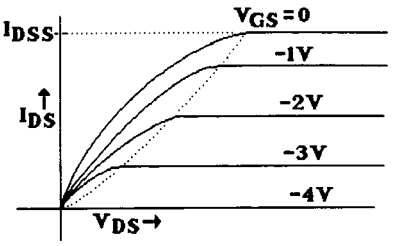


FIGURE 24.6

The upper boundary of the active region is marked by the onset of the avalanche breakdown of the gate-to-channel *pn* junction. When $V_{GS} = 0$, this occurs at $V_{DS} = BV_{DS} = BV_{DG}$. Since $V_{DG} = V_{DS} - V_{GS}$ and breakdown occurs when $V_{DG} = BV_{DG}$, as V_{GS} increases the breakdown voltage decreases, as given by $BV_{DG} = BV_{DS} - V_{GS}$. Thus $BV_{DS} = BV_{DG} + V_{GS}$. For example, if the gate-to-channel breakdown voltage is 50 V, the V_{DS} breakdown voltage will start off at 50 V when $V_{GS} = 0$ but decrease to 46 V when $V_{GS} = -4$ V.

In the nonsaturated region I_{DS} is a function of both V_{GS} and V_{DS}, and in the lower portion of the nonsaturated region where V_{DS} is small compared to V_P, I_{DS} becomes an approximately linear function of V_{DS}. This linear portion of the nonsaturated is called the *voltage-variable resistance* (VVR) region, for in this region the JFET acts like a linear resistance element between source and drain. The resistance is variable in that it is controlled by the gate voltage. This region and VVR application will be discussed in a later section. The JFET can also be operated in this region as a switch, and this will also be discussed in a later section.

JFET Biasing

Voltage Source Biasing

Now we will consider the biasing of JFETs for operation in the active region. The simplest biasing method is shown in Fig. 24.7, in which a voltage source V_{GG} is used to provide the quiescent gate-to-source bias voltage V_{GSQ}. In the active region the transfer equation for the JFET has been given as $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$, so for a quiescent drain current of I_{DSQ} the corresponding gate voltage will be given by $V_{GSQ} = V_P\,(1 - \sqrt{I_{DSQ}/I_{DSS}}$. For a Q point in the middle of the active region, we have that $I_{DSQ} = I_{DSS}/2$, so $V_{GSQ} = V_P\,(1 - \sqrt{1/2}) = 0.293\ V_P$.

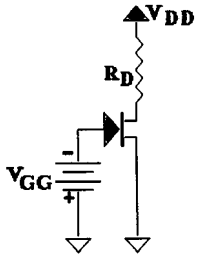


FIGURE 24.7 Voltage source biasing.

The voltage source method of biasing has several major drawbacks. Since V_P will have the opposite polarity of the drain supply voltage V_{DD}, the gate bias voltage will require a second power supply. For the case of an *n*-channel JFET, V_{DD} will come from a positive supply voltage and V_{GG} must come from a separate negative power supply voltage or battery. A second, and perhaps more serious, problem is the "open-loop" nature of this biasing method. The JFET parameters of I_{DDS} and V_P will exhibit very substantial unit-to-unit variations, often by as much as a 2:1 factor. There is also a significant temperature dependence of I_{DDS} and V_P. These variations will lead to major shifts in the position of the *Q* point and the resulting distortion of the signal. A much better biasing method is shown in Fig. 24.8.

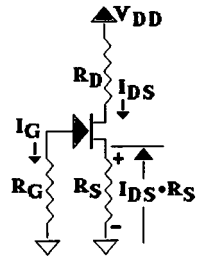


FIGURE 24.8 Self-biasing.

Self-Biasing

The biasing circuit of Fig. 24.8 is called a *self-biasing* circuit in that the gate-to-source voltage is derived from the voltage drop produced by the flow of drain current through the source biasing resistor R_S. It is a closed-loop system in that variations in the JFET parameters can be partially compensated for by the biasing circuit. The gate resistor R_G is used to provide a dc return path for the gate leakage current and is generally up in the megohm range.

The voltage drop across R_S is given by $V_S = I_{DS} \cdot R_S$. The voltage drop across the gate resistor R_G is $V_G = I_G \cdot R_G$. Since I_G is usually in the low nanoampere or even picoampere range, as long as R_G is not extremely large

the voltage drop across R_G can be neglected, so $V_G \cong 0$. Thus, we have that $V_{GS} = V_G - V_S \cong -V_S = -I_{DS} \cdot R_S$. For example, if $I_{DSS} = 10$ mA and $V_P = -4$ V, and for a Q point in the middle of the active region with $I_{DSQ} = I_{DSS}/2 = 5$ mA, we have that $V_{GSQ} = 0.293 V_P = -1.17$ V. Therefore the required value for the source biasing resistor is given by $R_S = -V_{GS}/I_{DSQ} = 1.17$ V/5 mA = 234 Ω. This produces a more stable quiescent point than voltage source biasing, and no separate negative power supply is required.

The closed-loop nature of this biasing circuit can be seen by noting that if changes in the JFET parameters were to cause I_{DS} to increase, the voltage drop across R_S would also increase. This will produce an increase in V_{GS} (in the negative direction for an *n*-channel JFET), which will act to reduce the increase in I_{DS}. Thus the net increase in I_{DS} will be less due to the feedback voltage drop produced by the flow of I_{DS} through R_S. The same basic action would, of course, occur for changes in the JFET parameters that would cause I_{DS} to decrease.

Bias Stability

Now let's examine the stability of the Q point. We will start again with the basic transfer equation as given by $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$. From this equation the change in the drain current, ΔI_{DS}, due to changes in I_{DSS}, V_{GS}, and V_P can be written as

$$\Delta I_{DS} = g_m \Delta V_{GS} - g_m \frac{V_{GS}}{V_P} \Delta V_P + \frac{I_{DS}}{I_{DSS}} \Delta I_{DSS}$$

Since $V_{GS} = -I_{DS} \cdot R_S$, $\Delta V_{GS} = -R_S \cdot \Delta I_{DS}$, we obtain that

$$\Delta I_{DS} = -g_m R_S \Delta I_{DS} - g_m \frac{V_{GS}}{V_P} \Delta V_P + \frac{I_{DS}}{I_{DSS}} \Delta I_{DSS}$$

Collecting terms in ΔI_{DS} on the left side gives

$$\Delta I_{DS}(1 + g_m R_S) = -g_m \frac{V_{GS}}{V_P} \Delta V_P + \frac{I_{DS}}{I_{DSS}} \Delta I_{DSS}$$

Now solving this for ΔI_{DS} yields

$$\Delta I_{DS} = \frac{-g_m (V_{GS}/V_P) \Delta V_P + \dfrac{I_{DS}}{I_{DSS}} \Delta I_{DSS}}{1 + g_m R_S}$$

From this we see that the shift in the quiescent drain current, ΔI_{DS}, is reduced by the presence of R_S by a factor of $1 + g_m R_S$.

If $I_{DS} = I_{DSS}/2$, then

$$g_m = \frac{2\sqrt{I_{DS} \cdot I_{DSS}}}{-V_P} = \frac{2\sqrt{I_{DS} \cdot 2I_{DS}}}{-V_P} = \frac{2\sqrt{2}\, I_{DS}}{-V_P}$$

Since $V_{GS} = 0.293 V_P$, the source biasing resistor will be $R_S = -V_{GS}/I_{DS} = -0.293\ V_P/I_{DS}$. Thus

$$g_m R_S = \frac{2\sqrt{2}\, I_{DS}}{-V_P} \times \frac{-0.293 V_P}{I_{DS}} = 2\sqrt{2} \times 0.293 = 0.83$$

so $1 + g_m R_S = 1.83$. Thus the sensitivity of I_{DS} due to changes in V_P and I_{DSS} is reduced by a factor of 1.83.

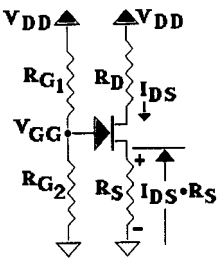


FIGURE 24.9

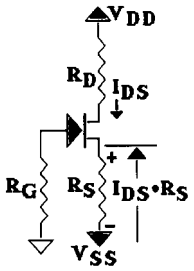


FIGURE 24.10

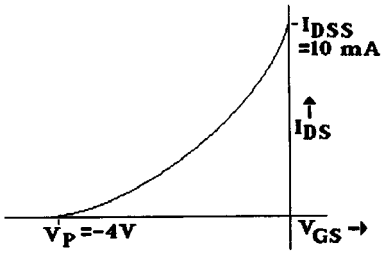


FIGURE 24.11 Transfer characteristic.

The equation for ΔI_{DS} can now be written in the following form for the fractional change in I_{DS}:

$$\frac{\Delta I_{DS}}{I_{DS}} = \frac{-0.83(\Delta V_P/V_P) + 1.41(\Delta I_{DSS}/I_{DSS})}{1.83}$$

so $\Delta I_{DS}/I_{DS} = -0.45\ (\Delta V_P/V_P) + 0.77\ (\Delta I_{DSS}/I_{DSS})$, and thus a 10% change in V_P will result in approximately a 4.5% change in I_{DS}, and a 10% change in I_{DSS} will result in an 8% change in I_{DS}. Thus, although the situation is improved with the self-biasing circuit using R_S, there will still be a substantial variation in the quiescent current with changes in the JFET parameters.

A further improvement in bias stability can be obtained by the use of the biasing methods of Figs. 24.9 and 24.10. In Fig. 24.9 a gate bias voltage V_{GG} is obtained from the V_{DD} supply voltage by means of the R_{G1}–R_{G2} voltage divider. The gate-to-source voltage is now $V_{GS} = V_G - V_S = V_{GG} - I_{DS}R_S$. So now for R_S we have $R_S = (V_{GG} - V_{GS})/I_{DS}$. Since V_{GS} is of opposite polarity to V_{GG}, this will result in a larger value for R_S than before. This in turn will result in a larger value for the $g_m R_S$ product and hence improved bias stability. If we continue with the preceding examples and now let $V_{GG} = V_{DD}/2 = +10$ V, we have that $R_S = (V_{GG} - V_{GS})/I_{DS} = [+10\text{V} - (-1.17\text{V})]/5\text{ mA} = 2.234\ \text{k}\Omega$, as compared to $R_S = 234\ \Omega$ that was obtained before. For g_m we have $g_m = 2\sqrt{I_{DS} \cdot I_{DSS}}/(-V_P) = 3.54$ mS, so $g_m R_S = 3.54\text{ mS} \cdot 2.234\ \text{k}\Omega = 7.90$. Since $1 + g_m R_S = 8.90$, we now have an improvement by a factor of 8.9 over the open-loop voltage source biasing and by a factor of 4.9 over the self-biasing method without the V_{GG} biasing of the gate.

Another biasing method that can lead to similar results is the method shown in Fig. 24.10. In this method the bottom end of the source biasing resistor goes to a negative supply voltage V_{SS} instead of to ground. The gate-to-source bias voltage is now given by $V_{GS} = V_G - V_S = 0 - (I_{DS} \cdot R_S + V_{SS})$ so that for R_S we now have $R_S = (-V_{GS} - V_{SS})/I_{DS}$. If $V_{SS} = -10$ V, and as before $I_{DS} = 5$ mA and $V_{GS} = -1.17$ V, we have $R_S = 11.7\text{ V}/5\text{ mA} = 2.34\ \text{k}\Omega$, and thus $g_m R_S = 7.9$ as in the preceding example. So this method does indeed lead to results similar to that for the R_S and V_{GG} combination biasing. With either of these two methods the change in I_{DS} due to a 10% change in V_P will be only 0.9%, and the change in I_{DS} due to a 10% change in I_{DSS} will be only 1.6%.

The biasing circuits under consideration here can be applied directly to the common-source (CS) amplifier configuration, and can also be used for the common-drain (CD), or source-follower, and common-gate (CG) JFET configurations.

Transfer Characteristics

Transfer Equation

Now we will consider the *transfer characteristics* of the JFET, which is a graph of the output current I_{DS} vs. the input voltage V_{GS} in the active region. In Fig. 24.11 a transfer characteristic curve for a JFET with $V_P = -4$ V and $I_{DSS} = +10$ mA is given. This is approximately a square-law relationship as given by $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$. This equation is not valid for V_{GS} beyond V_P (i.e., $V_{GS} < V_P$), for in this region the channel is pinched off and $I_{DS} \cong 0$.

At $V_{GS} = 0$, $I_{DS} = I_{DSS}$. This equation and the corresponding transfer curve can actually be extended up to the point where $V_{GS} \cong +0.5$ V. In the region where $0 < V_{GS} < +0.5$ V, the gate-to-channel *pn* junction is *forward-biased* and the depletion region width is reduced below the width under zero bias conditions. This reduction in the depletion region width leads to a corresponding expansion of the conducting channel and thus an increase in I_{DS} above I_{DSS}. As long as the gate-to-channel forward bias voltage is less than about 0.5 V, the *pn* junction will be essentially "off" and very little gate current will flow. If V_{GS} is increased much above +0.5 V, however, the gate-to-channel *pn* junction will turn "on" and there will be a substantial flow of gate voltage I_G. This gate current will load down the signal source and produce a voltage drop across the signal source resistance, as shown in Fig. 24.12. This voltage drop can cause V_{GS} to be much smaller than the signal source voltage V_{in}. As V_{in} increases, V_{GS} will ultimately level off at a forward bias voltage of about +0.7 V, and the signal source will lose control over V_{GS}, and hence over I_{DS}. This can result in severe distortion of the input signal in the form of clipping, and thus this situation should be avoided. Thus, although it is possible to increase I_{DS} above I_{DSS} by allowing the gate-to-channel junction to become forward-biased by a small amount ($\leq$0.5 V), the possible benefits are generally far outweighed by the risk of signal distortion. Therefore, JFETs are almost always operated with the gate-to-channel *pn* junction reverse-biased.

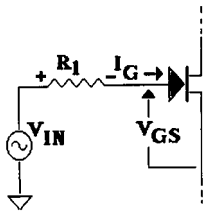


FIGURE 24.12 Effect of forward bias on V_{GS}.

Transfer Conductance

The slope of the transfer curve, dI_{DS}/dV_{GS}, is the *dynamic forward transfer conductance,* or *mutual transfer conductance,* g_m. We see that g_m starts off at zero when $V_{GS} = V_P$ and increases as I_{DS} increases, reaching a maximum when $I_{DS} = I_{DSS}$. Since $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$, g_m can be obtained as

$$g_m = \frac{dI_{DS}}{dV_{GS}} = 2I_{DSS}\frac{\left(1 - \dfrac{V_{GS}}{V_P}\right)}{-V_P}$$

Since

$$1 - \left(\frac{V_{GS}}{V_P}\right) = \sqrt{\frac{I_{DS}}{I_{DSS}}}$$

we have that

$$g_m = 2I_{DSS}\frac{\sqrt{I_{DS}/I_{DSS}}}{-V_P} = 2\frac{\sqrt{I_{DS} \cdot I_{DSS}}}{-V_P}$$

The maximum value of g_m is obtained when $V_{GS} = 0$ ($I_{DS} = I_{DSS}$) and is given by $g_m(V_{GS} = 0) = g_{m0} = 2I_{DS}/(-V_P)$.

Small-Signal AC Voltage Gain

Let's consider the CS amplifier circuit of Fig. 24.13. The input ac signal is applied between gate and source, and the output ac voltage is taken between drain and source. Thus the source electrode of this triode device is common to input and output, hence the designation of this JFET configuration as a CS amplifier.

A good choice of the dc operating point or quiescent point (Q point) for an amplifier is in the middle of the active region at $I_{DS} = I_{DSS}/2$. This allows for the maximum symmetrical drain current swing, from the quiescent level of $I_{DSQ} = I_{DSS}/2$, down to a minimum of $I_{DS} \cong 0$, and up to a maximum of $I_{DS} = I_{DSS}$. This choice for the Q point is also a good one from the standpoint of allowing for an adequate safety margin for the location

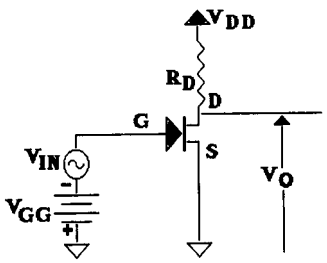


FIGURE 24.13 Common-source amplifier.

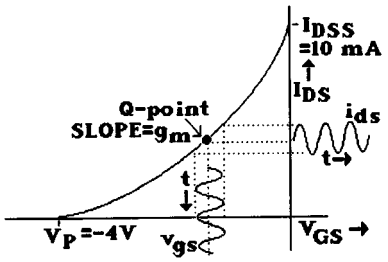


FIGURE 24.14 Transfer characteristic.

of the actual Q point due to the inevitable variations in device and component characteristics and values. This safety margin should keep the Q point well away from the extreme limits of the active region, and thus ensure operation of the JFET in the active region under most conditions. If $I_{DSS} = +10$ mA, then a good choice for the Q point would thus be around +5.0 mA. If $V_P = -4$ V, then

$$g_m = \frac{2\sqrt{I_{DS} \cdot I_{DSS}}}{-V_P} = \frac{2\sqrt{5\text{ mA} \cdot 10\text{ mA}}}{4\text{ V}} = 3.54\text{ mA/V} = 3.54\text{ mS}$$

If a small ac signal voltage v_{GS} is superimposed on the dc gate bias voltage V_{GS}, only a small segment of the transfer characteristic adjacent to the Q point will be traversed, as shown in Fig. 24.14. This small segment will be close to a straight line, and as a result the ac drain current i_{ds} will have a waveform close to that of the ac voltage applied to the gate. The ratio of i_{ds} to v_{GS} will be the slope of the transfer curve as given by $i_{ds}/v_{GS} \cong dI_{DS}/dV_{GS} = g_m$. Thus $i_{ds} \cong g_m v_{GS}$. If the net load driven by the drain of the JFET is the drain load resistor R_D, as shown in Fig. 24.13, then the ac drain current i_{ds} will produce an ac drain voltage of $v_{ds} = -i_{ds} \cdot R_D$. Since $i_{ds} = g_m v_{GS}$, this becomes $v_{ds} = -g_m v_{GS} \cdot R_D$. The ac small-signal voltage gain from gate to drain thus becomes $A_V = v_O/v_{\text{in}} = v_{ds}/v_{GS} = -g_m \cdot R_D$. The negative sign indicates signal inversion as is the case for a CS amplifier.

If the dc drain supply voltage is $V_{DD} = +20$ V, a quiescent drain-to-source voltage of $V_{DSQ} = V_{DD}/2 = +10$ V will result in the JFET being biased in the middle of the active region. Since $I_{DSQ} = +5$ mA in the example under consideration, the voltage drop across the drain load resistor R_D is 10 V. Thus $R_D = 10$ V/5 mA = 2 kΩ. The ac small-signal voltage gain A_V thus becomes $A_V = -g_m \cdot R_D = -3.54$ mS · 2 kΩ = –7.07. Note that the voltage gain is relatively modest as compared to the much larger voltage gains that can be obtained in a bipolar-junction transistor (BJT) common-emitter amplifier. This is due to the lower transfer conductance of both JFETs and MOSFETs (metal-oxide semiconductor field-effect transistors) as compared to BJTs. For a BJT the transfer conductance is given by $g_m = I_C/V_T$, where I_C is the quiescent collector current and $V_T = kT/q \cong 25$ mV is the thermal voltage. At $I_C = 5$ mA, $g_m = 5$ mA/25 mV = 200 mS, as compared to only 3.5 mS for the JFET in this example. With a net load of 2 kΩ, the BJT voltage gain will be –400 as compared to the JFET voltage gain of only 7.1. Thus FETs do have the disadvantage of a much lower transfer conductance, and therefore voltage gain, than BJTs operating under similar quiescent current levels, but they do have the major advantage of a much higher input impedance and a much lower input current. In the case of a JFET the input signal is applied to the *reverse-biased* gate-to-channel *pn* junction and thus sees a very high impedance. In the case of a common-emitter BJT amplifier, the input signal is applied to the *forward-biased* base-emitter junction, and the input impedance is given approximately by $r_{\text{in}} = r_{BE} \cong 1.5 \cdot \beta \cdot V_T/I_C$. If $I_C = 5$ mA and $\beta = 200$, for example, then $r_{\text{in}} \cong 1500\ \Omega$. This moderate input resistance value of 1.5 kΩ is certainly no problem if the signal source resistance is less than around 100 Ω. However, if the source resistance is above 1 kΩ, then there will be a substantial signal loss in the coupling of the signal from the signal source to the base of the transistor. If the source resistance is in the range of above 100 kΩ, and certainly if it is above 1 MΩ, then there will be severe signal attenuation due to the BJT input impedance, and the FET amplifier will probably offer a greater overall voltage gain. Indeed, when high-impedance signal sources are encountered, a multistage amplifier with a FET input stage followed by cascaded BJT stages is often used.

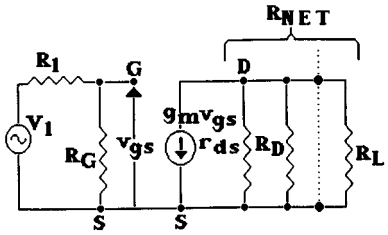


FIGURE 24.15 Effect of r_{ds} on R_{net}.

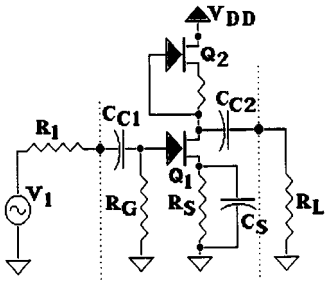


FIGURE 24.16 Active load circuit.

JFET Output Resistance

Dynamic Drain-to-Source Conductance

For the JFET in the active region the drain current I_{DS} is a strong function of the gate-to-source voltage V_{GS} but is relatively independent of the drain-to-source voltage V_{DS}. The transfer equation has previously been stated as $I_{DS} = I_{DSS}\,[1 - (V_{GS}/V_P)]^2$.

The drain current will, however, increase slowly with increasing V_{DS}. To take this dependence of I_{DS} on V_{DS} into account, the transfer equation can be modified to give

$$I_{DS} = I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2\left(1 + \frac{V_{DS}}{V_A}\right)$$

where V_A is a constant called the *Early voltage* and is a parameter of the transistor with units of volts. The early voltage V_A is generally in the range of 30 to 300 V for most JFETs. The variation of the drain current with drain voltage is the result of the *channel length modulation effect* in which the channel length decreases as the drain voltage increases. This decrease in the channel length results in an increase in the drain current. In BJTs a similar effect is the *base width modulation effect.*

The *dynamic drain-to-source conductance* is defined as $g_{ds} = dI_{DS}/dV_{DS}$ and can be obtained from the modified transfer equation $I_{DS} = I_{DSS}\,[1 - (V_{GS}/V_P)]^2\,[1 + V_{DS}/V_A]$ as simply $g_{ds} = I_{DS}/V_A$. The reciprocal of g_{ds} is *dynamic drain-to-source resistance* r_{ds}, so $r_{ds} = 1/g_{ds} = V_A/I_{DS}$. If, for example, $V_A = 100$ V, we have that $r_{ds} = 100\text{ V}/I_{DS}$. At $I_{DS} = 1$ mA, $r_{ds} = 100$ V/1 mA = 100 kΩ, and at $I_{DS} = 10$ mA, $r_{ds} = 10$ kΩ.

Equivalent Circuit Model of CS Amplifier Stage

A small-signal equivalent circuit model of a CS FET amplifier stage is shown in Fig. 24.15. The ac small-signal voltage gain is given by $A_V = -g_m \cdot R_{net}$, where $R_{net} = [r_{ds} \| R_D \| RL]$ is the net load driven by the drain for the FET and includes the dynamic drain-to-source resistance r_{ds}. Since r_{ds} is generally much larger than $[R_D \| R_L]$, it will usually be the case that $R_{net} \cong [R_D \| R_L]$, and r_{ds} can be neglected. There are, however, some cases in which r_{ds} must be taken into account. This is especially true for the case in which an active load is used, as shown in Fig. 24.16. For this case $R_{net} = [r_{ds1} \| r_{ds2} \| R_L]$, and r_{ds} can be a limiting factor in determining the voltage gain.

Consider an example for the active load circuit of Fig. 24.16 for the case of identical JFETs with the same quiescent current. Assume that $R_L >> r_{ds}$ so that $R_{net} \cong [r_{ds1} \| r_{ds2}] = \text{VA}/(2I_{DSQ})$. Let $I_{DSQ} = I_{DSS}/2$, so $g_m = -2\sqrt{I_{DSS} \cdot I_{DSQ}}/(-V_P) = 2\sqrt{2}I_{DSQ}/(-V_P)$. The voltage gain is

$$A_V = -g_m \cdot R_{net} = \frac{2\sqrt{2}I_{DSQ}}{V_P} \times \frac{V_A}{2I_{DSQ}} = \sqrt{2}\,\frac{V_A}{V_P}$$

If $V_A = 100$ V and $V_P = -2$ V, we obtain $A_V = -70$, so we see that with active loads relatively large voltage gains can be obtained with FETs.

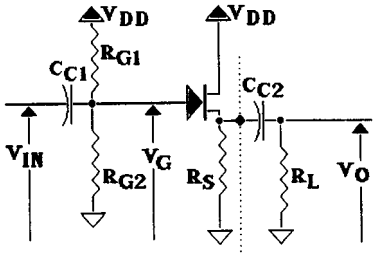


FIGURE 24.17 Source follower.

Another circuit in which the dynamic drain-to-source resistance r_{ds} is important is the constant-current source or current regulator diode. In this case the current regulation is directly proportional to the dynamic drain-to-source resistance.

Source Follower

Source-Follower Voltage Gain

We will now consider the CD JFET configuration, which is also known as the source follower. A basic CD circuit is shown in Fig. 24.17. The input signal is supplied to the gate of the JFET. The output is taken from the source of the JFET, and the drain is connected directly to the V_{DD} supply voltage, which is ac ground.

For the JFET in the active region we have that $i_{ds} = g_m v_{GS}$. For this CD circuit we also have that $v_{GS} = v_G - v_S$ and $v_S = i_{ds} R_{\text{net}}$, where $R_{\text{net}} = [R_S \| R_L]$ is the net load resistance driven by the transistor. Since $v_{GS} = i_{ds}/g_m$, we have that $i_{ds}/g_m = v_G - i_{ds} R_{\text{net}}$. Collecting terms in i_{ds} on the left side yields $i_{ds}[(1/g_m) + R_{\text{net}}] = v_G$, so

$$i_{ds} = \frac{v_G}{(1/g_m) + R_{\text{net}}} = \frac{g_m v_G}{1 + g_m R_{\text{net}}}$$

The output voltage is

$$v_O = v_S = i_{ds} R_{\text{net}} = \frac{g_m R_{\text{net}} v_G}{1 + g_m R_{\text{net}}}$$

and thus the ac small-signal voltage gain is

$$A_V = \frac{v_O}{v_G} = \frac{g_m R_{\text{net}}}{1 + g_m R_{\text{net}}}$$

Upon dividing through by g_m this can be rewritten as

$$A_V = \frac{R_{\text{net}}}{(1/g_m) + R_{\text{net}}}$$

From this we see that the voltage gain will be positive, and thus the source follower is a noninverting amplifier. We also note that A_V will always be less than unity, although for the usual case of $R_{\text{net}} >> 1/g_m$, the voltage gain will be close to unity.

The source follower can be represented as an amplifier with an open-circuit (i.e., no load) voltage transfer ratio of unity and an output resistance of $r_O = 1/g_m$. The equation for A_V can be expressed as $A_V = R_{\text{net}}/(R_{\text{net}} + r_O)$, which is the voltage division ratio of the $r_O = R_{\text{net}}$ circuit.

Source-Follower Examples

Let's consider an example of a JFET with I_{DSS} = 10 mA and V_P = –4 V. Let V_{DD} = +20 V and $I_{DSQ} = I_{DSS}/2$ = 5 mA. For $I_{DS} = I_{DSS}/2$ the value of V_{GS} is –1.17 V. To bias the JFET in the middle of the active region, we will let $V_{GQ} = V_{DD}/2$ = +10 V, so $V_{SQ} = V_{GQ} - V_{GS}$ = +10 V – (–1.17 V) = +11.17 V. Thus $R_S = V_{SQ}/I_{DSQ}$ = 11.17 V/5 mA = 2.23 kΩ.

The transfer conductance at I_{DS} = 5 mA is 3.54 mS so that $r_O = 1/g_m$ = 283 Ω. Since $g_m R_S$ = 7.9, good bias stability will be obtained. If $R_L >> R_S$, then $A_V \cong R_S/(r_O + R_S)$ = 2.23 kΩ/(283 Ω + 2.23 kΩ) = 0.887. If R_L = 1 kΩ, then R_{net} = 690 Ω, and A_V drops to 0.709, and if R_L = 300 Ω, R_{net} = 264 Ω and A_V is down to 0.483. A BJT emitter-follower circuit has the same equations for the voltage gain as the FET source follower. For the BJT case, $r_O = 1/g_m = V_T/I_O$ where V_T = thermal voltage = $kT/q \cong$ 25 mV and I_C is the quiescent collector current. For I_C = 5 mA, we get $r_O \cong$ 25 mV/5 mA = 5 Ω as compared to r_O = 283 Ω for the JFET case at the same quiescent current level. So the emitter follower does have a major advantage over the source follower since it has a much lower output resistance r_O and can thus drive very small load resistances with a voltage gain close to unity. For example, with R_L = 100 Ω, we get $A_V \cong$ 0.26 for the source follower as compared to $A_V \cong$ 0.95 for the emitter follower.

The FET source follower does, however, offer substantial advantages over the emitter follower of a much higher input resistance and a much lower input current. For the case in which a very high-impedance source, up in the megohm range, is to be coupled to a low-impedance load down in the range of 100 Ω or less, a good combination to consider is that of a cascaded FET source follower followed by a BJT emitter follower. This combination offers the very high input resistance of the source follower and the very low output resistance of the emitter follower.

For the source-follower circuit under consideration the input resistance will be $R_{in} = [R_{G1} \| R_{G2}]$ = 10 MΩ. If the JFET gate current is specified as 1 nA (max), and for good bias stability the change in gate voltage due to the gate current should not exceed $|V_P|/10$ = 0.4 V, the maximum allowable value for $[R_{G1} \| R_{G2}]$ is given by $I_G \cdot [R_{G1} \| R_{G2}] < 0.4$ V. Thus $[R_{G1} \| R_{G2}] < 0.4$ V/1 nA = 0.4 GΩ = 400 MΩ. Therefore R_{G1} and R_{G2} can each be allowed to be as large as 800 MΩ, and very large values for R_{in} can thus be obtained. At higher frequencies the input capacitance C_{in} must be considered, and C_{in} will ultimately limit the input impedance of the circuit. Since the input capacitance of the FET will be comparable to that of the BJT, the advantage of the FET source follower over the BJT emitter follower from the standpoint of input impedance will be obtained only at relatively low frequencies.

Source-Follower Frequency Response

The input capacitance of the source follower is given by $C_{in} = C_{GD} + (1 - A_V)C_{GS}$. Since A_V is close to unity, C_{in} will be approximately given by $C_{in} \cong C_{GD}$. The source-follower input capacitance can, however, be reduced below C_{GD} by a bootstrapping circuit in which the drain voltage is made to follow the gate voltage. Let's consider a representative example in which C_{GD} = 5 pF, and let the signal-source output resistance be R_1 = 100 kΩ. The input circuit is in the form of a simple *RC* low-pass network. The *RC* time constant is

$$\tau = [R \| R_{G1} \| R_{G2}] \cdot C_{in} \cong R_1 \cdot C_{in} \cong R_1 \cdot C_{GD}$$

Thus $\tau \cong$ 100 kΩ · 5 pF = 500 ns = 0.5 μs. The corresponding 3-dB or half-power frequency is $f_H = 1/(2\pi\tau)$ = 318 kHz. If R_1 = 1 MΩ, the 3-dB frequency will be down to about 30 kHz. Thus we see indeed the limitation on the frequency response that is due to the input capacitance.

Frequency and Time-Domain Response

Small-Signal CS Model for High-Frequency Response

We will now consider the frequency- and time-domain response of the JFET CS amplifier. In Fig. 24.18 an ac representation of a CS amplifier is shown, the dc biasing not being shown. In Fig. 24.19 the JFET small-signal ac equivalent circuit model is shown including the junction capacitances C_{GS} and C_{GD}. The gate-to-drain capacitance C_{GD} is a feedback capacitance in that it is connected between output (drain) and input (gate). Using

Miller's theorem for shunt feedback this feedback capacitance can be transformed into an equivalent input capacitance $C_{GD'} = (1 - A_V)C_{GD}$ and an equivalent output capacitance $C_{GD''} = (1 - 1/A_V)C_{GD}$, as shown in Fig. 24.20. The net input capacitance is now $C_{in} = C_{GS} + (1 - A_V)C_{GD}$ and the net output capacitance is $C_O = (1 - 1/A_V)C_{GD} + C_L$. Since the voltage gain A_V is given by $A_V = -g_m R_{net}$, where R_{net} represents the net load resistance, the equations for C_{in} and C_O can be written approximately as $C_{in} = C_{GS} + (1 + g_m R_{net})C_{GD}$ and $C_O = [1 + 1/(g_m R_{net})]C_{GD} + C_L$. Since usually $A_V = g_m R_{net} >> 1$, C_O can be written as $C_O \cong C_{GD} + C_L$. Note that the voltage gain given by $A_V = -g_m R_{net}$ is not valid in the higher frequency, where A_V will decrease with increasing frequency. Therefore the expressions for C_{in} and C_O will not be exact but will still be a useful approximation for the determination of the frequency- and time-domain responses. We also note that the contribution of C_{GD} to the input capacitance is increased by the Miller effect factor of $1 + g_m R_{net}$.

The circuit in Fig. 24.21 is in the form of two cascaded RC low-pass networks. The RC time constant on the input side is $\tau_1 = [R_1 \| R_G] \cdot C_{in} \cong R_1 \cdot C_{in}$, where R_1 is the signal-source resistance. The RC time constant on the output side is given by $\tau_2 = R_{net} \cdot C_O$. The corresponding breakpoint frequencies are

$$f_1 = \frac{1}{2\pi\tau_1} = \frac{1}{2\pi R_1 \cdot C_{in}}$$

and

$$f_2 = \frac{1}{2\pi\tau_2} = \frac{1}{2\pi R_{net} \cdot C_O}$$

The 3-dB or half-power frequency of this amplifier stage will be a function of f_1 and f_2. If these two breakpoint frequencies are separated by at least a decade (i.e., 10:1 ratio), the 3-dB frequency will be approximately equal to the lower of the two breakpoint frequencies. If the breakpoint frequencies are not well separated, then the 3-dB frequency can be obtained from the following approximate relationship: $(1/f_{3dB})^2 \cong (1/f_1)^2 + (1/f_2)^2$. The time-domain response as expressed in terms of the 10 to 90% rise time is related to the frequency-domain response by the approximate relationship that $t_{rise} \cong 0.35/f_{3dB}$.

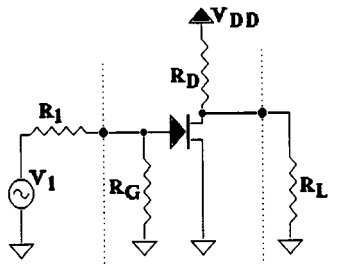


FIGURE 24.18 Common-source amplifier.

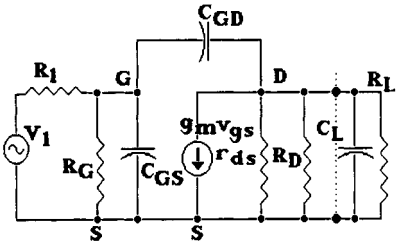


FIGURE 24.19 AC small-signal model.

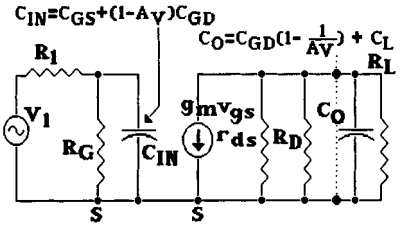


FIGURE 24.20

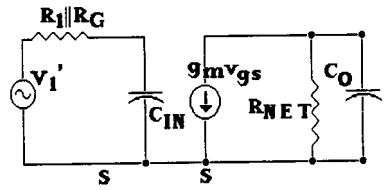


FIGURE 24.21

We will now consider a representative example. We will let $C_{GS} = 10$ pF and $C_{GD} = 5$ pF. We will assume that the net load driven by the drain of the transistors is $R_{net} = 2$ kΩ and $C_L = 10$ pF. The signal-source resistance $R_1 = 100$ Ω. The JFET will have $I_{DSS} = 10$ mA, $I_{DSQ} = I_{DSS}/2 = 5$ mA, and $V_P = -4$ V, so $g_m = 3.535$ mS. Thus the midfrequency gain is $A_V = -g_m R_{net} = -3.535$ mS · 2 kΩ = –7.07. Therefore we have that

$$C_{in} \cong C_{GS} + (1 + g_m R_{net})C_{GD} = 10 \text{ pF} + 8.07 \cdot 5 \text{ pF} = 50.4 \text{ pF}$$

and

$$C_O \cong C_{GD} + C_L = 15 \text{ pF}$$

Thus $\tau_1 = R_1 \cdot C_{in} = 100\ \Omega \cdot 50.4$ pF = 5040 ps = 5.04 ns, and $\tau_2 = R_{net} \cdot C_O = 2$ k$\Omega \cdot 15$ pF = 30 ns. The corresponding breakpoint frequencies are $f_1 = 1/(2\pi \cdot 5.04\text{ ns}) = 31.6$ MHz and $f_2 = 1/(2\pi \cdot 30\text{ ns}) = 5.3$ MHz. The 3-dB frequency of the amplifier can be obtained from $(1/f_{3dB})^2 \cong (1/f_1)^2 + (1/f_2)^2 = (1/31.6\text{ MHz})^2 + (1/5.3\text{ MHz})^2$, which gives $f_{3dB} \cong 5.2$ MHz. The 10 to 90% rise time can be obtained from $t_{rise} \cong 0.35/f_{3dB} = 0.35/5.2$ MHz = 67 ns.

In the preceding example the dominant time constant is the output circuit time constant of $\tau_2 = 30$ ns due to the combination of load resistance and output capacitance. If we now consider a signal-source resistance of 1 kΩ, the input circuit time constant will be $\tau_1 = R_1 \cdot C_{in} = 1000\ \Omega \cdot 50.4$ pF = 50.4 ns. The corresponding breakpoint frequencies are $f_1 = 1/(2\pi \cdot 50.4\text{ ns}) = 3.16$ MHz and $f_2 = 1/(2\pi \cdot 30\text{ ns}) = 5.3$ MHz. The 3-dB frequency is now $f_{3dB} \cong 2.7$ MHz, and the rise time is $t_{rise} \cong 129$ ns. If R_1 is further increased to 10 kΩ, we obtain $\tau_1 = R_1 \cdot C_{in} = 10$ k$\Omega \cdot 50.4$ pF = 504 ns, giving breakpoint frequencies of $f_1 = 1/(2\pi \cdot 504\text{ ns}) = 316$ kHz and $f_2 = 1/(2\pi \cdot 30\text{ ns}) = 5.3$ MHz. Now τ_1 is clearly the dominant time constant, the 3-dB frequency is now down to $f_{3dB} \cong f_1 = 316$ kHz, and the rise time is up to $t_{rise} \cong 1.1$ µs. Finally, for the case of $R_1 = 1$ MΩ, the 3-dB frequency will be only 3.16 kHz and the rise time will be 111 µs.

Use of Source Follower for Impedance Transformation

We see that large values of signal-source resistance can seriously limit the amplifier bandwidth and increase the rise time. In these cases, the use of an impedance transforming circuit such as an FET source follower or a BJT emitter follower can be very useful. Let's consider the use of a source follower as shown in Fig. 24.22. We will assume that both FETs are identical to the one in the preceding examples and are biased at $I_{DSQ} = 5$ mA. The source follower Q_1 will have an input capacitance of $C_{in} = C_{GD} + (1 - A_{V1})C_{GS} \cong C_{GD} = 5$ pF, since A_V will be very close to unity for a source follower that is driving a CS amplifier. The source-follower output resistance will be $r_O = 1/g_m = 1/3.535$ mS = 283 Ω. Let's again consider the case of $R_1 = 1$ MΩ. The time constant due to the combination of R_1 and the input capacitance of the source follower is $\tau_{SF} = 1$ M$\Omega \cdot 5$ pf = 5 µs. The time constant due to the combination of the source-follower output resistance r_O and the input capacitance of the CS stage is $\tau_1 = r_O \cdot C_{in} = 283\ \Omega \cdot 50.4$ pF = 14 ns, and the time constant of the output circuit is $\tau_2 = 30$ ns, as before. The breakpoint frequencies are $f_{SF} = 31.8$ kHz, $f_1 = 11$ MHz, and $f_2 = 5.3$ MHz. The 3-dB frequency of the system is now $f_{3dB} \cong f_{SF} = 31.8$ kHz, and the rise time is $t_{rise} \cong 11$ µs. The use of the source follower thus results in an improvement by a factor of 10:1 over the preceding circuit.

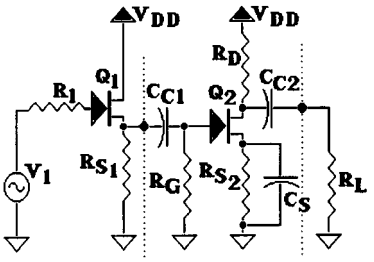


FIGURE 24.22

Voltage-Variable Resistor

Operation of a JFET as a Voltage-Variable Resistor

We will now consider the operation of a JFET as a voltage-variable resistor (VVR). A JFET can be used as a VVR in which the drain-to-source resistance r_{ds} of the JFET can be varied by variation of V_{GS}. For values of $V_{DS} << V_P$ the I_{DS} vs. V_{DS} characteristics are approximately linear, so the JFET looks like a resistor, the resistance value of which can be varied by the gate voltage as shown in Fig. 24.23.

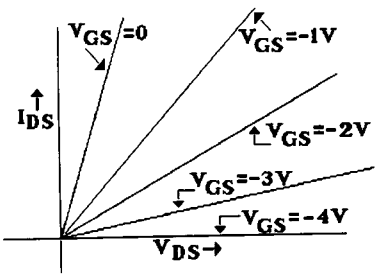
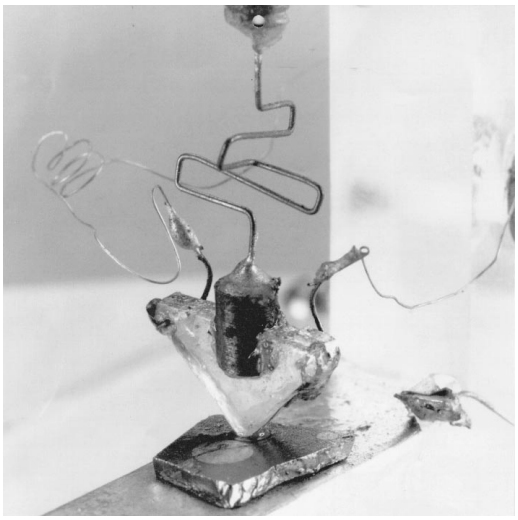


FIGURE 24.23

The channel conductance in the region where $V_{DS} << V_P$ is given by $g_{ds} = A\sigma/L = WH\sigma/L$, where the channel height H is given by $H = H_0 - 2W_D$. In this equation W_D is the depletion region width and H_0 is the value of H as $W_D \rightarrow 0$. The depletion region width is given by $W_D = K\sqrt{V_J} = K\sqrt{V_{GS} + \phi}$, where K is a constant, V_J is the junction voltage, and ϕ is the *pn*-junction contact potential (typically around 0.8 to 1.0 V). As V_{GS} increases, W_D increases and the channel height H decreases as given by $H = H_0 - 2K\sqrt{V_{GS} + \phi}$. When $V_{GS} = V_P$, the channel is completely pinched off, so $H = 0$ and thus $2K\sqrt{V_P + \phi} = H_0$. Therefore $2K = H_0/\sqrt{V_P + \phi}$, and thus

The Invention of the Transistor

In 1907, the American Telephone and Telegraph Company (AT&T) and the Western Electric Company combined their engineering departments and established the Bell Telephone Laboratory on West Street in New York City. By 1921, the laboratories constituted the largest industrial research organization in the country, occupying 400,000 square feet in a 13-story building in lower Manhattan and employing more than 1500 men and women. The organization was put on a more formal footing in 1925, when Frank B. Jewett was made President of Bell Telephone Laboratories, Inc. In the following decades, the labs distinguished themselves by contributions not only to communications technology, but to basic science as well. The awarding of the Nobel Prize in Physics to Clinton J. Davisson in 1937 was simply the most prominent recognition of the laboratories' scientific work.



The first point-contact transistor developed at Bell Labs in 1947 by John Bardeen, William Shockley, and Walter Brattain, had a thin gold foil along the sides of a polystyrene triangle. The foil was slit at the triangle's apex and was pressed against a piece of germanium by the metal piece at the top of the photo (Photo courtesy of AT&T Bell Laboratories.)

The true importance of the fusion of science and engineering in the industrial laboratory was made apparent to all in the years after World War II. In 1947, three Bell Labs physicist-engineers produced the single most significant electronic invention of the era—the transistor. John Bardeen, Walter Brattain, and William Shockley were consciously seeking to exploit new technology about the behavior of semiconducting materials when they devised a way to make a crystal of germanium do the work of a triode vacuum tube, the most basic of electronic components.

$$H = H_0 - H_0 \frac{\sqrt{V_{GS} + \phi}}{\sqrt{V_P + \phi}} = H_0 \left(1 - \frac{\sqrt{V_{GS} + \phi}}{\sqrt{V_P + \phi}} \right)$$

For g_{ds} we now have

$$g_{ds} = \frac{\sigma WH}{L} = \sigma \frac{WH_0}{L} \left(1 - \frac{\sqrt{V_{GS} + \phi}}{\sqrt{V_P + \phi}} \right)$$

When $V_{GS} = 0$, the channel is fully open or "on," and

$$g_{ds} = g_{ds}(\text{on}) = \sigma \frac{WH_0}{L} \left(1 - \frac{\sqrt{\phi}}{\sqrt{V_P + \phi}} \right)$$

AT&T Bell Laboratories

This photograph taken in 1948 is of the three Bell Labs physicist-engineers, John Bardeen, William Shockley, and Walter Brattain, who invented the first transistor. (Photo courtesy of AT&T Bell Laboratories.)

Their work built on the research of many before them, and much had to be done before the transistor and the solid-state devices that followed could become practical engineering tools, but in retrospect it is clear that the transistor gave the engineer the key to a whole new electronic world. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

The drain-to-source conductance can now be expressed as

$$g_{ds} = g_{ds}(\text{on}) \frac{1 - \left(\sqrt{V_{GS} + \phi} \Big/ \sqrt{V_P + \phi}\right)}{1 - \left(\sqrt{\phi} \Big/ \sqrt{V_P + \phi}\right)}$$

The reciprocal quantity is the drain-to-source resistance r_{ds} as given by $r_{ds} = 1/g_{ds}$, and $r_{ds}(\text{on}) = 1/g_{ds}(\text{on})$, so

$$r_{ds} = r_{ds}(\text{on}) \frac{1 - \left(\sqrt{\phi} \Big/ \sqrt{V_P + \phi}\right)}{1 - \left(\sqrt{V_{GS} + \phi} \Big/ \sqrt{V_P + \phi}\right)}$$

As $V_{GS} \rightarrow 0$, $r_{ds} \rightarrow r_{ds}(\text{on})$, and as $V_{GS} \rightarrow V_P$, $r_{ds} \rightarrow \infty$. This latter condition corresponds to the channel being pinched off in its entirety all the way from source to drain. This is like having a big block of insulator (i.e., the depletion region) between source and drain. When $V_{GS} = 0$, r_{ds} is reduced to its minimum value of $r_{ds}(\text{on})$,

which for most JFETs is in the 20- to 400-Ω range. At the other extreme, when $V_{GS} > V_P$, the drain-to-source current I_{DS} is reduced to a very small value, generally down into the low nanoampere or even picoampere range. The corresponding value of r_{ds} is not really infinite but is very large, generally well up into the gigaohm (1000 MΩ) range. Thus by variation of V_{GS}, the drain-to-source resistance can be varied over a very wide range. As long as the gate-to-channel junction is reverse-biased, the gate current will be very small, generally down in the low nanoampere or even picoampere range, so the gate as a control electrode draws very little current. Since V_P is generally in the 2- to 5-V range for most JFETs, the V_{DS} values required to operate the JFET in the VVR range are generally <0.1 V. In Fig. 24.23 the VVR region of the JFET I_{DS} vs. V_{DS} characteristics is shown.

VVR Applications

Applications of VVRs include automatic gain control (AGC) circuits, electronic attenuators, electronically variable filters, and oscillator amplitude control circuits.

When using a JFET as a VVR, it is necessary to limit V_{DS} to values that are small compared to V_P to maintain good linearity. In addition V_{GS} should preferably not exceed 0.8 V_P for good linearity, control, and stability. This limitation corresponds to an r_{ds} resistance ratio of about 10:1. As V_{GS} approaches V_P, a small change in V_P can produce a large change in r_{ds}. Thus unit-to-unit variations in V_P as well as changes in V_P with temperature can result in large changes in r_{ds} as V_{GS} approaches V_P.

The drain-to-source resistance r_{ds} will have a temperature coefficient (TC) due to two causes: (1) the variation of the channel resistivity with temperature and (2) the temperature variation of V_P. The TC of the channel resistivity is positive, whereas the TC of V_P is negative due to the negative TC of the contact potential ϕ. The positive TC of the channel resistivity will contribute to a positive TC of r_{ds}. The negative TC of V_P will contribute to a negative TC of r_{ds}. At small values of V_{GS}, the dominant contribution to the TC is the positive TC of the channel resistivity, so r_{ds} will have a positive TC. As V_{GS} gets larger, the negative TC contribution of V_P becomes increasingly important, and there will be a value of V_{GS} at which the net TC of r_{ds} is zero, and above this value of V_{GS} the TC will be negative. The TC of r_{ds}(on) is typically +0.3%/°C for *n*-channel JFETs and +0.7%/°C for *p*-channel JFETs. For example, for a typical JFET with an r_{ds}(on) = 500 Ω at 25°C and V_P = 2.6 V, the zero TC point will occur at V_{GS} = 2.0 V. Any JFET can be used as a VVR, although there are JFETs that are specifically made for this application.

A simple example of a VVR application is the electronic gain control circuit of Fig. 24.24. The voltage gain is given by $A_V = 1 + (R_F/r_{ds})$. If, for example, R_F = 19 kΩ and r_{ds}(on) = 1 kΩ, then the maximum gain will be $A_{V\max} = 1 + [R_F/r_{ds}(\text{on})] = 20$. As V_{GS} approaches V_P, r_{ds} will increase and become very large such that $r_{ds} >> R_F$, so that A_V will decrease to a minimum value of close to unity. Thus the gain can be varied over a 20:1 ratio. Note that $V_{DS} \cong V_{\text{in}}$, so to minimize distortion the input signal amplitude should be small compared to V_P.

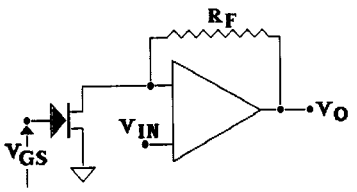


FIGURE 24.24 Electronic gain control.

Defining Terms

Active region: The region of JFET operation in which the channel is pinched off at the drain end but still open at the source end such that the drain-to-source current I_{DS} approximately saturates. The condition for this is that $|V_{GS}| < |V_P|$ and $|V_{DS}| > |V_P|$. The active region is also known as the saturated region.

Ohmic, nonsaturated, or triode region: The three terms all refer to the region of JFET operation in which a conducting channel exists all the way between source and drain. In this region the drain current varies with both V_{GS} and V_{DS}.

Drain saturation current, I_{DSS}: The drain-to-source current flow through the JFET under the conditions that $V_{GS} = 0$ and $|V_{DS}| > |V_P|$ such that the JFET is operating in the active or saturated region.

Pinch-off voltage, V_P: The voltage that when applied across the gate-to-channel *pn* junction will cause the conducting channel between drain and source to become pinched off. This is also represented as V_{GS} (off).

Related Topic

28.1 Large Signal Analysis

References

R. Mauro, *Engineering Electronics,* Englewood Cliffs, N.J.: Prentice-Hall, 1989, pp. 199–260.

J. Millman and A. Grabel, *Microelectronics,* 2nd ed., New York: McGraw-Hill, 1987, pp. 133–167, 425–429.

F. H. Mitchell, Jr. and F.H. Mitchell, Sr., *Introduction to Electronics Design,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1992, pp. 275–328.

C.J. Savant, M.S. Roden, and G.L. Carpenter, *Electronic Design,* 2nd ed., Menlo Park, Calif.: Benjamin-Cummings, 1991, pp. 171–208.

A.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., Philadelphia: Saunders, 1991, pp. 322–361.

24.2 Bipolar Transistors

Joseph Watson

Modern amplifiers abound in the form of *integrated circuits* (ICs), which contain transistors, diodes, and other structures diffused into single-crystal *dice.* As an introduction to these ICs, it is convenient to examine sing-transistor amplifiers, which in fact are also widely used in their own right as *discrete* circuits — and indeed much more complicated discrete signal-conditioning circuits are frequently found following sensors of various sorts.

There are two basic forms of transistor, the *bipolar* family and the *field-effect* family, and both appear in ICs. They differ in their modes of operation but may be incorporated into circuits in quite similar ways. To understand elementary circuits, there is no need to become too familiar with the physics of transistors, but some basic facts about their electrical properties must be known.

Consider the bipolar transistor, of which there are two types, *npn* and *pnp.* Electrically, they differ only in terms of current direction and voltage polarity. Figure 24.25(a) illustrates the idealized structure of an *npn* transistor, and diagram (b) implies that it corresponds to a pair of diodes with three leads. This representation does *not* convey sufficient information about the actual operation of the transistor, but it does make the point that the flow of conventional current (positive to negative) is easy from the *base* to the *emitter,* since it passes through a *forward-biased diode,* but difficult from the *collector* to the *base,* because flow is prevented by a *reverse-biased diode.*

Figure 24.25(c) gives the standard symbol for the *npn* transistor, and diagram (d) defines the direction of current flow and the voltage polarities observed when the device is in operation. Finally, diagram (e) shows that for the *pnp* transistor, all these directions are reversed and the polarities are inverted.

For a transistor, there is a main current flow between the collector and the emitter, and a very much smaller current flow between the base and the emitter. So, the following relations may be written:

$$I_E = I_C + I_B \tag{24.1}$$

(Note that the arrow on the transistor symbol defines the emitter and the direction of current flow—*out* for the *npn* device, and *in* for the *pnp.*) Also

$$I_C / I_B = h_{FE} \tag{24.2}$$

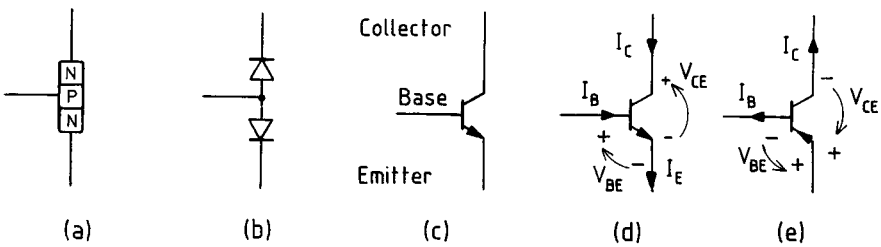


FIGURE 24.25 The bipolar transistor. (a) to (d) *npn* transistor; (e) *pnp* transistor.

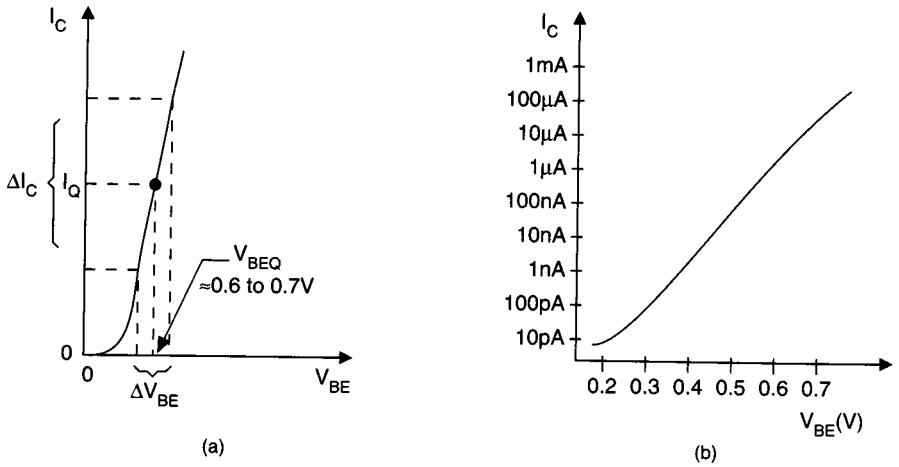


FIGURE 24.26 The transconductance curve for a transistor on (a) linear and (b) logarithmic axes.

Here, h_{FE} is called the *dc common-emitter current gain,* and because $I_C >> I_B$, then h_{FE} is large, typically 50 to 300. The implication of this may be seen immediately: if the small current I_B can be used to control the large current I_C, then the transistor may obviously be used as a current amplifier. [This is why Fig. 24.25(b) is inadequate—it completely neglects this all-important current-gain property of the transistor.] Furthermore, if a load resistance is connected into the collector circuit, it will become a voltage amplifier, too.

Unfortunately, h_{FE} is an ill-defined quantity and varies not only from transistor to transistor but also changes with temperature. The relationship between the base-emitter voltage V_{BE} and the collector current is much better defined and follows an exponential law closely over at least eight decades. This relationship is shown in both linear and logarithmic form in Fig. 24.26. Because the output current I_C is dependent upon the input voltage V_{BE}, the plot must be a transfer conductance or *transconductance* characteristic. The relevant law is

$$I_C = I_{ES}(e^{(q/kT)V_{BE}} - 1) \tag{24.3}$$

Here, I_{ES} is an extremely small leakage current internal to the transistor, q is the electronic charge, k is Boltzmann's constant, and T is the absolute temperature in kelvins. Usually, kT/q is called V_T and is about 26 mV at a room temperature of 25°C. This implies that for any value of V_{BE} over about 100 mV, then $\exp(V_{BE}/V_T) >> 1$, and for all normal operating conditions, Eq. (24.3) reduces to

$$I_C = I_{ES}e^{V_{BE}/V_T} \qquad \text{for } V_{BE} > 100 \text{ mV} \tag{24.4}$$

The term "normal operating conditions" is easily interpreted from Fig. 24.26(a), which shows that when V_{BE} has reached about 0.6 to 0.7 V, any small fluctuations in its value cause major fluctuations in I_C. This situation is illustrated by the dashed lines enclosing ΔV_{BE} and ΔI_C, and it implies that to use the transistor as an amplifier, working values of V_{BE} and I_C must be established, after which signals may be regarded as fluctuations around these values.

Under these *quiescent, operating,* or *working* conditions,

$$I_C = I_Q \qquad \text{and} \qquad V_{CE} = V_Q$$

and methods of defining these quiescent or operating conditions are called *biasing.*

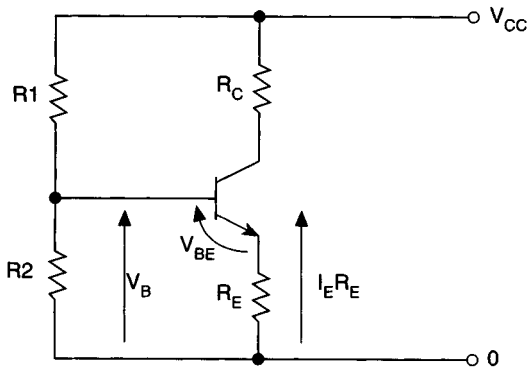


FIGURE 24.27 A transistor biasing circuit.

Biasing the Bipolar Transistor

A fairly obvious way to bias the transistor is to first establish a constant voltage V_B using a potential divider $R1$ and $R2$ as shown in the **biasing circuit** of Fig. 24.27. Here,

$$V_B \simeq \frac{V_{CC} R2}{R1 + R2}$$

if I_B is very small compared with the current through $R2$, which is usual. If it is not, this fact must be taken into account.

This voltage will be much greater than V_{BE} if a realistic power supply is used along with realistic values of $R1$ and $R2$. Hence, when the transistor is connected into the circuit, an emitter resistor must also be included so that

$$V_{BE} = V_B - I_E R_E \tag{24.5}$$

Now consider what happens when the power supply is connected. As V_B appears, a current I_B flows into the base and produces a much larger current $I_C = h_{FE} I_B$ in the collector. These currents add in the emitter to give

$$I_E = I_B + h_{FE} I_B = (1 + h_{FE}) I_B \simeq h_{FE} I_B \tag{24.6}$$

Clearly, I_E will build up until a fixed or quiescent value of base-emitter voltage V_{BEQ} appears. Should I_E try to build up further, V_{BE} will fall according to Eq. (24.5) and, hence, so will I_E. Conversely, should I_E not build up enough, V_{BE} will increase until it does so.

This is actually a case of current-derived negative feedback, and it successfully holds the collector current near the quiescent value I_Q. Furthermore, it does so in spite of different transistors with different values of h_{FE} being used and in spite of temperature variations. Actually, V_{BE} itself falls with temperature at about –2.2 mV/°C for constant I_C, and the circuit will compensate for this, too. The degree of success of the negative feedback in holding I_Q constant is called the *bias stability.*

This is one example of a **common-emitter** (CE) circuit, so-called because the emitter is the common terminal for both base and collector currents. The behavior of the transistor in such a circuit may be illustrated by superimposing a *load line* on the *output characteristics* of the transistor, as shown in Fig. 24.28.

If the collector current I_C is plotted against the collector-to-emitter voltage V_{CE}, a family of curves for various fixed values of V_{BE} or I_B results, as in Fig. 24.28. These curves show that as V_{CE} increases, I_C rises very rapidly and then turns over as it is limited by I_B. In the CE circuit, if I_B were reduced to zero, then I_C would also be

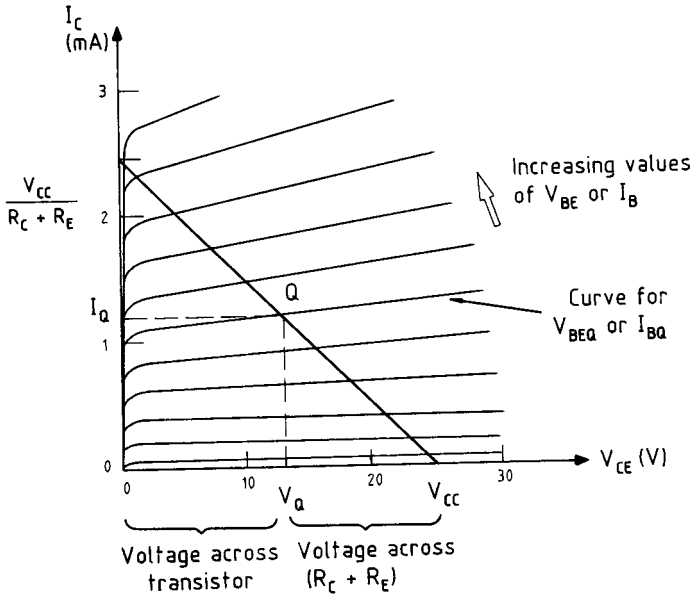


FIGURE 24.28 The load-line diagram.

zero (apart from a small leakage current I_{CE0}). Hence there would be no voltage drop in either R_C or R_E, and practically all of V_{CC} would appear across the transistor. That is, under *cut-off* conditions,

$$V_{CE} \rightarrow V_{CC} \qquad \text{for } I_B = 0 \tag{24.7}$$

Conversely, if I_B were large, I_C would be very large, almost all of V_{CC} would be dropped across $R_C + R_E$ and

$$I_C \rightarrow \frac{V_{CC}}{R_C + R_E} \qquad \text{for large } I_B \tag{24.8}$$

Actually, because the initial rise in I_C for the transistor is not quite vertical, there is always a small *saturation voltage* V_{CES} across the transistor under these conditions, where V_{CES} means the voltage across the transistor in the common-emitter mode when saturated. In this saturated condition $V_{CES} \simeq 0.3$ V for small silicon transistors. Both these conditions are shown in Fig. 24.28.

From the circuit of Fig. 24.27,

$$V_{CE} = V_{CC} - I_C(R_C + R_E) \tag{24.9a}$$

which may be rewritten as

$$I_C = -V_{CE}/(R_C + R_E) + V_{CC}/(R_C + R_E) \tag{24.9b}$$

This is the straight-line equation to the *dc load-line* (compare $y = mx + c$), showing that its slope is $-1/(R_C + R_E)$ and that it crosses the I_C axis at $V_{CC}/(R_C + R_E)$ as expected. The actual position of a point is determined by where this load line crosses the output characteristic in use, that is, by what value of V_{BE} or I_B is chosen. For example, the quiescent point for the transistor is where the load line crosses the output curve defined by $V_{BE} = V_{BEQ}$ (or $I_B = I_{BQ}$) to give $V_{CE} = V_Q$ and $I_C = I_Q$.

Note that because the transistor is *nonohmic* (that is, it does not obey Ohm's law), the voltage across it may only be determined by using the (ohmic) voltage drop across the resistors R_C and R_E according to Eq. (24.9). At the quiescent point this is

$$V_Q = V_{CC} - I_Q(R_C + R_E)$$

A design example will illustrate typical values involved with a small-transistor CE stage.

Example 1

A transistor is to be biased at a collector current of 1 mA when a 12-V power supply is applied. Using the circuit of Fig. 24.27, determine the values of $R1$, $R2$, and R_E if 3.4 V is to be dropped across R_E and if the current through $R2$ is to be 10 I_{BQ}. Assume that for the transistor used, $V_{BEQ} = 0.6$ V and $h_{FE} = 100$.

Solution. In this circuit $I_Q = 1$ mA $\simeq I_E$ (because $I_B << I_C$). Hence

$$R_E = \frac{V_{R_E}}{I_Q} = \frac{3.4}{1} = 3.4\ \text{k}\Omega$$

Also, $V_B = V_{RE} + V_{BE} = 3.4 + 0.6 = 4$ V. This gives

$$R2 = \frac{V_B}{10\ I_{BQ}}$$

where $I_{BQ} = I_Q/h_{FE} = 1/100 = 0.01$ mA, so

$$R2 = \frac{4}{10 \times 0.01} = 40\ \text{k}\Omega$$

Now $V_{R1} = V_{CC} - V_B = 12 - 4 = 8$ V, and the current through $R1$ is 10 $I_{BQ} + I_{BQ} = 11\ I_{BQ}$, so

$$R1 = \frac{V_{R1}}{I_{R1}} = \frac{8}{11 \times 0.01} = 72.7\ \text{k}\Omega$$

In the above design example, the base current I_{BQ} has been included in the current passing through $R1$. Had this not been done, $R1$ would have worked out at 80 kΩ. Usually, this difference is not very important because *discrete* (or individual) resistors are available only in a series of nominal values, and each of these is subject to a *tolerance,* including 10, 5, 2, and 1%.

In the present case, the following (5%) values could reasonably be chosen:

$$R_E = 3.3\ \text{k}\Omega \qquad R1 = 75\ \text{k}\Omega \qquad R2 = 39\ \text{k}\Omega$$

All this means that I_Q cannot be predetermined very accurately, but the circuit nevertheless settles down to a value close to the chosen one, and, most importantly, stays there almost irrespective of the transistor used and the ambient temperature encountered.

Having biased the transistor into an operating condition, it is possible to consider *small-signal operation.*

Small-Signal Operation

In the biasing circuit of Fig. 24.27, the collector resistor R_C had no discernible function, because it is simply the load resistor across which the signal output voltage is developed. However, it was included because it also drops a voltage due to the bias current flowing through it. This means that its value must not be so large that it robs the transistor of adequate operating voltage; that is, it must not be responsible for moving the operating point too far to the left in Fig. 24.28.

If the chosen bias current and voltage are I_Q and V_Q, then small signals are actually only fluctuations in these bias (or average) values that can be separated from them using coupling capacitors.

To inject an input signal to the base, causing V_{BE} and I_B to fluctuate by v_{be} and i_b, a signal source must be connected between the base and the common or zero line (also usually called ground or earth whether it is actually connected to ground or not!). However, most signal sources present a resistive path through themselves, which would shunt $R2$ and so change, or even destroy, the bias conditions. Hence, a coupling capacitor C_c must be included, as shown in Fig. 24.29, in series with a signal source represented by a Thévenin equivalent.

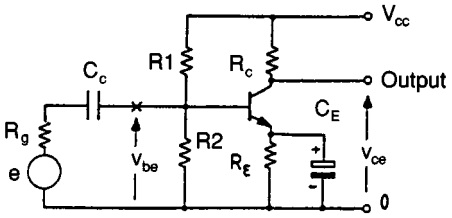


FIGURE 24.29 A complete common-emitter stage.

The emitter resistor R_E was included for biasing reasons (although there are other bias circuits that omit it), but for signal amplification purposes it must be shunted by a high-value capacitor C_E so that the signal current can flow down to ground without producing a signal voltage drop leading to negative feedback (as did the bias current). The value of C_E must be much greater than is apparent at first sight, and this point will be developed later; for the present, it will be assumed that it is large enough to constitute a short circuit at all the signal frequencies of interest. So, for ac signals R_E is short-circuited and only R_C acts as a load. This implies that a *signal* or *ac load line* comes into operation with a slope of $-1/R_C$, as shown in Fig. 24.30.

The ways in which the small-signal quantities fluctuate may now be examined. If v_{be} goes positive, this actually means that V_{BE} increases a little. This in turn implies that I_C increases by an amount i_c, so the voltage drop in R_C increases by v_{ce}. Keeping in mind that the top of R_C is held at a constant voltage, this means that the voltage at the bottom of R_C must fall by v_{ce}. This very important point shows that because v_{ce} falls as v_{be} rises, there is 180° phase shift through the stage. That is, the CE stage is an *inverting voltage amplifier.* However, because i_c increases into the collector as i_b increases into the base, it is also a *noninverting current amplifier.*

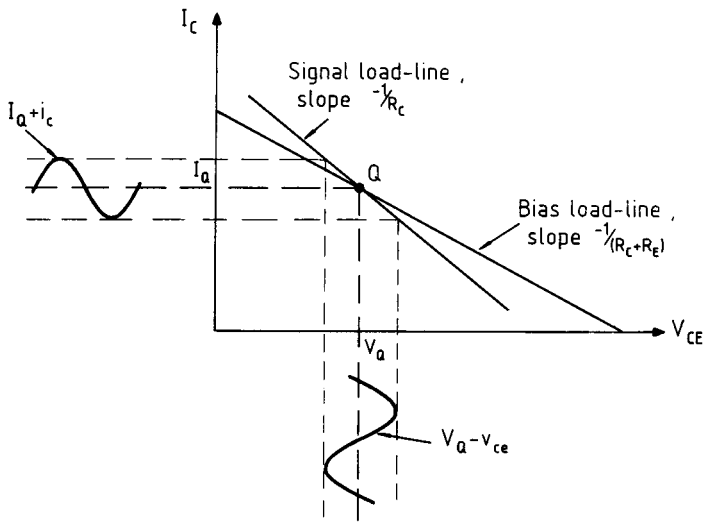


FIGURE 24.30 The signal or ac load line.

Now consider the amount by which v_{ce} changes with v_{be}, which is the *terminal voltage gain* of the stage. In Fig. 24.26, the slope of the transconductance curve at any point defines by how much I_C changes with a fluctuation in V_{BE}. That is, it gives the ratio i_c/v_{be} at any operating point Q. Equation (24.4) is

$$I_C = I_{ES} e^{V_{BE}/V_T}$$

so that

$$\frac{dI_C}{dV_{BE}} = \frac{1}{V_T} I_{ES} e^{V_{BE}/V_T}$$

or

$$\frac{i_c}{v_{be}} = \frac{I_C}{V_T} = g_m = \text{the transconductance} \tag{24.10}$$

Now the signal output voltage is

$$v_{ce} \simeq -i_c R_C$$

(Here, the approximation sign is because the collector-emitter path within the transistor does present a large resistance r_{ce} through which a very small part of i_c flows.)

The terminal voltage gain is therefore

$$A_v = \frac{v_{ce}}{v_{be}} \simeq \frac{-i_c R_C}{v_{be}} = -g_m R_C \tag{24.11}$$

where the negative sign implies signal inversion.

In practice, $V_T \simeq 26$ mV at room temperature, as has been mentioned, and this leads to a very simple numerical approximation. From Eq. (24.10) and using $I_C = I_Q$,

$$g_m = \frac{I_Q}{V_T} \simeq \frac{I_Q}{0.026} \simeq 39\, I_Q \qquad \text{mA/V}$$

if I_Q is in mA and at room temperature. This shows that irrespective of the transistor used, the transconductance may be approximated knowing only the quiescent collector current.

The magnitude and phase relationships between v_{ce} and i_c can easily be seen by including them on the signal load-line diagram as shown in Fig. 24.30, where the output characteristics of the transistor have been omitted for clarity. Sinusoidal output signals have been inserted, and either may be obtained from the other by following the signal load-line locus.

Now consider the small-signal current gain. Because the value of h_{FE} is not quite linear on the I_C/I_B graph, its slope too must be used for small-signal work. However, the departure from linearity is not great over normal working conditions, and the small-signal value h_{fe} is usually quite close to that of h_{FE}. Hence,

$$A_i = \frac{i_c}{i_b} \simeq h_{fe} \tag{24.12}$$

The small-signal or incremental input resistance to the base itself (to the right of point *X* in Fig. 24.29) may now be found:

$$R_{\text{in}} = \frac{v_{be}}{i_b} = \frac{v_{be}}{i_c}\frac{i_c}{i_b} \simeq \frac{h_{fe}}{g_m} \tag{24.13}$$

Three of the four main (midfrequency) parameters for the CE stage have now been derived, all from a rather primitive understanding of the transistor itself. The fourth, R_{out}, is the dynamic, incremental, or small-signal resistance of the transistor from collector to emitter, which is the slope of the output characteristic at the working point r_{ce}. Being associated with a reverse-biased (CB) junction, this is high—typically about 0.5 MΩ—so that the transistor acts as a current source feeding a comparatively low load resistance R_C. Summarizing, at mid frequency,

$$A_i \simeq h_{fe} \qquad A_v \simeq -g_m R_C \qquad R_{\text{in}} \simeq \frac{h_{fe}}{g_m} \qquad R_{\text{out}} \simeq r_{ce}$$

Example 2

Using the biasing values for *R*1, *R*2, and R_E already obtained in Example 1, calculate the value of R_C to give a terminal voltage gain of –150. Then determine the input resistance R_{in} if h_{fe} for the transistor is 10% higher than h_{FE}.

Solution. Because $I_Q = 1$ mA, $g_m \simeq 39 \times 1 = 39$ mA/V. Hence $A_v \simeq -g_m R_C$ or $-150 \simeq -39\, R_C$, giving

$$R_C = 150/39 \simeq 3.9\ \text{k}\Omega$$

(*Note:* This value *must* be checked to determine that it is reasonable insofar as biasing is concerned. In this case, it will drop $I_Q R_C = 1 \times 3.9 = 3.9$ V. Because $V_{RE} = 3.4$ V, this leaves 12 – 3.9 – 3.4 = 4.7 V across the transistor, which is reasonable.)

Finally,

$$R_{\text{in}} \simeq \frac{h_{fe}}{g_m} = \frac{110}{39} = 2.8\ \text{k}\Omega$$

A Small-Signal Equivalent Circuit

The conclusions reached above regarding the performance of the bipolar transistor are sufficient for the development of a basic equivalent circuit, or model, relevant *only* to small-signal operation. Taking the operating CE amplifier, this may be done by first "looking into" the base, shown as *b* in Fig. 24.31. Between this point and the actual active part of the base region *b′*, it is reasonable to suppose that the intervening (inactive) semiconductor material will present a small resistance $r_{bb'}$. This is called the *base spreading resistance*, and it is also shown in Fig. 24.31.

From *b′* to the emitter *e*, there will be a dynamic or incremental resistance given by

$$r_{b'e} = \frac{v_{b'e}}{i_b} = \frac{v_{b'e}}{i_c}\frac{i_c}{i_b} = \frac{h_{fe}}{g_m} \tag{24.14}$$

so that the full resistance from the base to the emitter must be

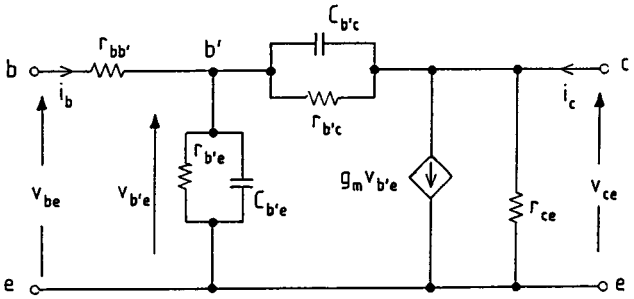


FIGURE 24.31 The hybrid-π small-signal transistor equivalent circuit or model.

$$R_{\text{in}} = r_{bb'} + r_{b'e} = r_{bb'} + \frac{h_{fe}}{g_m} \simeq \frac{h_{fe}}{g_m} \tag{24.15}$$

because $r_{bb'}$ is only about 10 to 100 Ω, which is small compared with $r_{b'e}$, this being several kilohms (as shown by the last example). It will now be understood why Eq. (24.13) gave $R_{\text{in}} \simeq h_{fe}/g_m$.

The reverse-biased junction that exists from b' to the collector ensures that the associated dynamic resistance $r_{b'c}$ will be very large indeed, which is fortunate, otherwise signal feedback from the output to the input would modify the gain characteristics of the amplifier. Typically, $r_{b'c}$ will be some tens of megohms.

However, because of transistor action, the dynamic resistance from collector to emitter, r_{ce}, will be smaller than $r_{b'c}$ and will typically be below a megohm. This "transistor action" may be represented by a current source from collector to emitter that is dependent upon either i_b or $v_{b'e}$. That is, it will be either $h_{fe}i_b$ or $g_m v_{b'e}$. The latter leads to the well-known hybrid-π model, and it is this which is shown in Fig. 24.31.

Where junctions or interfaces of any sort exist, there will always be distributed capacitances associated with them, and to make these easy to handle analytically, they may be "lumped" into single capacitances. In the present context, two lumped capacitances have been incorporated into the hybrid-π model, $C_{b'e}$ from base to emitter and $C_{b'c}$ from base to collector, respectively. These now complete the model, and it will be appreciated that they make it possible to analyze high-frequency performance. Typically, $C_{b'e}$ will be a few picofarads and will always be larger than $C_{b'c}$.

Figure 24.31 is the hybrid-π small-signal, dynamic, or incremental model for a bipolar transistor, and when external components are added and simplifications made, it makes possible the determination of the performance of an amplifier using that transistor not only at midfrequencies but at high and low frequencies, too.

Low-Frequency Performance

In Fig. 24.32 both a source and a load have been added to the hybrid-π equivalent circuit to model the complete CE stage of Fig. 24.29. Here, both $C_{b'e}$ and $C_{b'c}$ have been omitted because they are too small to affect the low-frequency performance, as has $r_{b'c}$ because it is large and so neither loads the source significantly compared to $r_{bb'} + r_{b'e}$ nor applies much feedback.

The signal source has been represented by a Thévenin equivalent that applies a signal via a coupling capacitor C_c. Note that this signal source has been returned to the emitter, which implies that the emitter resistor bypass capacitor C_E has been treated as a short circuit at all signal frequencies for the purposes of this analysis.

Because the top of biasing resistor $R1$ (Fig. 24.29) is taken to ground via the power supply insofar as the signal is concerned, it appears in parallel with $R2$, and the emitter is also grounded to the signal via C_E. That is, a composite biasing resistance to ground R_B appears:

$$R_B = \frac{R1 \cdot R2}{R1 + R2}$$

Finally, the collector load is taken to ground via the power supply and hence to the emitter via C_E.

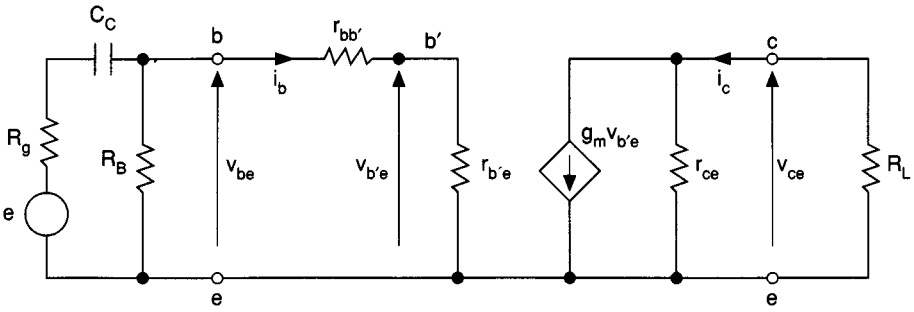


FIGURE 24.32 The loaded hybrid-π model for low frequencies.

Figure 24.32 shows that v_{be} is amplified independently of frequency, so the terminal voltage gain A_v may easily be determined:

$$A_v = \frac{v_{ce}}{v_{be}} = \frac{-i_c R_L}{v_{be}}$$

Now

$$v_{be} = \frac{v_{b'e}(r_{bb'} + r_{b'e})}{r_{b'e}} \simeq v_{b'e} \quad \text{and} \quad i_c = \frac{g_m v_{b'e} r_{ce}}{r_{ce} + R_L} \simeq g_m v_{b'e}$$

because $r_{b'e} >> r_{bb'}$ and $r_{ce} >> R_L$. So,

$$A_v \simeq -g_m R_L$$

which is as expected.

The model shows that v_{be} is amplified independently of frequency because there are no capacitances to its right, so an analysis of low-frequency response devolves down to determining v_{be} in terms of e. Here, part of e will appear across the capacitive reactance X_{Cc}, and the remainder is v_{be}. So, to make the concept of reactance valid, a sinusoidal signal E must be postulated, giving a sinusoidal value for $v_{be} = V_{be}$.

At midfrequencies, where the reactance of C_c is small, the signal input voltage is

$$V_{be}(f_m) = \frac{E \cdot R_{BP}}{R_g + R_{BP}} \tag{24.16}$$

where $R_{BP} = R_B R_{\text{in}}/(R_B + R_{\text{in}})$ and $R_{\text{in}} = r_{bb'} + r_{b'e}$ as before.

At low frequencies, where the reactance of C_c is significant,

$$V_{be}(f_{\text{low}}) = \frac{E \cdot R_{BP}}{\sqrt{(R_g + R_{BP})^2 + X_{C_c}^2}} \tag{24.17}$$

Dividing (24.16) by (24.17) gives

$$\frac{V_{be}(f_m)}{V_{be}(f_{\text{low}})} = \frac{\sqrt{(R_g + R_{BP})^2 + X_{C_c}^2}}{R_g + R_{BP}}$$

There will be a frequency f_L at which $|X_{Cc}| = R_g + R_{BP}$ given by

$$\frac{1}{2\pi f_L C_c} = R_g + R_{BP} \quad \text{or} \quad f_L = \frac{1}{2\pi C_c(R_g + R_{BP})} \tag{24.18}$$

At this frequency, $V_{be}(f_m)/V_{be}(f_L) = \sqrt{2}$ or $V_{be}(f_L)$ is 3 dB lower than $V_{be}(f_m)$.

Example 3

Using the circuit components of the previous examples along with a signal source having an internal resistance of $R_g = 5$ kΩ, find the value of a coupling capacitor that will define a low-frequency –3dB point at 42 Hz.

Solution. Using Eq. (24.18),

$$C_c = \frac{1}{2\pi(R_g + R_{BP})f_L}$$

where $R_{BP} = R1||R2||R_{\text{in}} = 75||39||2.8 = 2.5$ kΩ. That is,

$$C = \frac{10^6}{2\pi(5000 + 2500)(42)} \simeq 0.5\ \mu\text{F}$$

Since a single RC time constant is involved, the voltage gain of the CE stage will appear to fall at 6 dB/octave as the frequency is reduced because more and more of the signal is dropped across C_c. However, even if C_E is very large, it too will contribute to a fall in gain as it allows more and more of the output signal to be dropped across the $R_E||X_{CE}$ combination, this being applied also to the input loop, resulting in negative feedback. So, at very low frequencies, the gain roll-off will tend to 12 dB/octave. The question therefore arises of how large C_E should be, and this can be conveniently answered by considering a second basic form of transistor connection as follows.

The Emitter-Follower or Common-Collector (CC) Circuit

Suppose that R_C is short-circuited in the circuit of Fig. 24.29. This will not affect the biasing because the collector voltage may take any value (the output characteristic is nearly horizontal, as seen in Fig. 24.28). However, the small-signal output voltage ceases to exist because there is now no load resistor across which it can be developed, though the output current i_c will continue to flow as before.

If now C_E is removed, i_c flows entirely through R_E and develops a voltage which can be observed at the emitter $i_e R_E$ ($\simeq i_c R_E$). Consider the magnitude of this voltage. Figure 24.26(a) shows that for a normally operating transistor, the signal component of the base-emitter voltage ΔV_{BE} (or v_{be}) is very small indeed, whereas the constant component needed for biasing is normally about 0.6 to 0.7 V. That is, $v_{be} << V_{BE}$. This implies that the emitter voltage must always follow the base voltage but at a dc level about 0.6 to 0.7 V below it. So, if an output signal is taken from the emitter, it is almost the same as the input signal at the base. In other words, *the voltage gain of an* ***emitter follower*** *is almost unity.*

If this is the case, what is the use of the emitter follower? The answer is that because the signal *current gain* is unchanged at $i_e/i_b = (h_{fe} + 1) \simeq h_{fe}$, then the power gain must also be about h_{fe}. This means in turn that the output resistance must be the resistance "looking into" the transistor from the emitter, divided by h_{fe}. If the parallel combination of R_g and the bias resistors is R_G, then

$$R_{\text{out}(CC)} = \frac{R_G + r_{bb'} + r_{b'e}}{h_{fe}} \tag{24.19}$$

where $R_G = R_g||R1||R2$ (or $R_g||R_B$).

If a voltage generator with zero internal resistance ($R_g = 0$) were applied to the input, then this would become

$$R_{\text{out}(CC)} = \frac{r_{bb'} + r_{b'e}}{h_{fe}}$$

and if $r_{b'e} >> r_{bb'}$ (which is usual), then

$$R_{\text{out}(CC)} \simeq \frac{r_{b'e}}{h_{fe}} = \frac{1}{g_m} \tag{24.20}$$

Consider the numerical implications of this: if $I_C = 1$ mA, then $g_m \simeq 39$ mA/V (at room temperature), so $1/g_m \simeq 26\ \Omega$, which is a very low output resistance indeed. In fact, though it appears in parallel with R_E, it is unlikely that R_E will make any significant contribution because it is usually hundreds or thousands of ohms.

Example 4

Using the same bias resistors as for the CE examples, find the output resistance at the emitter of a CC stage.

Solution. The parallel resistances to the left of the base are

$$R_G = R_g||R1||R2 = 5||75||39 \approx 4.2\ \text{k}\Omega$$

Using Eq. (24.19),

$$R_{\text{out}} \approx \frac{R_G + r_{b'e}}{h_{fe}} = \frac{R_G}{h_{fe}} + \frac{1}{g_m} \qquad \text{(neglecting } r_{bb'}\text{)}$$

where $g_m \approx 39 I_C$, $I_Q = 1$ mA, and $h_{fe} = 110$, so

$$R_{\text{out}(CC)} \approx \frac{4200}{110} + \frac{1000}{39} \approx 63.8\ \Omega$$

From values like this, it is clear that the output of an emitter follower can be thought of as a good practical dependent voltage source of very low internal resistance.

The converse is also true: the input at the base presents a high resistance. This is simply because whereas much the same signal voltage appears at the base as at the emitter, the base signal current i_b is smaller than the emitter signal current i_e by a factor of $(h_{fe} + 1) \simeq h_{fe}$. Hence, the apparent resistance at the base must be at least $h_{fe} R_E$. To this must be added $r_{bb'} + r_{b'e}$ so that

$$R_{\text{in}(CC)} \simeq r_{bb'} + r_{b'e} + h_{fe} R_E \tag{24.21a}$$

Now h_{fe} is rarely less than about 100, so $h_{fe} R_E$ is usually predominant and

$$R_{\text{in}(CC)} \simeq h_{fe} R_E \tag{24.21b}$$

The emitter-follower circuit is therefore a *buffer stage* because it can accept a signal at a high resistance level without significant attenuation and reproduce it at a low resistance level and with *no phase shift* (except at high frequencies).

In this configuration, the unbypassed emitter resistor R_E is obviously in series with the input circuit as well as the output circuit. Hence, it is actually a feedback resistor and so may be given the alternative symbol R_F, as in Fig. 24.33. Because all the output signal voltage is fed back in series with the input, this represents 100% voltage-derived series negative feedback.

The hybrid-π model for the bipolar transistor may now be inserted into the emitter-follower circuit of Fig. 24.33, resulting in Fig. 24.34, from which the four midfrequency parameters may be obtained. As an example of the procedures involved, consider the derivation of the voltage gain expression.

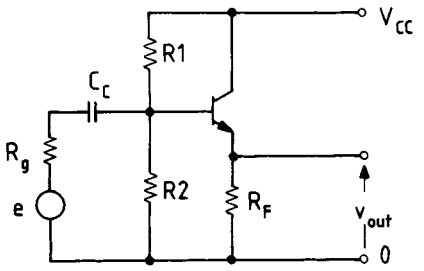


FIGURE 24.33 The emitter follower (or CC stage).

Summing signal currents at the emitter,

$$v_{\text{out}}\left(\frac{1}{R_F}+\frac{1}{r_{ce}}\right)=v_{b'e}\left(\frac{1}{r_{b'e}}+g_m\right)$$

Now $1/r_{ce} << 1/R_F$ and so may be neglected, and $v_{b'e} = v_{\text{in}} - v_{\text{out}}$, so

$$v_{\text{out}}\left(\frac{1}{R_F}\right)=(v_{\text{in}}-v_{\text{out}})\left(\frac{1}{r_{b'e}}+g_m\right)$$

or

$$v_{\text{out}}\left(\frac{1}{R_F}+\frac{1}{r_{b'e}}+g_m\right)=v_{\text{in}}\left(\frac{1}{r_{b'e}}+g_m\right)$$

giving

$$\begin{aligned}A_{v(CC)}&=\frac{v_{\text{out}}}{v_{\text{in}}}=\frac{1/r_{b'e}+g_m}{1/r_{b'e}+g_m+1/R_F}=\frac{1+g_m r_{b'e}}{1+g_m r_{b'e}+r_{b'e}/R_F}\\&\simeq\frac{g_m r_{b'e}}{g_m r_{b'e}+r_{b'e}/R_F}=\frac{g_m R_F}{g_m R_F+1}\end{aligned}\tag{24.22}$$

which is a little less than unity as expected.

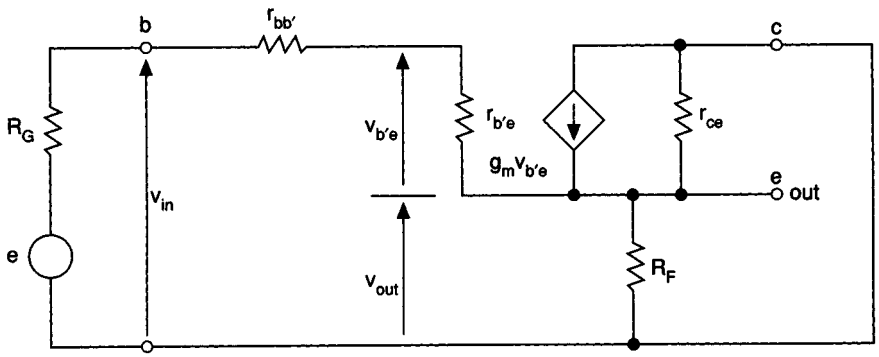


FIGURE 24.34 An emitter-follower equivalent circuit for low frequencies.

Similar derivations based on the equivalent circuit of Fig. 24.34 result in the other three basic midband operating parameters for the emitter follower, and all may be listed:

$$A_{i(CC)} \simeq h_{fe} \quad A_{v(CC)} \rightarrow +1$$
$$R_{\text{in}(CC)} \simeq r_{bb'} + r_{b'e} + h_{fe}R_F \simeq h_{fe}R_F$$

and

$$R_{\text{out}(CC)} \simeq \frac{R_G + r_{bb'} + r_{b'e}}{h_{fe}} \,||\, R_F \simeq \frac{R_G + r_{bb'} + r_{b'e}}{h_{fe}}$$
$$\simeq \frac{1}{g_m} \quad \text{if } R_g \rightarrow 0 \text{ and } r_{bb'} << r_{b'e}$$

The Common-Emitter Bypass Capacitor C_E

In a CE circuit such as that of Fig. 24.29, suppose C_c is large so that the low-frequency –3-dB point f_L is defined only by the parallel combination of the resistance at the emitter and C_E. It will now be seen why the emitter-follower work is relevant: the resistance appearing at the emitter of the CE stage is the same as the output resistance of the emitter-follower stage, and this will now appear in parallel with R_E. If this parallel resistance is renamed R_{emitter}, then, neglecting $r_{bb'}$,

$$R_{\text{emitter}} = R_{\text{out}(CC)} \,||\, R_E$$
$$\simeq \frac{R_G + r_{b'e}}{h_{fe}} \,||\, R_E$$
$$\simeq \frac{R_G + r_{b'e}}{h_{fe}}$$

and if C_E were to define f_L, then

$$f_L = \frac{1}{2\pi R_{\text{emitter}} C_E} \tag{24.23a}$$

For design purposes, C_E can be extracted for any given value of f_L:

$$C_E = \frac{1}{2\pi R_{\text{emitter}} f_L} \tag{24.23b}$$

Example 5

In Example 4, let C_c be large so that only C_E defines f_L at 42 Hz, and find the value of C_E.

Solution. In the emitter-follower example, where $R_g = 5$ kΩ, $R_{\text{out}(CC)}$ was found to be 63.8 Ω, and this is the same as R_{emitter} in the present case. Therefore,

$$C_E = \frac{10^6}{2\pi 63.8 \times 42} \simeq 60\ \mu\text{F}$$

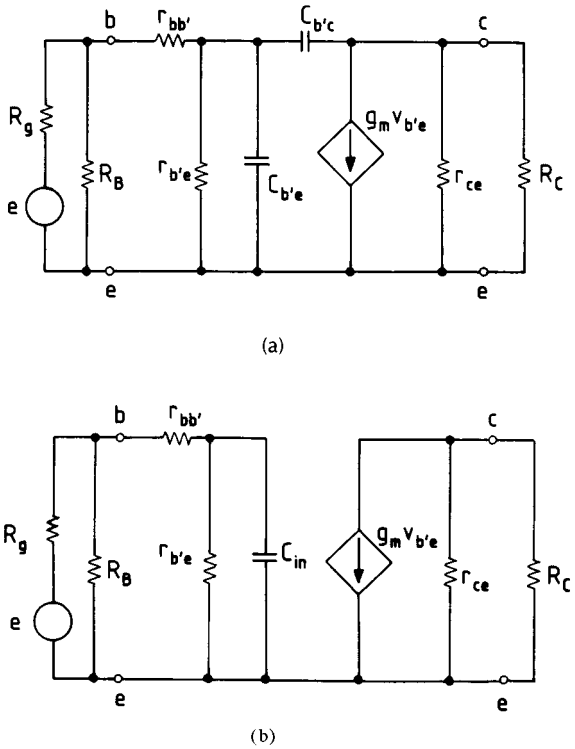


FIGURE 24.35 (a) The high-frequency hybrid-π model and (b) its simplification.

This is the value of C_E that would define f_L if C_c were large. However, if C_E is to act as a short circuit at this frequency, so allowing C_c to define f_L, then its value would have to be one or two orders of magnitude greater, that is, 600 to 6000 μF.

Summarizing, three possibilities exist:

1. If C_E is very large, C_c defines f_L and a 6-dB/octave roll-off results.
2. If C_c is large, C_E defines f_L and again a 6-dB/octave roll-off results.
3. If both C_c and C_E act together, a 12-dB/octave roll-off results.

In point of fact, at frequencies much less than f_L, both conditions (1) and (2) eventually produce 12-dB/octave roll-offs as the alternate "large" capacitors come into play at very low frequencies, but since the amplifier will not still have a useful gain at such frequencies, this is of little importance.

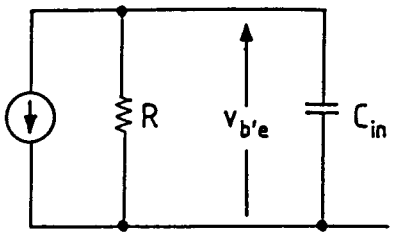
High-Frequency Response

Unlike the low-frequency response situation, the high-frequency response is governed by the small distributed capacitances inside the transistor structure, and these have been lumped together in the hybrid-π model of Fig. 24.31 as $C_{b'e}$ and $C_{b'c}$. At high frequencies, $r_{b'c}$ may be neglected in comparison with the reactance of $C_{b'c}$, so the model may be simplified as in Fig. 24.35(a). From this it will be seen that $C_{b'c}$ is a capacitance which appears from the output to the input so that it may be converted by the Miller Effect into a capacitance at the input of value:

$$C_{b'c}\,(1 - A_v) = C_{b'c}\,(1 + g_m R_C)$$

This will now add to $C_{b'e}$ to give C_{in}:

$$C_{\text{in}} = C_{b'e} + C_{b'c}\,(1 + g_m R_C) \tag{24.24}$$



$$R = \left(R_G + r_{bb'}\right) \| r_{b'e}$$

$$C_{in} = C_{b'e} + C_{b'c}\left(1 + g_m R_L\right)$$

FIGURE 24.36 Simplification of the input part of the high-frequency hybrid-π model.

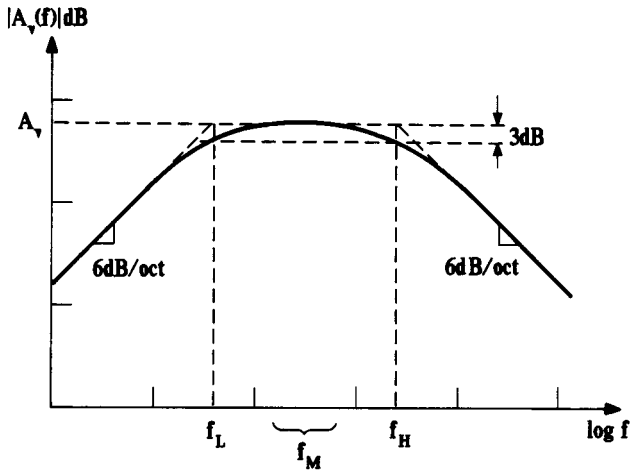


FIGURE 24.37 The complete frequency response.

This simplification is shown in Fig. 24.35(b), where C_{in} is seen to be shunted by the input parts of the model. These input parts may be reduced by sequential use of Thévenin–Norton transformations to result in Fig. 24.36, which is a simple parallel RC circuit driven by a current source. The actual value of this current source is immaterial—what matters is that the input signal to be amplified, $v_{b'e}$, will be progressively reduced as the frequency rises and the reactance of C_{in} falls.

Using a sinusoidal source, $V_{b'e}$ will be 3 dB down when $R = |X_{C_{in}}|$, which gives

$$R = \frac{1}{2\pi f_H C_{in}} \quad \text{or} \quad f_H = \frac{1}{2\pi R C_{in}} \tag{24.25}$$

where $R = (R_G + r_{bb'}) \| r_{b'e}$ from the circuit reduction.

Complete Response

Now that both the low- and high-frequency roll-offs have been related to single time constants (except when C_c and C_E act together), it is clear that the complete frequency response will look like Fig. 24.37, where the midband voltage gain is $A_v = -g_m R_C$.

Design Comments

The design of a simple single-transistor amplifier stage has now been covered in terms of both biasing and small-signal performance. These two concepts have been kept separate, but it will have been noticed that they

are bridged by the transconductance, because $g_m = (q/kT)I_Q$ ($\simeq 39I_Q$ at room temperature). That is, when I_Q has been determined, then the small-signal performance follows from expressions involving g_m.

In fact, once the quiescent voltage across the load resistor of a CE stage has been determined, the voltage gain follows from this irrespective of the values of I_Q and R_C.

If the quiescent voltage at the collector is V_{out}, then in dc biasing terms,

$$V_{RC} = I_Q R_C = (V_{CC} - V_{\text{out}})$$

and in small-signal terms,

$$A_v = -g_m R_C \cong -39 I_Q R_C \qquad \text{(at 25°C)}$$

$$= -39(V_{CC} - V_{\text{out}})$$

Thus, g_m really does act as a bridge between the bias and the small-signal conditions for the bipolar transistor amplifier stage.

Unfortunately, however, there are serious problems with such a stage from a practical viewpoint. For example, it cannot amplify down to dc because of the existence of C_c, and if a larger gain is needed, the cascading of such stages will present problems of phase shift and hence feedback stability. Furthermore, it cannot be produced in IC form because of the incorporation of large capacitances and somewhat critical and high-valued resistors. This leads to a reevaluation of the basic tenets of circuit design, and these may be summed up as follows: circuit design using *discrete* components is largely concerned with voltage drops across resistors (as has been seen), but the design of ICs depends extensively on *currents* and *current sources and sinks.*

Integrated Circuits

Monolithic ICs are fabricated on single chips of silicon or *dice* (the singular being *die*). This means that the active and passive structures on the chips are manufactured all at the same time, so it is easy to ensure that a large number of such structures are identical, or bear some fixed ratio to one another, but it is more difficult to establish precise values for such sets of structures. For example, a set of transistors may all exhibit almost the same values of h_{FE}, but the actual numerical value of h_{FE} may be subject to wider tolerances. Similarly, many pairs of resistors may bear a ratio n:1 to each other, but the actual values of these resistors are more difficult to define. So, in IC design, it is very desirable to exploit the close similarity of devices (or close ratios) rather than depend upon their having predictable absolute values. This approach has led to two ubiquitous circuit configurations, both of which depend upon device similarity: the **long-tailed pair or difference amplifier** (often called the *differential amplifier*), and the **current mirror**. This section will treat both, and the former is best introduced by considering the **degenerate common-emitter** stage.

The Degenerate Common-Emitter Stage

Consider two CE stages which are identical in every respect but which have no emitter resistor bypass capacitors, as shown in Fig. 24.38. Also, notice that in these diagrams, two power supply rails have been used, a positive one at V^+_{CC} and a negative one at V^-_{CC}. The reason for this latter, negative, rail is that the bases may be operated via signal sources referred to a common line or ground. (If, for example, V^+_{CC} and V^-_{CC} are obtained from batteries as shown, then the common line is simply the junction of the two batteries, as is also shown.) The absence of capacitors now means that amplification down to dc is possible.

It is now very easy to find the quiescent collector currents I_Q, because from a dc bias point of view the bases are connected to ground via resistances R_g, which will be taken as having low values so that they drop negligibly small voltages. Hence,

$$I_Q = \frac{|V^-_{CC}| - V_{BE}}{R_F} \tag{24.26}$$

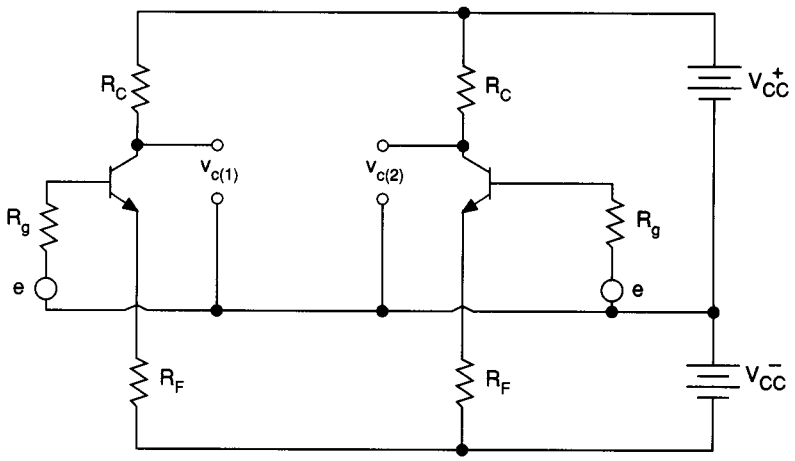


FIGURE 24.38 Two degenerate CE stages.

[For example, if industry-standard supplies of ±15 V are used, $V_{BE} = 0.6$ V, and for $R_F = 15$ kΩ, then $I_Q = (15 - 0.6)/15 = 0.96 \simeq 1$ mA.]

Now suppose that identical signals e are applied. At each collector, this will result in an output signal voltage v_c, where $v_c = -i_c R_C$. Also, at each emitter, the output signal voltage will be $v_e = i_e R_F \simeq i_c R_F$. That is,

$$\frac{v_c}{v_e} \simeq \frac{-i_c R_c}{i_c R_F} = -\frac{R_C}{R_F}$$

If the voltage gain from base to collector of a degenerate CE stage is $A_{v(dCE)}$ and the voltage gain from base to emitter is simply the emitter-follower gain $A_{v(CC)}$, then

$$v_c = A_{v(dCE)} e \qquad \text{and} \qquad v_e = A_{v(CC)} e$$

giving

$$\frac{A_{v(dCE)}}{A_{v(CC)}} \simeq -\frac{R_C}{R_F}$$

Now $A_{v(CC)}$ is known from Eq. (24.22) so that

$$A_{v(dCE)} \simeq -A_{v(CC)} \frac{R_C}{R_F} = -\frac{g_m R_C}{1 + g_m R_F} \simeq -\frac{R_C}{R_F} \tag{24.27}$$

Note that the input resistance to each base is as for the emitter-follower stage:

$$R_{\text{in}(dCE)} = r_{bb'} + r_{b'e} + h_{fe} R_F \simeq h_{fe} R_F$$

Now consider what happens if the emitters are connected together as in Fig. 24.39, where the two resistors R_F have now become R_X, where $R_X = \frac{1}{2} R_F$.

The two quiescent emitter currents now combine to give $I_X = 2I_E \simeq 2I_C$, and otherwise the circuit currents and voltages remain undisturbed. So, if the two input signals are identical, then the two output signals will also

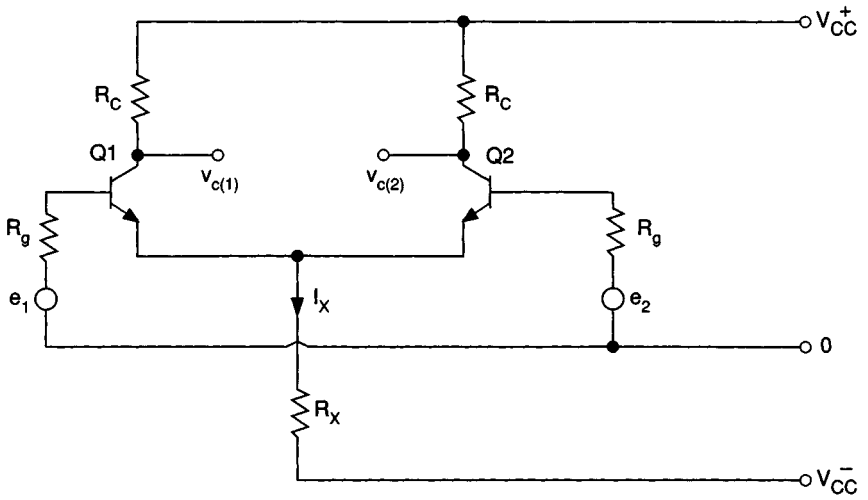


FIGURE 24.39 The difference amplifier.

be identical. This circuit is now called a *difference amplifier,* and the reason will become obvious as soon as the two input signals differ.

The Difference Amplifier

In Fig. 24.39, if $e_1 = e_2$, these are called common-mode input signals, $e_{in(CM)}$, and they will be amplified by $-R_C/R_F$ as for the degenerate CE stage. However, if $e_1 \neq e_2$, then $e_1 - e_2 = e_{in}$, the difference input signal. The following definitions now apply:

$$\frac{e_1 + e_2}{2} = e_{in(CM)} \quad \text{the common mode component}$$

and

$$\frac{\pm(e_1 - e_2)}{2} = e_{in(diff)} \quad \text{the difference component, or } \tfrac{1}{2} e_{in}$$

Hence, $e_1 = e_{in(CM)} + e_{in(diff)}$ and $e_2 = e_{in(CM)} - e_{in(diff)}$.

Consider the progress of a signal current driven by $e_1 - e_2$ and entering the base of $Q1$. It will first pass through R_g, then into the resistance R_{in} at the base of $Q1$, and will arrive at the emitter of $Q2$. Here, if R_X is large, most of this signal current will pass into the resistance presented by the $Q2$ emitter and eventually out of the $Q2$ base via another R_g to ground. The total series resistance is therefore

$$R_g + R_{in} = R_g + r_{bb'} + r_{b'e} + h_{fe} R_{emitter(2)}$$

But

$$R_{emitter(2)} = \frac{R_g + r_{bb'} + r_{b'e}}{h_{fe}}$$

so

$$R_g + R_{in} = 2(R_g + r_{bb'} + r_{b'e})$$

which is the resistance between the two signal sources. Hence,

$$i_{b(1)} = -i_{b(2)} = \frac{e_1 - e_2}{2(R_g + r_{bb'} + r_{b'e})}$$

giving

$$v_{c(1)} = -v_{c(2)} = \frac{h_{fe}R_C(e_1 - e_2)}{2(R_g + r_{bb'} + r_{b'e})}$$

so that the overall difference voltage gain to each collector is

$$A_{ov} = \frac{v_c}{e_1 - e_2} = \frac{\pm h_{fe}R_C}{2(R_g + r_{bb'} + r_{b'e})} \tag{24.28a}$$

If the voltage gain with the input signal measured between the actual bases is needed, R_g may be removed to give

$$A_v = \frac{\pm h_{fe}R_C}{2(r_{bb'} + r_{b'e})} \tag{24.28b}$$

Finally, if the output signal is measured between the collectors (which will be twice that at each collector because they are in antiphase), the difference-in–to–difference-out voltage gain will be

$$A_{v(\text{diff})} = \frac{h_{fe}R_C}{r_{bb'} + r_{b'e}} \simeq \frac{h_{fe}R_C}{r_{b'e}} = g_m R_C \tag{24.28c}$$

which is the same as for a single CE stage.

Note that this is considerably larger than the gain for a common-mode input signal; that is, the difference stage amplifies difference signals well but largely rejects common-mode signals. This common-mode rejection property is very useful, for often, small signals appear across leads, both of which may contain identical electrical noise. So, the difference stage tends to reject the noise while still amplifying the signal. Furthermore, the difference stage has the advantage that it needs no coupling or bypass capacitors and so will amplify frequencies down to zero (dc). Also, it is very stable biaswise and lends itself perfectly to realization on a monolithic IC.

To make the above derivation valid, the long-tail resistance R_X should be as large as possible so that most of the signal current enters the emitter of $Q2$. However, R_X must also carry the quiescent current, which would produce a very high quiescent voltage drop and so require a very high value of V_{CC-}. To overcome this, another transistor structure may be used within a configuration known as a current mirror.

The Current Mirror

The two transistors in Fig. 24.40 are assumed to be identical, and $Q1$ has its base and collector connected so that it acts simply as a diode (formed by the base-emitter junction). The current through it is therefore

$$I = \frac{V_{CC} - V_{BE}}{R} \tag{24.29}$$

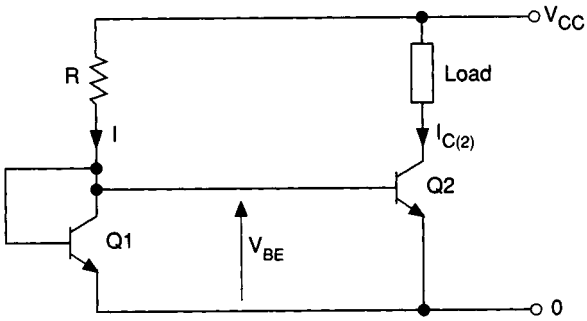


FIGURE 24.40 The current mirror.

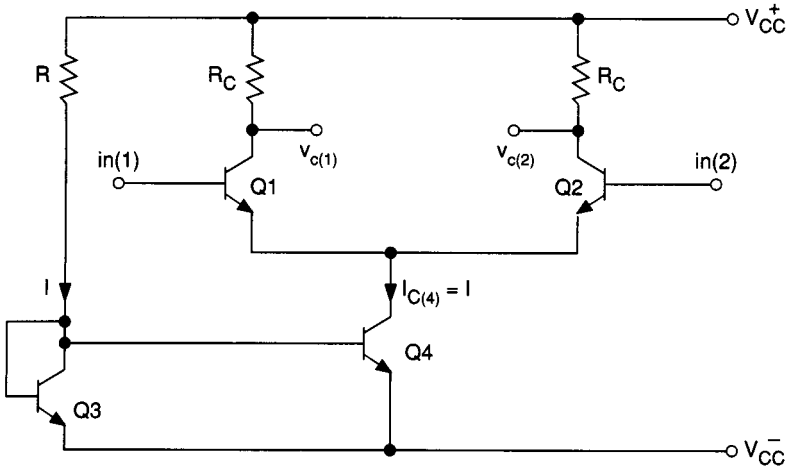


FIGURE 24.41 Current mirror biasing.

The voltage drop V_{BE} so produced is applied to $Q2$ as shown so that it is forced to carry the same collector current I; that is, it mirrors the current in $Q1$.

The transistor $Q2$ is now a device that carries a dc $I_{C(2)} = I$ but presents a large incremental resistance r_{ce} at its collector. This is exactly what is required by the difference amplifier pair, so it may be used in place of R_X.

The Difference Stage with Current Mirror Biasing

Figure 24.41 shows a complete difference stage complete with a current mirror substituting for the long-tail resistor R_X, where the emitter quiescent currents combine to give I_X:

$$I_X = \frac{V_{CC^+} + |V_{CC^-}| - V_{BE(3)}}{R} = I \tag{24.30}$$

This quiescent or bias current is very stable, because the change in $V_{BE(3)}$ due to temperature variations is exactly matched by that required by $Q4$ to produce the same current. The difference gain will be as discussed above, but the common-mode gain will be extremely low because of the high incremental resistance r_{ce} presented by the long-tail transistor.

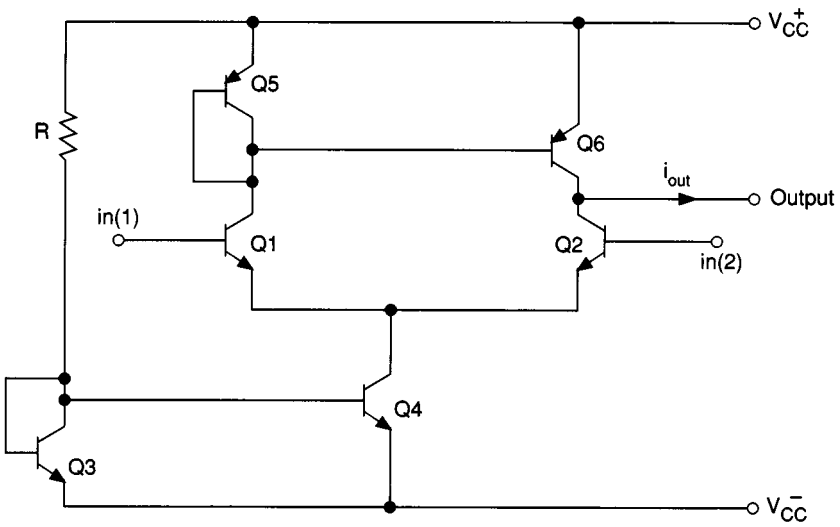


FIGURE 24.42 A complete difference amplifier stage.

The Current Mirror as a Load

A second current mirror may be used as a load for the difference amplifier, as shown in Fig. 24.42. This must utilize *pnp* transistor structures so that the *Q*6 collector loads the *Q*2 collector with a large incremental resistance $r_{ce}(6)$, making for an extremely high voltage gain. Furthermore, *Q*5 and *Q*6 combine the signal output currents of both *Q*1 and *Q*2 to perform a double-ended–to–single-ended conversion as follows. Taking signal currents,

$$i_{\text{out}} = i_c(6) - i_c(2)$$

But

$$i_c(6) = i_c(5)$$

by current mirror action, and

$$i_c(5) = i_c(1)$$

so

$$i_c(6) = i_c(1)$$

Also,

$$i_c(2) = -i_c(1)$$

by difference amplifier action, so

$$i_{\text{out}} = i_c(1) + i_c(1) = 2i_c(1) \tag{24.31}$$

Thus, both sides of the long-tailed pair are used to provide an output current that may then be applied to further stages to form a complete amplifier. Also, because no capacitors and only one resistor are needed, it is an easy circuit for monolithic integration on a single die.

Summary

It has been shown how a limited knowledge of bipolar operation can lead to properly biased amplifier stages using discrete transistors. An equivalent circuit—the hybrid-π model—was then derived, again from limited information, which made possible the analysis of such stages, and some purely practical design results were favorably compared with its predictions. Finally, the tenets of this equivalent circuit were used to evaluate the performance of the difference amplifier and current mirror circuits, which are the cornerstones of modern electronic circuit design in a very wide variety of its manifestations. These circuits are, in fact, the classic transconductance and translinear elements that are ubiquitous in modern IC signal conditioning and function networks.

It should be recognized that there are many models other than the one introduced here, from the simple but very common *h*-parameter version to complex and comprehensive versions developed for computer-aided design (CAD) methods. However, the present elementary approach has been from a design rather than an analytical direction, for it is obvious that powerful modern computer-oriented methods such as the SPICE variants become useful only when a basic circuit configuration has been established, and at the time of writing, this is still the province of the human designer.

Defining Terms

Biasing circuit: A circuit that holds a transistor in an operating condition ready to receive signals.

Common emitter: A basic transistor amplifier stage whose emitter is common to both input and output loops. It amplifies voltage, current, and hence power.

Current mirror: An arrangement of two (or more) transistors such that a defined current passing into one is mirrored in another at a high resistance level.

Degenerate common emitter: A combination of the common-emitter and emitter-follower stages with a very well-defined gain.

Difference amplifier or long-tailed pair: An arrangement of two transistors that amplifies difference signals but rejects common-mode signals. It is often called a differential pair.

Emitter follower or common collector: A basic transistor amplifier stage whose collector is common to both input and output loops. Its voltage gain is near unity, but it amplifies current and hence power. It is a high-input resistance, low-output resistance, or buffer, circuit.

Related Topic

28.2 Small Signal Analysis

Further Information

The following list of recent textbooks covers topics mainly related to analog circuitry containing both integrated and discrete semiconductor devices.

G. M. Glasford, *Analog Electronic Circuits,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
P. R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits,* 2nd ed., New York: Wiley, 1984.
J. Keown, *PSPICE and Circuit Analysis,* New York: Macmillan, 1991.
R.B. Northrop, *Analog Electronic Circuits,* Reading, Mass.: Addison-Wesley, 1990.
A.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., Philadelphia: Saunders, 1991.
T. Schubert and E. Kim, *Active and Non-Linear Electronics,* New York: Wiley, 1996.
J. Watson, *Analog and Switching Circuit Design,* New York: Wiley, 1989.

24.3 The Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET)

John R. Brews

The MOSFET is a transistor that uses a control electrode, the **gate**, to capacitively modulate the conductance of a surface **channel** joining two end contacts, the **source** and the **drain**. The gate is separated from the semiconductor **body** underlying the gate by a thin *gate insulator,* usually silicon dioxide. The surface channel is formed at the interface between the semiconductor body and the gate insulator (see Fig. 24.43).

The MOSFET can be understood by contrast with other field-effect devices, like the *JFET,* or junction field-effect transistor, and the *MESFET,* or metal semiconductor field-effect transistor [Hollis and Murphy, 1990]. These other transistors modulate the conductance of a *majority-carrier* path between two *ohmic* contacts by capacitive control of its cross section. (Majority carriers are those in greatest abundance in a field-free semiconductor, electrons in *n*-type material and holes in *p*-type material.) This modulation of the cross section can take place at any point along the length of the channel, so the gate electrode can be positioned anywhere and need not extend the entire length of the channel.

Analogous to these field-effect devices is the *buried-channel, depletion-mode,* or *normally on* MOSFET, which contains a surface layer of the same doping type as the source and drain (opposite type to the semiconductor body of the device). As a result, it has a built-in or normally on channel from source to drain with a conductance that is reduced when the gate depletes the majority carriers.

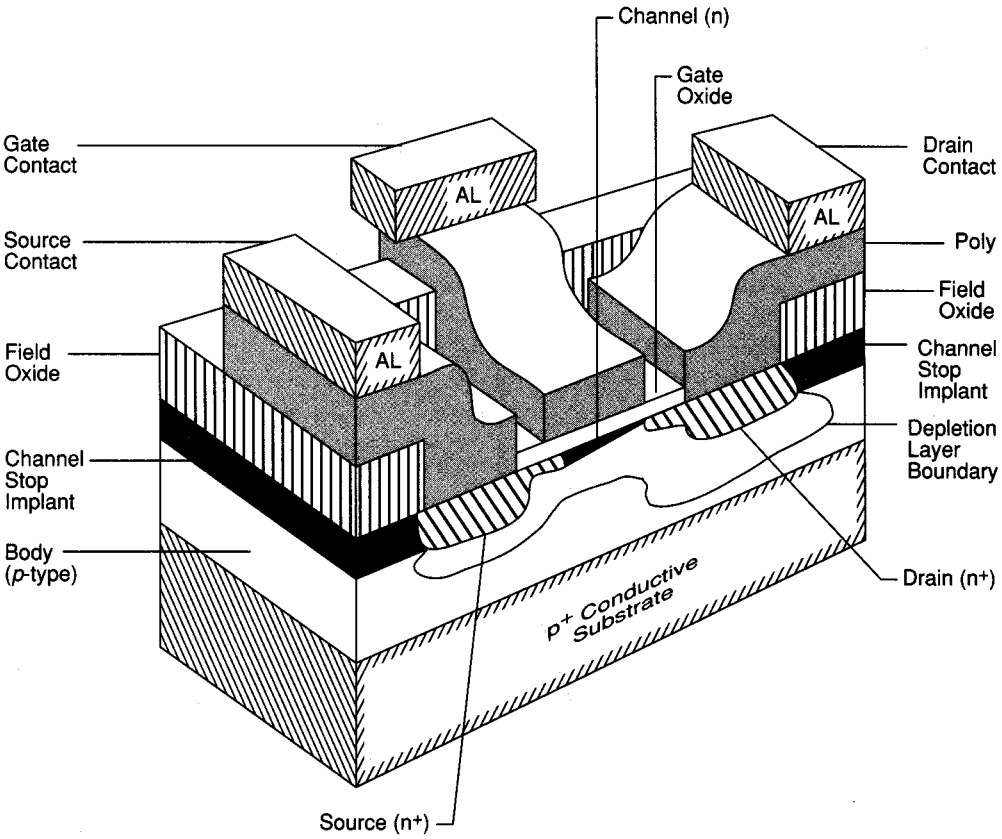


FIGURE 24.43 A high-performance *n*-channel MOSFET. The device is isolated from its neighbors by a surrounding thick *field oxide* under which is a heavily doped *channel stop* implant intended to suppress accidental channel formation that could couple the device to its neighbors. The drain contacts are placed over the field oxide to reduce the capacitance to the body, a parasitic that slows response times. These structural details are described later. (*Source:* After Brews, 1990.)

In contrast, the true MOSFET is an *enhancement-mode* or *normally off* device. The device is normally off because the body forms *pn* junctions with both the source and the drain, so no majority-carrier current can flow between them. Instead, *minority-carrier* current can flow, provided minority carriers are available. As discussed later, for gate biases that are sufficiently attractive, above **threshold**, minority carriers are drawn into a surface channel, forming a conducting path from source to drain. The gate and channel then form two sides of a capacitor separated by the gate insulator. As additional attractive charges are placed on the gate side, the channel side of the capacitor draws a balancing charge of minority carriers from the source and the drain. The more charges on the gate, the more populated the channel, and the larger the conductance. Because the gate *creates* the channel, to ensure electrical continuity the gate must extend over the entire length of the separation between source and drain.

The MOSFET channel is created by attraction to the gate and relies upon the insulating layer between the channel and the gate to prevent leakage of minority carriers to the gate. As a result, MOSFETs can be made only in material systems that provide very good gate insulators, and the best system known is the silicon–silicon dioxide combination. This requirement for a good gate insulator is not so important for JFETs and MESFETs, where the role of the gate is to *push away* majority carriers rather than to attract minority carriers. Thus, in GaAs systems where good insulators are incompatible with other device or fabricational requirements, MESFETs are used.

A more recent development in GaAs systems is the heterostructure field-effect transistor, or HFET [Pearton and Shaw, 1990], made up of layers of varying compositions of Al, Ga, and As or In, Ga, P, and As. These devices are made using molecular beam epitaxy or by organometallic vapor phase epitaxy, expensive methods still being refined for manufacture. HFETs include a variety of structures, the best known of which is the modulation doped FET, or MODFET. HFETs are field-effect devices, not MOSFETs, because the gate simply modulates the carrier density in a preexistent channel between ohmic contacts. The channel is formed spontaneously, regardless of the quality of the gate insulator as a condition of equilibrium between the layers, just as a depletion layer is formed in a *pn* junction. The resulting channel is created very near to the gate electrode, resulting in gate control as effective as in a MOSFET.

The silicon-based MOSFET has been successful primarily because the silicon–silicon dioxide system provides a stable interface with low trap densities, and because the oxide is impermeable to many environmental contaminants, has a high breakdown strength, and is easy to grow uniformly and reproducibly [Nicollian and Brews, 1982]. These attributes allow easy fabrication using lithographic processes, resulting in integrated circuits (ICs), with very small devices, very large device counts, and very high reliability at low cost. Because the importance of the MOSFET lies in this relationship to high-density manufacture, an emphasis of this article is to describe the issues involved in continuing miniaturization.

An additional advantage of the MOSFET is that it can be made using either electrons or holes as channel carrier. Using both types of devices in so-called complementary MOS (CMOS) technology allows circuits that draw no *dc* power if current paths include at least one series connection of both types of device because, in steady state, only one or the other type conducts, not both at once. Of course, in exercising the circuit, power is drawn during switching of the devices. This flexibility in choosing *n*- or *p*-channel devices has enabled large circuits to be made that use low power levels. Hence, complex systems can be manufactured without expensive packaging or cooling requirements.

Current-Voltage Characteristics

The derivation of the current-voltage characteristics of the MOSFET can be found in several sources [Annaratone, 1986; Brews, 1981; Pierret, 1990]. Here a qualitative discussion is provided.

Strong-Inversion Characteristics

In Fig. 24.44 the source-drain current I_D is plotted versus drain-to-source voltage V_D (the I-V curves for the MOSFET). At low V_D the current increases approximately linearly with increased V_D, behaving like a simple resistor with a resistance that is controlled by the gate voltage V_G: as the gate voltage is made more attractive for channel carriers, the channel becomes stronger, more carriers are contained in the channel, and its resistance R_{ch} drops. Hence, at larger V_G the current is larger.

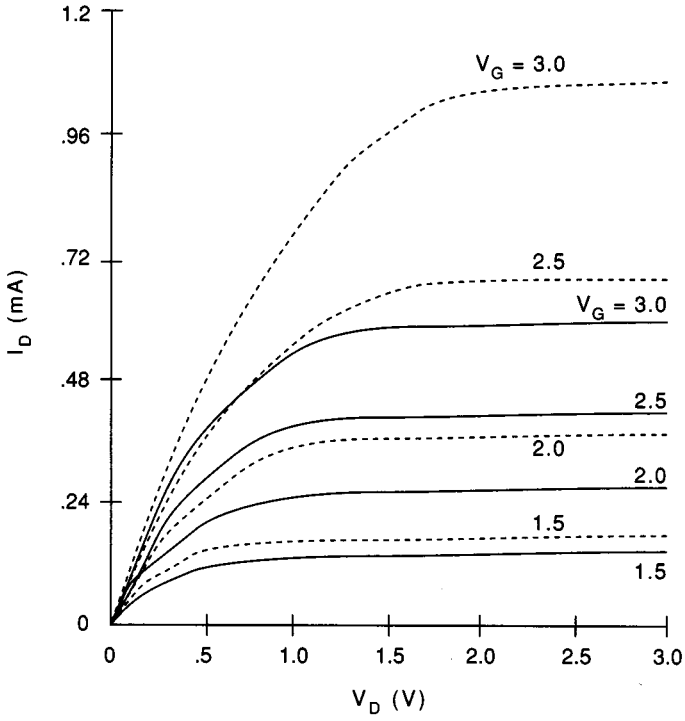


FIGURE 24.44 Drain current I_D versus drain voltage V_D for various choices of gate bias V_G. The dashed-line curves are for a long-channel device for which the current in saturation increases quadratically with gate bias. The solid-line curves are for a *short-channel* device that is approaching *velocity saturation* and thus exhibits a more linear increase in saturation current with gate bias, as discussed in the text.

At large V_D the curves flatten out, and the current is less sensitive to drain bias. The MOSFET is said to be in *saturation*. There are different reasons for this behavior, depending upon the field along the channel caused by the drain voltage. If the source-drain separation is short, near or below a micrometer, the usual drain voltage is sufficient to create fields along the channel of more than a few $\times 10^4$ V/cm. In this case the carrier energy is sufficient for carriers to lose energy by causing vibrations of the silicon atoms composing the crystal (optical phonon emission). Consequently, the carrier velocity does not increase much with increased field, saturating at a value $\upsilon_{sat} \approx 10^7$ cm/s in silicon MOSFETs. Because the carriers do not move faster with increased V_D, the current also saturates.

For longer devices the current-voltage curves saturate for a different reason. Consider the potential along the insulator-channel interface, the surface potential. Whatever the surface potential is at the source end of the channel, it varies from the source end to a value larger at the drain end by V_D because the drain potential is V_D higher than the source. The gate, on the other hand, is at the same potential everywhere. Thus, the difference in potential between the gate and the source is larger than that between the gate and the drain. Correspondingly, the oxide field at the source is larger than that at the drain, and as a result less charge can be supported at the drain. This reduction in attractive power of the gate reduces the number of carriers in the channel at the drain end, increasing channel resistance. In short, we have $I_D \approx V_D / R_{ch}$, but the channel resistance $R_{ch} = R_{ch}(V_D)$ is increasing with V_D. As a result, the current-voltage curves do not continue along the initial straight line, but bend over and saturate.

Another difference between the current-voltage curves for short devices and those for long devices is the dependence on gate voltage. For long devices, the current level in saturation, $I_{D,sat}$, increases quadratically with gate bias. The reason is that the number of carriers in the channel is proportional to $V_G - V_{TH}$ (where V_{TH} is the *threshold voltage*), as is discussed later, the channel resistance $R_{ch} \propto 1/(V_G - V_{TH})$, and the drain bias in saturation is approximately V_G. Thus, $I_{D,sat} = V_D/R_{ch} \propto (V_G - V_{TH})^2$, and we have quadratic dependence. When

the carrier velocity is saturated, however, the dependence of the current on drain bias is suppressed because the speed of the carriers is fixed at υ_{sat}, and $I_{D,sat} \propto \upsilon_{sat}/R_{ch} \propto (V_G - V_{TH})\,\upsilon_{sat}$, a linear gate-voltage dependence. As a result, the current available from a short device is not as large as would be expected if we assumed it behaved like a long device.

Subthreshold Characteristics

Quite different current-voltage behavior is seen in **subthreshold**, that is, for gate biases so low that the channel is in *weak inversion*. In this case the number of carriers in the channel is so small that their charge does not affect the potential, and channel carriers simply must adapt to the potential set up by the electrodes and the dopant ions. Likewise, in subthreshold any flow of current is so small that it causes no potential drop along the interface, which becomes an equipotential.

As there is no lateral field to move the channel carriers, they move by diffusion only, driven by a gradient in carrier density set up because the drain is effective in reducing the carrier density at the drain end of the channel. In subthreshold the current is then independent of drain bias once this bias exceeds a few tens of millivolts, enough to reduce the carrier density at the drain end of the channel to near zero.

In short devices, however, the source and drain are close enough together to begin to share control of the potential with the gate. If this effect is too strong, a drain-voltage dependence of the subthreshold characteristic then occurs, which is undesirable because it increases the MOSFET off current, and can cause a drain-bias dependent threshold voltage.

Although for a well-designed device there is no drain-voltage dependence in subthreshold, gate-bias dependence is exponential. The surface is lowered in energy relative to the semiconductor body by the action of the gate. If this *surface potential* is ϕ_S below that of the body, the carrier density is enhanced by a Boltzmann factor $\exp(q\phi_S/kT)$ relative to the body concentration, where kT/q = the thermal voltage ≈ 25 mV at 290 K. As ϕ_S is roughly proportional to V_G, this exponential dependence on ϕ_S leads to an exponential dependence upon V_G for the carrier density and, hence, for the current in subthreshold.

Important Device Parameters

A number of MOSFET parameters are important to the performance of a MOSFET. In this subsection some of these parameters are discussed, particularly from the viewpoint of digital ICs.

Threshold Voltage

The threshold voltage is vaguely defined as the gate voltage V_{TH} at which the channel begins to form. At this voltage devices begin to switch from "off" to "on," and circuits depend on a voltage swing that straddles this value. Thus, threshold voltage helps in deciding the necessary supply voltage for circuit operation, and also it helps in determining the leakage or "off" current that flows when the device is in the off state.

Threshold voltage is controlled by oxide thickness d and by body doping. To control the body doping, ion implantation is used so that the dopant-ion density is not simply a uniform extension of the bulk, background level N_B ions/unit volume but has superposed upon it an implanted-ion density. To estimate the threshold voltage, we need a picture of what happens in the semiconductor under the gate as the gate voltage is changed from its off level toward threshold.

If we imagine changing the gate bias from its off condition toward threshold, at first the result is to repel majority carriers, forming a surface *depletion layer* (refer to Fig. 24.43). In the depletion layer there are almost no carriers present, but there are dopant ions. In *n*-type material these dopant ions are positive donor impurities that cannot move under fields because they are locked in the silicon lattice, where they have been deliberately introduced to replace silicon atoms. In *p*-type material these dopant ions are negative acceptors. Thus, each charge added to the gate electrode to bring the gate voltage closer to threshold causes an increase in the depletion-layer width sufficient to balance the gate charge by an equal but opposite charge of dopant ions in the silicon depletion layer.

This expansion of the depletion layer continues to balance the addition of gate charge until threshold is reached. Then this charge response changes: above threshold any additional gate charge is balanced by an increasingly strong inversion layer or channel. The border between a depletion-layer and an inversion-layer response, threshold, should occur when

$$\frac{dqN_{\text{inv}}}{d\phi_S} = \frac{dQ_D}{d\phi_S} \tag{24.32}$$

where $d\phi_S$ is the small change in surface potential that corresponds to our incremental change in gate charge, qN_{inv} is the inversion-layer charge/unit area, and Q_D is the depletion-layer charge/unit area. According to Eq. (24.32), the two types of response are equal at threshold, so one is larger than the other on either side of this condition. To be more quantitative, the rate of increase in qN_{inv} is exponential; that is, its rate of change is proportional to qN_{inv}, so as qN_{inv} increases, so does the left side of Eq. (24.32). On the other hand, Q_D has a square-root dependence on ϕ_S, which means its rate of change becomes smaller as Q_D increases. Thus, as surface potential is increased, the left side of Eq. (24.32) increases proportional to qN_{inv} until, at threshold, Eq. (24.32) is satisfied. Then, beyond threshold, the exponential increase in qN_{inv} with ϕ_S swamps Q_D, making change in qN_{inv} the dominant response. Likewise, below threshold, the exponential decrease in qN_{inv} with decreasing ϕ_S makes qN_{inv} negligible and change in Q_D becomes the dominant response. The abruptness of this change in behavior is the reason for the term *threshold* to describe MOSFET switching.

To use Eq. (24.32) to find a formula for threshold voltage, we need expressions for N_{inv} and Q_D. Assuming the interface is held at a lower energy than the bulk due to the charge on the gate, the minority-carrier density at the interface is larger than in the bulk semiconductor, even below threshold. Below threshold and even up to the threshold of Eq. (24.32), the number of charges in the channel/unit area N_{inv} is given for *n*-channel devices approximately by [Brews, 1981]:

$$N_{\text{inv}} \approx d_{\text{INV}} \frac{n_i^2}{N_B} e^{q(\phi_S - V_S)/kT} \tag{24.33}$$

where the various symbols are defined as follows: n_i = intrinsic carrier density/unit volume $\approx 10^{10}/\text{cm}^3$ in silicon at 290 K, and V_S = body reverse bias, if any. The first factor, d_{INV}, is an effective depth of minority carriers from the interface given by

$$d_{\text{INV}} = \frac{\varepsilon_s kT/q}{Q_D} \tag{24.34}$$

where Q_D = depletion-layer charge/unit area due to charged dopant ions in the region where there are no carriers and ε_s is the dielectric permittivity of the semiconductor.

Equation (24.33) expresses the net minority-carrier density/unit area as the product of the bulk minority-carrier density/unit volume, n_i^2/N_B, with the depth of the minority-carrier distribution d_{INV} multiplied in turn by the customary Boltzmann factor $\exp[q(\phi_S - V_S)/kT]$ expressing the enhancement of the interface density over the bulk due to lower energy at the interface. The depth d_{INV} is related to the carrier distribution near the interface using the approximation (valid in *weak inversion*) that the minority-carrier density decays exponentially with distance from the oxide-silicon surface. In this approximation, d_{INV} is the *centroid* of the minority-carrier density. For example, for a uniform bulk doping of 10^{16} dopant ions/cm^3 at 290 K, using Eq. (24.33) and the surface potential at threshold from Eq. (24.38) below ($\phi_{TH} = 0.69$ V), there are $Q_D/q = 3 \times 10^{11}$ charges/cm^2 in the depletion layer at threshold. This Q_D corresponds to a $d_{\text{INV}} = 5.4$ nm and a carrier density at threshold of $N_{\text{inv}} = 5.4 \times 10^9$ charges/cm^2.

The next step in using the definition of threshold, Eq. (24.32), is to introduce the depletion-layer charge/unit area Q_D. For the ion-implanted case, Q_D is made up of two terms [Brews, 1981]:

$$Q_D = qN_B L_B \left\{ 2\left[q\phi_{TH}/(kT) - m_1 - 1 \right] \right\}^{1/2} + qD_I \tag{24.35}$$

where the first term is Q_B, the depletion-layer charge from bulk dopant atoms in the depletion layer with a width that has been reduced by the first moment of the implant, namely, m_1 given in terms of the centroid of the implant x_C by

$$m_1 = \frac{D_I x_C}{N_B L_B^2} \tag{24.36}$$

The second term is the additional charge due to the implanted-ion density within the depletion layer, D_I ions per unit area. The Debye length L_B is defined as

$$L_B^2 \equiv \frac{kT}{q} \frac{\varepsilon_s}{qN_B} \tag{24.37}$$

where ε_s is the dielectric permittivity of the semiconductor. The Debye length is a measure of how deeply a variation of surface potential penetrates into the body when $D_I = 0$ and the depletion layer is of zero width.

Approximating qN_{inv} by Eq. (24.33) and Q_D by Eq. (24.35), Eq. (24.32) determines the surface potential at threshold, ϕ_{TH}, to be

$$\phi_{TH} = 2\frac{kT}{q}\ln\frac{N_B}{n_i} + \frac{kT}{q}\ln\left(1 + \frac{qD_I}{Q_B}\right) \tag{24.38}$$

where the new symbols are defined as follows: Q_B = depletion-layer charge/unit area due to bulk body dopant N_B in the depletion layer, and qD_I = depletion-layer charge/unit area due to implanted ions in the depletion layer between the inversion-layer edge and the depletion-layer edge. Because even a small increase in ϕ_S above ϕ_{TH} causes a large increase in qN_{inv}, which can balance a rather large change in gate charge or gate voltage, ϕ_S does not increase much as $V_G - V_{TH}$ increases. Nonetheless, in strong inversion $N_{\text{inv}} \approx 10^{12}$ charges/cm^2, so in strong inversion ϕ_S will be about 10 kT/q larger than ϕ_{TH}.

Equation (24.38) indicates for uniform doping (no implant, $D_I = 0$) that threshold occurs approximately for $\phi_S = \phi_{TH} = 2(kT/q)\ln(N_B/n_i) \equiv 2\phi_B$, but for the nonuniformly doped case a larger surface potential is needed, assuming the case of a normal implant where D_I is positive, increasing the dopant density. The implant increases the required surface potential because the field at the surface is larger, narrowing the inversion layer, and reducing the channel strength for $\phi_S = 2\phi_B$. Hence, a somewhat larger surface potential is needed to increase qN_{inv} to the point that Eq. (24.32) is satisfied. Equation (24.38) would not apply if a significant fraction of the implant were confined to lie within the inversion layer itself. However, no realistic implant can be confined within a distance comparable to an inversion-layer thickness (a few tens of nanometers), so Eq. (24.38) covers practical cases.

With the surface potential ϕ_{TH} known, the potential on the gate at threshold Φ_{TH} can be found if we know the oxide field F_{ox} by simply adding the potential drop across the semiconductor to that across the oxide. That is, $\Phi_{TH} = \phi_{TH} + F_{\text{ox}}\, d$, with d = oxide thickness and F_{ox} given by Gauss's law as

$$\varepsilon_{\text{ox}} F_{\text{ox}} = Q_D \tag{24.39}$$

There are two more complications in finding the threshold voltage. First, the *gate voltage* V_{TH} usually differs from the gate potential Φ_{TH} at threshold because of a work-function difference between the body and the gate material. This difference causes a spontaneous charge exchange between the two materials as soon as the MOSFET is placed in a circuit allowing charge transfer to occur. Thus, even before any *voltage* is applied to the device, a *potential* difference exists between the gate and the body due to spontaneous charge transfer. The

second complication affecting threshold voltage is the existence of charges in the insulator and at the insulator-semiconductor interface. These nonideal contributions to the overall charge balance are due to traps and fixed charges incorporated during the device processing.

Ordinarily interface-trap charge is negligible ($<10^{10}$/cm^2 in silicon MOSFETs), and the other nonideal effects upon threshold voltage are accounted for by introducing the *flatband voltage* V_{FB}, which corrects the gate bias for these contributions. Then, using Eq. (24.39) with $F_{ox} = (V_{TH} - V_{FB} - \phi_{TH})/d$ we find

$$V_{TH} = V_{FB} + \phi_{TH} + Q_D \frac{d}{\varepsilon_{ox}} \tag{24.40}$$

which determines V_{TH} even for the nonuniformly doped case, using Eq. (24.38) for ϕ_{TH} and Q_D at threshold from Eq. (24.35). If interface-trap charge/unit area is not negligible, then terms in the interface-trap charge/unit area Q_{IT} must be added to Q_D in Eq. (24.40).

From Eqs. (24.35) and (24.38), the threshold voltage depends upon the implanted dopant-ion profile only through two parameters, the net charge introduced by the implant in the region between the inversion layer and the depletion-layer edge qD_p and the centroid of this portion of the implanted charge x_C. As a result, a variety of implants can result in the same threshold, ranging from the extreme of a δ-function spike implant of dose D_I/unit area located at the centroid x_O to a box-type rectangular distribution with the same dose and centroid, namely, a rectangular distribution of width $x_W = 2x_C$ and volume density D_I/x_W. (Of course, x_W must be no larger than the depletion-layer width at threshold for this equivalence to hold true, and x_C must not lie within the inversion layer.) This weak dependence on the details of the profile leaves flexibility to satisfy other requirements, such as control of off current.

As already said, for gate biases $V_G > V_{TH}$ any gate charge above the threshold value is balanced mainly by inversion-layer charge. Thus, the additional oxide field, given by $(V_G - V_{TH})/d$, is related by Gauss's law to the inversion-layer carrier density approximately by

$$\varepsilon_{ox} \frac{V_G - V_{TH}}{d} \approx qN_{inv} \tag{24.41}$$

which shows that channel strength above threshold is proportional to $V_G - V_{TH}$, an approximation often used in this article. Thus, the switch in balancing gate charge from the depletion layer to the inversion layer causes N_{inv} to switch from an exponential gate-voltage dependence in subthreshold to a linear dependence above threshold.

For circuit analysis Eq. (24.41) is a convenient *definition* of V_{TH} because it fits current-voltage curves. If this definition is chosen instead of the charge-balance definition of Eq. (24.32), then Eqs. (24.32) and (24.38) result in an *approximation* to ϕ_{TH}.

Driving Ability and $I_{D,sat}$

The driving ability of the MOSFET is proportional to the current it can provide at a given gate bias. One might anticipate that the larger this current, the faster the circuit. Here this current is used to find some response times governing MOSFET circuits.

MOSFET current is dependent upon the carrier density in the channel, or upon $V_G - V_{TH}$, see Eq. (24.41). For a long-channel device, driving ability depends also on channel length. The shorter the channel length, L, the greater the driving ability, because the channel resistance is directly proportional to the channel length. Although it is an oversimplification, let us suppose that the MOSFET is primarily in saturation during the driving of its load. This simplification will allow a clear discussion of the issues involved in making faster MOSFET's without complicated mathematics. Assuming the MOSFET to be saturated over most of the switching period, driving ability is proportional to current in saturation, or to

$$I_{D,sat} = \frac{\varepsilon_{ox} Z\mu}{2dL} \left(V_G - V_{TH}\right)^2 \tag{24.42}$$

where the factor of two results from the saturating behavior of the *I-V* curves at large drain biases and Z is the width of the channel normal to the direction of current flow. Evidently, for long devices driving ability is quadratic in $V_G - V_{TH}$, and inversely proportional to d.

The result of Eq. (24.42) holds for long devices. For short-channel devices, as explained for Fig. 24.44, the larger fields exerted by the drain electrode cause *velocity saturation* and, as a result, $I_{D,\text{sat}}$ is given roughly by [Einspruch and Gildenblat, 1989]

$$I_{D,\text{sat}} \approx \frac{\varepsilon_{\text{ox}} Z \upsilon_{\text{sat}}}{d} \frac{(V_G - V_{TH})^2}{V_G - V_{TH} + F_{\text{sat}} L} \tag{24.43}$$

where υ_{sat} is the carrier saturation velocity, about 10^7 cm/s for silicon at 290 K, F_{sat} is the field at which velocity saturation sets in, about 5×10^4 V/cm for electrons and not well established as $\gtrsim 10^5$ V/cm for holes in silicon MOSFETs. For Eq. (24.43) to agree with Eq. (24.42) at long L, we need $\mu \approx 2\upsilon_{\text{sat}}/F_{\text{sat}} \approx 400$ cm^2(V·s) for electrons in silicon MOSFETs, which is only roughly correct. Nonetheless, we can see that for devices in the submicron channel length regime, $I_{D,\text{sat}}$ tends to become independent of channel length L and becomes more linear with $V_G - V_{TH}$ and less quadratic (see Fig. 24.44). Equation (24.43) shows that velocity saturation is significant when $V_G/L \gtrsim F_{\text{sat}}$ for example, when $L \lesssim 0.5$ μm if $V_G - V_{TH} = 2.5$ V.

To relate $I_{D,\text{sat}}$ to a gate response time, τ_G, consider one MOSFET driving an identical MOSFET as load capacitance. Then the current from Eq. (24.43) charges this capacitance to a voltage V_G in a gate response time τ_G given by [Shoji, 1988]

$$\begin{aligned} \tau_G &= \frac{C_G V_G}{I_{D,\text{sat}}} \\ &= \frac{L}{\upsilon_{\text{sat}}} \left(1 + \frac{C_{\text{par}}}{C_{\text{ox}}}\right) \frac{V_G (V_G - V_{TH} + F_{\text{sat}} L)}{(V_G - V_{TH})^2} \end{aligned} \tag{24.44}$$

where C_G is the MOSFET gate capacitance $C_G = C_{\text{ox}} + C_{\text{par}}$, with $C_{\text{ox}} = \varepsilon_{\text{ox}} ZL/d$ the MOSFET oxide capacitance, and C_{par} the parasitic component of the gate capacitance [Chen, 1990]. The parasitic capacitance C_{par} is due mainly to overlap of the gate electrode over the source and drain and partly to fringing-field and channel-edge capacitances. For short-channel lengths, C_{par} is a significant part of C_G, and keeping C_{par} under control as L is reduced is an objective of gate-drain alignment technology. Typically, $V_{TH} \approx V_G/4$, so

$$\tau_G \approx \left(\frac{L}{\upsilon_{\text{sat}}}\right)\left(1 + \frac{C_{\text{par}}}{C_{\text{ox}}}\right)\left(1.3 + 1.8 \frac{F_{\text{sat}} L}{V_G}\right) \tag{24.45}$$

Thus, on an intrinsic level, the gate response time is closely related to the transit time of an electron from source to drain, which is L/υ_{sat} in velocity saturation. At shorter L, a linear reduction in delay with L is predicted, while for longer devices the improvement can be quadratic in L, depending upon how V_G is scaled as L is reduced.

The gate response time is not the only delay in device switching, because the drain-body *pn* junction also must charge or discharge for the MOSFET to change state [Shoji, 1988]. Hence, we must also consider a drain response time τ_D. Following Eq. (24.44), we suppose that the drain capacitance C_D is charged by the supply voltage through a MOSFET in saturation so that

$$\tau_D = \frac{C_D V_G}{I_{D,\text{sat}}} = \frac{C_D}{C_G} \tau_G \tag{24.46}$$

Equation (24.46) suggests that τ_D will show a similar improvement to τ_G as L is reduced, provided that C_D/C_G does not increase as L is reduced. However, $C_{ox} \propto L/d$, and the major component of C_{par}, namely, the overlap capacitance contribution, leads to $C_{par} \propto L_{ovlp}/d$ where L_{ovlp} is roughly three times the length of overlap of the gate over the source or drain [Chen, 1990]. Then $C_G \propto (L + L_{ovlp})/d$ and, to keep the C_D/C_G ratio from increasing as L is reduced, either C_D or oxide-thickness d must be reduced along with L.

Clever design can reduce C_D. For example, various *raised-drain* designs reduce the drain-to-body capacitance by separating much of the drain area from the body using a thick oxide layer. The contribution to drain capacitance stemming from the sidewall depletion-layer width next to the channel region is more difficult to handle, because the sidewall depletion layer is deliberately reduced during miniaturization to avoid *short-channel* effects, that is, drain influence upon the channel in competition with gate control. As a result this sidewall contribution to the drain capacitance tends to increase with miniaturization unless junction depth can be shrunk.

Equations (24.45) and (24.46) predict reduction of response times by reduction in channel length L. Decreasing oxide thickness leads to no improvement in τ_G, but Eq. (24.46) shows a possibility of improvement in τ_D, because C_D is independent of d while C_G increases as d decreases. The *ring oscillator*, a closed loop of an odd number of inverters, is a test circuit whose performance depends primarily on τ_G and τ_D. Gate delay/stage for ring oscillators is found to be near 12 ps/stage at 0.1 μm channel length, and 60 ps/stage at 0.5 μm.

For circuits, interconnection capacitances and fan-out (multiple MOSFET loads) will increase response times beyond the device response time, even when parasitics are taken into account. Thus, we are led to consider interconnection delay, τ_{INT}. Although a lumped model suggests, as with Eq. (24.46), that $\tau_{INT} \approx (C_{INT}/C_G)\,\tau_G$, the length of interconnections requires a *distributed* model. Interconnection delay is then

$$\tau_{INT} = \frac{R_{INT}C_{INT}}{2} + R_{INT}C_G + \left(1 + \frac{C_{INT}}{C_G}\right)\tau_G \tag{24.47}$$

where the new symbols are R_{INT} = interconnection resistance, C_{INT} = interconnection capacitance, and we have assumed that the interconnection joins a MOSFET driver in saturation to a MOSFET load C_G. For small R_{INT}, τ_{INT} is dominated by the last term, which resembles Eqs. (24.44) and (24.46). However, unlike the ratio C_D/C_G in Eq. (24.46), it is difficult to reduce or even maintain the ratio C_{INT}/C_G in Eq. (24.47) as L is reduced. Remember, $C_G \propto Z\,(L + L_{ovlp})/d$. Reduction of L therefore tends to increase C_{INT}/C_G, especially because interconnect cross sections cannot be reduced without impractical increases in R_{INT}. What is worse, along with reduction in L, chip sizes usually increase, making line lengths longer, increasing R_{INT} even at constant cross section. As a result, interconnection delay becomes a major problem as L is reduced. The obvious way to keep C_{INT}/C_G under control is to increase the device width Z so that $C_G \propto Z\,(L + L_{ovlp})/d$ remains constant as L is reduced. A better way is to cascade drivers of increasing Z [Chen, 1990; Shoji, 1988]. Either solution requires extra area, however, reducing the packing density that is a major objective in decreasing L in the first place. An alternative is to reduce the oxide thickness d, a major technology objective today.

Transconductance

Another important device parameter is the small-signal transconductance g_m [Sedra and Smith, 1991; Haznedar, 1991], which determines the amount of output current swing at the drain that results from a given input voltage variation at the gate, that is, the small-signal gain:

$$g_m = \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D=\text{const}} \tag{24.48}$$

Using the chain rule of differentiation, the transconductance in saturation can be related to the small-signal *transition* or *unity-gain frequency*, which determines at how high a frequency ω the small-signal current gain $|\imath_{out}/\imath_{in}| = g_m/(\omega C_G)$ drops to unity. Using the chain rule,

$$g_m = \frac{\partial I_{D,\text{sat}}}{\partial Q_G}\frac{\partial Q_G}{\partial V_G} = \omega_T C_G \tag{24.49}$$

where C_G is the oxide capacitance of the device, $C_G = \partial Q_G/\partial V_G|V_D$ with Q_G = the charge on the gate electrode. The frequency ω_T is a measure of the small-signal, high-frequency speed of the device, neglecting parasitic resistances. Using Eq. (24.43) in Eq. (24.49) we find that the transition frequency also is related to the transit time L/υ_{sat} of Eq. (24.45), so that both the digital and small-signal circuit speeds are related to this parameter.

Output Resistance and Drain Conductance

For small-signal circuits the output resistance r_o of the MOSFET [Sedra and Smith, 1991] is important in limiting the gain of amplifiers. This resistance is related to the small-signal drain conductance g_D in saturation by

$$r_o = \frac{1}{g_D} = \frac{\partial V_D}{\partial I_{D,\text{sat}}}\bigg|_{V_G = \text{const}} \tag{24.50}$$

If the MOSFET is used alone as a simple amplifier with a load line set by a resistor R_L, the gain becomes

$$\left|\frac{\upsilon_o}{\upsilon_{\text{in}}}\right| = g_m \frac{R_L r_o}{R_L + r_o} \le g_m R_L \tag{24.51}$$

showing how gain is reduced if r_o is reduced to a value approaching R_L.

As devices are miniaturized, r_o is decreased, g_D increased, due to several factors. At moderate drain biases, the main factor is channel-length modulation, the reduction of the channel length with increasing drain voltage that results when the depletion region around the drain expands toward the source, causing L to become drain-bias dependent. At larger drain biases, a second factor is drain control of the inversion-layer charge density, which can compete with gate control in short devices. This is the same mechanism discussed later in the context of subthreshold behavior. At rather high drain bias, carrier multiplication further lowers r_o.

In a digital inverter, a lower r_o widens the voltage swing needed to cause a transition in output voltage. This widening increases power loss due to current spiking during the transition, and reduces noise margins [Annaratone, 1986]. It is not, however, a first-order concern in device miniaturization for digital applications. Because small-signal circuits are more sensitive to r_o than digital circuits, MOSFETs designed for small-signal applications cannot be made as small as those for digital applications.

Limitations upon Miniaturization

A major factor in the success of the MOSFET has been its compatibility with processing useful down to very small dimensions. Today channel lengths (source-to-drain spacings) of 0.5 μm are manufacturable, and further reduction to 0.1 μm has been achieved for limited numbers of devices in test circuits such as ring oscillators. In this section some of the limits that must be considered in miniaturization are outlined [Brews, 1990].

Subthreshold Control

When a MOSFET is in the "off" condition, that is, when the MOSFET is in *subthreshold,* the off current drawn with the drain at supply voltage must not be too large in order to avoid power consumption and discharge of ostensibly isolated nodes [Shoji, 1988]. In small devices, however, the source and drain are closely spaced, so there exists a danger of direct interaction of the drain with the source, rather than an interaction mediated by the gate and channel. In an extreme case, the drain may draw current directly from the source, even though the gate is "off" (*punchthrough*). A less extreme but also undesirable case occurs when the drain and gate jointly control the carrier density in the channel (*drain-induced barrier lowering*, or drain control of threshold voltage).

In such a case, the on–off behavior of the MOSFET is not controlled by the gate alone, and switching can occur over a range of gate voltages dependent on the drain voltage. Reliable circuit design under these circumstances is very complicated, and testing for design errors is prohibitive. Hence, in designing MOSFETs, a drain-bias independent subthreshold behavior is necessary.

A measure of the range of influence of the source and drain is the depletion-layer width of the associated *pn* junctions. The depletion layer of such a junction is the region in which all carriers have been depleted, or pushed away, due to the potential drop across the junction. This potential drop includes the applied bias across the junction and a spontaneous *built-in* potential drop induced by spontaneous charge exchange when *p* and *n* regions are brought into contact. The depletion-layer width W of an abrupt junction is related to potential drop V and dopant-ion concentration/unit volume N by

$$W = \left(\frac{2\varepsilon_s V}{qN} \right)^{1/2} \tag{24.52}$$

To avoid subthreshold problems, a commonly used rule of thumb is to make sure that the channel length is longer than a minimum length $L_{\min}$ related to the junction depth r_j, the oxide thickness d, and the depletion-layer widths W_S and W_D of the source and drain by [Brews, 1990]

$$L_{\min} = A[r_j d(W_S + W_D)^2]^{1/3} \tag{24.53}$$

where the empirical constant $A = 0.88\ \text{nm}^{-1/3}$ if r_j, W_S, and W_D are in micrometers and d is in nanometers.

Equation (24.53) shows that smaller devices require shallower junctions (smaller r_j), thinner oxides (smaller d), or smaller depletion-layer widths (smaller voltage levels or heavier doping). These requirements introduce side effects that are difficult to control. For example, if the oxide is made thinner while voltages are not reduced proportionately, then oxide fields increase, requiring better oxides. If junction depths are reduced, better control of processing is required, and the junction resistance is increased due to smaller cross sections. To control this resistance, various *self-aligned contact* schemes have been developed to bring the source and drain contacts closer to the gate [Brews, 1990; Einspruch and Gildenblat, 1989], reducing the resistance of these connections. If depletion-layer widths are reduced by increasing the dopant-ion density the *driving ability* of the MOSFET suffers because the threshold voltage increases. That is, Q_D increases in Eq. (24.40), reducing $V_G - V_{TH}$. Thus, for devices that are not velocity-saturated, that is, devices where $V_G/L \lesssim F_{\text{sat}}$, increasing V_{TH} results in slower circuits.

As secondary consequences of increasing dopant-ion density, channel conductance is further reduced due to the combined effects of increased scattering of electrons from the dopant atoms and increased oxide fields that pin carriers in the inversion layer closer to the insulator–semiconductor interface, increasing scattering at the interface. These effects also reduce driving ability, although for shorter devices they are important only in the linear region (that is, below saturation), assuming that mobility μ is more strongly affected than saturation velocity υ_{sat}.

Hot-Electron Effects

Another limit upon how small a MOSFET can be made is a direct result of the larger fields in small devices. Let us digress to consider why proportionately larger voltages, and thus larger fields, are used in smaller devices. First, according to Eq. (24.45), τ_G is shortened if voltages are increased, at least so long as $V_G/L \lesssim F_{\text{sat}}$ 5×10^4 V/cm. If τ_G is shortened this way, then so are τ_D and τ_{INT}, Eqs. (24.46) and (24.47). Thus, faster response is gained by increasing voltages into the velocity saturation region. Second, the fabricational control of smaller devices has not improved proportionately as L has shrunk, so there is a larger percentage variation in device parameters with smaller devices. Thus, disproportionately larger voltages are needed to ensure that all devices operate in the circuit, to overcome this increased fabricational "noise." Thus, to increase speed and to cope with fabricational variations, fields go up in smaller devices.

As a result of these larger fields along the channel direction, a small fraction of the channel carriers have enough energy to enter the insulating layer near the drain. In silicon-based *p*-channel MOSFETs, energetic holes can become trapped in the oxide, leading to a positive oxide charge near the drain that reduces the strength of the channel, degrading device behavior. In *n*-channel MOSFETs, energetic electrons entering the oxide create interface traps and oxide wear-out, eventually leading to gate-to-drain shorts [Pimbley et al., 1989].

To cope with these problems "drain-engineering" has been tried, the most common solution being the *lightly doped drain* [Chen, 1990; Einspruch and Gildenblat, 1989; Pimbley et al., 1989]. In this design, a lightly doped extension of the drain is inserted between the channel and the drain proper. To keep the field moderate and reduce any peaks in the field, the lightly doped drain extension is designed to spread the drain-to-channel voltage drop as evenly as possible. The aim is to smooth out the field at a value close to F_{sat} so that energetic carriers are kept to a minimum. The expense of this solution is an increase in drain resistance and a decreased gain. To increase packing density, this lightly doped drain extension can be stacked vertically alongside the gate, rather than laterally under the gate, to control the overall device area.

Thin Oxides

According to Eq. (24.53), thinner oxides allow shorter devices and therefore higher packing densities for devices. In addition, driving ability is increased, shortening response times for capacitive loads, and output resistance and transconductance are increased. There are some basic limitations upon how thin the oxide can be made. For instance, there is a maximum oxide field that the insulator can withstand. It is thought that the intrinsic breakdown voltage of SiO_2 is of the order of 10^7 V/cm, a field that can support $\approx 2 \times 10^{13}$ charges/cm^2, a large enough value to make this field limitation secondary. Unfortunately, as they are presently manufactured, the intrinsic breakdown of MOSFET oxides is much less likely to limit fields than defect-related leakage or breakdown, and control of these defects has limited reduction of oxide thicknesses in manufacture to about 5 nm to date.

If defect-related problems could be avoided, the thinnest useful oxide would probably be about 3 nm, limited by direct tunneling of channel carriers to the gate. This tunneling limit is not well established, and also is subject to oxide-defect enhancement due to tunneling through intermediate defect levels. Thus, the manufacture of thin oxides is a very active area of exploration.

Dopant-Ion Control

As devices are made smaller, the precise positioning of dopant inside the device is critical. At high temperatures during processing, dopant ions can move. For example, source and drain dopants can enter the channel region, causing position dependence of threshold voltage. Similar problems occur in isolation structures that separate one device from another [Pimbley et al., 1989; Einspruch and Gildenblat, 1989; Wolf, 1995].

To control these thermal effects, process sequences are carefully designed to limit high-temperature steps. This design effort is shortened and improved by the use of computer modeling of the processes. Dopant-ion movement is complex, however, and its theory is made more difficult by the growing trend to use *rapid thermal processing* that involves short-time heat treatments. As a result, dopant response is not steady state, but transient. Computer models of transient response are primitive, forcing further advance in small-device design to be more empirical.

Other Limitations

Besides limitations directly related to the MOSFET, there are some broader difficulties in using MOSFETs of smaller dimension in chips involving even greater numbers of devices. Already mentioned is the increased delay due to interconnections that are lengthening due to increasing chip area and increasing complexity of connection. The capacitive loading of MOSFETs that must drive signals down these lines can slow circuit response, requiring extra circuitry to compensate. Another limitation is the need to isolate devices from each other [Brews, 1990; Chen 1990; Einspruch and Gildenblat, 1989; Pimbley et al., 1989; Wolf, 1995], so their actions remain uncoupled by parasitics. As isolation structures are reduced in size to increase device densities, new parasitics are discovered. A developing solution to this problem is the manufacture of circuits on insulating substrates, silicon-on-insulator technology [Colinge, 1991]. To succeed, this approach must deal with new problems, such as the electrical quality of the underlying silicon–insulator interface and the defect densities in the silicon layer on top of this insulator.

Defining Terms

Channel: The conducting region in a MOSFET between source and drain. In an *enhancement*-mode (or normally off) MOSFET, the channel is an inversion layer formed by attraction of minority carriers toward the gate. These carriers form a thin conducting layer that is prevented from reaching the gate by a thin *gate-oxide* isulating layer when the gate bias exceeds *threshold*. In a *buried-channel*, or *depletion-mode* (or normally on) MOSFET, the channel is present even at zero gate bias, and the gate serves to increase the channel resistance when its bias is nonzero. Thus, this device is based on majority-carrier modulation, like a MESFET.

Gate: The control electrode of a MOSFET. The voltage on the gate capacitively modulates the resistance of the connecting channel between the source and drain.

Source, drain: The two output contacts of a MOSFET, usually formed as *pn* junctions with the *substrate* or *body* of the device.

Strong inversion: The range of gate biases corresponding to the "on" condition of the MOSFET. At a fixed gate bias in this region, for low drain-to-source biases the MOSFET behaves as a simple gate-controlled resistor. At larger drain biases, the channel resistance can increase with drain bias, even to the point that the current *saturates*, or becomes independent of drain bias.

Substrate or body: The portion of the MOSFET that lies between the *source* and *drain* and under the *gate*. The gate is separated from the body by a thin *gate insulator*, usually silicon dioxide. The gate modulates the conductivity of the body, providing a gate-controlled resistance between the source and drain. The body is sometimes *dc*-biased to adjust overall circuit operation. In some circuits the body voltage can swing up and down as a result of input signals, leading to "body-effect" or "back-gate bias" effects that must be controlled for reliable circuit response.

Subthreshold: The range of gate biases corresponding to the "off" condition of the MOSFET. In this regime the MOSFET is not perfectly "off" but conducts a leakage current that must be controlled to avoid circuit errors and power consumption.

Threshold: The gate bias of a MOSFET that marks the boundary between "on" and "off" conditions.

Related Topic

13.2 Parameter Extraction for Analog Circuit Simulation

References

The following references are not to the original sources of the ideas discussed in this article, but have been chosen to be generally useful to the reader.

M. Annaratone, *Digital CMOS Circuit Design,* Boston: Kluwer Academic, 1986.

J. R. Brews, "Physics of the MOS transistor" in *Applied Solid State Science*, *Supplement 2A*, D. Kahng, Ed., New York: Academic, 1981.

J. R. Brews, "The submicron MOSFET" in *High-Speed Semiconductor Devices,* S. M. Sze, Ed., New York: Wiley, 1990, pp. 139–210.

J. Y. Chen, *CMOS Devices and Technology for VLSI,* Englewood Cliffs, N.J.: Prentice-Hall, 1990.

J.-P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI,* Boston: Kluwer Academic, 1991.

H. Haznedar, *Digital Microelectronics,* Redwood City, Calif.: Benjamin-Cummings, 1991.

M.A. Hollis and R.A. Murphy, "Homogeneous field-effect transistors," in *High-Speed Semiconductor Devices,* S. M. Sze, Ed., New York: Wiley, 1990, pp. 211–282.

N.G. Einspruch and G. Sh. Gildenblat, Eds., *VLSI Microstructure Science*, vol. 18, *Advanced MOS Device Physics*, New York: Academic, 1989.

N.R. Malik, *Electronic Circuits: Analysis, Simulation, and Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1995.

E.H. Nicollian and J.R. Brews, *MOS Physics and Technology*, New York: Wiley, 1982, chap. 1.

S.J. Pearton and N.J. Shaw, "Heterostructure field-effect transistors" in *High-Speed Semiconductor Devices,* S.M. Sze, Ed., New York: Wiley, 1990, pp. 283–334.

R.F. Pierret, *Modular Series on Solid State Devices, Field Effect Devices,* 2nd ed., vol. 4, Reading, Mass.: Addison-Wesley, 1990.
J.M. Pimbley, M. Ghezzo, H.G. Parks, and D.M. Brown, *VLSI Electronics Microstructure Science, Advanced CMOS Process Technology,* vol. 19, N. G. Einspruch, Ed., New York: Academic, 1989.
S.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., Philadelphia: Saunders, 1991.
M. Shoji, *CMOS Digital Circuit Technology,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.
S. Wolf, Silicon Processing for the VLSI era: volume 3 — the submicron MOSFET, Sunset Beach, CA: Lattice Press, 1995.

Further Information

The references given in this section have been chosen to provide more detail than is possible to provide in the limited space of this article. In particular, Annaratone [1986] and Shoji [1988] provide much more detail about device and circuit behavior. Chen [1990], Pimbley et al. [1989], and Wolf [1995] provide many technological details of processing and its device impact. Haznedar [1991], Sedra and Smith [1991], and Malik [1995] provide much information about circuits. Brews [1981] and Pierret [1990] provide good discussions of the derivation of the device current-voltage curves and device behavior in all bias regions.

Brewer, J.E., Zargham, M.R., Tragoudas, S., Tewksbury, S. “Integrated Circuits”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

25

Integrated Circuits

25.1 Integrated Circuit Technology
Technology Perspectives • Technology Generations • National Technology Roadmap for Semiconductors

25.2 Layout, Placement, and Routing
What Is Layout? • Floorplanning Techniques • Placement Techniques • Routing Techniques

25.3 Application-Specific Integrated Circuits
Introduction • Primary Steps of VLSI ASIC Design • Increasing Impact of Interconnection Delays on Design • General Transistor-Level Design of CMOS Circuits • ASIC Technologies • Interconnection Performance Modeling • Clock Distribution • Power Distribution • Analog and Mixed-Signal ASICs

Joe E. Brewer
Northrop Grumman Corporation

Medhi R. Zargham and Spyros Tragoudas
Southern Illinois University

Stuart Tewksbury
West Virginia University

25.1 Integrated Circuit Technology

Joe E. Brewer

Integrated circuit (IC) technology, the cornerstone of the modern electronics industry, is subject to rapid change. Electronic engineers, especially those engaged in research and development, can benefit from an understanding of the structure and pattern of growth of the technology.

Technology Perspective

A solid state IC is a group of interconnected circuit elements formed on or within a continuous substrate. While an integrated circuit may be based on many different material systems, silicon is by far the dominant material. More than 98% of contemporary electronic devices are based on silicon technology. On the order of 85% of silicon ICs are complementary metal oxide semiconductor (CMOS) devices.

From an economic standpoint the most important metric for an IC is the "level of functional integration." Since the invention of the IC by Jack Kilby in 1958, the level of integration has steadily increased. The pleasant result is that cost and physical size per function reduce continuously, and we enjoy a flow of new, affordable information processing products that pervade all aspects of our day-to-day lives. The historical rate of increase is a doubling of functional content per chip every 18 months.

For engineers who work with products that use semiconductor devices, the challenge is to anticipate and make use of these enhanced capabilities in a timely manner. It is not an overstatement to say that survival in the marketplace depends on rapid "design-in" and deployment.

For engineers who work in the semiconductor industry, or in its myriad of supporting industries, the challenge is to maintain this relentless growth. The entire industry is marching to a drumbeat. The cost of technology development and the investment in plant and equipment have risen to billions of dollars. Companies that lag behind face a serious loss of market share and, possibly, dire economic consequences.

Technology Generations

The concept of a technology generation emerged from analysis of historical records, was clearly defined by Gordon Moore in the 1960s, and codified as Moore's law. The current version of the law is that succeeding generations will support a four times increase in circuit complexity, and that new generations emerge at approximately 3-year intervals. The associated observations are that linear dimensions of device features change by a factor of 0.7, and the economically viable die size grows by a factor of 1.6.

Minimum feature size stated in microns (micrometers) is the term used most frequently to label a technology generation. "Feature" refers to a geometric object in the mask set such as a linewidth or a gate length. The "minimum feature" is the smallest dimension that can be reliably used to form the entity.

Figure 25.1 displays the technology evolution sequence. In the diagram succeeding generations are numbered using the current generation as the "0" reference. Because this material was written in 1996, the "0" generation is the 0.35 μm minimum feature size technology that began volume production in 1995.

An individual device generation has been observed to have a reasonably well-defined life cycle which covers about 17 years. The first year of volume manufacture is the reference point for a generation, but its lifetime actually extends further in both directions. As shown in Fig. 25.2, one can think of the stages of maturity as ranging over a linear scale which measures years to production in both the plus and minus directions. The 17-year life cycle of a single generation, with new generations being introduced at 3-year intervals, means that at any given time up to six generations are being worked on. This tends to blur the significance of research news and company announcements unless the reader is sensitive to the technology overlap in time.

To visualize this situation, consider Fig. 25.3. The top row lists calendar years. The second row shows how the life cycle of the 0.35 μm generation relates to the calendar. The third row shows the life cycle of the 0.25 μm generation vs. the calendar. Looking down any column corresponding to a specific calendar year, one can see which generations are active and identify their respective life cycle year.

generation 0	generation 1	generation 2	generation 3	generation 4	generation 5
1995	1998	2001	2004	2007	2010
0.35 μ	0.25 μ	0.18 μ	0.1 μ	0.07 μ	0.05 μ
production	development	research	research	research	research

FIGURE 25.1 Semiconductor technology generation time sequence.

			INDUSTRIAL RESEARCH				DEVELOPMENT					MANUFACTURING				
		UNIVERSITY RESEARCH					feasibility		productization							
-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5

FIGURE 25.2 Life cycle of a semiconductor technology generation.

95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11
0.35 μm	1	2	3	4	5											
-3	-2	-1	0.25 μm	1	2	3	4	5								
-6	-5	-4	-3	-2	-1	0.18 μm	1	2	3	4	5					
-9	-8	-7	-6	-5	-4	-3	-2	-1	0.13 μm	1	2	3	4	5		
	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0.10 μm	1	2	3	4
				-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0.07 μm	1

FIGURE 25.3 Time overlap of semiconductor technology generations.

One should not interpret the 17-year life cycle as meaning that no work is being performed that is relevant to a generation before the 17-year period begins. For example, many organizations are conducting experiments directed at transistors with gate lengths smaller than 0.1 μm. This author's interpretation is that when basic research efforts have explored technology boundary conditions, the conditions are ripe for a specific generation to begin to coalesce as a unique entity. When a body of research begins to seek compatible materials and processes to enable design and production at the target feature size, the generation life cycle begins. This is a rather diffused activity at first, and it becomes more focused as the cycle proceeds.

National Technology Roadmap for Semiconductors

The National Technology Roadmap for Semiconductors (NTRS) is an almost 200-page volume distributed by the Semiconductor Industry Association (SIA). Focused on mainstream leading edge technology, the roadmap provides a common vision for the industry. It enables a degree of cooperative precompetitive research and development among the fiercely competitive semiconductor device manufacturers. It is a dynamic document which will be revised and reissued to reflect learning on an as-needed basis.

The NTRS is compiled by engineers and scientists from all sectors of the U.S. IC technology base. Industry, academia, and government organizations participate in its formulation. Key leaders are the Semiconductor Research Corporation (SRC) and SEMATECH industry consortia. The roadmap effort is directed by the Roadmap Coordinating Group (RCG) of the SIA.

The starting assumption of the NTRS is that Moore's law will continue to describe the growth of the technology. The overall roadmap comprises many individual roadmaps which address defined critical areas of semiconductor research, development, engineering, and manufacturing. In each area, needs and potential solutions for each technology generation are reviewed. Of course, this process is more definitive for the early generations because knowledge is more complete and the range of alternatives is restricted.

The NTRS document provides a convenient summary table which presents some of the salient characteristics of the six technology generations ranging from 1995 to 2010. That summary is reproduced (with minor variations in format) as Table 25.1.

TABLE 25.1 Overall Roadmap Technology Characteristics

	Year of First DRAM Shipment/Minimum Feature (μm)					
	1995/0.35	1998/0.25	2001/0.18	2004/0.13	2007/0.10	2010/0.07
Memory						
Bits/chip (DRAM/Flash)	64M	256M	1G	4G	16G	64G
Cost/bit @ volume (millicents)	0.017	0.007	0.003	0.001	0.0005	0.0002
Logic (high-volume microprocessor)						
Logic transistors/cm^2 (packed)	4M	7M	13M	25M	50M	90M
Bits/cm^2 (cache SRAM)	2M	6M	20M	50M	100M	300M
Cost/transistor @ volume (millicents)	1	0.5	0.2	0.1	0.05	0.02
Logic (low-volume ASIC)						
Transistors/cm^2 (auto layout)	2M	4M	7M	12M	25M	40M
Non-recurring engineering Cost/transistor (millicents)	0.3	0.1	0.05	0.03	0.02	0.01
Number of chip I/Os						
Chip to package (pads) high performance	900	1350	2000	2600	3600	4800
Number of package pins/balls						
Microprocessor/controller	512	512	512	512	800	1024
ASIC (high performance)	750	1100	1700	2200	3000	4000
Package cost (cents/pin)	1.4	1.3	1.1	1.0	0.9	0.8
Chip frequency (MHz)						
On-chip clock, cost performance	150	200	300	400	500	625
On-chip clock, high performance	300	450	600	800	1000	1100
Chip-to-board speed, high performance	150	200	250	300	375	475
Chip size (mm^2)						
DRAM	190	280	420	640	960	1400
Microprocessor	250	300	360	430	520	620

TABLE 25.1 (continued) Overall Roadmap Technology Characteristics

	Year of First DRAM Shipment/Minimum Feature (μm)					
	1995/0.35	1998/0.25	2001/0.18	2004/0.13	2007/0.10	2010/0.07
ASIC	450	660	750	900	1100	1400
Max number wiring levels (logic)						
On-chip	4–5	5	5–6	6	6–7	7–8
Electrical defect density (d/m^2)	240	160	140	120	100	25
Minimum mask count	18	20	20	22	22	24
Cycle time days (theoretical)	9	10	10	11	11	12
Maximum substrate diameter (mm)						
Bulk or epitaxial or SOI wafer	200	200	300	300	400	400
Power supply voltage (V)						
Desktop	3.3	2.5	1.8	1.5	1.2	0.9
Battery	2.5	1.8–2.5	0.9–1.8	0.9	0.9	0.9
Maximum power						
High performance with heatsink (W)	80	100	120	140	160	180
Logic without heatsink (W)	5	7	10	10	10	10
Battery (W)	2.5	2.5	3.0	3.5	4.0	4.5
Design and test						
Volume tester cost/pin ($K)	3.3	1.7	1.3	0.7	0.5	0.4
Number of test vectors (μP/M)	16–32	16–32	16–32	8–16	4–8	4
% IC function with BIST/DFT	25	40	50	70	90	90+

Related Topics

1.1 Resistors • 23.1 Processes

Further Information

The NTRS is available from the SIA, 181 Metro Drive, Suite 450, San Jose, CA 95110, telephone 408-436-6600, fax 408-436-6646. The document can also be accessed via the SEMATECH home page at <http://www.sematech.org>.

Information concerning the IC life cycle can be found in Larrabee, G. B. and Chatterjee, P. "DRAM Manufacturing in the 90's — Part 1: The History Lesson" and "Part 2: The Roadmap," Semiconductor International, pp. 84–92, May 1991.

25.2 Layout, Placement, and Routing

Mehdi R. Zargham and Spyros Tragoudas

Very large scale integrated (VLSI) electronics presents a challenge, not only to those involved in the development of fabrication technology, but also to computer scientists, computer engineers, and electrical engineers. The ways in which digital systems are structured, the procedures used to design them, the trade-offs between hardware and software, and the design of computational algorithms will all be greatly affected by the coming changes in integrated electronics.

A VLSI chip can today contain millions of transistors and is expected to contain more than 100 million transistors in the year 2000. One of the main factors contributing to this increase is the effort that has been invested in the development of computer-aided design (CAD) systems for VLSI design. The VLSI CAD systems are able to simplify the design process by hiding the low-level circuit theory and device physics details from the designer, and allowing him or her to concentrate on the functionality of the design and on ways of optimizing it.

A VLSI CAD system supports descriptions of hardware at many levels of abstraction, such as system, subsystem, register, gate, circuit, and layout levels. It allows designers to design a hardware device at an abstract level and progressively work down to the layout level. A layout is a complete geometric representation (a set of rectangles) from which the latest fabrication technologies directly produce reliable, working chips. A VLSI

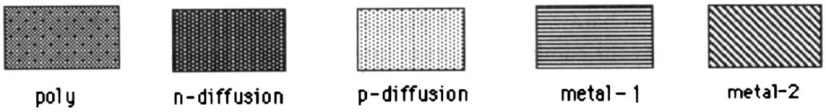


FIGURE 25.4 Different layers.

CAD system also supports verification, synthesis, and testing of the design. Using a CAD system, the designer can make sure that all of the parts work before actually implementing the design.

A variety of VLSI CAD systems are commercially available that perform all or some of the levels of abstraction of design. Most of these systems support a *layout editor* for designing a circuit **layout.** A layout-editor is software that provides commands for drawing lines and boxes, copying objects, moving objects, erasing unwanted objects, and so on. The output of such an editor is a design file that describes the layout. Usually, the design file is represented in a standard format, called Caltech Intermediate Form (CIF), which is accepted by the fabrication industry.

What Is Layout?

For a specific circuit, a layout specifies the position and dimension of the different layers of materials as they would be laid on the silicon wafer. However, the layout description is only a symbolic representation, which simplifies the description of the actual fabrication process. For example, the layout representation does not explicitly indicate the thickness of the layers, thickness of oxide coating, amount of ionization in the transistors channels, etc., but these factors are implicitly understood in the fabrication process. Some of the main layers used in any layout description are *n*-diffusion, *p*-diffusion, poly, metal-1, and metal-2. Each of these layers is represented by a polygon of a particular color or pattern. As an example, Fig. 25.4 presents a specific pattern for each layer that will be used through the rest of this section.

As is shown in Fig. 25.5, an *n*-diffusion layer crossing a poly layer implies an nMOS transistor, and a *p*-diffusion crossing poly implies a pMOS transistor.

Note that the widths of diffusion and poly are represented with a scalable parameter called *lambda.* These measurements, referred to as *design rules,* are introduced to prevent errors on the chip, such as preventing thin lines from opening (disconnecting) and short circuiting.

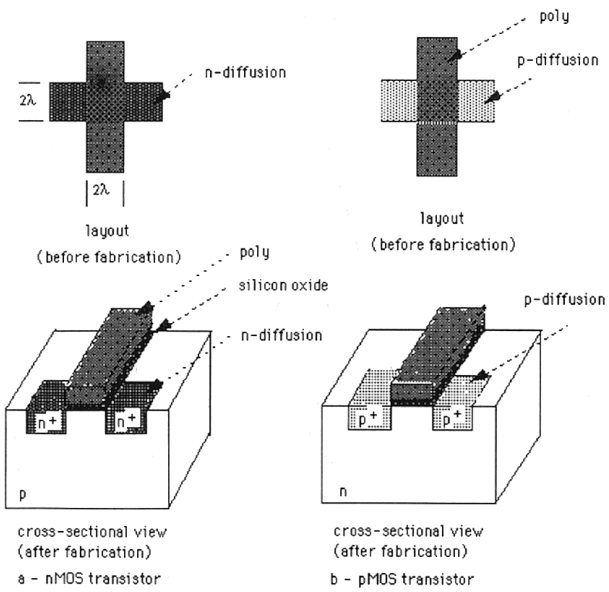


FIGURE 25.5 Layout and fabrication of MOS transistors.

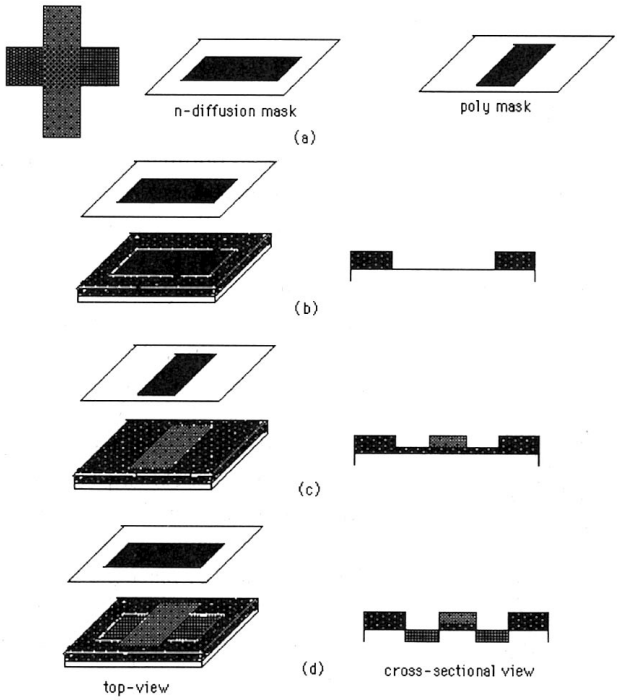


FIGURE 25.6 Fabrication steps for an nMOS transistor.

Implementing the design rules based on lambda makes the design process independent of the fabrication process. This allows the design to be rescaled as the fabrication process improves.

Metal layers are used as wires for connections between the components. This is because metal has the lowest propagation delay compared to the other layers. However, sometimes a poly layer is also used for short wires in order to reduce the complexity of the wire routing. Any wire can cross another wire without getting electrically affected as long as they are in different layers. Two different layers can be electrically connected together using *contacts.* The fabrication process of the contacts depends on types of the layers that are to be connected. Therefore, a layout editor supports different types of contacts by using different patterns.

From the circuit layout, the actual chip is fabricated. Based on the layers in the layout, various layers of materials, one on top of the others, are laid down on a silicon wafer. Typically, the processing of laying down each of these materials involves several steps, such as masking, oxide coating, lithography and etching [Mead and Conway, 1980]. For example, as shown in Fig. 25.6(a), for fabricating an nMOS transistor, first two masks, one for poly and one for *n*-diffusion, are obtained from the circuit layout. Next, the *n*-diffusion mask is used to create a layer of silicon oxide on the wafer [see Fig. 25.6(b)]. The wafer will be covered with a thin layer of oxide in places where the transistors are supposed to be placed as opposed to a thick layer in other places. The poly mask is used to place a layer of polysilicon on top of the oxide layer to define the gate terminals of the transistor [see Fig. 25.6(c)]. Finally, the *n*-diffusion regions are made to form the source and drain terminals of the transistor [see Fig. 25.6(d)].

To better illustrate the concept of layout design, the design of an inverter in the CMOS technology is shown in Fig. 25.7. An inverter produces an output voltage that is the logical inverse of its input. Considering the circuit diagram of Fig. 25.7(a), when the input is 1, the lower nMOS is on, but the upper pMOS is off. Thus, the output becomes 0 by becoming connected to the ground through the nMOS. On the other hand, if the input is 0, the pMOS is on and the nMOS is off, so the output must find a charge-up path through the pMOS to the supply and therefore becomes 1. Figure 25.7(b) represents a layout for such an inverter. As can be seen from this figure, the problem of a layout design is essentially reduced to drawing and painting a set of polygons. Layout editors provide commands for drawing such polygons. The commands are usually entered at the keyboard or with a mouse and, in some menu-driven packages, can be selected as options from a pull-down menu.

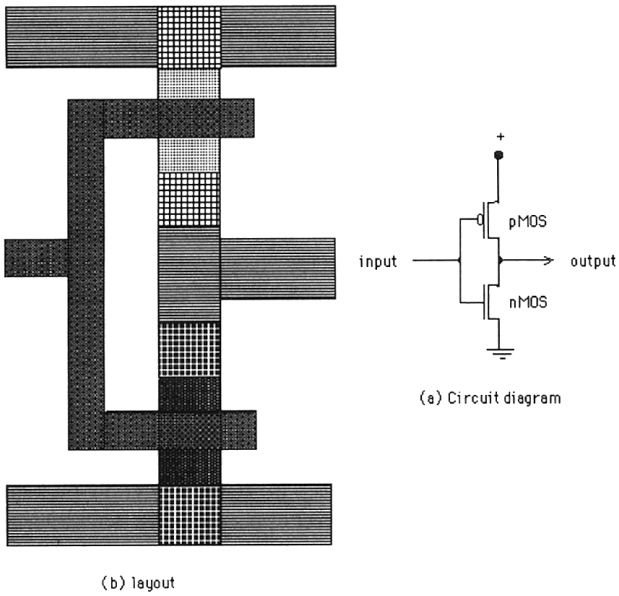


FIGURE 25.7 An inverter.

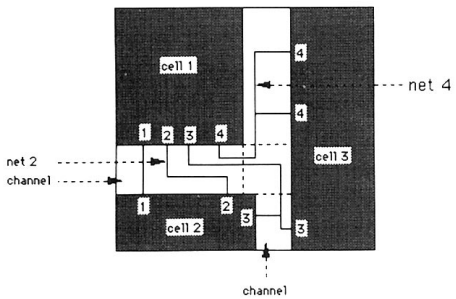


FIGURE 25.8 Placement and routing.,

In addition to the drawing commands, often a layout system provides tools for minimizing the overall area of the layout (i.e., size of the chip). Today a VLSI chip consists of a lot of individual cells, with each one laid out separately. A cell can be an inverter, a NAND gate, a multiplier, a memory unit, etc. The designer can make the layout of a cell and then store it in a file called the *cell library*. Later, each time the designer wants to design a circuit that requires the stored cell, he or she simply copies the layout from the cell library. A layout may consist of many cells. Most of the layout systems provide routines, called **floorplanning, placement** and **routing** routines, for placing the cells and then interconnecting them with wires in such a way that minimizes the layout area. As an example, Fig. 25.8 presents the placement of three cells. The area between the cells is used for routing. The entire routing surface is divided into a set of rectangular routing areas called channels. The sides of each channel consist of a set of terminals. A wire that connects the terminals with the same ID is called a net. The router finds a location for the wire segments of each net within the channel. The following sections classify various types of placement and routing techniques and provide an overview of the main steps of some of these techniques.

Floorplanning Techniques

The floorplanning problem in Computer Aided Design of Integrated Circuits is similar to that in Architecture and the goal is to find a location for each cell based on proximity (layout adjacency) criteria to other cells. We

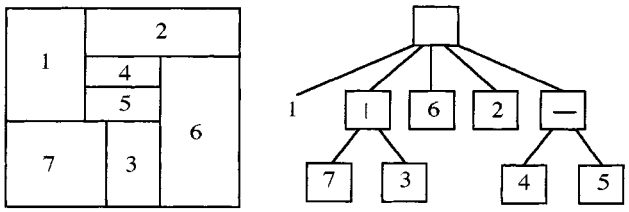


FIGURE 25.9 A hierarchical floorplan and its associated tree. The root node has degree 5. The internal node labeled with | indicates a vertical slicing. The internal node labeled with — indicates a horizontal slicing.

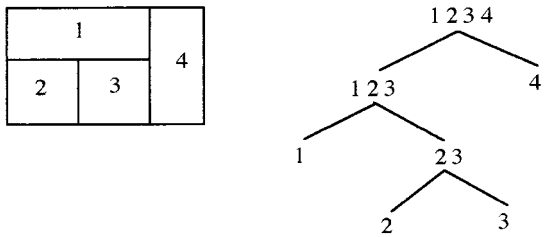


FIGURE 25.10 A sliceable floorplan and its associated binary tree.

consider rectangular floorplans whose boundaries are rectangles. It is desirable to obtain a floorplan that minimizes the overall area of the layout.

An important goal in floorplanning is the cell sizing problem where the goal is to determine the dimensions of variable cells whose area is invariant. All cells are assumed to be rectangular, and in the cell sizing problem the goal is to determine the width and height of each cell subject to predetermined upper and lower bounds on their ratio, and to their product being equal to its area, so that the final floorplan has optimal area.

One of the early approaches in floorplanning is the hierarchical, where recursive bipartition or partition into more than two parts is recursively employed and a floorplan tree is constructed. The tree simply reflects the hierarchical construction of the floorplan. Figure 25.9 shows a hierarchical floorplan and its associated tree. The partitioning problem and related algorithms are discussed extensively later in this section.

Many early hierarchical floorplanning tools insist that the floorplan be sliceable. A sliceable floorplan is recursively defined as follows: (a) a cell or (b) a floorplan that can be bipartitioned into two sliceable floorplans with either a horizontal or vertical line. Figure 25.10 shows a sliceable floorplan whose tree is binary.

Many tools that produce sliceable floorplans are still in use because of their simplicity. In particular, many problems arising in sliceable floorplanning are solvable optimally in polynomial time [Sarrafzadeh and Wong, 1996]. Unfortunately, sliceable floorplans are rarely optimal (in terms of their area), and they often result in layouts with very difficult routing phases. (Routing is discussed later in this section.) Figure 25.11 shows a compact floorplan that is not sliceable.

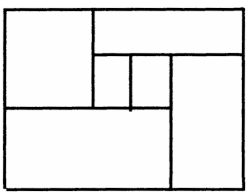


FIGURE 25.11 A compact layout that is not sliceable.

Hierarchical tools that produce nonsliceable floorplans have also been proposed [Sarrafzadeh and Wong, 1996]. The major problem in the development of such tools is that we are often facing problems that are intractable and thus we have to rely on heuristics in order to obtain fast solutions. For example, the cell sizing problem can be tackled optimally in sliceable floorplans [Otten, 1983 and Stockmeyer, 1983] but the problem is intractable for general nonsliceable floorplans.

A second approach to floorplanning is the rectangular dual graph. The idea here is to use duality arguments and express the cell adjacency constraints in terms of a graph, and then use an algorithm to translate the graph into a rectangular floorplan. A rectangular dual graph of a rectangular floorplan is a planar graph G = (V,E), where V is the set of cells and E is the set of edges, and an edge (C_1,C_2) is in E if and only if cells C_1 and C_2 are adjacent in the floorplan. See Fig. 25.12 for a rectangular floorplan and its rectangular dual graph G.

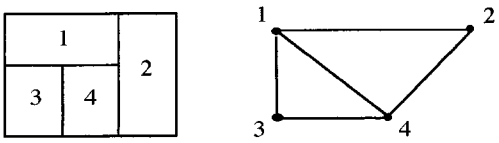


FIGURE 25.12 A rectangular floorplan and its associated dual planer graph.

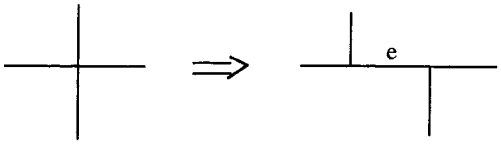


FIGURE 25.13 A cross junction can be replaced by 2 T-junctions.

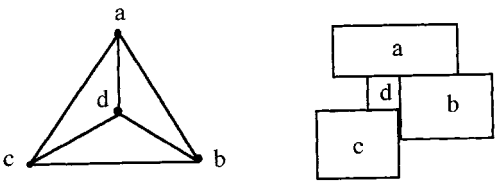


FIGURE 25.14 For a cycle of size 3 that is not a face we cannot satisfy all constraints.

Let us assume that the floorplan does not contain cross junctions. Figure 25.13 shows a cross junction. This restriction does not significantly increase the area of a floorplan because, as Fig. 25.13 shows, a cross junction can be replaced by two T-junctions by simply adding a short edge e.

It has been shown that in the absence of cross junctions the dual graph is planar triangulated (PT), and every T-junction corresponds to a triangulated face of the dual PT graph. Unfortunately, not all PT graphs have a rectangular floorplan. For example, in the graph of Fig. 25.14 we cannot satisfy the adjacency requirements of edges (a,b), (b,c) and (c,a) at the same time. Note that the later edges form a cycle of length three that is not a face. It has been shown that a PT graph has a rectangular floorplan if and only if it does not contain such cycles of length three. Moreover, a linear time algorithm to obtain such a floorplan has been presented [Sarrafzadeh and Wong, 1996]. The rectangular dual graph approach is a new method for floorplanning, and many floorplanning problems, such as the sizing problem, have not been tackled yet.

Rectangular floorplans can be obtained using simulated annealing and genetic algorithms. Both techniques are used to solve general optimization problems for which the solution space is not well understood. The approaches are easy to implement, but the algorithms have many parameters which require empirical adjustments, and the results are usually unpredictable.

A final approach to floorplanning, which unfortunately requires substantial computational resources and results to an intractable problem, is to formulate the problem as a mixed-integer linear programming (LP). Consider the following definitions:

W_i,H_i,R_i : width, height and area of cell C_i
X_i,Y_i : coordinates of lower left corner of cell C_i
X,Y : the width and height of the final floorplan
A_i,B_i : lower and upper bound for the ratio W_i/H_i of cell C_i
P_{ij}, Q_{ij} : variables that take 0/1 values for each pair of cells C_i and C_j

The goal is to find X_i,Y_i,W_i, and H_i for each cell so that all constraints are satisfied and XY is minimized. The latter is a nonlinear constraint. However, we can fix the width W and minimize the height of the floorplan as follows:

$$\min Y$$
$$X_i + W_i \le W$$
$$Y \ge Y_i + H_i$$

The complete mixed-integer LP formulation is [Sutanthavibul et al., 1991]:

$$\min Y$$
$$X_i, Y_i, W_i \ge 0$$
$$P_{ij}, Q_{ij} = 0 \text{ or } 1$$
$$X_i + W_i \le W$$
$$Y \ge Y_i + H_i$$
$$X_i + W_i \le X_j + W(P_{ij} + Q_{ij})$$
$$X_j + W_j \le X_i + W(1 - P_{ij} + Q_{ij})$$
$$Y_i + H_i \le Y_j + H(1 + P_{ij} - Q_{ij})$$
$$Y_j + H_j \le Y_i + H(2 - P_{ij} - Q_{ij})$$

When H_i appears in the above equations, it must be replaced (using first-order approximation techniques) by $H_i = D_i W_i + E_i$ where D_i and E_i are defined below:

$$W_{min} = \sqrt{R_i A_i}$$

$$W_{max} = \sqrt{R_i B_i}$$

$$H_{min} = \sqrt{R_i / B_i}$$

$$H_{max} = \sqrt{R_i / A_i}$$

$$D_i = (H_{max} - H_{min})/(W_{min} - W_{max})$$

$$E_i = H_{max} - D_i W_{min}$$

The unknown variables are X_i, Y_i, W_i, P_{ij}, and Q_{ij}. All other variables are known. The equations can then be fed into an LP solver to find a minimum cost solution for the unknowns.

Placement Techniques

Placement is a restricted version of floorplanning where all cells have fixed dimension. The objective of a placement routine is to determine an optimal position on the chip for a set of cells in a way that the total occupied area and total estimated length of connections are minimized. Given that the main cause of delay in a chip is the length of the connections, providing shorter connections becomes an important objective in placing a set of cells. The placement should be such that no cells overlap and enough space is left to complete all the connections.

All exact methods known for determining an optimal solution require a computing effort that increases exponentially with number of cells. To overcome this problem, many heuristics have been proposed [Preas and Lorenzetti, 1988]. There are basically three strategies of heuristics for solving the placement problem, namely, *constructive, partitioning,* and *iterative* methods. Constructive methods create placement in an incremental manner where a complete placement is only available when the method terminates. They often start by placing a *seed* (a seed can be a single cell or a group of cells) on the chip and then continuously placing other cells based on some heuristics such as size of cells, connectivity between the cells, design condition for connection lengths, or size of chip. This process continues until all the cells are placed on the chip. Partitioning methods divide the cells into two or more partitions so that the number of connections that cross the partition boundaries

is minimized. The process of dividing is continued until the number of cells per partition becomes less than a certain small number. Iterative methods seek to improve an initial placement by repeatedly modifying it. Improvement might be made by transforming one cell to a new position or switching positions of two or more cells. After a change is made to the current placement configuration based on some cost function, a decision is made to see whether to accept the new configuration. This process continues until an optimal (in most cases a near optimal) solution is obtained. Often the constructive methods are used to create initial placement on which an iterative method subsequently improves.

Constructive Method

In most of the constructive methods, at each step an unplaced cell is selected and then located in the proper area. There are different strategies for selecting a cell from the collection of unplaced cells [Wimer and Koren, 1988]. One strategy is to select the cell that is most strongly connected to already placed cells. For each unplaced cell, we find the total of its connections to all of the already placed cells. Then we select the unplaced cell that has the maximum number of connections. As an example consider the cells in Fig. 25.15. Assume that cells c_1 and c_2 are already placed on the chip. In Fig. 25.16 we see that cell c_5 has been selected as the next cell to be placed. This is because cell c_5 has the largest number of connections (i.e., three) to cells c_1 and c_2.

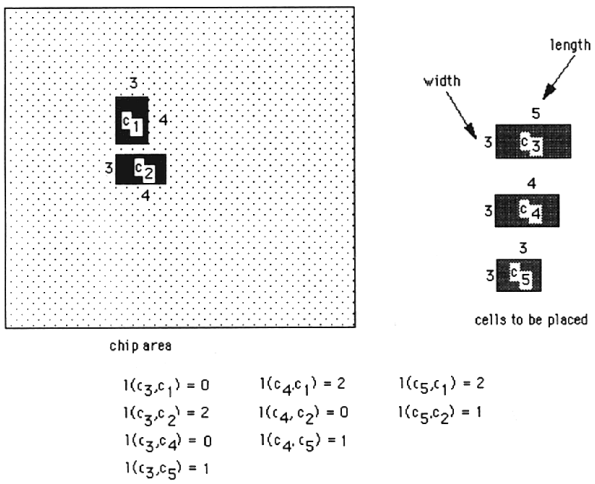


FIGURE 25.15 Initial configuration.

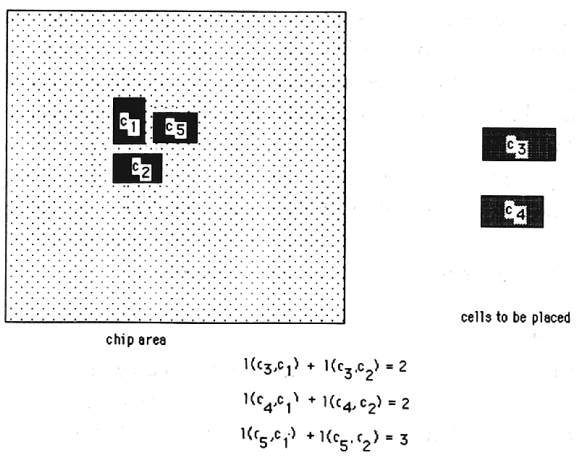


FIGURE 25.16 Selection based on the number of connections.

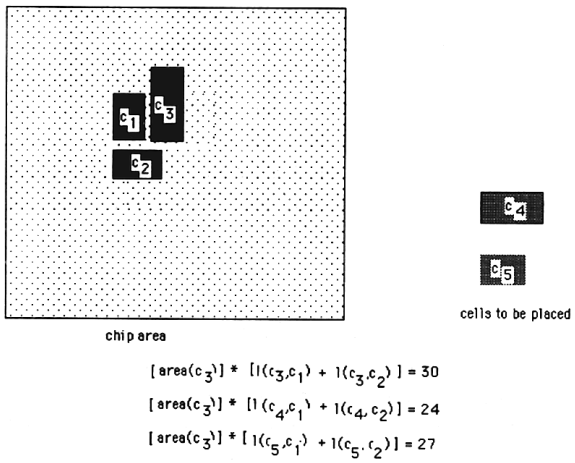


FIGURE 25.17 Selection based on the number of connections and area.

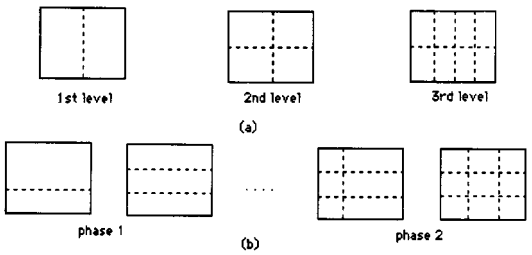


FIGURE 25.18 Partitioning.

The foregoing strategy does not consider area as a factor and thus results in fragmentation of the available free area; this may make it difficult to place some of the large unplaced cells later. This problem can be overcome, however, by considering the product of the number of connections and the area of the cell as a criteria for the selection. Figure 25.17 presents an example of such a strategy. Cell c_3 is selected as the next choice since the product of its area and its connections to c_1 and c_2 combine to associate with the maximum value.

Partitioning Method

The approaches for the partitioning method can be classified as quadratic and sliced bisection. In both approaches the layout is divided into two subareas, A and B, each having a size within a predefined range. Each cell is assigned to one of these subareas. This assignment is such that the number of interconnections between the two subareas is minimal. For example, Fig. 25.18 presents successive steps for the quadratic and sliced-bisection methods. As shown in Fig. 25.18(a), in the first step of the quadratic method the layout area is divided into two almost equal parts; in the second step the layout is further divided into four almost equal parts in the opposite direction. This process continues until each subarea contains only one cell. Similar to the quadratic method, the sliced-bisection also divides the layout area into several subareas.

The sliced-bisection method has two phases. In the first phase, the layout area is iteratively divided into a certain number of almost equal subareas in the same direction. In this way, we end up with a set of slices [see Fig. 25.18(b)]. Similarly, the second phase divides the area into a certain number of subareas; however, the slicing is done in the opposite direction.

Several heuristics have been proposed for each of the preceding partitioning methods. Here, for example, we emphasize the work of Fiduccia–Mattheyses [Fiduccia and Mattheyses, 1982], which uses the quadratic method. For simplicity, their algorithm is only explained for one step of this method. Initially the set of cells is randomly divided into two sets, A and B. Each set represents a subarea of the layout and has size equal to the area it represents. A cell is selected from one of these sets to be moved to the other set. The selection of the

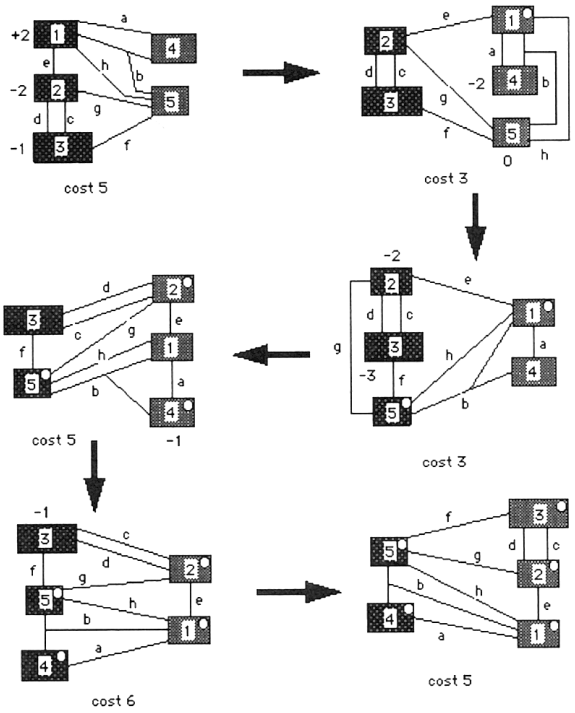


FIGURE 25.19 Illustration of a pass.

cell depends on three criteria. The cell should be free, i.e., the cell must have a minimum gain among the gains of all other free cells. A cell c has gain g if the number of interconnections between the cells of the two sets decreases by g units when c is moved from its current set to the other. Finally, the selected set's move should not violate a predefined balancing criterion that guarantees that the sizes of the two sets are almost equal. After moving the selected cell from its current set to the complementary set, it is no longer free. A new partition, which corresponds to the new instance of the two sets A and B, is created. The cost of a partition is defined as the number of interconnections between cells in the two sets of the partition. The Fiduccia–Mattheyses algorithm keeps track of the best partition encountered so far, i.e., the partition with the minimum cost. The algorithm will move the selected cell to the complementary set even if its gain is negative. In this case, the new partition is worse than the previous one, but the move can eventually lead to better partitions.

This process of generating new partitions and keeping track of the best encountered partition is repeated until no free cells are left. At this point of the process, which is called a pass, the algorithm returns the best partition and terminates. To obtain better partitions, the algorithm can be modified such that more passes occur. This can easily be done by selecting the best partition of a pass as the initial partition of the next pass. In this partition, however, all cells are free. The modified algorithm terminates whenever a new pass returns a partition that is no better than the partition returned by the previous pass. This way, the number of passes generated will never be more than the number of the interconnections in the circuit and the algorithm always terminates.

The balancing criterion in the Fiduccia–Mattheyses algorithm is easily maintained when the cells have uniform areas. The only way a pass can be implemented to satisfy the criterion is to start with a random initial partition in which the two sets differ by one cell, each time select the cell of maximum gain from the larger sized set, move the cell and generate a new partition, and repeat until no free cells are left.

In the example of Fig. 25.19, the areas of the cells are nonuniform. However, the assigned cell areas ensure that the previously described operation occurs so that the balancing criterion is satisfied. (The cell areas are omitted in this figure, but they correspond to the ones given in Fig. 25.15.) Figure 23.19 illustrates a pass of the Fiduccia–Mattheyses algorithm. During this pass, six different partitions are generated, i.e., the initial partition and five additional ones. Note that according to the description of the pass, the number of additional partitions equals the number of cells in the circuit.

Each partition consists of the cells of the circuit (colored according to the set to which they belong and labeled with an integer), the gain value associated with each free cell in the set from which the selected cell will be moved (this value can be a negative number), and the nets (labeled with letters). In the figure a cell that is no longer free is distinguished by an empty circle placed inside the rectangle that represents that cell.

The initial partition has cost 5 since nets *a*, *b*, *h*, *g*, *f* connect the cells in the two sets. Then the algorithm selects cell 1, which has the maximum gain. The new partition has cost 3 (nets *e*, *g*, *f*), and cell 1 is no longer free. The final partition has no free cells. The best partition in this pass has cost 3.

Iterative Method

Many iterative techniques have been proposed. Here, we emphasize one of these techniques called simulated annealing. Simulated annealing, as proposed by Kirkpatrick et al. [1983], makes the connection between statistical mechanics and combinatorial optimization problems. The main advantage with simulated annealing is its hill-climbing ability, which allows it to back out of inferior local solutions and find better solutions.

Sechen has applied simulated annealing to the placement problem and has obtained good solutions [Sechen, 1990]. The method basically involves the following steps:

BEGIN

1. Find an initial configuration by placing the cells randomly. Set the initial temperature, *T*, and the maximum number of iterations.
2. Calculate the cost of the initial configuration.
   A general form of the cost function may be: $Cost = c_1 * Area\ of\ layout + c_2 * Total\ interconnection\ length$ where c_1 and c_2 are tuning factors.
3. While (stopping criterion is not satisfied)
   {587
   a. For (maximum number of iteration)
      {
      1.Transform the old configuration into a new configuration.
      This transformation can be in the form of exchange of positions of two randomly selected cells or change of position of a randomly selected cell.
      2.Calculate the cost of the new configuration.
      3.If (new cost < old cost) accept the iteration, else check if the new iteration could be accepted with the probability: e ($|$new cost – old cost$|$ / *T*). There are also other options for the probability function.
      }
   b. Update the temperature.
      }

END

The parameter *T* is called *temperature;* it is initially set to a very large value, so that the probability of accepting "uphill" moves is very close to 1, that it is slowly decreasing toward zero, according to a rule called the cooling schedule. Usually, the new reduced temperature is calculated as follows:

$$\text{New temperature} = (\text{cooling rate}) \times (\text{Old temperature})$$

Using a faster cooling rate can result in getting stuck at local minima; however, a cooling rate that is too slow can pass over the possible global minima. In general, the cooling rate is taken from approximately 0.80 to 0.95.

Usually, the stopping criterion for the while-loop is implemented by recording the cost function's value at the end of each temperature stage of the annealing process. The stopping criterion is satisfied when the cost function's value has not changed for a number of consecutive stages.

Though simulated annealing is not the ultimate solution to placement problems, it gives very good results compared to the other popular techniques. The long execution time of this algorithm is its major disadvantage. Although a great deal of research has been done in improving this technique, substantial improvements have not been achieved.

Routing Techniques

Given a collection of cells placed on a chip, the routing problem is to connect the terminals (or ports) of these cells for a specific design requirement. The routing problem is often divided into three subproblems: global, detailed, and specialized routers. The global router considers the overall routing region in order to distribute the nets over the channels based on their capacities while keeping the length of each net as short as possible. For every channel that a net passes through, the net's id is placed on the sides of the channel. Once the terminals of each channel are determined, the detailed router connects all the terminals with the same id by a set of wire segments. (A wire segment is a piece of material described by a layer, two end-points, and a width.) The specialized router is designed to solve a specific problem such as routing of power and ground wires. These wires require special attention for two reasons: 1) they are usually routed in one layer in order to reduce the parasitic capacitance of contacts, and 2) they are usually wider than other wires (signal and data) since they carry more current.

Detailed routers further divide into general-purpose and restricted routers. The general-purpose routers impose very few constraints on the routing problem and operate on a single connection at a time. Since these routers work on the entire design in a serial fashion, the size of the problems they can attempt is limited. On the other hand, the restricted routers require some constraints on the routing problem, such as empty rectangular areas with all of the pins on the periphery. Because of their limited scope, these routers can do a better job of modeling the contention of nets for the routing resources and therefore can be viewed as routing the nets in parallel.

To reduce the complexity of restricted routers, this type of router often uses a rectangular grid on which trunks (horizontal wire segments) and branches (vertical wire segments) are placed on different layers. In other words, the layers supported by the technology are divided into two groups, horizontal and vertical. This is known as the Manhattan model. On the other hand, in a non-Manhattan model, the assignment of a layer to a vertical or horizontal direction is not enforced. Given the freedom of direction, assignment to layers reduces the channel width and vias in many cases; the latter model usually produces a better result than the former.

In the literature, many different techniques have been proposed for restricted routers. In general these techniques can be grouped into four different approaches: 1) algorithms (such as left-edge, maze, greedy, hierarchical); 2) expert systems; 3) neural networks; and 4) genetic algorithms [Zobrist, 1994; Sarrafzadeh, 1996; and Lengauer, 1990]. As an example, we consider here only one of the techniques, called a maze router which is widely known. The maze router can be used as a global router and/or detailed router. It finds the shortest rectilinear path by propagating a wavefront from a source point toward a destination point [Lee, 1969]. Considering the routing surface as a rectangular array of cells, the algorithm starts by marking the source cell as visited. In successive steps, it visits all the unvisited neighbors of visited cells. This continues until the destination cell is visited. For example, consider the cell configuration given in Fig. 25.20.

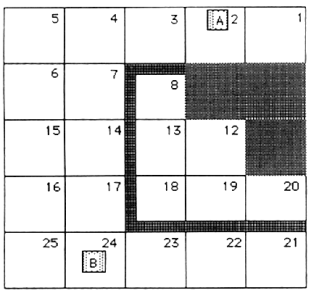


FIGURE 25.20 Initial configuration.

We would like to find a minimal-crossing path from source cell A (cell 2) to destination cell B (cell 24). (The minimal-crossing path is defined as a path that crosses over the fewest number of existing paths.) The algorithm begins by assigning the start cell 2 to a list, denoted as L, i.e., L consists of the single entry {2}. For each entry in L, its immediate neighbors (which are not blocked) will be added to an auxiliary list L′. (The auxiliary cell list L′ is provided for momentary storage of names of cells.) Therefore, list L′ contains entries {1,3}. To these cells a chain coordinate and a weight are assigned, as denoted in Fig. 25.21. For example, in cell 3, we have the pair (0, →), meaning that the chain coordinate is toward right and the cell weight is 0. The weight for a cell represents the number of wires that should be crossed in order to reach that cell from the source cell. The cells with minimum weight in list L′ are appended

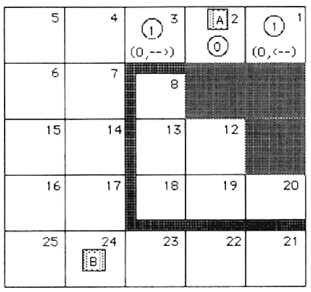


FIGURE 25.21 First step.

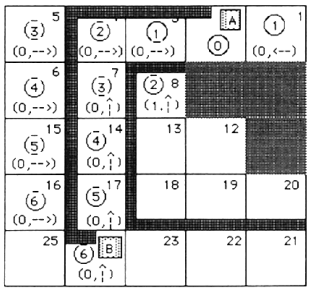


FIGURE 25.22 Final step.

to list L. Thus, cells 1 and 3 are appended to list L. Moreover, cell 2 is erased from list L. Appending the immediate neighbors of the cells in L to L′, we find that list L′ now contains entries 4 and 8. Note that cell 8 has a weight of 1; this is because a wire must be crossed in order to reach to this cell. Again the cells with minimum weight in list L′ are appended to list L, and cell 3 and cell 1 are erased from L. Now L contains entry {4}. The above procedure is repeated until it reaches to the final cell B. Then a solution is found by tracing the chain coordinated from cell B to cell A as shown in Fig. 25.22.

The importance of Lee's algorithm is that it always finds the shortest path between two points. Since it routes one net at a time, however, there is a possibility of having some nets unrouted at the end of the routing process. The other weak points of this technique are the requirements of a large memory space and long execution time. For this reason, the maze router is often used as a side router for the routing of critical nets and/or routing of leftover unrouted nets.

Defining Terms

Floorplanning: A floorplan routine determines an approximate position for each cell so that the total area is minimized.

Layout: Specifies the position and dimension of the different layers of materials as they would be layed on the silicon wafer.

Placement: A placement routine determines an optimal position on the chip for a set of cells with fixed dimensions in a way that the total occupied area and the total estimated length of connections are minimized.

Routing: Given a collection of cells placed on a chip, the routing routine connects the terminals of these cells for a specific design requirement.

Related Topic

23.1 Processes

References

C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," *Proceedings of the 19th Annual Design Automation Conference,* (July), pp. 175–181, 1982.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science,* vol. 220, no. 4598 (May), pp. 671–680, 1983.

C. Y. Lee, "An algorithm for path connections and its application," *IRE Transactions on Electronic Computers,* (Sept.), pp. 346–365, 1969.

T. Lengauer, *Combinatorial Algorithms for Integrated Circuit Layout,* New York: John Wiley & Sons, 1990.

C. A. Mead and L. A. Conway, *Introduction to VLSI Systems,* Reading, Mass.: Addison-Wesley, 1980.

R. H. J. M. Otten, "Efficient floorplan optimization," *International Journal on Computer Design,* pp. 499–503, IEEE/ACM, 1983.

B. Preas and M. Lorenzetti, *Physical Design Automation of VLSI Systems,* Menlo Park, Calif.: Benjamin/Cummings, 1988.

M. Sarrafzadeh and C. K. Wong, *An Introduction to VLSI Physical Design,* New York: McGraw-Hill, 1996.
C. Sechen, "Chip-planning, placement and global routing of macro-cell integrated circuits using simulated annealing," *International Journal of Computer Aided VLSI Design 2,* pp. 127–158, 1990.
L. Stockmeyer, "Optimal orientation of cells in slicing floorplan designs," *Information and Control,* 57 (2) pp. 91–101, 1983.
S. Sutanthavibul, E. Shargowitz, and J. B. Rosen, "An analytical approach to floorplan design and optimization," *IEEE Transactions on Computer Aided-Design,* 10 (6) pp. 761–769, 1991.
S. Wimer and I. Koren, "Analysis of strategies for constructive general block placement," *IEEE Transactions on Computer-Aided Design,* vol. 7, no. 3 (March), pp. 371–377, 1988.
G. W. Zobrist, Editor, *Routing, Placement, and Partitioning,* Ablex Publishing, 1994.

Further Information

Other recommended layout design publications include Weste and Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective,* Reading, Mass.: Addison-Wesley, 1988, and the book by B. Preas and M. Lorenzetti, *Physical Design Automation of VLSI Systems,* Menlo Park, Calif.: Benjamin/Cummings, 1988. The first book describes the design and analysis of a layout. The second book describes different techniques for development of CAD systems.

Another source is *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* which is published monthly by the Institute of Electrical and Electronics Engineers.

25.3 Application-Specific Integrated Circuits

S. K. Tewksbury

Introduction

Present-day silicon very large scale integration complementary metal-oxide semiconductor (**VLSI CMOS**) technologies can place and interconnect several million transistors (representing over a million gates) on a single integrated circuit (**IC**) approximately 1 cm square. Provided with such a vast number of gates, a digital system designer can implement very sophisticated and complex system functions (including full systems) on a single IC. However, efficient design (including optimized performance) of such functions using all these gates is a complex puzzle of immense complexity. If this technology were to have been provided to the world overnight, it is doubtful that designers could in fact make use of this vast amount of logic on an IC.

However, this technology instead has evolved over a long period of time (about three decades), starting with only a few gates per IC, with the number of gates per IC doubling consistently about every 18 months (a progression referred to as Moore's Law), and evolving to the present high gate densities per IC. Projections [The National Technology Roadmap for Semiconductors, 1994] of this evolution over the next 15 years, as shown in Table 25.2, promise continued dramatic advances in the amount of logic and memory which will be provided on individual ICs. Paralleling the technology evolution, computer-aided design (**CAD**) tools and electronics design automation (**EDA**) tools [Rubin, 1987; Sherwani, 1993; Hill and Peterson, 1993; Banerjee, 1994] have evolved to assist designers of the increasingly complex ICs. With these CAD tools, today's design teams effectively have an army of very "experienced" designers embedded in the tools, capable of applying the knowledge gained over the long and steady history of IC evolution. This prior experience captured in CAD/EDA tools includes the ability to convert a high-level description [Camposano and Wolf, 1991; Gajski et al., 1992] of a specific function (e.g., ALU, register, control unit, microcontroller, etc.) into an efficient physical implementation of that function.

Figure 25.23 is a photomicrograph of a contemporary, high-performance VLSI application-specific IC (**ASIC**) circuit, the ADSP-1060 (SHARC) digital signal processor (DSP) from Analog Devices, Inc. The right two thirds of the IC is 4 Mbit of SRAM, providing considerable on-chip memory. The DSP is on the left third of the IC. A 0.5-μm CMOS technology with two levels of metal was used, with a total of about 20 million transistors on the IC. The IC provides 120 MFLOP of performance.

TABLE 25.2 Prediction of VLSI Evolution by Semiconductor Industries Association

Year	1995	1998	2001	2004	2007	2010
Feature size (μm)	0.35	0.25	0.18	0.13	0.10	0.07
DRAM bits/chip	64M	256M	1G	4G	16G	64G
ASIC gates/chip	5M	14M	26M	50M	210M	430M
Chip size (ASIC) (mm^2)	450	660	750	900	1100	1400
Maximum number of wiring levels (logic)	4–5	5	5–6	6	6–7	7–8
On-chip speed (MHz)	300	450	600	800	1000	1100
Chip-to-board speed (MHz)	150	200	250	300	375	475
Desktop supply voltage (V)	3.3	2.5	1.8	1.5	1.2	0.9
Maximum power (W), heatsink	80	100	120	140	160	180
Maximum power (W), no heatsink	5	7	10	10	10	10
Power (W), battery systems	2.5	2.5	3.0	3.5	4.0	4.5
Number of I/Os	900	1350	2000	2600	3600	4800

Adapted from *The National Technology Roadmap for Semiconductors,* Semiconductor Industry Association, San Jose, Calif., 1994.

This section summarizes the design process, the gate-level physical design, and several issues which have become particularly important with today's VLSI technologies, with a focus on ASICs. An important and growing issue is that of testing, including built-in testing, design for testability, and related topics (see e.g., Jha and Kundu [1990] and Parker [1992]). This is a broad topic, beyond the scope of this section.

Primary Steps of VLSI ASIC Design

The VLSI IC design process consists of a sequence of well-defined steps [Preas and Lorenzetti, 1988; Hill et al., 1989; DeMicheli, 1994a and b] related to the definition of the functions to be designed; organization of the circuit blocks implementing these logic functions within the area of the IC; verification and simulation at several stages of design (e.g., behavioral simulation, gate-level simulation, circuit simulation [White and Sangiovanni-Vincentelli, 1987; Lee et al., 1993], etc.); routing of physical interconnections among the blocks, and final detailed placement and transistor-level layout of the VLSI circuit. This process can also be used hierarchically to design one of the blocks making up the overall IC, representing that circuit block in terms of simpler blocks. This establishes the "top-down" hierarchical approach, extendable to successively lower-level elements of the overall design.

These general steps are illustrated in Fig. 25.24(a), roughly showing the basic steps taken. A representative example [Lipman, 1995] of a contemporary design approach is illustrated in Fig. 25.24(b). Design approaches are continually changing and Fig. 25.24(b) is merely one of several current design sequences. Below, we summarize the general steps highlighted in Fig. 25.24(a).

A. *Behavioral Specification of Function:* The behavioral specification is essentially a description of the function expected to be performed by the IC. The design can be represented by schematic capture, with the designer representing the design using block diagrams. High-level description languages (**HDLs**) such as VHDL [Armstrong, 1989; Lipsett et al., 1990; Mazor and Langstraat, 1992] and Verilog [Thomas and Moorby, 1991] are increasingly used to provide a detailed specification of the function in a manner which is largely independent of the physical design of the function. VHDL and Verilog are "hardware description languages," representing designs from a variety of viewpoints including behavioral descriptions, structural descriptions, and logical descriptions. Figure 25.25(a) illustrates the specification of the overall function in terms of subfunctions (A(s), B(s), …, E(s)) as well as expansion of one subfunction (C(s)) into still simpler functions (c1, c2, c3, …, c6).

B. *Verification of Function's Behavior:* It is important to verify that the behavior specification of today's complex ICs properly represents the behavior desired. In the case of HDL languages, there may be software "debugging" of the "program" until the desired behavior is obtained, much as programming languages need to be debugged until correct operation is obtained. Early verification is important since any errors in the specification of the function will lead to an IC which is faulty as a result of the design, rather than of physical defects.

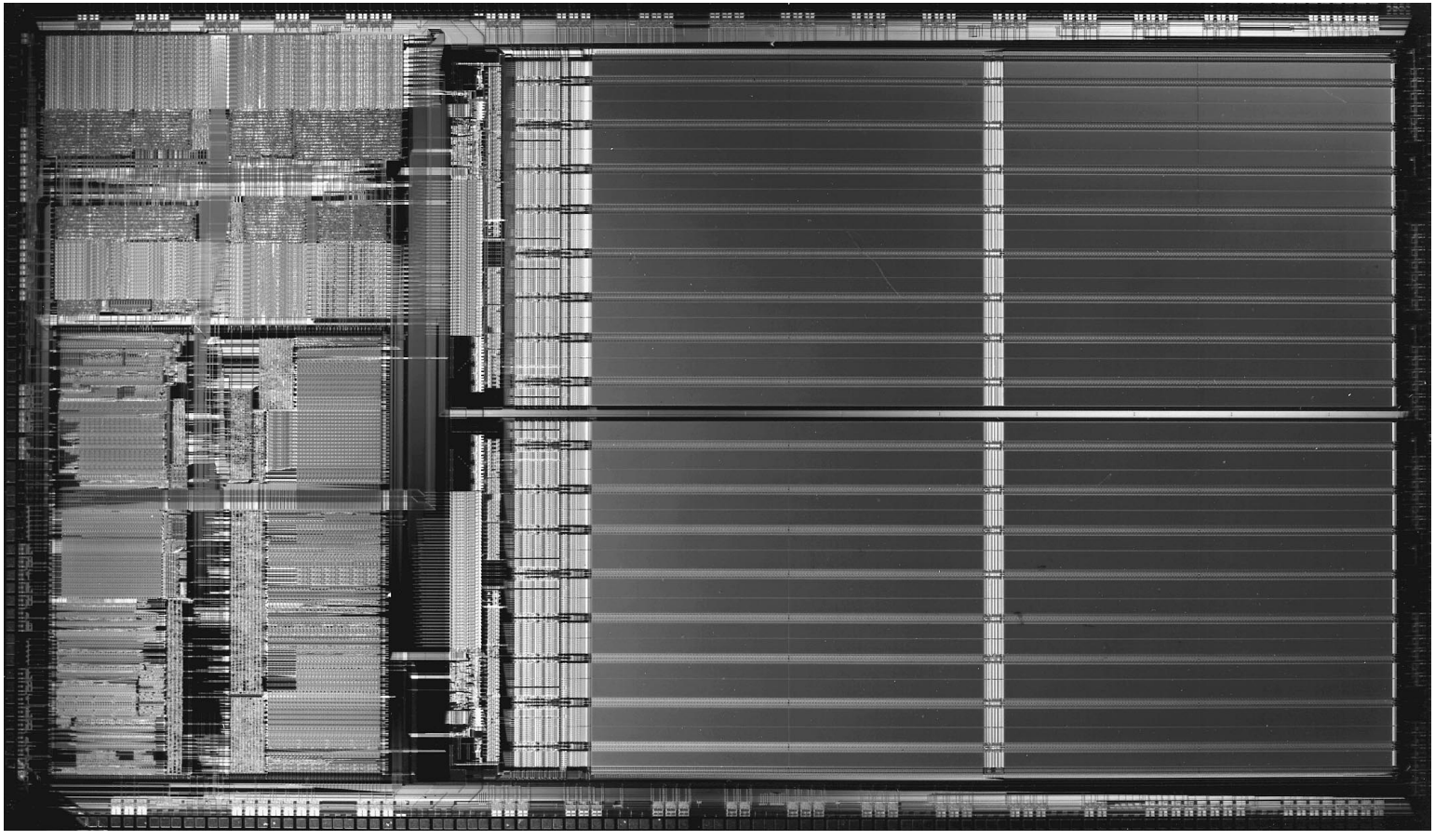


FIGURE 25.23 Photomicrograph of the SHARC DSP of Analog Devices, Inc. (Courtesy of Douglas Garde, Analog Devices, Inc., Norwood, Mass.)

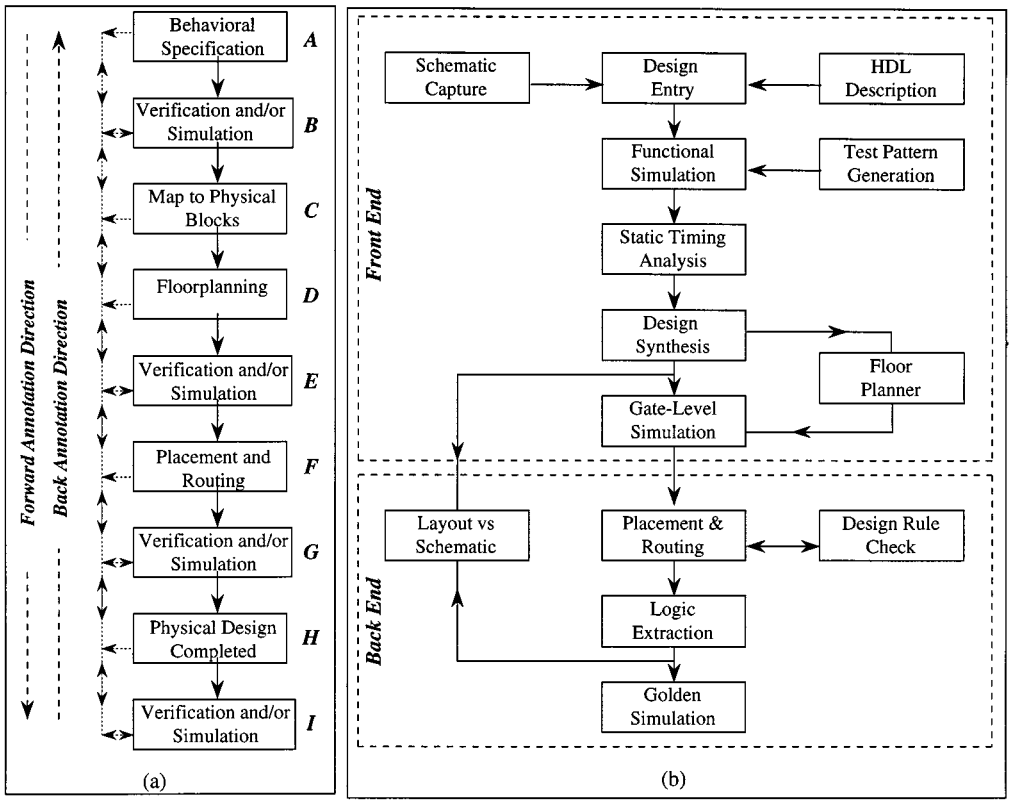


FIGURE 25.24 Representative VLSI design sequences. (a) Simplified but representative sequence. (b) Example design approach from recent trade journal [Lipman, 1995].

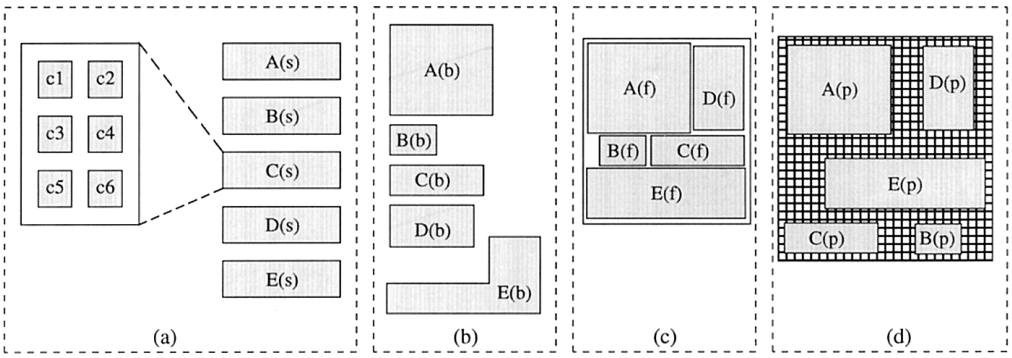


FIGURE 25.25 Circuit stages of design. (a) Initial specification (e.g., HDL, schematic, etc.) of ASIC function in terms of functions, with the next lower level description of function C illustrated. (b) Estimated size of physical blocks implementing functions. (c) Floorplanning to organize blocks on an IC. (d) Placement and routing of interconnections among blocks.

C. *Mapping of Logical Function into Physical Blocks:* Next, the logical functions, e.g., A(s), B(s), …, E(s) in Fig. 25.25(a), are converted into physical circuit blocks, e.g., blocks A(b), B(b), …, E(b) in Fig. 25.25(b). Each physical block represents a logic function as a set of interconnected gates. Although details of the physical layout are not known at this point, the estimated area and aspect ratio (ratio of height to width) of each circuit block is needed to organize these blocks within the area of the IC.

D. *Floorplanning:* Next, the individual circuit blocks must be compactly arranged to fit within the minimum area. Floorplanning establishes this organization, as illustrated in Fig. 25.25(c). At this stage, routing of interconnections among blocks has not been performed, perhaps requiring modifications to the floor plan after routing. During floor planning, the design of a logic function in terms of a block function can be modified to achieve shapes which better match the available IC area. For example, the block E(b) in Fig. 25.25(b) has been redesigned to provide a different geometric shape for block E(f) in the floor plan in Fig. 25.25(c).

E. *Verification/Simulation of Function Performance:* Given the floorplan, it is possible to estimate the average length of interconnections among the blocks (with the actual length not known until after interconnection routing in step F below. Signal timing throughout the IC is estimated, allowing verification that the various circuit blocks interact within timing margins.

F. *Placement and Routing:* When an acceptable floorplan has been established, the next step is to complete the routing of interconnections among the various blocks of the design. As interconnections are routed, the overall IC area expands to provide the area for the wiring, with the possibility that the original floorplan (which ignored interconnections) is not optimal. Placement [Shahookar and Mazumder, 1991] considers various rearrangements of the circuit blocks, without changing their internal design but allowing rotations, etc. For example, the arrangement of some of the blocks in Fig. 25.25(c) have been changed in the arrangement in Fig. 25.25(d).

G. *Verification/Simulation of Performance:* Following placement and routing, the detailed layout of the circuit blocks and interconnections has been established and more accurate simulations of signal timing and circuit behavior can be performed, verifying that the circuit behaves as desired with the interblock interconnections in place.

H. *Physical Design at Transistor/Cell Level:* As the above steps are completed, the design is moving closer toward a full definition of the ASIC circuit in terms of a set of physical masks precisely specifying placement of all the transistors and interconnections. In this step, that process is completed.

I. *Verification/Simulation of Performance:* Before fabricating the masks and proceeding with manufacture of the ASIC circuit, a final verification of the ASIC is normally performed. Figure 25.24(b) represents this step as "golden simulation," a process based on detailed and accurate simulation tools, tuned to the process of the foundry and providing the final verification of desired performance.

Increasing Impact of Interconnection Delays on Design

In earlier generations of VLSI technology (with larger transistors and wider interconnection lines/spacings), delays through the low-level logic gates greatly dominated delays along interconnection lines. This was largely the result of the lower resistance of the larger cross-section interconnections. Under these conditions, a single pass through a design sequence such as shown in Fig. 25.24 was often adequate. In particular, the placement of blocks on the ICs, although impacting the lengths of interconnections among blocks, did not have a major impact on performance since the gate delays within the blocks were substantially larger than interconnection delays between blocks. Under such conditions, the steps of floorplanning and of placement and routing focused on such objectives as minimum overall area and minimum interconnection area.

However, as feature sizes have decreased below about 0.5 μm, this condition has changed and current VLSI technology has interconnection delays substantially larger than logic delays. The increasing importance of interconnection delays is driven by several effects. The smaller feature size leads to interconnections with a higher resistance R^* per unit length and with a higher capacitance C^* per unit area (the capacitance increase also reflecting additional metal layers and coupling capacitances). For interconnection lines among high-level blocks (spanning the IC), the result is a larger RC time constant ($R^*LC \times {}^*L$), with L the line length. While interconnect delays are increasing, gate delays are decreasing. Figure 25.26 illustrates the general behavior on technology scaling to smaller features. A logic function F in a previous-generation technology requires a smaller physical area and has a higher speed in a later, scaled technology (i.e., a technology with feature sizes decreased). Although the intrablock line lengths decrease (relaxing the impact within the block of higher R^*C^*), the interblock lines continue to have lengths proportional to the overall IC size (which is increasing), with the larger R^*C^* leading to increased RC delays on such interconnections.

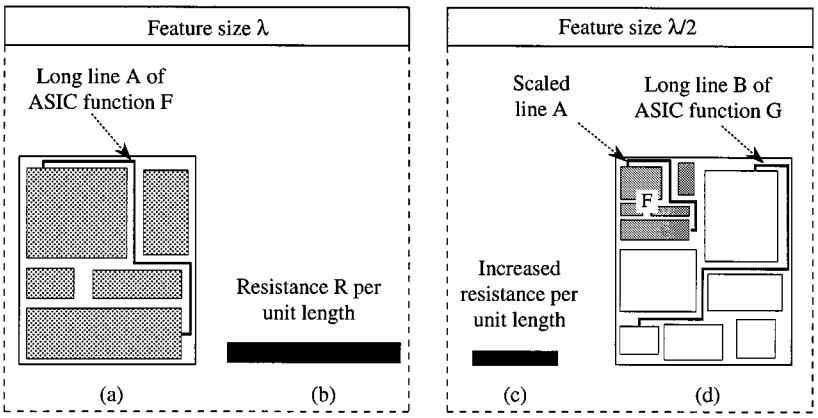


FIGURE 25.26 Interconnect lengths under scaling of feature size. (a) Initial VLSI ASIC function F with line A extending across IC. (b) Interconnection cross section with resistance R^* per unit length. (c) Interconnection cross section, in scaled technology, with increased resistance per unit length. (d) VLSI ASIC function G in scaled technology containing function F but reduced in size (including interconnection A) and containing a long line B extending across the IC.

As the interconnection delays have become increasingly dominant, the design process has evolved into an iterative process through the design steps, as illustrated by the back-and-forward annotation arrows in Fig. 25.24(a). Initial estimates of delays in step B need to be refined through back-annotation of interconnection delay parameters obtained after floorplanning and/or after placement and routing to reflect the actual interconnection characteristics, perhaps requiring changes in the initial specification of the desired function in terms of logical and physical blocks. This iterative process moving between the logical design and the physical design of an ASIC has been problematic since often the logical design is performed by the company developing the ASIC, whereas the physical design is performed by the company (the "foundry") fabricating the ASIC. CAD tools are an important vehicle for coordination of the interface between the designer and the foundry.

General Transistor Level Design of CMOS Circuits

The previous section has emphasized the CAD tools and general design steps involved in designing an ASIC. A top-down approach was emphasized, with the designer addressing successively more-detailed portions of the overall design through a hierarchical organization of the overall description of the function. However, the design process also presumes a considerable understanding of the bottom-up principles through which the overall IC function will eventually appear as a fully detailed specification of the transistor and interconnection structures throughout the overall IC [Dillinger, 1988; Weste and Eshraghian, 1993; Wolf, 1994; Rabaey, 1996; Kang and Leblebici, 1996].

Figure 25.27 illustrates the transistor-level description of a simple three-input, NAND gate. VLSI ASIC logic circuits are dominated by this general structure, with the **PMOS transistors** (making up the pull-up section) connected to the supply voltage V_{dd} and the **NMOS transistors** (making up the pull-down section) connected to the ground return *GND*. When the logic function generates a logic "1" output, the pull-up section is shorted through its PMOS transistors to *Vdd* leading to a high output voltage while the pull-down section is open, with no connection of *GND* to the output. When generating a logic "0" output, the pull-down section is shorted to *GND* while the pull-up section is open (no path to V_{dd}). Since the output is either a 1 or a 0, only one of the sections (pull-up or pull-down) is shorted, with no dc current flowing directly from V_{dd} to *GND* through the logic circuit pull-up and pull-down sections.

The PMOS transistors used in the pull-up section are fabricated with P-type source and drain regions on N-type substrates. The NMOS transistors used in the pull-down section, on the other hand, are fabricated with N-type source and drain regions on P-type substrates. Since a given silicon wafer is either N-type or P-type, a deep, opposite doping-type region must be placed in the silicon wafer for those transistors needing a substrate of the opposite type. The shaded regions in Fig. 25.27(a) represent the "substrate types" within which the

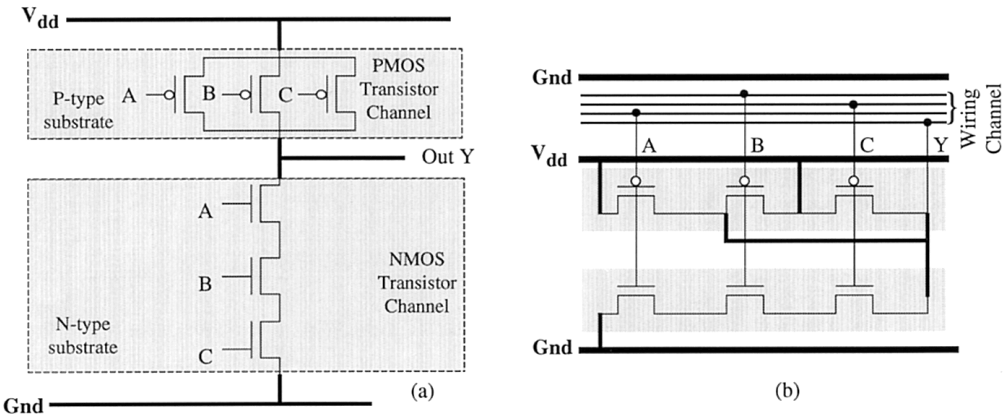


FIGURE 25.27 Transistor representation of three-input NAND gate. (a) Transistor representation without regard to layout. (b) Transistor representation using parallel rows of PMOS and NMOS transistors, with interconnections connected from a wiring channel.

transistors are fabricated. A deep doping of the desired substrate type is provided, with transistors fabricated within such deep, substrate doping "wells." To allow tight packing of transistors, large substrate doping wells are used, with a large number of transistors placed in each well.

Each logic cell must be connected to power (V_{dd}) and ground (*Gnd*), requiring that external power and ground connections to the IC be routed (on continuous metal lines to avoid resistive voltage drops) to each logic cell on the IC. Early IC technologies provided only a single level of metallization on which to route power and ground and an interdigitated layout, illustrated in Fig. 25.28(a), was adopted. Given this power and ground layout approach, channels of pull-up sections and channels of pull-down sections were placed between the power and ground interconnections, as illustrated in Fig. 25.28(b).

Bottom-up IC design is also hierarchical, with the designer completing detailed layout of a specific logic function (e.g., a binary adder), placing that detailed layout (a *cell*) in a library of physical designs and then reusing that library cell when other instantiations of the cell are required elsewhere in the IC. Figure 25.29 illustrates this cell-based approach, with cells of common height but varying width placed in rows between the power and ground lines. The straight power and ground lines shown in Fig. 25.29 allow tight packing of adjacent rows.

To achieve tight packing of cells within a row, adjacent cells (Fig. 25.29) are abutted against each other. Since metal interconnections are used within cells (and most basic cells are reused throughout the design), it is generally not possible to route intercell metal interconnections over cells. This constraint can be relaxed, as shown in Fig. 25.30(b), when additional metal layers are available, restricting a subset of the layers for intracell use and allowing over-the-cell routing [Sherwani et al., 1995] with the other layers.

When metal intercell interconnections cannot be safely routed over cells, interconnection channels must be provided between rows of logic cells, leading to the ***wiring channels*** above and/or below rows of logic cells as shown in Fig. 25.29. In this approach, all connections to and from a logic cell are fed from the top and/or bottom wiring channel. The width of the wiring channel is adjusted to provide space for the number of intercell interconnections required in the channel. Special cells providing through-cell routing can be used to support short interconnections between adjacent rows of cells. Given this layout style at the lowest level of cells, larger functions can be readily constructed, as illustrated in Fig. 25.30(a). Figure 25.29(b) illustrates interconnections (provided in the polysilicon layer under the metal layers) to the logic cells from the wiring channel. For classical CMOS logic cells, the same set of input signals are applied to both the pull-up and pull-down sections, as in the example in Fig. 25.27(b). By organizing the sequence of transistors along the pull-up and pull-down sections properly, inputs can extend vertically between a PMOS transistor in the pull-up section and a corresponding NMOS transistor in the pull-down section. Algorithms to determine the appropriate ordering of transistors evolved early in CAD tools.

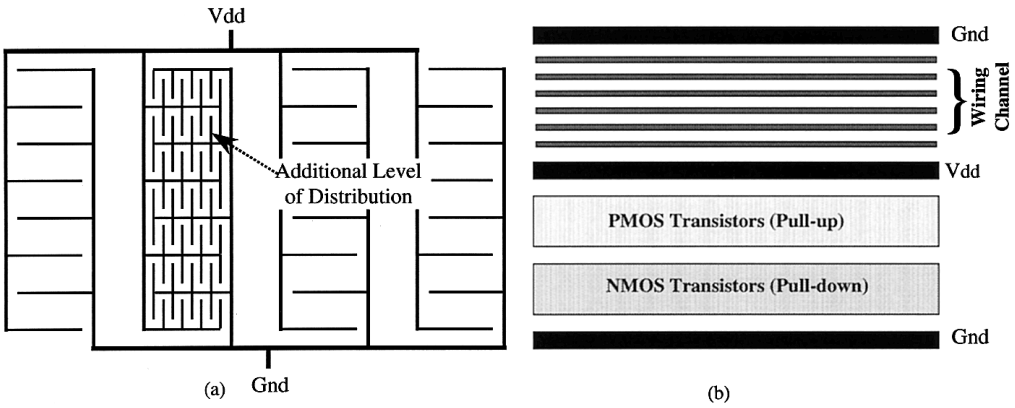


FIGURE 25.28 Power and ground distribution (interdigitated lines) with rows of logic cells and rows of wiring channels. (a) Overall power distribution and organization of logic cells and wiring channels. (b) Local region of power distribution network.

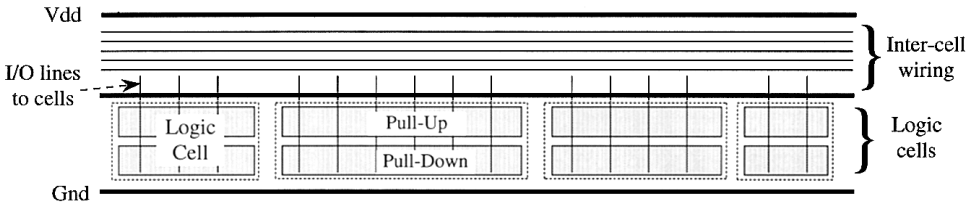


FIGURE 25.29 Cell-based logic design, with cells organized between power and ground lines and with intercell wiring in channels above (and/or below, also) the cell row.

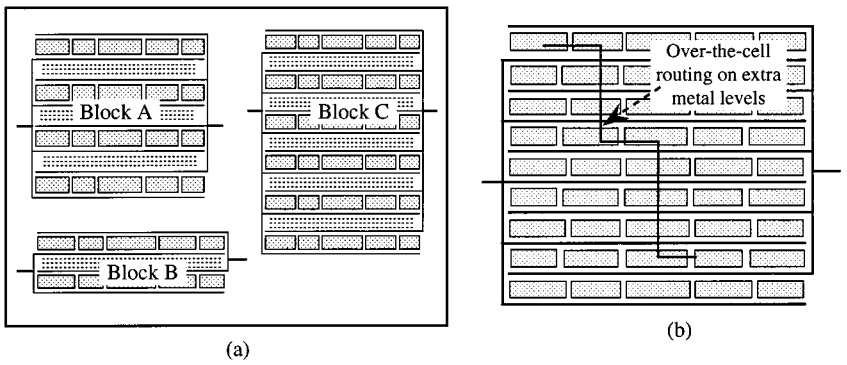


FIGURE 25.30 (a) Construction of larger blocks from cells with wiring channels between rows of cells within the block. (b) Over-the-cell routing on upper metal levels which are not used within the cells.

ASIC Technologies

Drawing on the discussion above, the primary ASIC technologies (gate arrays, sea-of-gate arrays, standard cell ASICs, ASICs with "megacells," and field-programmable gate arrays) can be easily summarized. For comparison, full custom VLSI is briefly described first.

Full Custom Design

In *full custom design,* custom logic cells are designed, starting at the lowest level (i.e., transistor-based cell design) and extending to higher levels (e.g., combinations of cells for higher-level functions) to create the overall IC function. Figure 25.31(a) illustrates the general layout at the cell level. The designer can exploit new cell designs

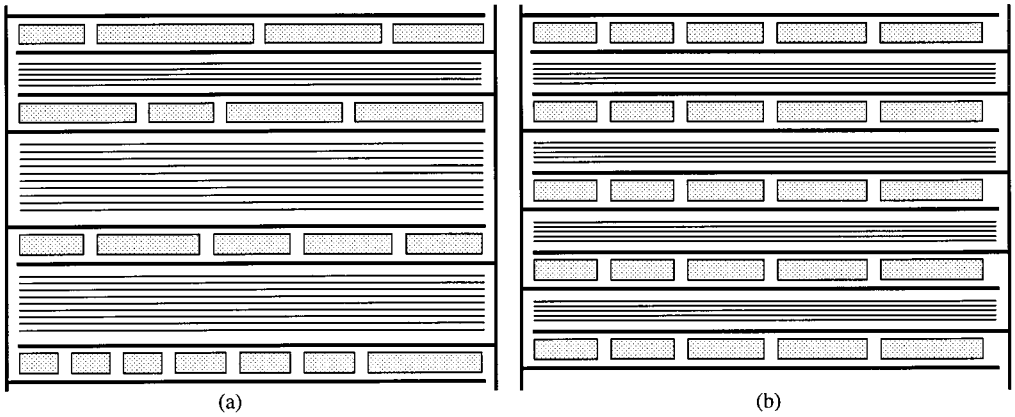


FIGURE 25.31 (a) Full custom layout (with custom cells) or standard cell ASIC layout (using library cells). (b) Gate array layout, with fixed width wiring channels (in prefabricated, up to metallization, wafers).

which improve the performance of the specific function being designed, can provide interconnections through wiring areas between logic cells to create compact functions, can use the full variety of CMOS circuit designs (e.g., dynamic logic, pass-transistor logic, etc.), can use cells previously developed privately for earlier ICs, and can use cells from a standard library provided by the foundry.

Standard Cell ASIC Technology

Consider a full custom circuit design completed using only predefined logic cells from a foundry-provided specific library and physical design processes provided by standard EDA/CAD tools. This approach, *standard cell design* [Heinbuch, 1987, *SCMOS Standard Cell Library,* 1989], is one of the primary ASIC technologies. A critical issue impacting standard cell ASIC design is the quality of the standard cell library provided by the foundry. By providing a rich set of library cells, the coordination between the logical design and the physical design is substantially more effective. Figure 25.31(a) also illustrates the general standard cell approach, using standard library cells of design-specified height (according to cells used), design-specified width logic rows, and varying width of the wiring channel to accommodate the number of interconnection wires determined during place and route.

Gate Array ASIC Technology

The *gate array technology* [Hollis, 1987] is based on partially prefabricated (up to but not including the final metallization layer) wafers with simple gate cells. Such *noncustomized* wafers are stockpiled and the ASIC designer specifies the final metallization layer added to *customize* the gate array. Gate array cells draw on the general cell design shown earlier, Fig. 25.27(b). Figure 25.32 illustrates representative non-customized transistor-level cells although different foundries use different physical layouts) with the dashed lines representing metal layers (including power and ground) added during the final metallization process.

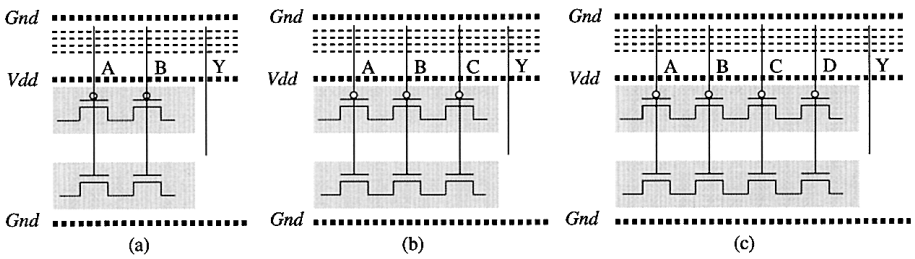


FIGURE 25.32 Basic noncustomized gate array cells (two input, three and four input examples). The dashed lines represent the power and ground lines, as well as interconnections in the wiring channel, which are placed on the IC during customization.

The ASIC designer's task is to translate the desired VLSI logic function into a physical design using the basic gate cells provided on the noncustomized IC. To avoiding the routing of intercell interconnections around long rows of logic cells, some of the "cells" along the row are feedthrough cells, allowing routing of interconnections to other logic cell rows. Included in the designer's toolset are library functions predefining the construction of more-complex logic functions (e.g., adders, registers, multipliers, etc.) from the gate cells on the noncustomized gate array IC.

The gate array technology shares the costs of masks among all ASIC customers and exploits high-volume production of the noncustomized wafers. On the other hand, construction of higher-level functions must adhere to the predefined positions and type of gate cells, leading to less-efficient and lower-performance designs than in the standard cell approach. In addition, the width of the wiring channel is fixed, limiting the number of parallel interconnections in the channel and imposing wasted area if all available wires are not used.

Sea-of-Gates ASIC Technology

The *sea-of-gates* technology also uses premanufactured, noncustomized gate arrays. However, additional metallization layers are provided, with lower-level metallization layer(s) used to program the internal function of the cells and the upper-level layer(s) used for over-the-cell routing [Sherwani et al., 95] of signals among cells, such as illustrated earlier in Fig. 25.30(b). This eliminates the need for wiring channels and feedthrough cells, leading to denser arrays of transistors.

CMOS Circuits Using Megacell Elements

The examples above have focused on low-level logic functions. However, as the complexity of VLSI ICs has increased, it has become increasingly important to include standard, high-level functions (e.g., microprocessors, DSPs, PCI interfaces, MPEG coders, RAM arrays, etc.) within an ASIC. For example, an earlier generation of microprocessor may offer the necessary performance and would occupy only a small portion of the area of a present-generation ASIC. Including such a standard microprocessor has the advantages of allowing the ASIC design to be completed more quickly as well as providing users with the microprocessor standard instruction set and software development tools. Such large cells are called *megacells*. As VLSI technologies advance and as standards increasingly impact design decisions, the use of standard megacells will become increasingly common.

Field-Programmable Gate Arrays: Evolving to an ASIC Technology

The *field-programmable gate array* (FPGA), like the gate array, places fixed cells on the wafer, and the FPGA designer constructs more-complex functions from these cells. However, the cells provided on the FPGA can be substantially more complex than the simple gates provided on the gate array. In addition, the term *field programmable* highlights the customizing of the ASIC by the user, rather than by the foundry manufacturing the FPGA. The *mask-programmable gate array* (MPGA) is similar to the FPGA (using more-complex cells than the gate array), but the programming is performed by addition of the metal layer by the FPGA manufacturer.

Figure 25.33 shows an example of cells and programmable interconnections for a representative FPGA technology [Actel, 1995]. The array of cells is constructed from two types of cell, which alternate along the logic cell rows of the FPGA. The combinational logic cell — "C-module" in Fig. 25.33(a) — provides a ROM-based lookup table (LUT) able to efficiently implement a complex logic function (with four data inputs and two control signals). The sequential cell (S-module) adds a flip-flop to the combinational module, allowing efficient realization of sequential circuits. The interconnection approach illustrated in Fig. 25.33(c) is based on (1) short vertical interconnections directly connecting adjacent modules, (2) long vertical interconnections extending through the overall array, (3) long horizontal interconnections extending across the overall array, (4) points at which the long vertical interconnections can be connected to cells, and (5) points at which the long vertical and horizontal lines can be used for general routing. The long horizontal and vertical lines are broken into segments, with programmable links between successive segments. The programmer can then connect a set of adjacent line segments to create the desired interconnection line. In addition, programmable connection points allow the programmer to make transitions between the long vertical lines and the long horizontal lines. By connecting various inputs the cell of an FPGA to either V_{dd} or *GND*, the cell can be "programmed" to perform one of its possible functions. The basic array is complemented by additional driver and other circuitry around the perimeter of the FPGA for interfacing to the "external world."

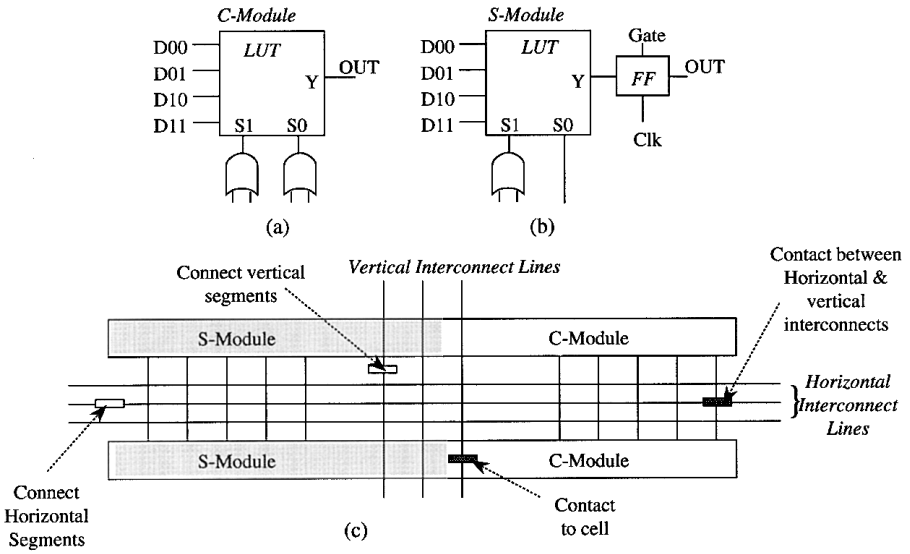


FIGURE 25.33 Example of FPGA elements (Actel FPGA family [Actel, 1995]). Combinational cells (a) and sequential cells (b). (c) Programmable wiring organization.

Different FPGA manufacturers have developed different basic cells, seeking to provide the most useful functionality in the cells for generation of overall FPGA functions. Cells range from fine-grained cells consisting of basic gates through medium-grained cells providing more complex programmable functions to large-grained cells. Different FPGA manufacturers also provide different approaches to the programming step. FPGA programming technologies include one-time programming or multiple-time programming capabilities with the programming either nonvolatile (i.e., programming is retained when power is turned off) or volatile (i.e., programming is lost when power is turned off). Physical programming includes antifuse (fuse) approaches in which the programming nodes are normally off (on) and are "blown" into a permanently on (off) state, providing one-time, nonvolatile programming. Electrical switches also can be used for programming, with an electrical control signal setting the state of the switch. The state control signal can be provided by an EPROM (one-time, nonvolatile), an EEPROM (multiple-time, nonvolatile), or an SRAM (multiple-time, volatile), with different approaches having different advantages and disadvantages.

Interconnection Performance Modeling

Accurate estimation of signal timing is increasingly important in contemporary VLSI ASICs and will become even more important as feature sizes decrease further. Higher clock rates impose tighter timing margins, requiring more accurate modeling of the signal delays. In addition, more-sophisticated models are required for smaller feature size VLSI.

Perhaps of greatest impact is the rapid increase in the importance of interconnect delay relative to gate delay. In the earlier 1 μm VLSI technologies, typical gate delays were about six times the average interconnection delays. For the 0.5 μm technologies, gate delays had decreased while interconnect delays had increased, becoming approximately equal. At 0.3 μm, the decreasing gate delay and increasing interconnect delays have led to average interconnect delays about six times greater than typical gate delays. Accurate estimation of signal delays early in the design is therefore increasingly difficult, since the designer does not have a detailed knowledge of interconnection lengths and nearby lines, which can cause coupling noise, until much of the design has been completed. As the design proceeds and the interconnection lengths become better specified, parameters related to the signal performance can be fed back (back-annotated) to the earlier design steps, allowing the design to be adapted to reflect necessary changes to achieve desired performance.

In earlier VLSI technologies, a linear delay model was adequate, representing the overall delay τ from the input to one cell (cell A in Fig. 25.34) to the input to the connected cell (cell B in Fig. 25.34) by an analytic

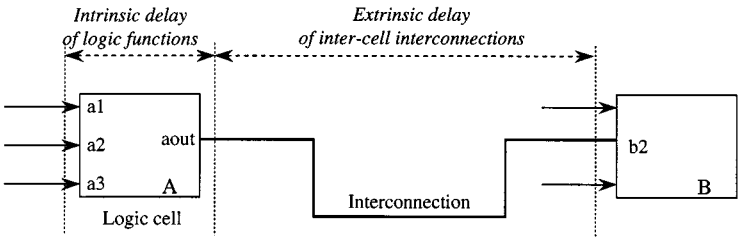


FIGURE 25.34 Logic cell delay (intrinsic delay) and intercell interconnection delay (extrinsic delay).

form such as $\tau = \tau(0) + k1 \cdot C(\text{out}) + k2 \cdot t(s)$, where $\tau(0)$ is the intrinsic (internal) delay of the cell with no loading, $C(\text{out})$ is the output capacitance seen by the output driver of the cell, $t(s)$ is the **rise/fall time** of the signal, and the parameters $k1$ and $k2$ are constants. In the case of deep submicron CMOS technologies, the overall delay must be divided into the **intrinsic delay** of the gate and the **extrinsic delay** of the interconnect, each having substantially more-complex models than the linear model.

Factors impacting the intrinsic delay of gates include the following, with the input and output signals referring to logic cell A in Fig. 25.34.

1. A 0-to-1 change in an input to cell A may cause a different delay to the output of cell A than a 1-to-0 change in that input.
2. Starting from the time when the input starts to change, slower transition times lead to a longer delays before the threshold voltage is reached, leading to longer delays to the output.
3. Once the input passes the threshold voltage, a slower changing input may lead to a longer delay to the output transition.
4. The delay from a change in a given input to a change in the output may depend on the state of other inputs to the circuit.

These are merely representative examples of the more-complex behavior seen in the gates as the feature sizes decrease. Together, those complexities lead to a nonlinear delay model, which is typically implemented as a LUT, rather than with an analytic expression.

The models used for interconnections [Tewksbury, 1994] have also changed, reflecting the changing interconnection parameters and increasing clock rates. The three primary models are as follows.

Lumped RC Model: If the rise/fall times of the signal are substantially greater than the round-trip propagation delay of the signal, then the voltage and current are approximately constant across the length of the interconnection. The interconnection is modeled as a single lumped resistance and a single lumped capacitance. The signal does not incur a propagation delay, and at all points along the line the signal has the same rise/fall times.

Distributed RC Model: If the line is sufficiently long, the signal sees a decreasing resistance and capacitance toward the destination as the signal traverses the line. To represent this changing *RC,* the distributed RC model divides the overall line into shorter segments, each of which can be represented by the lumped RC model above. The transfer function of the overall line is then the product of the transfer functions of the sections. The propagation delay is negligible, though the rise/fall times increase as the signal propagates toward the far-end gates (significant if the line is tapped along its length to drive multiple gates).

Distributed RLC Model: As the rise/fall times become shorter, the relative contributions of capacitance and inductance change. In particular, the impedance of the capacitance is inversely proportional to frequency while that of the inductance is proportional to frequency. At sufficiently high data rates, the inductance effects become significant. In this case, the signal is delayed as it propagates toward the far-end gates, with the rise/fall times increasing along the line. Different terminal points of a net will see the signal at different times with different rise/fall times.

Given the wide range of lengths of signal interconnections, all three models above are relevant; the lumped RC model suitable for short interconnections, the distributed RC model for low-to-moderate speed signals on longer-length interconnections, and the distributed RLC model for high-speed signals on longer interconnections.

Accurate modeling of signals propagating on an interconnection requires detailed knowledge of the capacitance C^* and inductance L^* per unit length along the length of the line. As additional metal layers have been provided, capacitance to neighboring lines (on different or the same metal layer) has become increasingly important, even exceeding the capacitance to ground in some cases. Extraction of accurate interconnect delay parameters may require the use of three-dimensional field solvers, with two-dimensional analysis used for less-accurate modeling of signal behavior.

In addition to the effects noted above, crosstalk (increasingly problematic for buses whose parallel lines run long distances) and reflections (of increasing importance as the signal frequencies increase) degrade signals. This broad range of effects impacting signal delay, distortion, and noise have made *signal integrity* an increasingly important issue in VLSI design. Signal integrity effects also appear on the "dc" power and ground lines, due to large transient currents caused by switching gates and switching drivers of output lines.

Clock Distribution

Signals are increasingly distorted not only by long line lengths, but also by the higher clock frequency in present-day VLSI circuits, with the combination of long lines and high clock rates of particular concern. Present-day VLSI circuits include a vast number of flip-flops (often as registers) distributed across the area of the VLSI circuit. *Synchronous ASICs* use a common clock, distributed to each of these flip-flops. With the clock signal being the longest interconnection on the VLSI circuit and the highest-frequency signal, design of the clock distribution network is critical for highest performance.

Complex synchronous ASICs are designed assuming that all flip-flops are clocked simultaneously. *Clock skew* is the maximum difference between the times of clock transitions at any two flip-flops of the overall VLSI circuit. The clock network must deliver clock signals to each of the flip-flops within the margins set by the allowed clock skew, margins which are substantially less than the clock period. For example, part of the 2-ns clock period of a high-speed VLSI circuit operating with a 500-MHz clock is consumed by the rise/fall times of the signals appearing at the input to the flip-flop and by the specified setup and hold times of the flip-flop. The result is that the clock must be applied to the flip-flop within a time interval small as compared with the clock period.

The distance over which the clock signal can travel before incurring a delay greater than the clock skew defines *isochronous* regions (illustrated in Fig. 25.35(a) as shaded regions) within the IC. If the external clock can be provided to such regions with zero clock skew, then clock routing within the isochronous region is not critical. Figure 25.35(a) illustrates the H-tree approach, whose clock paths have equal lengths to terminal points, ideally delivering clock pulses to each of the terminal points (leaf nodes) of the tree simultaneously (zero skew). In a real circuit, precisely zero clock skew is not achieved since different network segments encounter different environments of data lines coupled electrically to the clock line segment.

In Fig. 25.35(a), a single buffer drives the entire H-tree network, requiring a large area buffer and wide clock lines toward the connection of the clock line to the external clock signal. Such a large buffer can account for up to 30% or more of the total VLSI circuit power dissipation. Figure 25.35(b) illustrates a distributed buffer approach, with a given buffer only having to drive those clock line segments to the next level of buffers. In this case, the buffers can be smaller and the clock lines can be narrower. The 300-MHz DEC Alpha microprocessor, for example, uses an H-tree clock distribution network with multiple stages of buffering extending to the final legs of the H-tree network. Another approach to relax the clock distribution problem uses multiple input/output (I/O) pins for the clock. In this case, a number of smaller H-trees can be driven separately, one starting at each clock I/O pin.

The constraint on clock timing is a bound on clock skew, not a requirement for zero clock skew. In Fig. 25.35(c), the clock network uses multiple buffers but allows different path lengths consistent with clock skew margins. For tight margins, an H-tree can be used to deliver clock pulses to local regions in which distribution proceeds using a different buffered network approach such as that in Fig. 25.35(c).

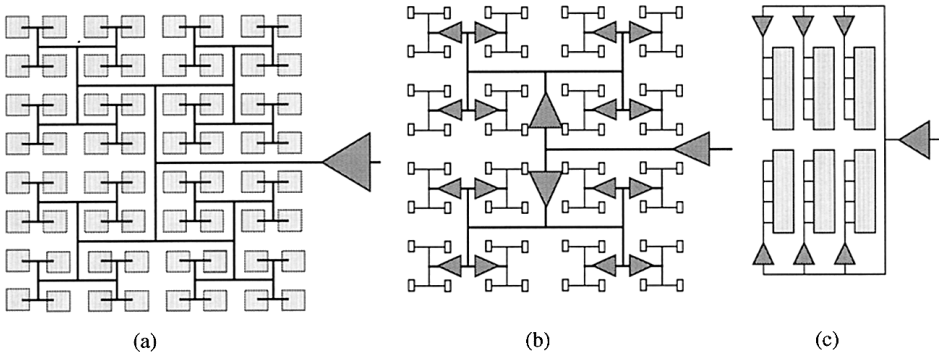


FIGURE 25.35 H-tree clock distribution. (a) Example of single driver and isochronous (shaded) regions. (b) Example of distributed drivers. (c) Example of clock distribution with unequal line lengths but within skew tolerances.

Other approaches for clock distribution are gaining in importance. For example, a lower frequency clock can be distributed across the VLSI circuit, with phase-locked loops (PLLs) used to multiply the clock rate at various sites on the circuit. In addition to multiplying the low clock rate, the PLL can also adjust the phase of the high rate clock, correcting clock skew which may have been introduced in the routing of the low rate clock to that PLL. Another approach is to generate a local clock at a local register when the register input data changes. This *self-timed circuit* approach leads to *asynchronous circuits* but can be quite effective for register-transfer logic (generating a local clock for a large register).

Power Distribution

Present-day VLSI ASICs consume considerable power, unless specifically designed for battery-operated, low-power portable electronics [Chandrakasan and Broderson, 1995]. A 40-W IC operating at 3.3 V requires a current of 12 A, with currents increasing in future generations of high-power VLSI (due not only to the higher power dissipation but also to lower V_{dd}).

Voltage and ground lines must be properly sized to prevent the peak current density exceeding the level at which the power lines will be physically "blown out," leading to rapid and catastrophic failure of the circuit. An equally serious problem is gradual deterioration of a voltage or ground line, eventually leading to failure as a result of *electromigration*. Electromigration failure affects both signal and power lines, but is particularly important in power lines because of the constant direction of the current. As current flows through an aluminum interconnection, the average force exerted on the metal atoms by the electrons leads to a slow migration of those atoms in the direction of electron flow, causing the line to migrate in the direction of the electrons (opposite to the current flow). In regions of the metal line where discontinuities occur (e.g., at the naturally occurring grain boundaries), a void can develop, creating an open in the line. Fortunately, there is a current density threshold level (about 1 mA/μm) below which electromigration is insignificant. Notably, copper, in addition to having a lower resistivity than aluminum, has greater resistance to electromigration. Accurate estimates of power dissipation due to logic switching within logic blocks of the ASIC are also necessary to assess thermal heating within the IC.

Another issue in power distribution concerns *ground bounce* (or *simultaneous switching noise*), which is increasingly problematic as the number of ASIC I/O data pins increases. Consider M output lines switching simultaneously to the 1 state, each of those lines outputting a current transient I(out) within a time t(out). (If M output lines switch simultaneously to the 0 state, then a corresponding input current transient is produced.) The net output current $M \cdot I$(out) is fed through the V_{dd} pin (returned to ground in the case of outputs switching to 0). With an inductance L associated with the V_{dd} pin, a transient voltage $DV \approx L \cdot M \cdot I(\text{out})/t(\text{out})$ is imposed on V_{dd}. A similar effect occurs on the ground connection for outputs switching to 0. With a total output current of 200 mA/ns and a ground pin inductance of 5 nH, the voltage transient is about 1 V. The voltage transient propagates through the IC, potentially causing logic blocks to fail to produce the correct outputs. The transient voltage can be reduced by reducing the power line inductance L, for example by replacing

the single V_{dd} and *Gnd* pins by multiple V_{dd} and *Gnd* pins, with *K* voltage pins reducing the inductance by a factor of *K*.

With power distribution and power line noise problems growing in importance, EDA/CAD tools are rapidly evolving to provide the designer with early estimates and final accurate assessments of various measures of current and of power dissipation.

Analog and Mixed-Signal ASICs

One of the exciting ASIC areas undergoing rapid development is the addition of analog integrated circuits [Geiger et al., 1990; Ismail and Fiez, 1994; Laker and Sansen, 1994] to the standard digital VLSI ASIC, corresponding to *mixed-signal* VLSI. **Mixed-signal ICs** allow the IC to interact directly with the real physical (and analog) world. Library cells representing various analog circuit functions supplement the usual digital circuit cells of the library, allowing the ASIC designer to add needed analog circuits within the same general framework as the addition of digital circuit cells. Automotive electronics is a representative example, with many sensors providing analog information which is converted into a digital format and analyzed using microcomputers or other digital circuits. Mixed-signal library cells include A/D and D/A converters, comparators, analog switches, sample-and-hold circuits, etc., while analog library cells include op amps, precision voltage sources, and phase-locked loops.

As such mixed-signal VLSI ASICs evolve, EDA/CAD tools will also evolve to address the performance and design issues related to analog circuits and their behavior in a digital circuit environment. In addition, analog high-level description languages (AHDLs) are being developed to support high-level specifications of both analog and mixed-signal circuits.

Summary

For about three decades, microelectronics technologies have been evolving, starting with primitive digital logic functions and evolving to the extraordinary capabilities available in present-day VLSI ASICs. This evolution promises to continue for at least another decade, leading to VLSI ICs containing complex systems and vast memory on a single IC. ASIC technologies (including the EDA/CAD tools which guide design to a final IC) deliver this complex technology to the systems designers, including those not associated with a company having a microfabrication facility. This delivery of a highly complex technology to the average electronic systems designer is the result of a steady migration of specialized skill to very powerful EDA/CAD tools which control the complexity of the design process and the result of a need to provide a wide variety of electronics designers with access to technologies which earlier had been available only within large, vertically integrated companies.

Defining Terms

ASIC: Application-specific integrated circuit — an integrated circuit designed for a specific application.

CAD: Computer-aided design — software programs which assist the design of electronic, mechanical, and other components and systems.

CMOS: Complementary metal-oxide semiconductor transistor circuit composed of PMOS and NMOS transistors.

EDA: Electronics design automation — software programs which automate various steps in the design of electronics components and systems.

Extrinsic Delay: Also called *point-to-point delay,* the delay from the transition of output of a logic cell to the transition at the input of another logic cell.

HDL: High-level description language — a software "language" used to describe the function performed by a circuit (or collection of circuits).

IC: Integrated circuit — a normally silicon substrate in which electronic devices and interconnections have been fabricated.

Intrinsic delay: Also called *pin-to-pin delay,* the delay between the transition of an input to a logic cell to the transition at the output of that logic cell.

Mixed-signal ICs: Integrated circuits including circuitry performing digital logic functions as well as circuitry performing analog circuit functions.

NMOS transistor: A metal-oxide semiconductor transistor which is in the on state when the voltage input is high and in the off state when the voltage input is low.

PMOS Transistor: A metal-oxide semiconductor transistor which is in the off state when the voltage input is high and in the on state when the voltage input is low.

Rise/(fall) time: The time required for a signal (normally voltage) to change from a low (high) value to a high (low) value.

V_{dd}**:** The supply voltage used to drive logic within an IC.

VLSI: Very large scale integration — microelectronic integrated circuits containing large (presently millions) of transistors and interconnections to realize a complex electronic function.

Wiring channel: A region extending between the power and ground lines on an IC and dedicated for placement of interconnections among logic cells.

Related Topic

79.1 IC Logic Family Operation and Characteristics

References

Actel, 1995. *Actel FPGA Data Book and Design Guide,* Sunnyvale, Calif.: Actel Corp., 1995.

J. R. Armstrong, *Chip-Level Modeling with VHDL,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

P. Banerjee, 1994. *Parallel Algorithms for VLSI Computer-Aided Design,* Englewood Cliffs, N.J.: Prentice-Hall.

R. Camposano, and W. Wolf (Eds.), *High Level VLSI Synthesis,* Norwell, Mass.: Kluwer Academic Publishers, 1991.

A. Chandrakasan and R. Broderson, *Low Power Digital CMOS Design,* Norwell, Mass.: Kluwer Academic Publishers, 1995.

G. De Micheli, *Synthesis of Digital Circuits,* New York: McGraw-Hill, 1994a.

G. De Micheli, *Synthesis and Optimization of Digital Circuits,* New York: McGraw-Hill, 1994b.

T. E. Dillinger, *VLSI Engineering,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.

D. Gajski, N. Dutt, A. Wu, and S. Lin, *High-Level Synthesis: Introduction to Chip and Systems Design,* Norwell, Mass.: Kluwer Academic Publishers, 1992.

R. L. Geiger, P. E. Allen, and N. R. Strader, *VLSI Design Techniques for Analog and Digital Circuits,* New York: McGraw-Hill, 1990.

D. V. Heinbuch, *CMOS Cell Library,* New York: Addison-Wesley, 1987.

F. J. Hill, and G. R. Peterson, *Computer Aided Logical Design with Emphasis on VLSI,* New York: John Wiley & Sons, 1993.

D. Hill, D. Shugard, J. Fishburn, and K. Keutzer, *Algorithms and Techniques for VLSI Layout Synthesis,* Norwell, Mass.: Kluwer Academic Publishers, 1989.

E. E. Hollis, *Design of VLSI Gate Array ICs,* Englewood Cliffs, N.J.: Prentice-Hall., 1987.

M. Ismail, and T. Fiez, *Analog VLSI: Signal and Information Processing,* New York: McGraw-Hill, 1994.

N. Jha, and S. Kundu, *Testing and Reliable Design of CMOS Circuits,* Norwell, Mass.: Kluwer Academic Publishers, 1990.

S.-M. Kang, and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design,* New York: McGraw-Hill, 1996.

K. R. Laker, and W. M. C. Sansen, *Design of Analog Integrated Circuits and Systems,* New York: McGraw-Hill, 1994.

K. Lee, M. Shur, T. A. Fjeldly, and Y. Ytterdal, *Semiconductor Device Modeling for VLSI,* Englewood Cliffs, N.J.: Prentice-Hall, 1993.

J. Lipman, "EDA tools put it together," *Electronics Design News* (EDN), Oct 26, 1995, pp. 81–92.

R. Lipsett, C. Schaefer, and C. Ussery, *VHDL: Hardware Description and Design,* Norwell, Mass.: Kluwer Academic Publishers, 1990.

S. Mazor, and P. Langstraat, *A Guide to VHDL,* Norwell, Mass.: Kluwer Academic Publishers, 1992.

The National Technology Roadmap for Semiconductors, San Jose, Calif.: Semiconductor Industry Association, 1994.
K. P. Parker, *The Boundary-Scan Handbook,* Norwell, Mass.: Kluwer Academic Publishers, 1992.
B. Preas, and M. Lorenzetti, *Physical Design Automation of VLSI Systems,* Menlo Park, Calif.: Benjamin-Cummings, 1988.
J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective,* Englewood Cliffs, N.J.: Prentice-Hall, 1996.
S. Rubin, *Computer Aids for VLSI Design,* New York: Addison-Wesley, 1987.
SCMOS Standard Cell Library, Center for Integrated Systems, Mississippi State University, 1989.
K. Shahookar, and P. Mazumder, "VLSI placement techniques," *ACM Computing Surveys,* vol. 23(2), pp. 143–220, 1991.
N. A. Sherwani, *Algorithms for VLSI Design Automation,* Norwell, Mass.: Kluwer Academic Publishers, 1993.
N. A. Sherwani, S. Bhingarde, and A. Panyam, *Routing in the Third Dimension: From VLSI Chips to MCMs,* Piscataway, N.J.: IEEE Press., 1995.
S. Tewksbury (Ed.), *Microelectronic Systems Interconnections: Performance and Modeling,* Piscataway, N.J.: IEEE Press, 1994.
D. E. Thomas, and P. Moorby, *The Verilog Hardware Description Language,* Norwell, Mass.: Kluwer Academic Publishers, 1991.
N. H. E. Weste, and K. Eshraghian, *Principles of CMOS VLSI Design,* New York: Addison-Wesley, 1993.
J. White, and A. Sangiovanni-Vincentelli, *Relaxation Methods for Simulation of VLSI Circuits,* Norwell, Mass.: Kluwer Academic Publishers, 1987.
W. Wolf, *Modern VLSI Design: A Systems Approach,* Englewood Cliffs, N.J.: Prentice-Hall, 1994.

Further Information

The *Institute for Electrical and Electronics Engineers, Inc.* (IEEE) publishes several professional journals which describe the broad range of issues related to contemporary VLSI circuits, including the *IEEE Journal of Solid-State Circuits* and the *IEEE Transactions on Very Large Scale Integration Systems.* Other applications-related journals from the IEEE cover VLSI-related topics. Representative examples include the *IEEE Transactions on Signal Processing,* the *IEEE Transactions on Computers,* the *IEEE Transactions on Image Processing,* and the *IEEE Transactions on Communications.* Several conferences also highlight VLSI circuits, including the *IEEE Solid-State Circuits Conference,* the *IEEE Custom Integrated Circuits Conference,* and several others.

Commercial software tools and experiences related to ASIC designs are changing rapidly, but are well covered in several trade journals including *Integrated System Design* (the Verecom Group, Los Altos, Calif.), *Computer Design* (PennWell Publishing Co., Nashua, N.H.), *Electronics Design News* (Cahners Publishing Co., Highlands Ranch, Colo.), and *Electronic Design* (Penton Publishing Inc., Cleveland, Ohio).

There are also many superb books covering the many topics related to IC design and VLSI design. The references for this chapter consist mainly of books, all of which are well-established treatments of various aspects of VLSI circuits and VLSI design automation.

Blackwell, G.R. “Surface Mount Technology”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

26

Surface Mount Technology

26.1 Introduction
26.2 Definition and Considerations
Considerations in the Implementation of SMT
26.3 SMT Design, Assembly, and Test Overview
26.4 Surface Mount Device (SMD) Definitions
26.5 Substrate Design Guidelines
26.6 Thermal Design Considerations
26.7 Adhesives
26.8 Solder Paste and Joint Formation
26.9 Parts Inspection and Placement
Parts Placement
26.10 Reflow Soldering
Post-Reflow Inspection
26.11 Cleaning
26.12 Prototype Systems

Glenn R. Blackwell
Purdue University

26.1 Introduction

This section on surface mount technology (SMT) will familiarize the reader with the process steps in a successful SMT design. The new user of SMT is referred to Mims [1987] and Leibson [1987] for introductory material. Being successful with the implementation of SMT means the engineers involved must commit to the principles of concurrent engineering. It also means that a continuing commitment to a quality techniques is necessary, whether that is Taguchi, TQM, SPC, DOE, another technique, or a combination of several quality techniques, lest you too have quality problems with SMT (Fig. 26.1).

26.2 Definition and Considerations

SMT is a collection of scientific and engineering methods needed to design, build, and test products made with electronic components that mount to the surface of the printed circuit board without holes for leads [Higgins, 1991]. This definition notes the breadth of topics necessary to understand SMT, and also clearly says that the successful implementation of SMT will require the use of concurrent engineering [Classon, 1993; Shina, 1991]. Concurrent engineering means that a team of design, manufacturing, test, and marketing people will concern themselves with board layout, parts and parts placement issues, soldering, cleaning, test, rework, and packaging, before any product is made. The careful control of all these issues improves both yield and reliability of the final product. In fact, SMT cannot be reasonably implemented without the use of concurrent engineering, and/or the principles contained in Design for Manufacturability (DFM) and Design for Testability (DFT), and therefore any facility that has not embraced these principles should do so if implementation of SMT is its goal.

Considerations in the Implementation of SMT

Main reasons to consider implementation of SMT include:

- reduction in circuit board size
- reduction in circuit board weight
- reduction in number of layers in the circuit board
- reduction in trace lengths on the circuit board, with correspondingly shorter signal transit times and potentially higher-speed operation

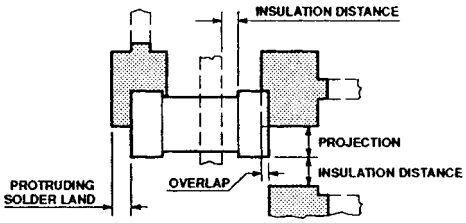


FIGURE 26.1 Placement misalignment of an SMT chip resistor. (*Source:* Phillips Semiconductors, *Surface Mount Process and Application Notes,* Sunnyvale, Calif.: Phillips Semiconductors, 1991. With permission.)

However, not all these reductions may occur in any given product redesign from **through-hole technology** (THT) to SMT.

Most companies that have not converted to SMT are considering doing so. All is of course not golden in SMT Land. During the assembly of a **through-hole** board, either the component leads go through the holes or they do not, and the component placement machines can typically detect the difference in force involved. During SMT board assembly, the placement machine does not have such direct feedback, and accuracy of final soldered placement becomes a stochastic (probability-based) process, dependent on such items as component pad design, accuracy of the PCB artwork and fabrication which affects the accuracy of trace location, accuracy of solder paste deposition location and deposition volume, accuracy of adhesive deposition location and volume if adhesive is used, accuracy of placement machine vision system(s), variations in component sizes from the assumed sizes, and thermal issues in the solder reflow process. In THT test, there is a through-hole at every potential test point, making it easy to align a bed-of-nails tester. In SMT designs, there are not holes corresponding to every device lead. The design team must consider form, fit and function, time-to-market, existing capabilities, testing, rework capabilities, and the cost and time to characterize a new process when deciding on a change of technologies.

26.3 SMT Design, Assembly, and Test Overview

- Circuit design (not covered in this chapter)
- Substrate [typically Printed Circuit Board (PCB)] design
- Thermal design considerations
- Bare PCB fabrication and tests (not covered in this chapter)
- Application of adhesive, if necessary
- Application of solder paste
- Placement of components in solder paste
- Reflowing of solder paste
- Cleaning, if necessary
- Testing of populated PCB (not covered in this chapter)

Once circuit design is complete, substrate design and fabrication, most commonly of a printed circuit board (PCB), enters the process. Generally, PCB assembly configurations using surface mount devices (SMDs) are classified as shown in Fig. 26.2.

Type I — only SMDs are used, typically on both sides of the board. No through-hole components are used. Top and bottom may contain both large and small active and passive SMDs. This type board uses reflow soldering only.

Type II — a double-sided board, with SMDs on both sides. The top side may have all sizes of active and passive SMDs, as well as through-hole components, while the bottom side carries passive SMDs and

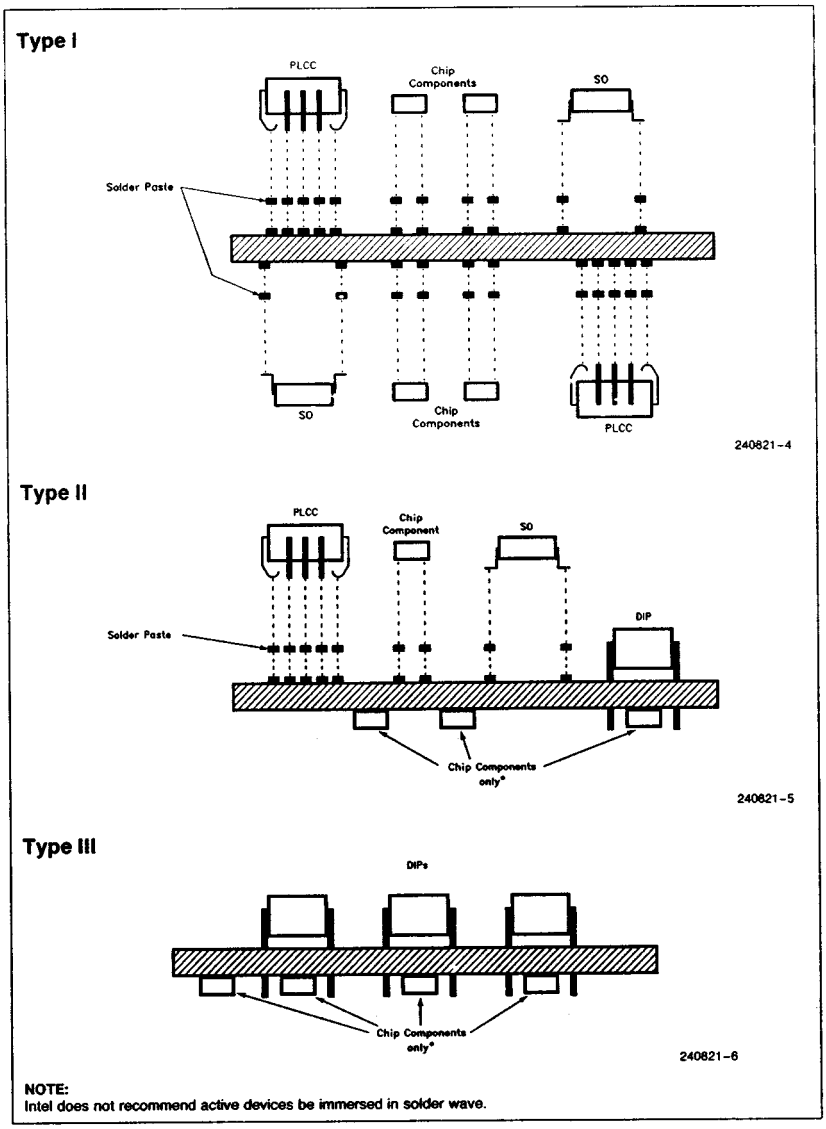


FIGURE 26.2 Type I, II, and III SMT circuit boards. (*Source:* Intel Corporation, *Packaging*, Santa Clara, Calif.: Intel Corporation, 1994. With permission.)

small active components such as transistors. This type board requires both reflow and wave soldering, and will require placement of bottom-side SMDs in adhesive.

Type III — top side has only through-hole components, which may be active and/or passive, while the bottom side has passive and small active SMDs. This type board uses wave soldering only, and also requires placement of the bottom-side SMDs in adhesive.

It should be noted that with the ongoing increase in usage of various techniques to place IC dice directly on circuit boards, Type III in some articles means a mix of packaged SMT ICs and bare die on the same board.

A Type I bare board will first have solder paste applied to the component pads on the board. Once solder paste has been deposited, active and passive parts are placed in the paste. For prototype and low-volume lines this can be done with manually guided X–Y tables using vacuum needles to hold the components, while in medium and high-volume lines automated placement equipment is used. This equipment will pick parts from

reels, sticks, or trays, then place the components at the appropriate pad locations on the board, hence the term "pick and place" equipment.

After all parts are placed in the solder paste, the entire assembly enters a reflow oven to raise the temperature of the assembly high enough to reflow the solder paste and create acceptable solder joints at the component lead/pad transitions. Reflow ovens most commonly use convection and IR heat sources to heat the assembly above the point of solder liquidus, which for 63/37 tin-lead eutectic solder is 183°C. Due to the much higher thermal conductivity of the solder paste compared to the IC body, reflow soldering temperatures are reached at the leads/pads before the IC chip itself reaches damaging temperatures. The board is inverted and the process repeated.

If mixed-technology Type II is being produced, the board will then be inverted, an adhesive will be dispensed at the centroid of each SMD, parts placed, the adhesive cured, the assembly re-righted, through-hole components mounted, and the circuit assembly will then be wave-soldered which will create acceptable solder joints for both the through-hole components and bottom-side SMDs.

A Type III board will first be inverted, adhesive dispensed, SMDs placed on the bottom-side of the board, the adhesive cured, the board re-righted, through-hole components placed, and the entire assembly wave-soldered. It is imperative to note that only passive components and small active SMDs can be successfully bottom-side wave-soldered without considerable experience on the part of the design team and the board assembly facility. It must also be noted that successful wave soldering of SMDs requires a dual-wave machine with one turbulent wave and one laminar wave.

It is common for a manufacturer of through-hole boards to convert first to a Type II or Type III substrate design before going to an all-SMD Type I design. This is especially true if amortization of through-hole insertion and wave-soldering equipment is necessary. Many factors contribute to the reality that most boards are mixed-technology Type II or Type III boards. While most components are available in SMT packages, through-hole connectors are still commonly used for the additional strength the through-hole soldering process provides, and high-power devices such as three-terminal regulators are still commonly through-hole due to off-board heat-sinking demands. Both of these issues are actively being addressed by manufacturers and solutions exist which allow Type I boards with connectors and power devices [Holmes, 1993].

Again, it is imperative that all members of the design, build, and test teams be involved from the design stage. Today's complex board designs mean that it is entirely possible to exceed the ability to adequately test a board if the test is not designed-in, or to robustly manufacture the board if in-line inspections and handling are not adequately considered. Robustness of both test and manufacturing are only assured with full involvement of all parties to overall board design and production.

It cannot be overemphasized that the speed with which packaging issues are moving requires anyone involved in SMT board or assembly issues to stay current and continue to learn about the processes. Subscribe to one or more of the industry-oriented journals noted in the "Further Information" section at the end of this Chapter, obtain any IC industry references, and purchase several SMT reference books.

26.4 Surface Mount Device (SMD) Definitions

The new user of SMDs must rapidly learn the packaging sizes and types for SMDs. Resistors, capacitors, and most other passive devices come in two-terminal packages which have end-terminations designed to rest on substrate pads/lands (Fig. 26.3).

SMD ICs come in a wide variety of packages, from 8-pin Small Outline Packages (SOLs) to 1000+ connection packages in a variety of sizes and lead configurations, as shown in Fig. 26.4. The most common commercial packages currently include **Plastic Leaded Chip Carriers** (PLCCs), Small Outline packages (SOs), **Quad Flat Packs** (QFPs), and Plastic Quad Flat Packs (PQFPs) also know as Bumpered Quad Flat Packs (BQFPs). Add in Tape Automated Bonding (TAB), Ball Grid Array (BGA) and other newer technologies, and the IC possibilities become overwhelming. Space prevents examples of all these technologies from being included here. The reader is referred to the standards of the Institute for Interconnecting and Packaging Electronic Circuits (IPC)[1] to find the latest package standards, and to the proceedings of the most recent National Electronics Production and

[1]IPC, 7380 N. Lincoln Ave, Lincolnwood, IL 60646-1705, 708-677-2850.

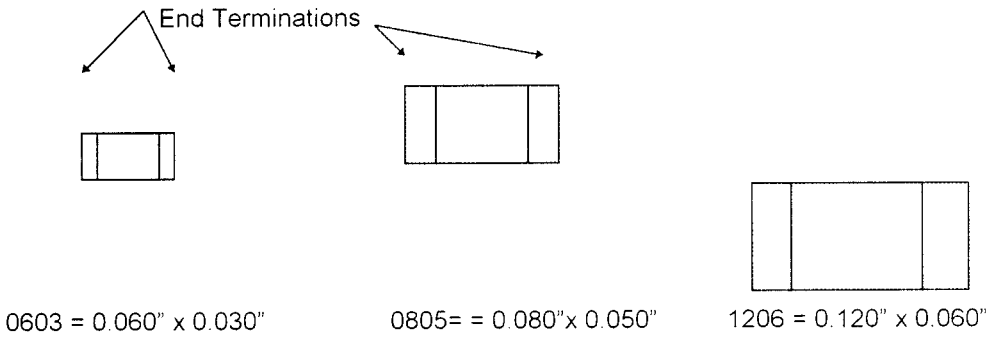


FIGURE 26.3 Example of passive component sizes (top view)(not to scale).

Productivity (NEPCON) Conference[1] for information on industry uses of the latest SMT packages. A good overview of package styles is found in Appendix A of Hollomon [1995] and (for ICs only) in Signetics [1991b] which is being updated as of this writing.

Each IC manufacturer's data books will have packaging information for their products. The engineer should be familiar with the term "lead pitch", which means the center-to-center distance between IC leads. Pitch may be in thousandths of an inch, also known as mils, or in millimeters. Common pitches are 0.050 in. (50 mil pitch), 0.025 in. (25 mil pitch) frequently called "fine pitch", and 0.020 in. and smaller frequently called "ultra-fine pitch". Metric equivalents are 1.27 mm, 0.635 mm, and 0.508 mm and smaller. Conversions from metric to inches are easily approximated if one remembers that 1 mm approximately equals 40 mils.

For process control, design teams must consider the minimum and maximum package size variations allowed by their part suppliers, the moisture content of parts as-received, and the relative robustness of each lead type. Incoming inspection should consist of both electrical and mechanical tests. Whether these are spot checks, lot checks, or no checks will depend on the relationship with the vendor.

26.5 Substrate Design Guidelines

As noted previously, substrate (typically PCB) design has an effect not only on board/component layout, but also on the actual manufacturing process. Incorrect land design or layout can negatively affect the placement process, the solder process, the test process or any combination of the three. Substrate design must take into account the mix of SMDs that are available for use in manufacturing.

The considerations noted here as part of the design process are neither all-encompassing, nor in sufficient detail for a true SMT novice to adequately deal with all the issues involved in the process. They are intended to guide an engineer through the process, allowing him/her to access more detailed information as necessary. General references are noted at the end of this chapter, and specific references will be noted as applicable. In addition, conferences such as the NEPCON, and SMI[2] are invaluable sources of information for both the beginner and the experienced SMT engineer. Although these guidelines are noted as "steps", they are not necessarily in an absolute order, and may require several iterations back-and-forth among the steps to result in a final satisfactory process and product.

After the circuit design (schematic capture) and analysis, Step 1 in the process is to determine whether all SMDs will be used in the final design making a Type I board, or whether a mix of SMDs and through-hole parts will be used, leading to a Type II or Type III board. This decision will be governed by some or all of the following considerations:

- Current parts stock
- Existence of current through-hole placement and/or wave solder equipment

[1]NEPCON, rep. by Reed Exhibition Co., Norwalk, CT.

[2]Surface Mount Technology Association, 5200 Wilson Rd., Suite 100, Minneapolis, MN 55424.

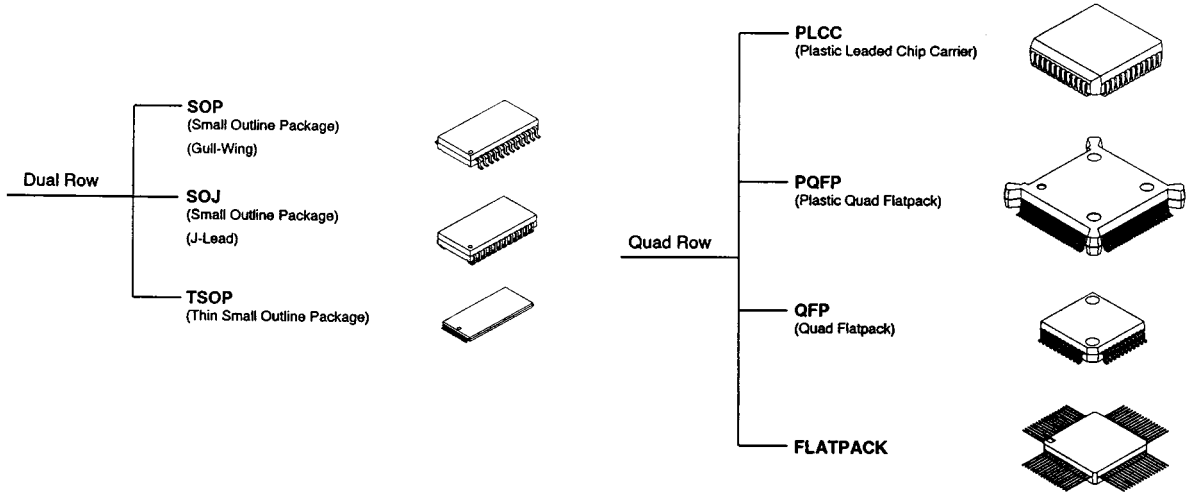


FIGURE 26.4 Examples of SMT plastic packages. (*Source:* Intel Corporation, *Packaging,* Santa Clara, Calif.: Intel Corporation, 1994. With permission.)

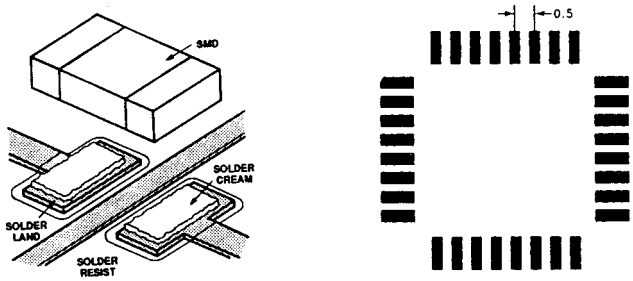


FIGURE 26.5 (**a**) Footprint land and resist. (*Source:* Phillips Semiconductors, *Surface Mount Process and Application Notes,* Sunnyvale, Calif.: Phillips Semiconductors, 1991. With permission.) (**b**) QFP footprint. (*Source:* Intel Corporation, *Packaging,* Santa Clara, Calif.: Intel Corporation, 1994. With permission.)

- Amortization of current through-hole placement and solder equipment
- Existence of reflow soldering equipment, or cost of new reflow soldering equipment
- Desired size of the final product
- Panellization of smaller Type I boards
- Thermal issues related to high power circuit sections on the board

It may be desirable to segment the board into areas based on function: RF, low power, high power, etc. using all SMDs where appropriate, and mixed-technology components as needed. Power and connector portions of the circuit may point to the use of through-hole components, although as mentioned both these issues are being addressed by circuit board material and connector manufacturers. Using one solder technique (reflow or wave) simplifies processing, and may outweigh other considerations.

Step 2 in the SMT process is to define all the footprints of the SMDs under consideration for use in the design. The footprint is the copper pattern or "land", on the circuit board upon which the SMD will be placed. Footprint examples are shown in Figs. 26.5a and 26.5b, and footprint recommendations are available from IC manufacturers and in the appropriate data books. They are also available in various ECAD packages used for the design process, or in several references that include an overview of the SMT process [Electronic Packaging and Production, 1994]. However, the reader is seriously cautioned about using the general references for anything other than the most common passive and active packages. Even the position of pin 1 may be different among IC manufacturers of the "same" chip. The footprint definition may also include the position of the solder resist pattern surrounding the copper pattern. Footprint definition sizing will vary depending on whether reflow or wave solder process is used. Wave solder footprints will require recognition of the direction of travel of the board through the wave, to minimize solder shadowing in the final fillet, as well as requirements for solder thieves. The copper footprint must allow for the formation of an appropriate, inspectable solder fillet.

If done as part of the EDA process (electronic design automation, using appropriate electronic CAD software), the software will automatically assign copper directions to each component footprint, as well as appropriate coordinates and dimensions. These may need adjustment based on considerations related to wave soldering, test points, RF and/or power issues, and board production limitations. Allowing the software to select 5 mil traces when the board production facility to be used can only reliably do 10 mil traces would be inappropriate. Likewise, the solder resist patterns must be governed by the production capabilities.

Final footprint and trace decisions will:

- allow for optimal solder fillet formation
- minimize necessary trace and footprint area
- allow for adequate test points
- minimize board area, if appropriate
- set minimum inter-part clearances for placement and test equipment to safely access the board (Fig. 26.6)

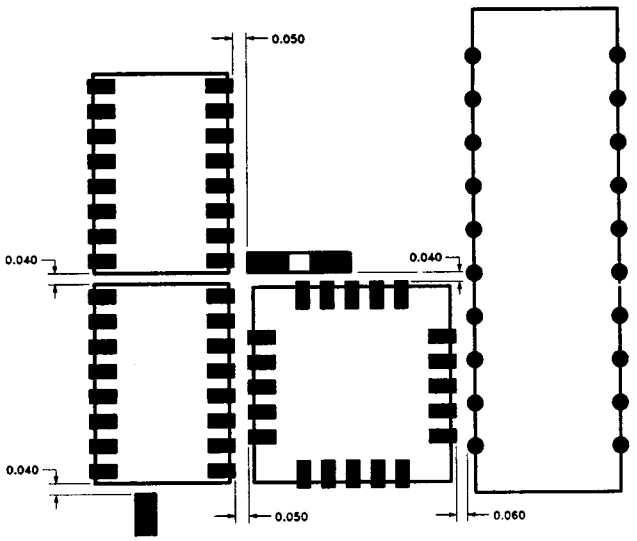


FIGURE 26.6 Minimum land-to-land clearance examples. (*Source:* Intel Corporation, *Packaging*, Santa Clara, Calif.: Intel Corporation, 1994. With permission.)

- allow adequate distance between components for post-reflow operator inspections
- allow room for adhesive dots on wave-soldered boards
- minimize solder bridging

Decisions that will provide optimal footprints include a number of mathematical issues, including:

- component dimension tolerances
- board production capabilities, both artwork and physical tolerances across the board relative to a 0–0 fiducial
- how much artwork/board shrink or stretch is allowable
- solder deposition volume consistencies with respect to fillet sizes
- placement machine accuracies
- test probe location controls and bed-of-nails grid pitch

Design teams should restrict wave-solder-side SMDs to passive components and transistors. While small SMT ICs can be successfully wave-soldered, this is inappropriate for an initial SMT design, and is not recommended by some IC manufacturers (Fig. 26.2).

These decisions may require a statistical computer program, if available to the design team. The stochastic nature of the overall process suggests a statistical programmer will be of value.

26.6 Thermal Design Considerations

Thermal management issues remain major concerns in the successful design of an SMT board and product. Consideration must be taken of the variables affecting both board temperature and junction temperature of the IC. The reader is referred to Chapter 33 in this Handbook for the basics of Thermal Management, and to Bar-Cohen and Kraus [1988] for a more detailed treatment on thermal issues affecting ICs and PCB design.

The design team must understand the basic heat transfer characteristics of most SMT IC packages [Capillo, 1993]. Since the silicon chip of an SMD is equivalent to the chip in an identical-function DIP package, the smaller SMD package means the internal lead frame metal has a smaller mass than the lead frame in a DIP package. This lesser ability to conduct heat away from the chip is somewhat offset by the leadframe of many SMDs being constructed of copper, which has a lower thermal resistance than the Kovar and Alloy 42 materials

commonly used for DIP packages. However, with less metal and shorter lead lengths to transfer heat to ambient air, more heat is typically transferred to the circuit board itself. Several board thermal analysis software packages are available, and are highly recommended for boards that are expected to develop high thermal gradients [Flotherm, 1995].

Since all electronics components generate heat in use, and elevated temperatures negatively affect the reliability and failure rate of semiconductors, it is important that heat generated by SMDs be removed as efficiently as possible. The design team needs to have expertise with the variables related to thermal transfer:

- junction temperature: T_j
- thermal resistances: Θ_{jc}, Θ_{ca}, Θ_{cs}, Θ_{sa}
- temperature sensitive parameter (TSP) method of determining Θs
- power dissipation: P_D
- thermal characteristics of substrate material

SMT packages have been developed to maximize heat transfer to the substrate. These include PLCCs with integral heat spreaders, the SOT-89 power transistor package, the DPAK power transistor package, and many others. Analog ICs are also available in power packages. Note that all of these devices are designed primarily for processing with the solder paste process, and some specifically recommend against their use with wave-solder applications. Heat sinks and heat pipes should also be considered for high-power ICs.

In the conduction process, heat is transferred from one element to another by direct physical contact between the elements. Ideally the material to which heat is being transferred should not be adversely affected by the transfer. As an example, the **glass transition temperature** T_g of FR-4 is 125°C. Heat transferred to the board has little or no detrimental affect as long as the board temperature stays at least 50°C below T_g. Good heat sink material exhibits high thermal conductivity, which is not a characteristic of fiberglass. Therefore, the traces must be depended on to provide the thermal transfer path [Choi et al., 1994]. Conductive heat transfer is also used in the transfer of heat from IC packages to heat sinks, which also requires use of thermal grease to fill all air gaps between the package and the "flat" surface of the sink.

The previous discussion of lead properties of course does not apply to leadless devices such as Leadless Ceramic Chip Carriers (LCCCs). Design teams using these and similar packages must understand the better heat transfer properties of the alumina used in ceramic packages, and must match TCEs between the LCCC and the substrate, since there are no leads to bend and absorb mismatches of expansion.

Since the heat transfer properties of the system depend on substrate material properties, it is necessary to understand several of the characteristics of the most common substrate material, FR-4 fiberglass. The glass transition temperature has already been noted, and board designers must also understand that multi-layer FR-4 boards do not expand identically in the X-, Y-, and Z-directions as temperature increases. Plate-through-holes will constrain z-axis expansion in their immediate board areas, while non-through-hole areas will expand further in the z-axis, particularly as the temperature approaches and exceeds T_g [Lee et al., 1984]. This unequal expansion can cause delamination of layers and plating fracture.

If the design team knows that there will be a need for higher abilities to dissipate heat and/or a need for higher glass transition temperatures and lower coefficients of thermal expansion (TCE) than FR-4 possesses, many other materials are available, examples of which will follow.

Note in Table 26.1 that copper-clad Invar has both variable T_g and variable thermal conductivity depending on the volume mix of copper and Invar in the substrate. Copper has a high TCE and Invar has a low TCE, so the TCE increases with the thickness of the copper layers. In addition to heat transfer considerations, board material decisions must also be based on the expected vibration, stress, and humidity in the application.

Convective heat transfer involves transfer due to the motion of molecules, typically airflow over a heat sink, and depends on the relative temperatures of the two media involved. It also depends on the velocity of air flow over the boundary layer of the heat sink. Convective heat transfer is primarily effected when forced air flow is provided across a substrate, and when convection effects are maximized through the use of heat sinks. The rules that designers are familiar with when designing THT heat-sink device designs also apply to SMT design.

The design team must consider whether passive conduction and convection will be adequate to cool a populated substrate or whether forced-air cooling or liquid cooling will be needed. Passive conductive cooling

TABLE 26.1

Substrate Material (Units)	T_g – Glass Transition Temperature (°C)	TCE – Thermal Coefficient of X–Y Expansion (PPM/°C)	Thermal Conductivity (W/M°C)	Moisture Absorption (%)
FR-4 Epoxy glass	125	13–18	0.16	0.10
Polymide glass	250	12–16	0.35	0.35
Copper-clad invar	Depends on resin	5–7	160XY — 15–20Z	NA
Poly Aramid fiber	250	3–8	0.15	1.65
Alumina/ceramic	NA	5–7	20–45	NA

is enhanced with thermal layers in the substrate, such as the previously mentioned copper/Invar. There will also be designs that will rely on the traditional through-hole device with heat sink to maximize heat transfer. An example of this would be the typical three-terminal voltage regulator mounted on a heat sink or directly to a metal chassis for heat conduction, for which standard calculations apply [Lee et al., 1993].

Many specific examples of heat transfer may need to be considered in board design, and of course most examples involve both conductive and convective transfer. For example, the air gap between the bottom of a standard SMD and the board effects the thermal resistance from the case to ambient, Θ_{ca}. A wider gap will result in a higher resistance, due to poorer convective transfer, whereas filling the gap with a thermal-conductive epoxy will lower the resistance by increasing conductive heat transfer. Thermal-modeling software is the best way to deal with these types of issues, due to the need for rigorous application of computational fluid dynamics (CFD) [Lee, 1994].

26.7 Adhesives

In the surface mount assembly process, Type II and Type III boards will always require adhesive to mount the SMDs for passage through the solder wave. This is apparent when one envisions components on the bottom side of the substrate with no through-hole leads to hold them in place. Adhesives will stay in place after the soldering process, and throughout the life of the substrate and the product, since there is no convenient means for adhesive removal once the solder process is complete. This means the adhesive used must meet a number of both physical and chemical characteristics that should be considered during the three phases of adhesive use in SMT production: pre-application properties relating to storage and dispensing issues, curing properties relating to time and temperature needed for cure, and post-curing properties: relating to final strength, mechanical stability, and reworkability. Among these characteristics are:

- electrically non-conductive
- thermal coefficient of expansion similar to the substrate and the components
- stable in both storage and after application, prior to curing
- stable physical drop shape — retains drop height and fills z-axis distance between the board and the bottom of the component; thixotropic with no adhesive migration
- non-corrosive to substrate and component materials
- chemically inert to flux, solder, and cleaning materials used in the process
- cureable as appropriate to the process: UV, oven, or air-cure
- removable for rework and repair
- once cured, unaffected by temperatures in the solder process
- adhesive color, for easy identification by operators

One-part adhesives are easier to work with than two-part adhesives because an additional process step is not required. The user must verify that the adhesive has sufficient shelf life and pot life for the user's perceived process requirements. Both epoxy and acrylic adhesives are available as one- or two-part systems, and must be cured thermally. Generally, epoxy adhesives are cured by oven-heating, while acrylics may be formulated to be cured by long-wave UV light or heat.

Adhesive can be applied by screening techniques similar to solder paste screen application, by pin transfer techniques, and by syringe deposition. Screen and pin-transfer techniques are suitable for high-volume production lines with few product changes over time. Syringe deposition using an X–Y table riding over the board with a volumetric pump and syringe tip is more suitable for lines with a varying product mix, prototype lines, and low-volume lines where the open containers of adhesive necessary in pin-transfer and screen techniques are avoided. Newer syringe systems are capable of handling high-volume lines. See Fig. 26.8 for methods of adhesive deposition.

If Type II or Type III assemblies are used, and thermal transfer between components and the substrate is a concern, the design team should consider thermally conductive adhesives.

Regardless of the type of assembly, the type of adhesive used, or the curing technique used, adhesive volume and height must be carefully controlled. Slump of adhesive after application is undesirable because the adhesive must stay high enough to solidly contact the bottom of the component, and must not spread and contaminate any pad associated with the component.

If adhesive dot height = X, substrate metal height = Y, and SMD termination thickness = Z, then X>Y+Z, allowing for all combinations of potential errors, e.g., end termination min and max thickness, adhesive dot min and max height, and substrate metal min and max height:

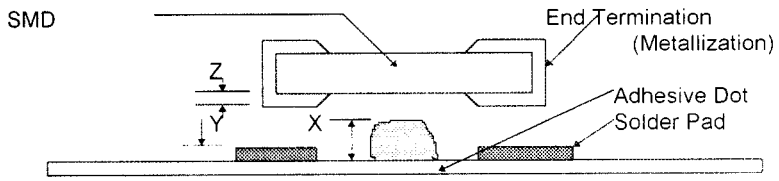


FIGURE 26.7 Relation of adhesive dot, substrate, and component. (*Source:* Phillips Semiconductors, *Surface Mount Process and Application Notes,* Sunnyvale, Calif.: Phillips Semiconductors, 1991. With permission.)

Typically, end termination thickness variations are available from the part manufacturer. Solder pad thickness variations are a result of the board manufacturing process, and will vary not only on the type of board metallization (standard etch vs. plated-through-hole) but also on the variations within each type. For adequate dot height, which will allow for some dot compression by the part, X should be between 1.5X and 2.5X of the total Y+Z, or just Z when dummy tracks are used. If adhesive dots are placed on masked areas of the board, mask thickness must also be considered.

A common variation on the above design is to place "dummy" copper pads under the center of the part. Since these pads are etched and plated at the same time as the actual solder pads, the variation in metal height Y is eliminated as an issue. Adhesive dots are placed on the dummy pads and X > Z is the primary concern.

Adhesive dispensing quality issues are addressed by considerations of:

- type of adhesive to be used
- process-area ambient temperature and humidity
- incoming quality control
- no voids in cured adhesive to prevent trapping of flux, dirt, etc.
- volume control
- location control
- as in Fig. 26.7, all combinations of termination, dot, and substrate height/thicknesses

Prasad [1997] has an excellent in-depth discussion of adhesives in SMT production.

26.8 Solder Paste and Joint Formation

Solder joint formation is the culmination of the entire process. Regardless of the quality of the design, or any other single portion of the process, if high-quality reliable solder joints are not formed, the final product is not reliable. It is at this point that PPM levels take on their finest meaning. For a medium-size substrate (nominal

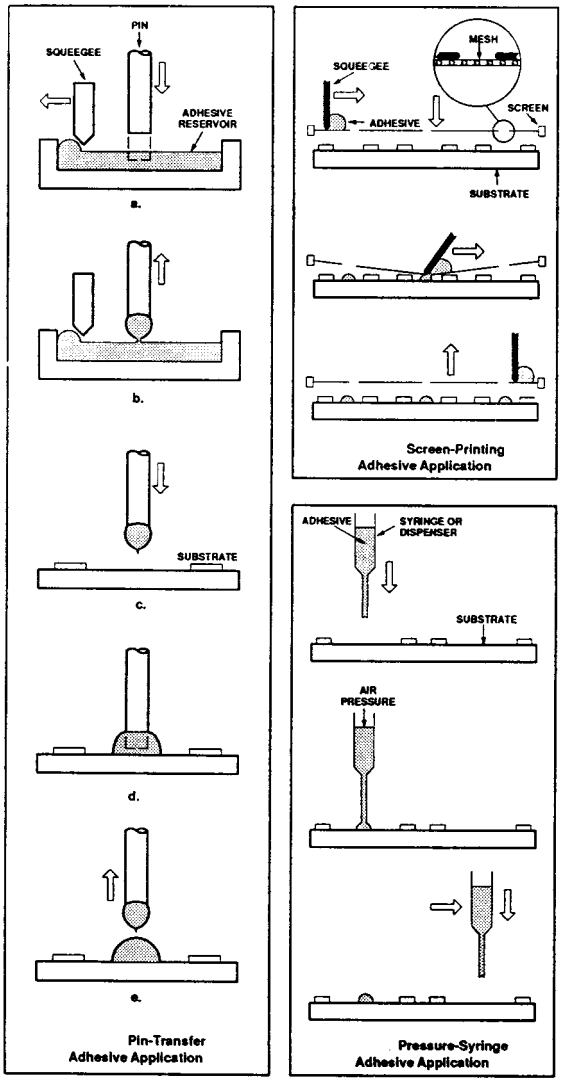


FIGURE 26.8 Methods of adhesive deposition.

6" X 8"), with a medium density of components, a typical mix of active and passive parts on the topside and only passive and 3- or 4-terminal active parts on bottomside, there may be in excess of 1000 solder joints per board. If solder joints are manufactured at the 3 sigma level (99.73% good joints, or 0.27% defect rate, or 2700 defects per 1 million joints), *there will be 2.7 defects per board*!! At the 6 sigma level, of 3.4 PPM, there will be a defect on 1 board out of every 294 boards produced. If your anticipated production level is 1000 units per day, you will have 3.4 rejects based solely on solder joint problems, not counting other sources of defects.

Solder paste may be deposited by syringe, or by screen or stencil printing techniques. Stencil techniques are best for high-volume/speed production although they do require a specific stencil for each board design. Syringe and screen techniques may be used for high-volume lines and are also suited to mixed-product lines where only small volumes of a given board design are to have solder paste deposited. Syringe deposition is the only solder paste technique that can be used on boards which already have some components mounted. It is also well suited for prototype lines and for any use requires only software changes to develop a different deposition pattern.

Solder joint defects have many possible origins:

- poor or inconsistent solder paste quality
- inappropriate solder pad design/shape/size/trace connections
- substrate artwork or production problems, e.g., mismatch of copper and mask, warped substrate
- solder paste deposition problems, e.g., wrong volume or location
- component lead problems, e.g., poor coplanarity or poor tinning of leads
- placement errors, e.g., part rotation or X–Y offsets
- reflow profile, e.g., preheat ramp too fast or too slow; wrong temperatures created on substrate
- board handling problems, e.g., boards getting jostled prior to reflow.

Once again, a complete discussion of all of the potential problems that can affect solder joint formation is beyond the scope of this chapter. Many references are available which address the issues. An excellent overview of solder joint formation theory is found in Lau [1991]. Update information this and all SMT topics is available each year at conferences such as SMI and NEPCON.

While commonly used solder paste for both THT and SMT production contains 63-37 eutectic tin-lead solder, other metal formulations are available, including 96-4 tin-silver (a.k.a. silver solder). The fluxes available include RMA, water-soluble, and no-clean. The correct decision rests as much on the choice of flux as it does on the proper metal mixture. A solder paste supplier can best advise on solder pastes for specific needs. Many studies are in process to determine a no-lead replacement for lead-based solder in commercial electronic assemblies. The design should investigate the current status of these studies as well as the status of no-lead legislation as part of the decision-making process.

To better understand solder joint formation, one must understand the make-up of solder paste used for SMT soldering. The solder paste consists of microscopic balls of solder, most commonly tin-lead with the accompanying oxide film, flux, and activator and thickener solvents as shown in Fig. 26.9.

The fluxes are an integral part of the solder paste, and are discussed further in Section 26.11. RMA, water soluble, and no-clean flux/pastes are available. An issue directly related to fluxes, cleaning and fine-pitch components (25 mil pitch and less) is reflowing in an inert environment. Inert gas blanketing the oven markedly reduces the development of oxides in the elevated temperatures present. Oxide reduction needs are greater with the smaller metal balls in paste designed for fine-pitch parts because there is more surface area on which oxides can form. No-clean fluxes are not as active as other fluxes and therefore have a lesser ability to reduce the oxides formed on both the paste metal and substrate metallizations. Inerting the oven tends to solve these problems. However, it brings with it control issues that must be considered.

Regardless of the supplier, frequent solder paste tests are advisable, especially if the solder is stored for prolonged periods before use. At a minimum, viscosity, percent metal, and solder sphere formation should be tested [Capillo, 1990]. Solder sphere formation is particularly important because acceptable particle sizes will vary depending on the pitch of the smallest-pitch part to be used, and the consistency of solder sphere formation will effect the quality of the final solder joint. Round solder spheres have the smallest surface area for a given

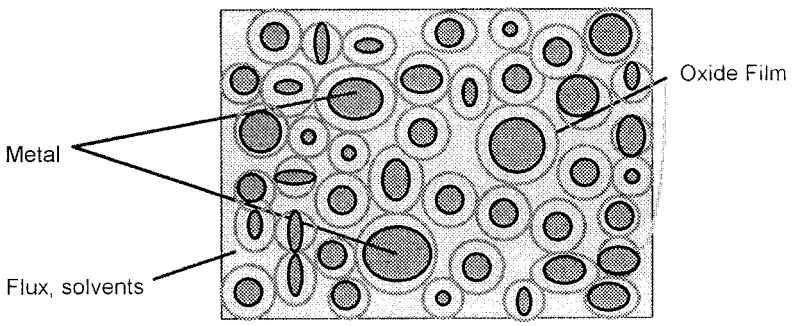


FIGURE 26.9 Make-up of SMT solder paste.

volume. Therefore, they will have the least amount of oxide formation. Uneven distribution of sphere sizes within a given paste can lead to uneven heating during the reflow process, with the result that the unwanted solder balls will be expelled from the overall paste mass at a given pad/lead site. Fine-pitch paste has smaller ball sizes and consequently more surface area on which oxides can form.

It should be noted at this point that there are three distinctly different "solder balls" referred to in this chapter and in publications discussing SMT. The solder sphere test refers to the ability of a volume of solder to form a ball shape due to its inherent surface tension when reflowed (melted) on a non-wettable surface. This ball formation is dependent on minimum oxides on the microscopic metal balls which make up the paste — the second type of "solder ball". It is also dependent on the ability of the flux to reduce the oxides that are present, as well the ramp-up of temperature during the preheat and drying phases of the reflow oven profile. Too steep a time/temperature slope can cause rapid escape of entrapped volatile solvents, resulting in expulsion of small amounts of metal that will form undesirable "solder balls" of the third type, small metal balls scattered around the solder joint(s) on the substrate itself rather than on the tinned metal of the joint. This third type of ball can also be formed by excess solder paste on the pad, and by mis-deposition on non-wettable areas of the substrate.

The reader is referred to Lau [1991] for discussions of finite element modeling of solder joints, and detailed analytical studies of most aspects of basic joints and of joint failures. Various articles by Engelmaier et al. also address many solder joint reliability issues and their analytical analysis. These and other sources will discuss in detail the quality issues that effect solder paste:

- viscosity and its measurement
- printability
- open time
- slump
- metal content
- particle/ball size in mesh
- particle/ball size consistency
- wetting
- storage conditions

Note with regard to viscosity measurements, some paste manufacturers will prefer the spindle technique and some the spiral technique. To properly compare the paste manufacturer's readings with your tests, the same technique must be used.

26.9 Parts Inspection and Placement

Briefly, all parts must be inspected prior to use. Functional parts testing should be performed on the same basis as for through-hole devices. Each manufacturer of electronic assemblies is familiar with the various processes used on through-hole parts, and similar processes must be in place on SMDs. Problems with solderability of leads and lead planarity are two items that can lead to the largest number of defects in the finished product. Solderability is even more important with SMDs than with through- hole parts because all electrical and mechanical strength rests within the solder joint, there being no hole-with-lead to add mechanical strength.

Lead coplanarity is defined as follows. If a multi-lead part, e.g., IC, is placed on a planar surface, lack of ideal coplanarity exists if the lowest solderable part of any lead does not touch that surface. **Coplanarity** requirements vary depending on the pitch of the component leads and their shape, but generally out-of-plane measurements should not exceed 4 mils (0.004 in.) for 50-mil pitch devices, and 2 mils for 25-mil pitch devices.

All SMDs undergo thermal shocking during the soldering process, particularly if the SMDs are to be wave-soldered (Type II or Type III boards), which means they will be immersed in the molten solder wave for 2 to 4 s. Therefore, all plastic-packaged parts must be controlled for moisture content. If the parts have not been stored in a low-humidity environment (<25%RH), then the absorbed moisture will expand during the solder process and crack the package — a phenomenon know as "popcorning" because the crack is accompanied by a loud "pop" and the package expands due to the expansion of moisture, just like real popcorn. IC suppliers have strict recommendations on storage ambient humidity and temperature, and also on the baking procedures

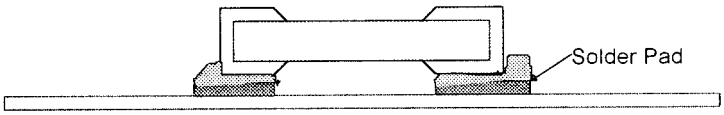


FIGURE 26.10a Part placed *into* solder paste with a passive part.

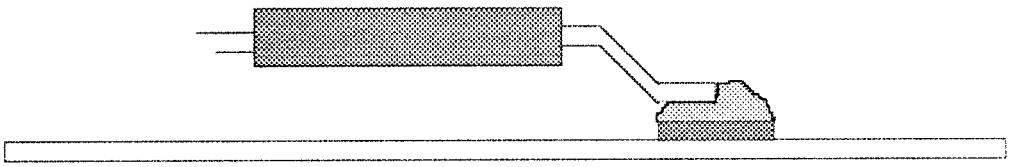


FIGURE 26.10b Part placed *into* solder paste with an active part.

necessary to allow safe reflow soldering if those recommendations are not heeded. Follow them carefully. Typically if storage RH is above 20%, baking must be considered prior to reflow.

Parts Placement

Proper parts placement not only places the parts within an acceptable window relative to the solder pad pattern on the substrate, but the placement machine will apply enough downward pressure on the part to force it halfway into the solder paste as well (Fig. 26.10a and 26.10b). This assures both that the part will sit still when the board is moved, and that coplanarity offsets within limits will still result in an acceptable solder joint. The effects of coplanarity can be done mathematically by considering:

- the thickness of the solder paste deposit
- maximum coplanarity offsets among leads
- lead penetration in paste

If T is the thickness of paste, C is maximum allowable coplanarity, and P is penetration in paste (as a percentage of overall average paste thickness), then:

$$P = \frac{C}{T} \times 100\%$$

Parts placement may be done manually for prototype or low-volume operations, although this author suggests the use of guided X–Y tables with vacuum part pickup for even the smallest operation. Manual placement of SMDs does not lend itself to repeatable work. For medium and high volume work, a multitude of machines are available. See Fig. 26.11 for the four general categories of automated placement equipment.

One good source for manufacturer's information on placement machines and most other equipment used in the various SMT production and testing phases is the annual "Directory of Suppliers to the Electronics Manufacturing Industry", published by Electronic Packaging and Production. Among the elements to consider in the selection of placement equipment, whether fully automated or X–Y-vacuum assist tables, are:

- volume of parts to be placed per hour
- conveyorized links to existing equipment
- packaging of components to be handled: tubes, reels, trays, bulk, etc.
- ability to download placement information from CAD/CAM systems
- ability to modify placement patterns by the operator
- vision capability needed, for board fiducials and/or fine-pitch parts

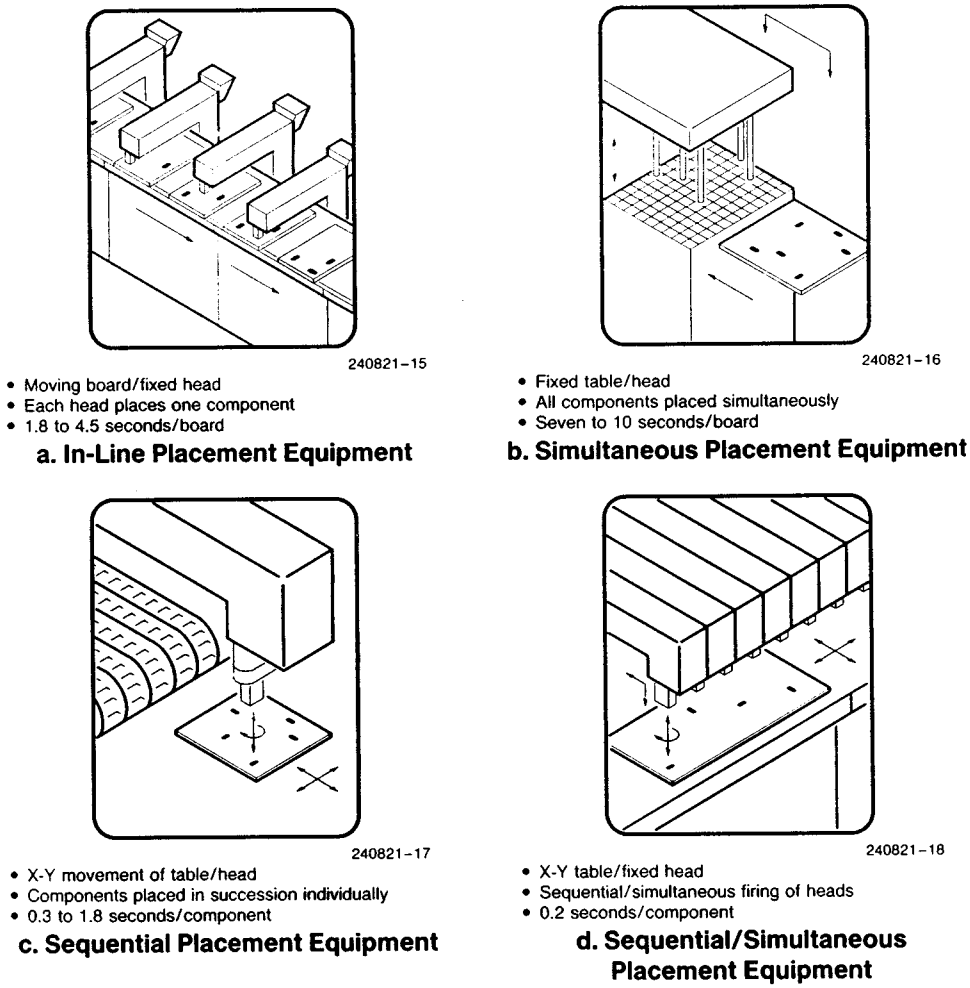


FIGURE 26.11 Four major categories of placement equipment. (*Source:* Intel Corporation, *Packaging,* Santa Clara, Calif.: Intel Corporation, 1994. With permission.)

Checks of new placement equipment, or in-place equipment when problems occur, should include:

- X-accuracy
- Y-accuracy
- Z-accuracy
- placement pressure
- vision system checks both downward and upward

It must be emphasized that placement accuracy checks cannot be made using standard circuit boards. Special glass plates with engraved measurement patterns and corresponding glass parts must be used because standard circuit boards and parts vary too widely to allow accurate X-, Y-, and Θ-measurements.

26.10 Reflow Soldering

Once SMDs have been placed in solder paste, the assembly will be reflow soldered. This can be done in either batch-type ovens or conveyorized continuous-process ovens. The choice depends primarily on the board throughput per hour required. While many early ovens were of the vapor phase type, most ovens today use

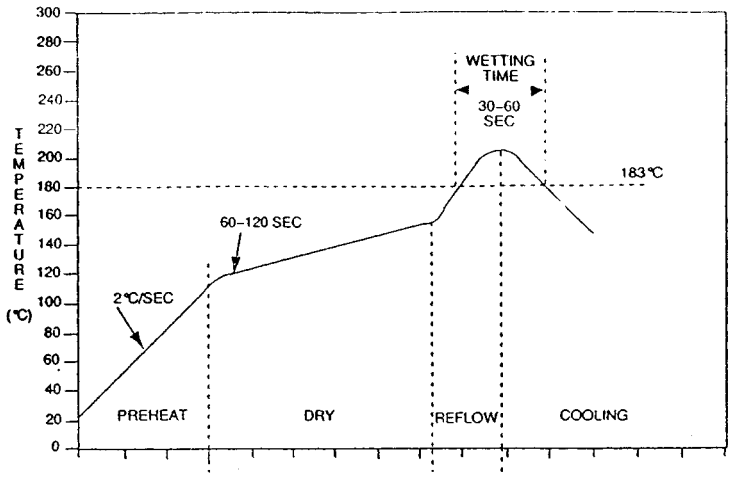


FIGURE 26.12 Typical thermal profile for SMT reflow soldering (Type I or II assemblies). (*Source:* Cox, N.R., *Reflow Technology Handbook,* Minneapolis, Minn.: Research, Inc., 1992. With permission.)

infrared (IR) heating, convection heating, or a combination of the two. In IR ovens the absorbance of the paste, parts, glue, etc. as a function of color should considered. Convection ovens tend to be more forgiving with variations in color and thermal masses on the substrates. This author does not recommend vapor phase (condensation heating) ovens to new users of SMT. All ovens are zoned to provide a thermal profile necessary for successful SMD soldering. An example of an oven profile is shown in Fig. 26.12, and the phases of reflow soldering that are reflected in that example include:

- Preheat: The substrate, components and solder paste preheat.
- Dry: Solvents evaporate from the solder paste. Flux activates, reduces oxides, and evaporates. Both low- and high-mass components have enough soak time to reach temperature equilibrium.
- Reflow: The solder paste temperature exceeds the liquidus point and reflows, wetting both the component leads and the board pads. Surface tension effects occur, minimizing wetted volume.
- Cooling: The solder paste cools below the liquidus point, forming acceptable (shiny and appropriate volume) solder joints.

The setting of the reflow profile is not trivial. It will vary on whether the flux is RMA, water soluble, or no clean, and it will vary depending on both the mix of low- and high-thermal mass components, and on how those components are laid out on the board. The profile should exceed the liquidus temperature of the solder paste by 20 to 25°C. While final setting of the profile will depend on actual quality of the solder joints formed

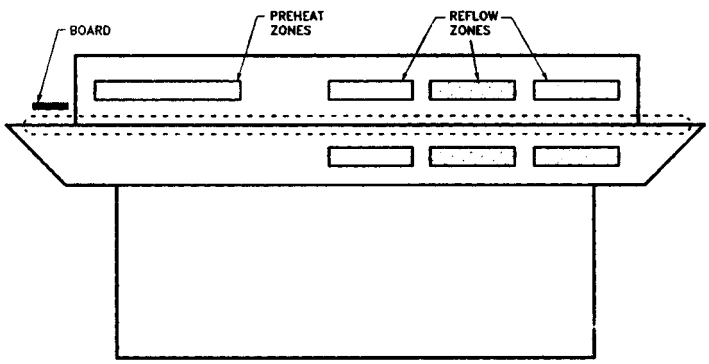


FIGURE 26.13 Conveyorized reflow oven showing zones which create the profile. (*Source:* Intel Corporation, *Packaging,* Santa Clara, Calif.: Intel Corporation, 1994. With permission.)

in the oven, initial profile setting should rely heavily on information from the solder paste vendor, as well as the oven manufacturer. Remember that the profile shown is the profile to be developed on the substrate, and the actual control settings in various stages of the oven itself may be considerably different, depending on the thermal inertia of the product in the oven and the heating characteristics of the particular oven being used. This should be determined not by the oven settings but by instrumenting actual circuit boards with thermocouples and determining that the profiles at various locations on the circuit board meet the specifications necessary for good soldering.

Defects as a result of poor profiling may include:

- component thermal shock
- solder splatter
- solder balls formation
- dewetted solder
- cold or dull solder joints

It should be noted that many other problems may contribute to defective solder joint formation. One example would be placement misalignment which contributes to the formation of solder bridges, as shown in Fig. 26.14.

Other problems that may contribute to defective solder joints include poor solder mask adhesion, and unequal solder land areas at opposite ends of passive parts, which creates unequal moments as the paste liquifies and develops surface tension. Wrong solder paste volumes, whether too much or too little, will create defects, as will board shake in placement machines and coplanarity problems in IC components. Many of these problems should be covered and compensated for during the design process and the qualification of SMT production equipment.

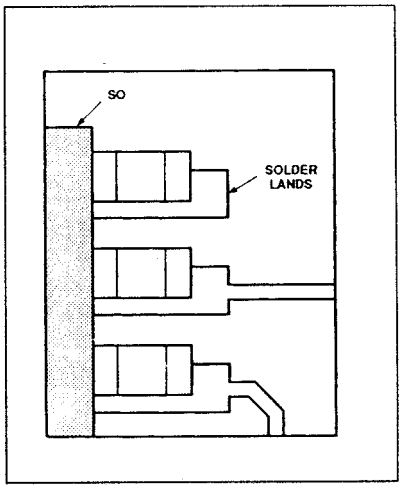


FIGURE 26.14 Solder bridge risk due to misalignment. (*Source:* Phillips Semiconductors, *Surface Mount Process and Application Notes,* Sunnyvale, Calif.: Phillips Semiconductors, 1991. With permission.)

Post-Reflow Inspection

Final analysis of the process is performed based on the quality of the solder joints formed in the reflow process. Whatever criteria may have been followed during the overall process, solder joint quality is the final determining factor of the correctness of the various process steps. As noted earlier, the quality level of solder joint production is a major factor in successful board assembly. A primary criteria is the indication of wetting at the junction of the reflowed solder and the part termination. This same criteria shown in Fig. 26.15 applies to both through-hole and SMDs, with only the inspection location being different.

Note that criteria shown in Fig. 26.15 are for any solderable surface, whether component or board, SMT or THT. Some lead surfaces are defined as not solderable, e.g., the cut and not-tinned end of an SO or QFP lead is not considered solderable. Parts manufacturers will define whether a given surface is designed to be solderable.

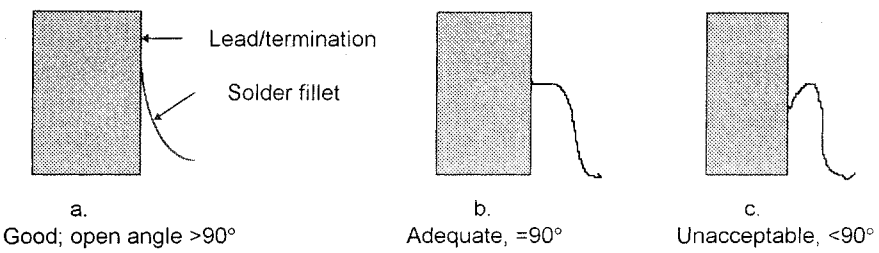


FIGURE 26.15 Solder joint inspection criteria.

Presentation of criteria for all the various SMD package types and all the possible solder joint problems is beyond the scope of this chapter. The reader is directed to Hollomon [1995], Hwang [1989], Lau [1991], Klein-Wassink [1989], and Prasad [1997] for an in-depth discussion of these issues.

26.11 Cleaning

Cleaning, like all other parts of the process, should be considered during the design phase. Cleaning requirements are determined largely by the flux in the solder paste, and should be determined before production is ever started. Design issues may effect the choice of flux, which determines cleaning. For example, low clearance (close to the substrate) parts are difficult to clean under, and the use of no-clean flux may be suggested.

Rosin Mildly Activated (RMA) is the old standard, and if cleaning is needed with RMA, either solvent-based cleaners or water with saponifiers must be used. (Saponifiers are alkaline materials that react with the rosin so that it becomes water-soluble.) RMA tends to be non-active at "room" temperatures, and may not need to be cleaned on commercial products designed for indoor use. The major limitation on RMA at this point is the need for chemicals in the cleaning process.

Water-soluble fluxes are designed to be cleaned with pure water. They remain active at room temperatures, and therefore *must* be cleaned when used. Beyond their activity, the other disadvantage to water soluble fluxes is that their higher surface tension relative to solvent-based cleaners means there is more difficulty cleaning under low-clearance components.

No-clean fluxes are designed to not be cleaned, and this means they are of low activity (they don't reduce oxides as well as other types of flux) and when used they should *not* be cleaned. No-clean fluxes are designed so that after reflow they microscopically encapsulate themselves, sealing in any active components. If the substrate is subsequently cleaned, the encapsulants may be destroyed, leaving the possibility of active flux components remaining on the board.

26.12 Prototype Systems

Systems for all aspects of SMT assembly are available to support low-volume/prototype needs. These systems will typically have manual solder-paste deposition and parts placement systems, with these functions being assisted for the user. Syringe solder paste deposition may be as simple as a manual medical-type syringe dispenser which must be guided and squeezed freehand. More sophisticated systems will have the syringe mounted on an X–Y arm to carry the weight of the syringe, and will apply air pressure to the top of the syringe with a foot-pedal control, freeing the operator's arm to guide the syringe to the proper location on the substrate and perform the negative z-axis maneuver which will bring the syringe tip into the proper location and height above the substrate. Dispensing is then accomplished by a timed air pressure burst applied to the top of the syringe under foot-pedal control. Paste volume is likewise determined by trial-and-error with the time/pressure relation and depends on the type and manufacturer of paste being dispensed.

Parts placement likewise may be as simple as tweezers and may progress to hand-held vacuum probes to allow easier handling of the components. As mentioned in the Section 26.9, X–Y arm/tables are available which have vacuum-pick nozzles to allow the operator to pick a part from a tray, reel, or stick, and move the part over the correct location on the substrate. The part is then moved down into the solder paste, the vacuum is turned off manually or automatically, and the nozzle is raised away from the substrate.

Soldering of prototype/low-volume boards may be done by contact soldering of each component, by a manually guided hot-air tool, or in a small batch or conveyorized oven. Each step up in soldering sophistication is, of course, accompanied by an increase in the investment required.

For manufacturers with large prototype requirements, it is possible to set up an entire line that would involve virtually no hardware changes from one board to another:

- ECAD design and analysis, producing Gerber files.
- CNC circuit board mill takes Gerber files and mills out two-sided boards.
- Software translation package generates solder-pad centroid information.

- Syringe solder paste deposition system takes translated Gerber file and dispenses appropriate amount at each pad centroid.
- Software translation package generates part centroid information.
- Parts placement equipment places parts based on translated part centroid information
- Assembly is reflow soldered.

The only manual process in the above system is adjustment of the reflow profile based on the results of soldering an assembly. The last step in the process would be to test the finished prototype board. This system could also be used for very small volume production runs, and all components as described are available. With a change from milled boards to etched boards, the system can be used as a flexible assembly system.

Defining Terms

Coefficient of thermal expansion (CTE, a.k.a. TCE): A measure of the ratio between the measure of a material and its expansion as temperature increases. May be different in X-, Y-, and Z-axes. Expressed in PPM/°C. A measure that allows comparison of materials that are to be joined.

Coplanarity: A simplified definition of planarity, which is difficult to measure. Coplanarity is the distance between the highest and lowest leads, and is easily measured by placing the IC on a flat surface such as a glass plate. The lowest leads will then rest on the plate, and the measured difference to the lead highest above the plate is the measurement of coplanarity.

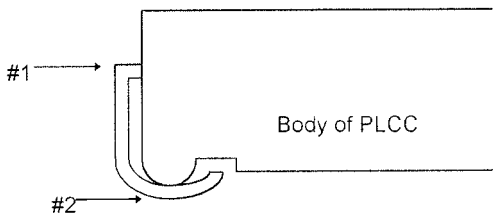
Glass transition temperature (T_g): Below T_g a polymer substance, such as fiberglass, is relatively linear in its expansion/contraction due to temperature changes. Above T_g, the expansion rate increases dramatically and becomes non-linear. The polymer will also lose its stability, i.e., an FR-4 board "droops" above T_g.

Gull wing: An SMD lead shape as shown in Fig. 26.4 for SOPs and QFPs, and in Fig. 26.11b. So called because it looks like a gull's wing in flight.

J-Lead: An SMD lead shape as shown in the PLCC definition. So called because it is in the shape of the capital letter J.

Land: A metallized area intended for the placement of one termination of a component. Lands may be tinned with solder, or be bare copper in the case of SMOBC circuit board fabrication.

PLCC: Plastic leaded chip carrier. Shown in Fig. 26.4, it is a common SMT IC package and is the only package that has the leads bent back under the IC itself.



Planarity: Lying in the same plane. A plane is defined by the exit of the leads from the body of the IC (arrow #1 above). A second plane is defined as the average of the lowest point all leads are below the first plane (arrow #2 above). Non-planarity is the maximum variation in mils or mm of any lead of an SMD from the lowest point plane.

Quad flat pack: Any flat pack IC package that has gull-wing leads on all four sides.

Through-hole: Also a plate-through-hole (PTH). A via that extends completely through a substrate and is solder-plated.

SMOBC: An acronym for "solder mask on bare copper", a circuit board construction technique that does not tin the copper traces with solder prior to the placement of the solder mask on the board.

Through-Hole Technology (THT): The technology of using leaded components that require holes through the substrate for their mounting (insertion) and soldering.

Related Topics

25.2 Layout, Placement, and Routing • 33.3 Chip Module Thermal Resistance

References

F. M. Mims, III, "Surface mount technology: an introduction to the packaging revolution," *Radio Electronics,* pp 58-90, 1987.

S. H. Leibson, "The promise of surface mount technology," *EDN* Magazine, pp 165-174, 1987.

C. Higgins, Signetics Corp., presentation, November 1991.

J. G. Holmes, "Surface mount solution for power devices," *Surface Mount Technol.,* 18-20, 1993.

IPC, Proceedings, National Electronics Production and Packaging Conference-West (NEPCON West), Reed Exhibition Companies, Norwallk, Conn., 1996.

J. K. Hollomon, Jr., *Surface Mount Technology for PC Board Design,* Indianapolis, Ind.: Prompt Publishing, 1995, chapt. 2.

C. Capillo, *Surface Mount Technology, Materials, Processes and Equipment*, New York: McGraw-Hill, 1990, chapt. 3.

C. Capillo, "Conduction heat transfer measurements for an array of surface mounted heated components," *Am. Soc. Mech. Eng., Heat Transfer Div.,* Proceedings of the 1993 ASME Annual Meeting, 263, 69-78, 1993.

Flotherm, "Advanced thermal analysis of packaged electronic systems," Westborough, Mass.: Flomerics, Inc., 1995.

C. Y. Choi, S. J. Kim, A. Ortega, "Effects of substrate conductivity on convective cooling of electronic components," *J. Electron. Packaging,* 116(3), 198-205, 1994.

L. C. Lee et al., "Micromechanics of multilayer printed circuit board," *IBM J. Res. Dev.,* 28(6), 1984.

L. C. Lee et al., *Linear/Interface IC Device Databook,* vol. 1, Section 3 Addendum, Motorola, 1993.

T. Y. Lee, "Application of a CFD tool for system-level thermal simulation," *IEEE Trans. Components, Packaging, and Manufacturing Technol.,* Part A, 17(4), 564-571, 1994.

J. H. Lau, *Solder Joint Reliability,* New York: Van Nostrand Reinhold, 1991.

C. Capillo, *Surface Mount Technology Materials, Processes and Equipment.* New York: McGraw Hill, 1990, chapts. 7 and 8.

Electronic Packaging and Production, Highlands Ranch, Colo.: Cahners Publishing Co.

Intel, *Packaging,* Santa Clara, Calif.: intel Corporation, 1994.

N. R. Cox, *Reflow Technology Handbook,* Minneapolis, Minn.: Research, Inc., 1992.

F. Classon, *Surface Mount Technology for Concurrent Engineering and Manufacturing,* New York: McGraw-Hill, 1993.

J. S. Hwang, *Solder Paste in Electronics Packaging,* New York: Van Nostrand Reinhold, 1989.

C. Lea, *A Scientific Guide to SMT,* Electrochemical Publishing Co. Ltd, 1988.

R. J. Klein-Wassink, *Soldering in Electronics,* Electrochemical Publishing Co. Ltd, 1989.

P. P. Marcoux, *Fine Pitch Surface Mount Technology,* New York: Van Nostrand Reinhold, 1992.

R. P. Prasad, *Surface Mount Technology Principles and Practice,* 2nd ed., New York: Van Nostrand Reinhold, 1997.

R. Rowland, *Applied Surface Mount Assembly,* New York: Van Nostrand Reinhold, 1993.

S. G. Shina, *Concurrent Engineering and Design for Manufacture of Electronic Products,* New York: Van Nostrand Reinhold, 1991.

Phillips Semiconductor, Signetics Surface Mount Process and Application Notes, Sunnyvale, Calif.: Phillips Semiconductor, 1991.

A. Bar-Cohen, A. D. Kraus, *Advances in Thermal Modelling of Electronic Components and Systems,* New York: ASME Press, 1988.

Further Information

Specific journal references are available on any aspect of SMT. A search of the COMPENDEX Engineering Index 1987–present will show over 1500 references specifically to SMT topics.

Education/Training: a partial list of organizations that specialize in education and training directly related to issues in surface mount technology:

Electronic Manufacturing Productivity Facility (a joint operation of the U.S. Navy-Naval Avionics Center and Purdue University in Indianapolis). 714 North Senate Ave., Indianapolis, IN 46202-3112. 317-226-5607

SMT Plus, Inc. 5403-F Scotts Valley Drive, Scotts Valley, CA 95066; 408-438-6116

Surface Mount Technology Association (SMTA), 5200 Wilson Rd, Ste 100, Edina, MN 55424-1338. 612-920-7682.

Conferences directly related to SMT:

Surface Mount International (SMI). Sponsered by SMTA.

National Electronics Packaging and Production Conference (NEPCON). Coordinated by Reed Exhibition Co., P.O. Box 5060, Des Plaines, IL 60017-5060. 708-299-9311.

Kennedy, E.J., Wait, J.V. “Operational Amplifiers”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

27

Operational Amplifiers

E.J. Kennedy
University of Tennessee

John V. Wait
University of Arizona (Retired)

27.1 Ideal and Practical Models
The Ideal Op Amp • Practical Op Amps • SPICE Computer Models
27.2 Applications
Noninverting Circuits

27.1 Ideal and Practical Models

E.J. Kennedy

The concept of the **operational amplifier** (usually referred to as an *op amp*) originated at the beginning of the Second World War with the use of vacuum tubes in dc amplifier designs developed by the George A. Philbrick Co. [some of the early history of operational amplifiers is found in Williams, 1991]. The op amp was the basic building block for early electronic servomechanisms, for synthesizers, and in particular for analog computers used to solve differential equations. With the advent of the first monolithic integrated-circuit (IC) op amp in 1965 (the μA709, designed by the late Bob Widlar, then with Fairchild Semiconductor), the availability of op amps was no longer a factor, while within a few years the cost of these devices (which had been as high as $200 each) rapidly plummeted to close to that of individual discrete transistors.

Although the digital computer has now largely supplanted the analog computer in mathematically intensive applications, the use of inexpensive operational amplifiers in instrumentation applications, in pulse shaping, in filtering, and in signal processing applications in general has continued to grow. There are currently many commercial manufacturers whose main products are high-quality op amps. This competitiveness has ensured a marketplace featuring a wide range of relatively inexpensive devices suitable for use by electronic engineers, physicists, chemists, biologists, and almost any discipline that requires obtaining quantitative analog data from instrumented experiments.

Most operational amplifier circuits can be analyzed, at least for first-order calculations, by considering the op amp to be an "ideal" device. For more quantitative information, however, and particularly when frequency response and dc offsets are important, one must refer to a more "practical" model that includes the internal limitations of the device. If the op amp is characterized by a really complete model, the resulting circuit may be quite complex, leading to rather laborious calculations. Fortunately, however, computer analysis using the program **SPICE** significantly reduces the problem to one of a simple input specification to the computer. Today, nearly all the op amp manufacturers provide SPICE models for their line of devices, with excellent correlation obtained between the computer simulation and the actual measured results.

The Ideal Op Amp

An **ideal operational amplifier** is a dc-coupled amplifier having two inputs and normally one output (although in a few infrequent cases there may be a differential output). The inputs are designated as noninverting (designated + or NI) and inverting (designated – or Inv.). The amplified signal is the *differential* signal, v_e, between the two inputs, so that the output voltage as indicated in Fig. 27.1 is

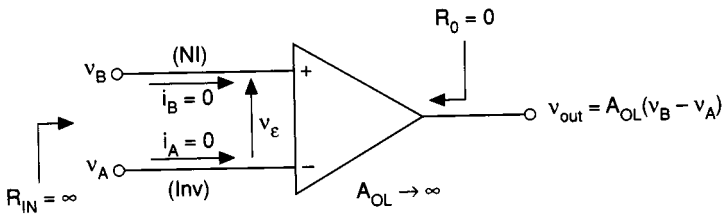


FIGURE 27.1 Configuration for an ideal op amp.

$$v_{out} = A_{OL}(v_B - v_A) \tag{27.1}$$

The general characteristics of an ideal op amp can be summarized as follows:

1. The open-loop gain A_{OL} is infinite. Or, since the output signal v_{out} is finite, then the differential input signal v_ε must approach zero.
2. The input resistance R_{IN} is infinite, while the output resistance R_O is zero.
3. The amplifier has zero current at the input (i_A and i_B in Fig. 27.1 are zero), but the op amp can either sink or source an infinite current at the output.
4. The op amp is not sensitive to a common signal on both inputs (i.e., $v_A = v_B$); thus, the output voltage change due to a common input signal will be zero. This common signal is referred to as a common-mode signal, and manufacturers specify this effect by an op amp's *common-mode rejection ratio* (CMRR), which relates the ratio of the open-loop gain (A_{OL}) of the op amp to the common-mode gain (A_{CM}). Hence, for an ideal op amp CMRR = ∞.
5. A somewhat analogous specification to the CMRR is the *power-supply rejection ratio* (PSRR), which relates the ratio of a power supply voltage change to an equivalent input voltage change produced by the change in the power supply. Because an ideal op amp can operate with any power supply, without restriction, then for the ideal device PSRR = ∞.
6. The gain of the op amp is not a function of frequency. This implies an infinite bandwidth.

Although the foregoing requirements for an ideal op amp appear to be impossible to achieve practically, modern devices can quite closely approximate many of these conditions. An op amp with a field-effect transistor (FET) on the input would certainly not have zero input current and infinite input resistance, but a current of <10 pA and an $R_{IN} = 10^{12}\ \Omega$ is obtainable and is a reasonable approximation to the ideal conditions. Further, although a CMRR and PSRR of infinity are not possible, there are several commercial op amps available with values of 140 dB (i.e., a ratio of 10^7). Open-loop gains of several precision op amps now have reached values of $>10^7$, although certainly not infinity. The two most difficult ideal conditions to approach are the ability to handle large output currents and the requirement of a gain independence with frequency.

Using the ideal model conditions it is quite simple to evaluate the two basic op amp circuit configurations, (1) the inverting amplifier and (2) the noninverting amplifier, as designated in Fig. 27.2.

For the ideal inverting amplifier, since the open-loop gain is infinite and since the output voltage v_o is finite, then the input differential voltage (often referred to as the *error signal*) v_ε must approach zero, or the input current is

$$i_I = \frac{v_I - v_\varepsilon}{R_1} = \frac{v_I - 0}{R_1} \tag{27.2}$$

The feedback current i_F must equal i_I, and the output voltage must then be due to the voltage drop across R_F, or

$$v_o = -i_F R_F + v_\varepsilon = -i_I R_F = -\left(\frac{R_F}{R_1}\right) v_I \tag{27.3}$$

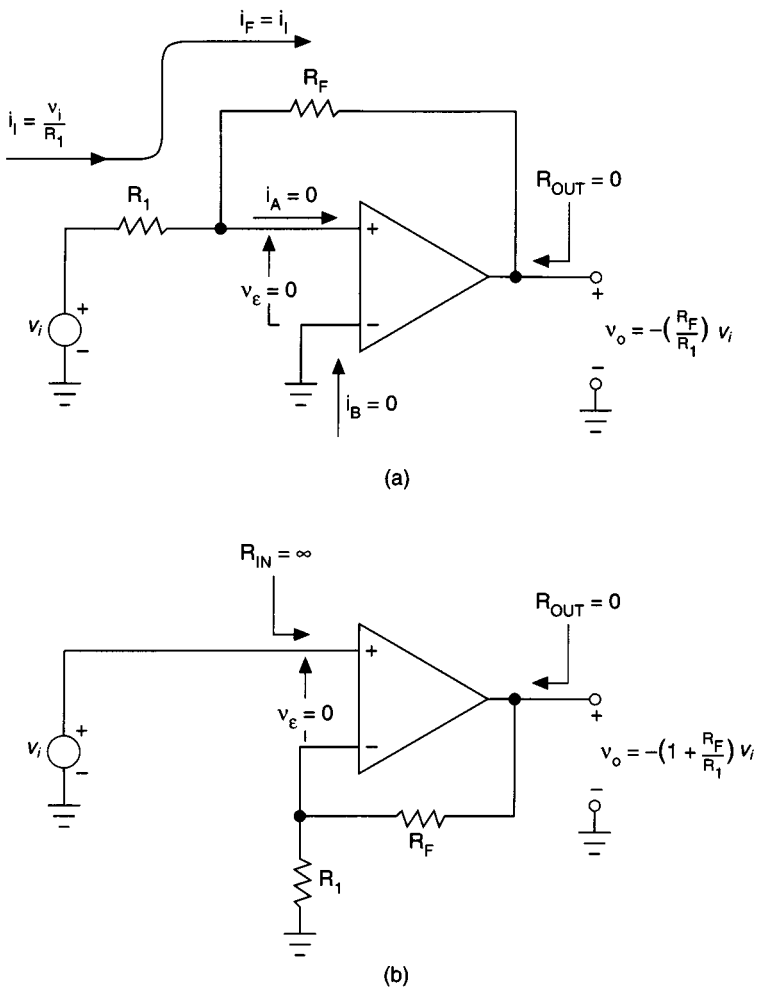


FIGURE 27.2 Illustration of (a) the inverting amplifier and (b) the noninverting amplifier. (*Source:* E.J. Kennedy, *Operational Amplifier Circuits, Theory and Applications,* New York: Holt, Rinehart and Winston, 1988, pp. 4, 6. With permission.)

The inverting connection thus has a voltage gain v_o/v_I of $-R_F/R_1$, an input resistance seen by v_I of R_1 ohms [from Eq. (27.2)], and an output resistance of 0 Ω. By a similar analysis for the noninverting circuit of Fig. 27.2(b), since v_ε is zero, then signal v_I must appear across resistor R_1, producing a current of v_I/R_1, which must flow through resistor R_F. Hence the output voltage is the sum of the voltage drops across R_F and R_1, or

$$v_o = R_F\left(\frac{v_I}{R_1}\right) + v_I = \left(1 + \frac{R_F}{R_1}\right)v_I \tag{27.4}$$

As opposed to the inverting connection, the input resistance seen by the source v_I is now equal to an infinite resistance, since R_{IN} for the ideal op amp is infinite.

Practical Op Amps

A nonideal op amp is characterized not only by finite open-loop gain, input and output resistance, finite currents, and frequency bandwidths, but also by various nonidealities due to the construction of the op amp circuit or external connections. A complete model for a practical op amp is illustrated in Fig. 27.3. The nonideal

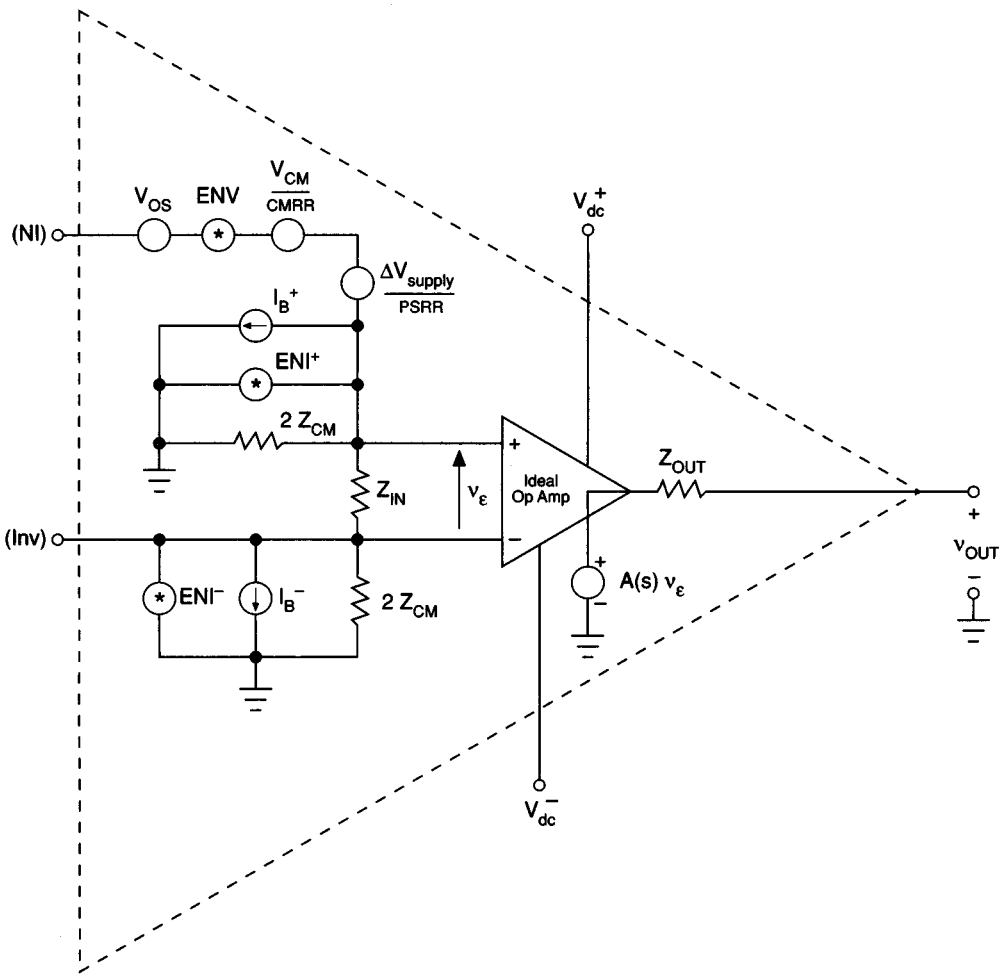


FIGURE 27.3 A model for a practical op amp illustrating nonideal effects. (*Source:* E.J. Kennedy, *Operational Amplifier Circuits, Theory and Applications,* New York: Holt, Rinehart and Winston, 1988, pp. 53, 126. With permission.)

effects of the PSRR and CMRR are represented by the input series voltage sources of ΔV_{supply}/PSRR and V_{CM}/CMRR, where ΔV_{supply} would be any total change of the two power supply voltages, V_{dc}^+ and V_{dc}^-, from their nominal values, while V_{CM} is the voltage common to both inputs of the op amp. The open-loop gain of the op amp is no longer infinite but is modeled by a network of the output impedance Z_{out} (which may be merely a resistor but could also be a series *R-L* network) in series with a source $A(s)$, which includes all the open-loop poles and zeroes of the op amp as

$$A(s) = \frac{A_{OL}\left(1 + \dfrac{s}{\omega_{Z_1}}\right)(1 + \cdots)}{\left(1 + \dfrac{s}{\omega_{p1}}\right)\left(1 + \dfrac{s}{\omega_{p2}}\right)(1 + \cdots)} \tag{27.5}$$

where A_{OL} is the finite dc open-loop gain, while poles are at frequencies ω_{p1}, ω_{p2}, . . . and zeroes are at ω_{Z1}, etc. The differential input resistance is Z_{IN}, which is typically a resistance R_{IN} in parallel with a capacitor C_{IN}. Similarly, the common-mode input impedance Z_{CM} is established by placing an impedance $2Z_{CM}$ in parallel

with each input terminal. Normally, Z_{CM} is best represented by a parallel resistance and capacitance of $2R_{CM}$ (which is >> R_{IN}) and $C_{CM}/2$. The dc bias currents at the input are represented by I_B^+ and I_B^- current sources that would equal the input base currents if a differential bipolar transistor were used as the input stage of the op amp, or the input gate currents if FETs were used. The fact that the two transistors of the input stage of the op amp may not be perfectly balanced is represented by an equivalent input *offset voltage* source, V_{OS}, in series with the input.

The smallest signal that can be amplified is always limited by the inherent random noise internal to the op amp itself. In Fig. 27.3 the noise effects are represented by an *equivalent input voltage source* (ENV), which when multiplied by the gain of the op amp would equal the total output noise present if the inputs to the op amp were shorted. In a similar fashion, if the inputs to the op amp were open circuited, the total output noise would equal the sum of the noise due to the *equivalent input current sources* (ENI^+ and ENI^-), each multiplied by their respective current gain to the output. Because noise is a random variable, this summation must be accomplished in a squared fashion, i.e.,

$$E_O^2\left(\text{rms volt}^2/\text{Hz}\right) = (\text{ENV})^2 A_v^2 + (\text{ENI}^+)^2 A_{I1}^2 + (\text{ENI}^-)^2 A_{12}^2 \tag{27.6}$$

Typically, the correlation (C) between the ENV and ENI sources is low, so the assumption of $C \approx 0$ can be made.

For the basic circuits of Fig. 27.2(a) or (b), if the signal source v_I is shorted then the output voltage due to the nonideal effects would be (using the model of Fig. 27.3)

$$v_o = \left(V_{OS} + \frac{V_{CM}}{\text{CMRR}} + \frac{\Delta V_{\text{supply}}}{\text{PSRR}}\right)\left(1 + \frac{R_F}{R_1}\right) + I_B^- R_F \tag{27.7}$$

provided that the loop gain (also called loop transmission in many texts) is related by the inequality

$$\left(\frac{R_1}{R_1 + R_F}\right) A(s) >> 1 \tag{27.8}$$

Inherent in Eq. (27.8) is the usual condition that $R_1 \ll Z_{IN}$ and Z_{CM}. If a resistor R_2 were in series with the noninverting input terminal, then a corresponding term must be added to the right hand side of Eq. (27.7) of value $-I_B^+ R_2 (R_1 + R_F)/R_1$. On manufacturers' data sheets the individual values of I_B^+ and I_B^- are not stated; instead the average input bias current and offset current are specified as

$$I_B = \frac{I_B^+ + I_B^-}{2}; \qquad I_{\text{offset}} = \left|I_B^+ - I_B^-\right| \tag{27.9}$$

The output noise effects can be obtained using the model of Fig. 27.3 along with the circuits of Fig. 27.2 as

$$\begin{aligned} E_{\text{out}}^2(\text{rms volts}^2/\text{Hz}) = {} & E_1^2\left(\frac{R_F}{R_1}\right)^2 + E_F^2 + (\text{ENV}^2 + E_2^2) \times \\ & \left(1 + \frac{R_F}{R_1}\right)^2 + (\text{ENI}^-)^2 R_F^2 + (\text{ENI}^+)^2 R_2^2 \left(1 + \frac{R_F}{R_1}\right)^2 \end{aligned} \tag{27.10}$$

where it is assumed that a resistor R_2 is also in series with the noninverting input of either Fig. 27.2(a) or (b). The thermal noise (often called Johnson or Nyquist noise) due to the resistors R_1, R_2, and R_F is given by (in rms volt^2/Hz)

$$E_1^2 = 4kT\,R_1$$
$$E_2^2 = 4kT\,R_2 \qquad (27.11)$$
$$E_F^2 = 4kT\,R_F$$

where k is Boltzmann's constant and T is absolute temperature (°Kelvin). To obtain the total output noise, one must multiply the E_{out}^2 expression of Eq. (27.10) by the noise bandwidth of the circuit, which typically is equal to $\pi/2$ times the –3 dB signal bandwidth, for a single-pole response system [Kennedy, 1988].

SPICE Computer Models

The use of op amps can be considerably simplified by computer-aided analysis using the program SPICE. SPICE originated with the University of California, Berkeley, in 1975 [Nagel, 1975], although more recent user-friendly commercial versions are now available such as HSPICE, HPSPICE, IS-SPICE, PSPICE, and ZSPICE, to mention a few of those most widely used. A simple macromodel for a near-ideal op amp could be simply stated with the SPICE subcircuit file (* indicates a comment that is not processed by the file)

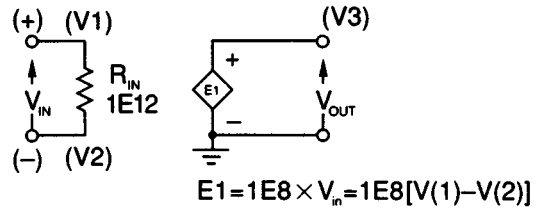
```
.SUBCKT IDEALOA 1 2 3
*A near-ideal op amp: (1) is noninv, (2) is inv, and (3) is output.
RIN 1 2 1E12
E1 (3, 0) (1, 2) 1E8
.ENDS IDEALOA
```
(27.12)

The circuit model for IDEALOA would appear as in Fig. 27.4(a). A more complete model, but not including nonideal offset effects, could be constructed for the 741 op amp as the subcircuit file OA741, shown in Fig. 27.4(b).

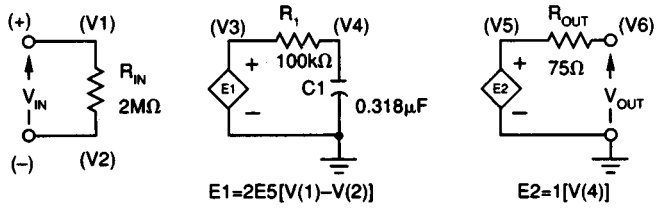
```
.SUBCKT OA741 1 2 6
*A linear model for the 741 op amp: (1) is noninv, (2) is inv, and
*(6) is output. RIN = 2MEG, AOL = 200,000, ROUT = 75 ohm,
*Dominant open - loop pole at 5 Hz, gain - bandwidth product
*is 1 MHz.
RIN 1 2 2MEG
E1 (3, 0) (1, 2) 2E5
R1 3 4 100K
C1 4 0 0.318UF ; R1 × C1 = 5HZPOLE
E2 (5, 0) (4, 0) 1.0
ROUT 5 6 75
.ENDS OA741
```
(27.13)

The most widely used op amp macromodel that includes dc offset effects is the **Boyle model** [Boyle et al., 1974]. Most op amp manufacturers use this model, usually with additions to add more poles (and perhaps zeroes). The various resistor and capacitor values, as well as transistor, and current and voltage generator, values are intimately related to the specifications of the op amp, as shown earlier in the nonideal model of Fig. 27.3. The appropriate equations are too involved to list here; instead, the interested reader is referred to the article by Boyle in the listed references. The Boyle model does not accurately model noise effects, nor does it fully model PSRR and CMRR effects.

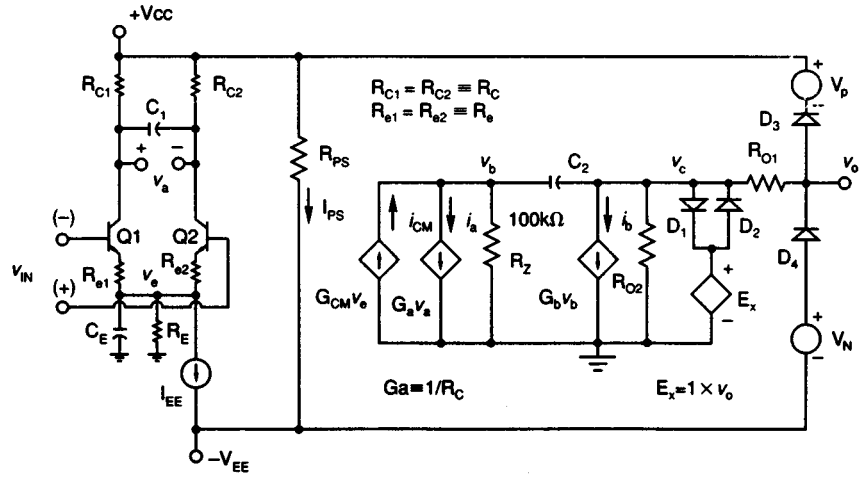
A more circuits-oriented approach to modeling op amps can be obtained if the input transistors are removed and a model formed by using passive components along with both fixed and dependent voltage and current sources. Such a model is shown in Fig. 27.5. This model not only includes all the basic nonideal effects of the op amp, allowing for multiple poles and zeroes, but can also accurately include ENV and ENI noise effects.



(a)



(b)



(c)

FIGURE 27.4 Some simple SPICE macromodels. (a) A near ideal op amp. (b) A linear model for a 741 op amp. (c) The *Boyle* macromodel.

The circuits-approach macromodel can also be easily adapted to current-feedback op amp designs, whose input impedance at the noninverting input is much greater than that at the inverting input [see Williams, 1991]. The interested reader is referred to the text edited by J. Williams, listed in the references, as well as the SPICE modeling book by Connelly and Choi [1992].

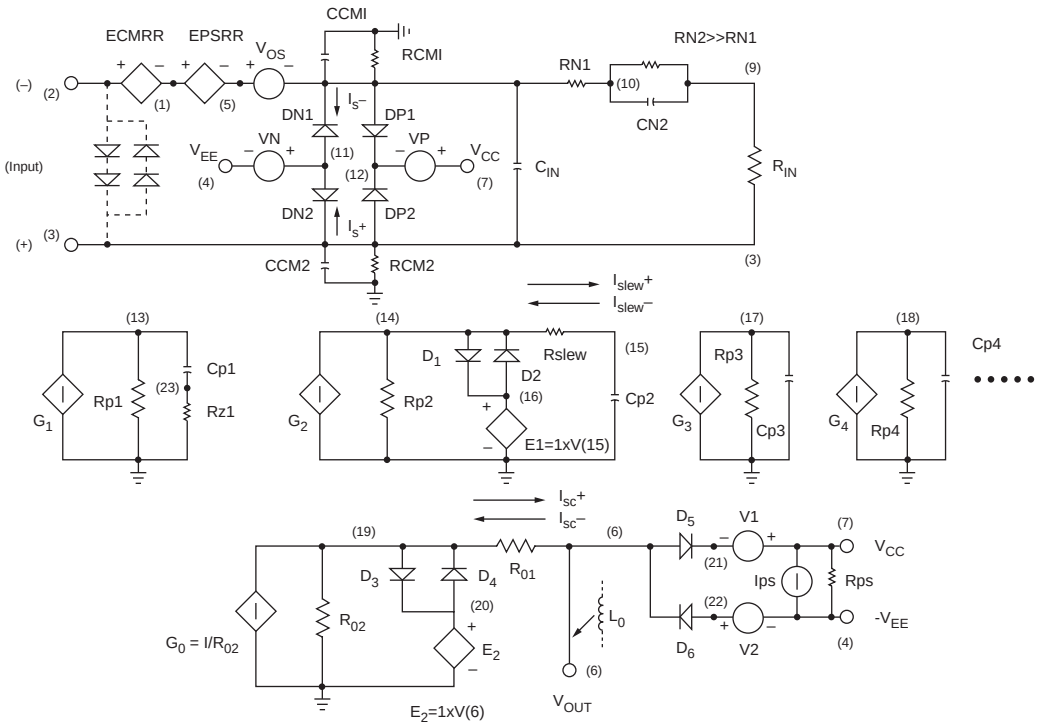


FIGURE 27.5 A SPICE circuits-approach macromodel.

A comparison of the SPICE macromodels with actual manufacturer's data for the case of an LM318 op amp is demonstrated in Fig. 27.6, for the open-loop gain versus frequency specification.

Defining Terms

Boyle macromodel: A SPICE computer model for an op amp. Developed by G.R. Boyle in 1974.

Equivalent noise current (ENI): A noise current source that is effectively in parallel with either the noninverting input terminal (ENI^+) or the inverting input terminal (ENI^-) and represents the total noise contributed by the op amp if either input terminal is open circuited.

Equivalent noise voltage (ENV): A noise voltage source that is effectively in series with either the inverting or noninverting input terminal of the op amp and represents the total noise contributed by the op amp if the inputs were shorted.

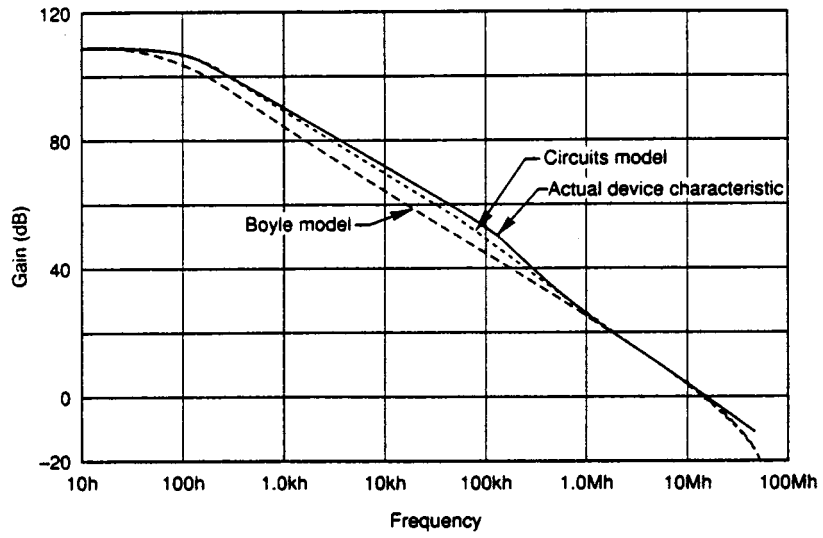
Ideal operational amplifier: An op amp having infinite gain from input to output, with infinite input resistance and zero output resistance and insensitive to the frequency of the signal. An ideal op amp is useful in first-order analysis of circuits.

Operational amplifier (op amp): A dc amplifier having both an inverting and noninverting input and normally one output, with a very large gain from input to output.

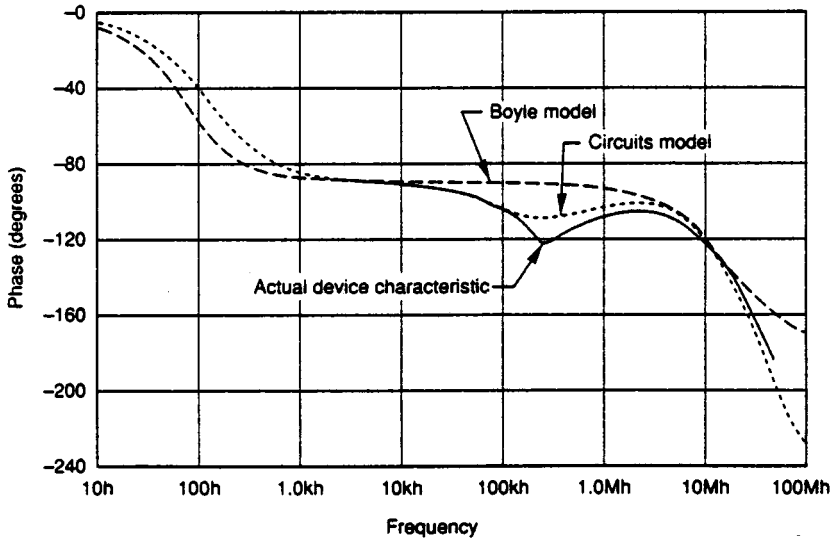
SPICE: A computer simulation program developed by the University of California, Berkeley, in 1975. Versions are available from several companies. The program is particularly advantageous for electronic circuit analysis, since dc, ac, transient, noise, and statistical analysis is possible.

Related Topic

13.1 Analog Circuit Simulation



(a)



(b)

FIGURE 27.6 Comparison between manufacturer's data and the SPICE macromodels.

References

G.R. Boyle et al., "Macromodeling of integrated circuit operational amplifiers," *IEEE J. S. S. Circuits,* pp. 353–363, 1974.

J.A. Connelly and P. Choi, *Macromodeling with SPICE,* Englewood Cliffs, N.J.: Prentice-Hall, 1992.

E.J. Kennedy, *Operational Amplifier Circuits, Theory and Applications,* New York: Holt, Rinehart and Winston, 1988.
L.W. Nagel, *SPICE 2: A Computer Program to Simulate Semiconductor Circuits,* ERL-M520, University of California, Berkeley, 1975.
J. Williams (ed.), *Analog Circuit Design,* Boston: Butterworth-Heinemann, 1991.

27.2 Applications

John V. Wait

In microminiature form (epoxy or metal packages or as part of a VLSI mask layout) the **operational amplifier** (op amp) is usually fabricated in integrated circuit (IC) form. The general environment is shown in Fig. 27.7. A pair of + and – regulated power supplies (or batteries) may supply all of the op amp in a system, typically with $\pm 10 - \pm 15$ V. The ground and power supply buses are usually assumed, and an individual op-amp symbol is shown in Fig. 27.8. Such amplifiers feature:

1. A high voltage gain, down to and including dc, and a dc open loop gain of perhaps 10^5 (100 dB) or more
2. An inverting (–) and noninverting (+) symbol
3. Minimized dc offsets, a high input impedance, and a low output impedance
4. An output stage able to deliver or absorb currents over a dynamic range approaching the power supply voltages

It is important *never* to use the op amp without feedback between the output and inverting terminals at all frequencies. A simple inverting amplifier is shown in Fig. 27.9. Here the voltage gain is

$$V_{\text{out}} / V_{\text{in}} = -K = -R_F / R_1$$

The circuit gain is determined essentially by the external resistances, within the bandwidth and output-driving capabilities of the op amp (more later). If $R_F = R_1 = R$, we have the simple *unity gain inverter* of Fig. 27.10.

Figure 27.11 shows a more flexible *summer-inverter* circuit with

$$v_0 = -(K_1 v_1 + K_2 v_2 + \ldots + K_n v_n)$$

where $K_i = R_F / R_i$.

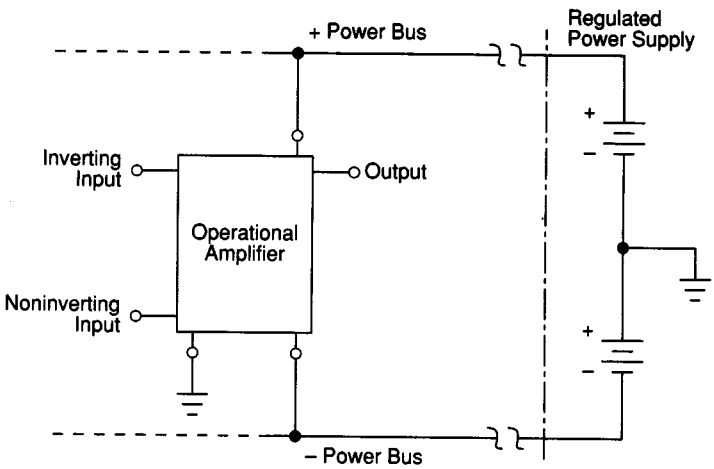


FIGURE 27.7 Typical operational amplifier environment.

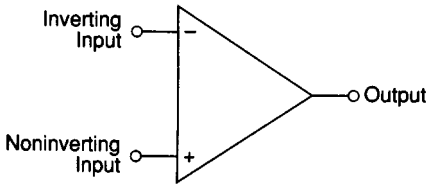


FIGURE 27.8 Conventional operational amplifier symbol. Only active signal lines are shown, and all signals are referenced to ground.

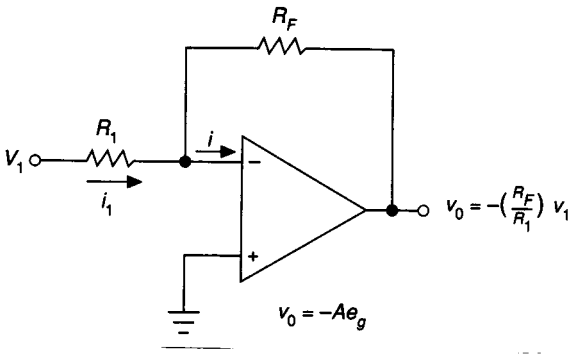


FIGURE 27.9 Simple resistive inverter-amplifier.

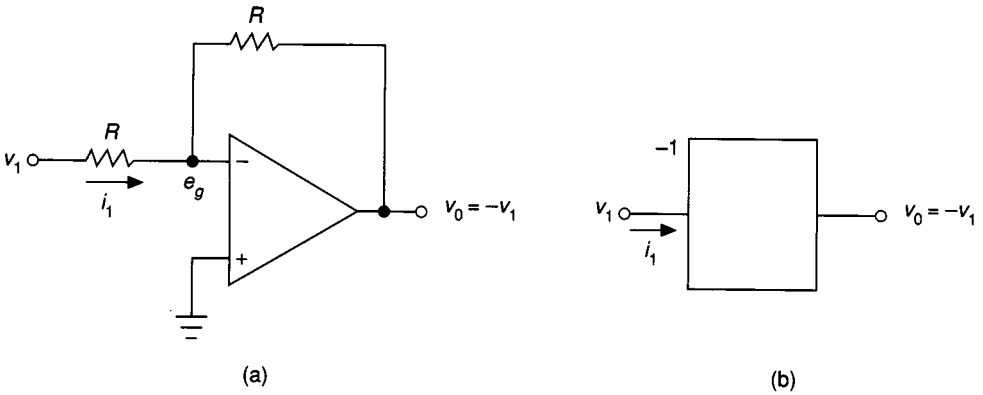


FIGURE 27.10 A simple unity gain inverter, showing (a) detailed circuit; (b) block-diagram symbol.

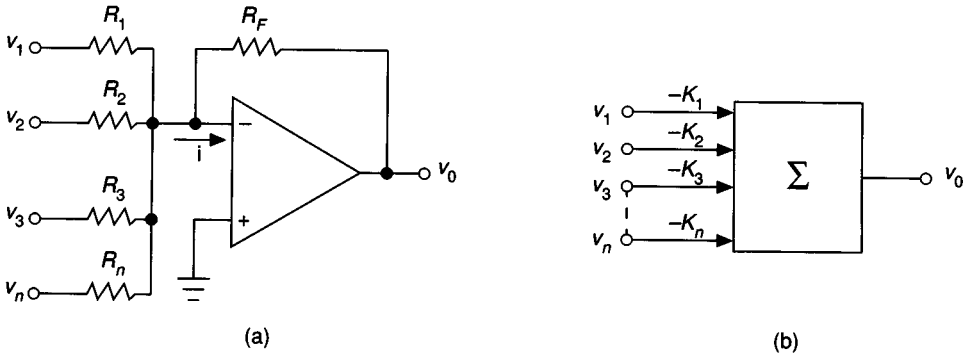


FIGURE 27.11 The summer-inverter circuit, showing (a) complete circuit; (b) block-diagram symbol.

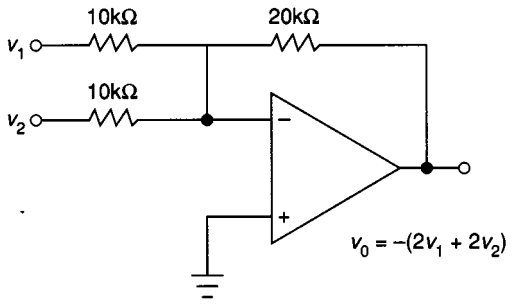


FIGURE 27.12 Simple summer-inverter.

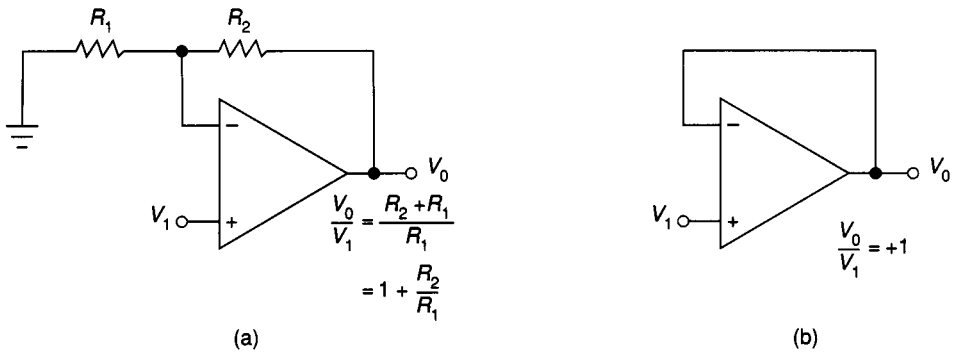


FIGURE 27.13 Noninverting amplifier circuit with resistive elements. (a) General circuit; (b) simple unity gain follower.

The summer-inverter is generally useful for precisely combining or mixing signals, e.g., summing and inverting. The signal levels must be appropriately limited but may generally be *bipolar* (+/–).

The resistance values should be in a proper range since (a) too low resistance values draw excessive current from the signal source, and (b) too high resistance values make the circuit performance too sensitive to stray capacitances and dc offset effects.

Typical values are from 1 MΩ and 10 kΩ. The circuit of Fig. 27.12 shows a circuit to implement

$$v_0 = -4v_1 - 2v_2$$

Noninverting Circuits

Figure 27.13(a) shows the useful *noninverting* amplifier circuit. It has a voltage gain

$$V_0/V_1 = (R_2 + R_1)/R_1$$
$$= 1 + (R_2/R_1)$$

Figure 27.13(b) shows the important unity gain follower circuit, which has a *very high input impedance,* which lightly loads the signal source but which can provide a reasonable amount of output current milliamps.

It is fairly easy to show that the *inverting first-order low-pass filter* of Fig. 27.14 has a dc gain or $-R_2/R_1$ and a –3-dB frequency = $1/(2\pi R_2 C)$.

Figure 27.15 shows a two-amplifier differentiator and high-pass filter circuit with a resistive input impedance and a low-frequency cutoff determined by R_1 and C.

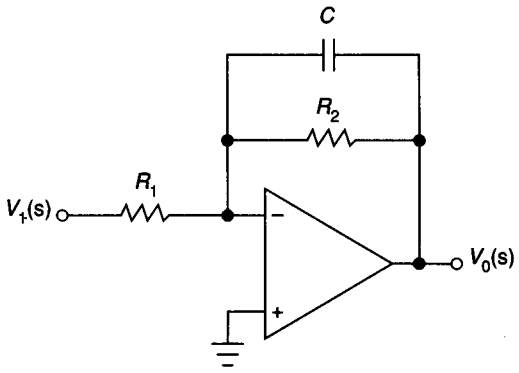


FIGURE 27.14 First-order low-pass filter circuit.

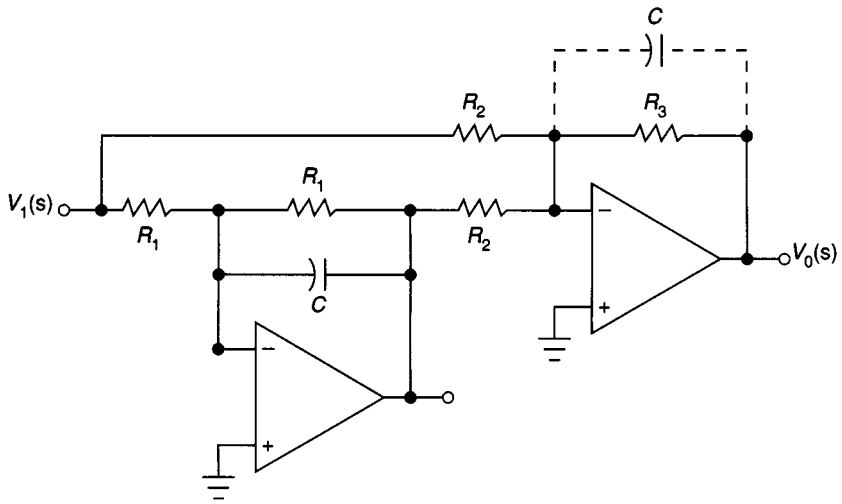


FIGURE 27.15 A two-amplifier high-pass circuit.

Op amps provide good differential amplifier circuits. Figure 27.16 is a single amplifier circuit with a differential gain

$$A_d = R_0/R_1$$

Good resistance matching is required to have good common-mode rejection of unwanted common-mode signals (static, 60-Hz hum, etc.). The one-amplifier circuit of Fig. 27.16 has a differential input impedance of $2R_1$. R_1 may be chosen to provide a good load for a microphone, phono-pickup, etc.

The improved three-amplifier instrumentation amplifier circuit of Fig. 27.17, which several manufacturers provide in a single module, provides

1. Very high voltage gain
2. Good common-mode rejection
3. A differential gain
4. High input impedance

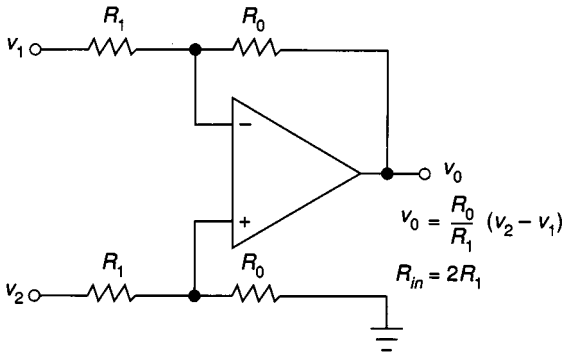


FIGURE 27.16 Single-output differential-input amplifier circuit.

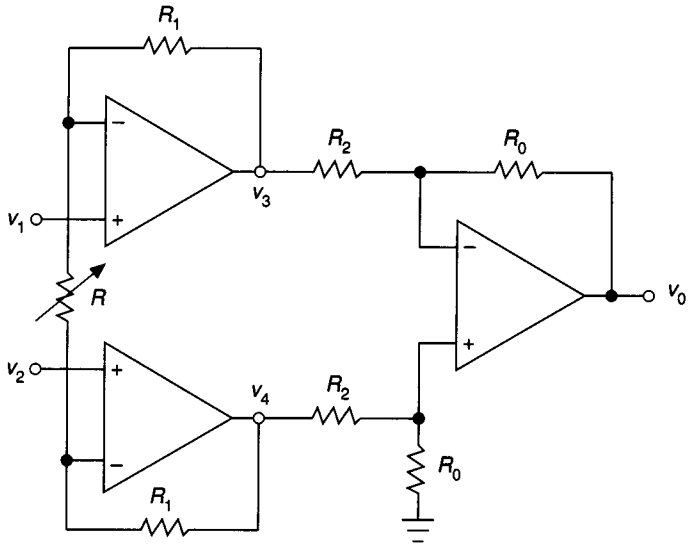


FIGURE 27.17 A three-amplifier differential-input instrumentation amplifier featuring high input impedance and easily adjustable gain.

$$A_d = -\frac{R_0}{R_2}\left(1 + \frac{2R_1}{R}\right)(V_2 - V_1)$$

Operational amplifier circuits form the heart of many precision circuits, e.g., regulated power supplies, precision comparators, peak-detection circuits, and waveform generators [Wait et al., 1992]. Another important area of application is **active RC filters** [Huelsman and Allen, 1980]. Microminiature electronic circuits seldom use inductors. Through the use of op amps, resistors, and capacitors, one can implement precise filter circuits (low-pass, high-pass, and bandpass). Figures 27.18 and 27.19 show second-order low-pass and bandpass filter circuits that feature relatively low sensitivity of filter performance to component values. Details are provided in Wait et al. [1992] and Huelsman and Allen [1980].

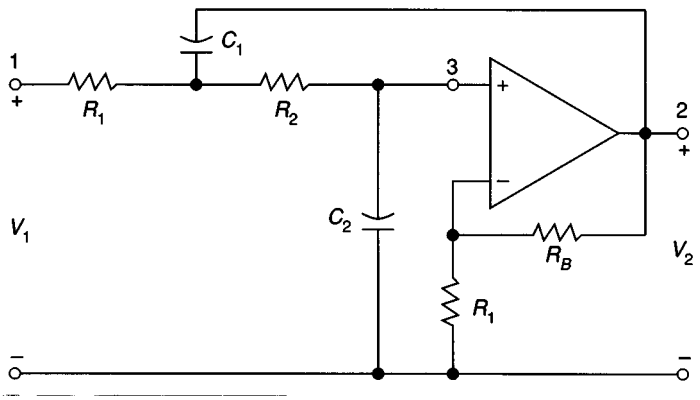


FIGURE 27.18 Sallen and Key low-pass filter.

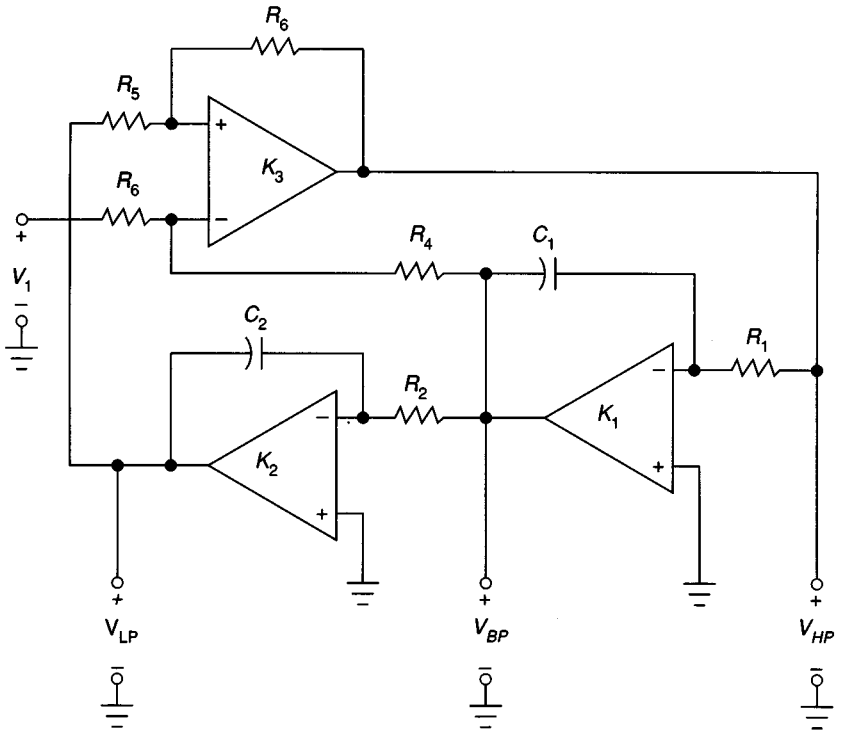


FIGURE 27.19 State-variable filter.

Of course, the op amp does not have infinite bandwidth and gain. An important op-amp parameter is the unity-gain frequency, f_u. For example, it is fairly easy to show the actual bandwidth of a constant gain amplifier of nominal gain G is approximately

$$f_{-3\text{ dB}} = f_u/G$$

Thus, an op amp with f_u = 1 MHz will provide an amplifier gain of 20 up to about 50 kHz.

When a circuit designer needs to accurately explore the performance of an op-amp circuit design, modern circuit simulation programs (SPICE, PSPICE, and MICRO-CAP) permit a thorough study of circuit design, as related to op-amp performance parameters. We have not here treated nonlinear op-amp performance limitations such as *slew rate, full-power bandwidth,* and *rated output.* Surely, the op-amp circuit designer must be careful not to exceed the output rating of the op amp, as related to *maximum output voltage* and *current* and *output rate-of-change.*

Nevertheless, op-amp circuits provide the circuit designer with a handy and straightforward way to complete electronic system designs with the use of only a few basic circuit components plus, of course, the operational amplifier.

Defining Terms

Active RC filter: An electronic circuit made up of resistors, capacitors, and operational amplifiers that provide well-controlled linear frequency-dependent functions, e.g., low-, high-, and bandpass filters.

Analog-to-digital converter (ADC): An electronic circuit that receives a magnitude-scaled analog voltage and generates a binary-coded number proportional to the analog input, which is delivered to an interface subsystem to a digital computer.

Digital-to-analog converter (DAC): An electronic circuit that receives an *n*-bit digital word from an interface circuit and generates an analog voltage proportional to it.

Electronic switch: An electronic circuit that controls analog signals with digital (binary) signals.

Interface: A collection of electronic modules that provide data transfer between analog and digital systems.

Operational amplifier: A small (usually integrated circuit) electronic module with a bipolar (+/–) output terminal and a pair of differential input terminals. It is provided with power and external components, e.g., resistors, capacitors, and semiconductors, to make amplifiers, filters, and wave-shaping circuits with well-controlled performance characteristics, relatively immune to environmental effects.

Related Topic

29.1 Synthesis of Low-Pass Forms

References

Electronic Design, Hasbrook Heights, N.J.: Hayden Publishing Co.; a biweekly journal for electronics engineers. (In particular, see the articles in the Technology section.)

Electronics, New York: McGraw-Hill; a biweekly journal for electronic engineers. (In particular, see the circuit design features.)

J.G. Graeme, *Applications of Operational Amplifiers,* New York: McGraw-Hill, 1973.

L.P. Huelsman, and P.E. Allen, *Introduction to the Theory and Design of Active Filters.* New York: McGraw-Hill, 1980.

J. Till, "Flexible Op-Amp Model Improves SPICE," *Electronic Design,* June 22, 1989.

G.E. Tobey, J.G. Graeme, and L.P. Huelsman, *Operational Amplifiers,* New York: McGraw-Hill, 1971.

J.V. Wait, L.P. Huelsman, and G.A. Korn, *Introduction to Operational Amplifier Theory and Applications,* 2nd ed., New York: McGraw-Hill, 1992.

Further Information

For further information see J.V. Wait, L.P. Huelsman, and G.A. Korn, *Introduction to OperationalAmplifier Theory and Applications,* 2nd ed., New York: McGraw-Hill, 1992, a general textbook on the design of operational amplifier circuits, including the SPICE model of operational amplifiers; and L.P. Huelsman and P.E. Allen, *Introduction to the Theory and Design of Active Filters,* New York: McGraw-Hill, 1980, a general textbook of design considerations and configurations of active RC filters.

Carpenter, G.L., Choma, Jr., J. “Amplifiers”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

28

Amplifiers

28.1 Large Signal Analysis
DC Operating Point • Graphical Approach • Power Amplifiers
28.2 Small Signal Analysis
Hybrid-Pi Equivalent Circuit • Hybrid-Pi Equivalent Circuit of a Monolithic BJT • Common Emitter Amplifier • Design Considerations for the Common Emitter Amplifier • Common Base Amplifier • Design Considerations for the Common Base Amplifier • Common Collector Amplifier

Gordon L. Carpenter
California State University, Long Beach

John Choma, Jr.
University of Southern California

28.1 Large Signal Analysis

Gordon L. Carpenter

Large signal amplifiers are usually confined to using bipolar transistors as their solid state devices because of the large linear region of amplification required. One exception to this is the use of VMOS for large power outputs due to their ability to have a large linear region. There are three basic configurations of amplifiers: common emitter (CE) amplifiers, common base (CB) amplifiers, and common collection(CC) amplifiers. The basic configuration of each is shown in Fig. 28.1.

In an amplifier system, the last stage of a voltage amplifier string has to be considered as a large signal amplifier, and generally EF amplifiers are used as large signal amplifiers. This then requires that the dc bias or dc operating point (quiescent point) be located near the center of the load line in order to get the maximum output voltage swing. Small signal analysis can be used to evaluate the amplifier for voltage gain, current gain, input impedance, and output impedance, all of which are discussed later.

DC Operating Point

Each transistor connected in a particular amplifier configuration has a set of characteristic curves, as shown in Fig. 28.2.

When amplifiers are coupled together with capacitors, the configuration is as shown in Fig. 28.3. The load resistor is really the input impedance of the next stage. To be able to evaluate this amplifier, a dc equivalent circuit needs to be developed as shown in Fig. 28.4. This will result in the following dc bias equation:

$$I_{CQ} = \frac{V_{BB} - V_{BE}}{R_B/\text{beta} + R_E} \qquad \text{Assume } h_{FE} >> 1$$

where beta (h_{FE}) is the current gain of the transistor and V_{BE} is the conducting voltage across the base-emitter junction. This equation is the same for all amplifier configurations. Looking at Fig. 28.3, the input circuit can be reduced to the dc circuit shown in Fig. 28.4 using circuit analysis techniques, resulting in the following equations:

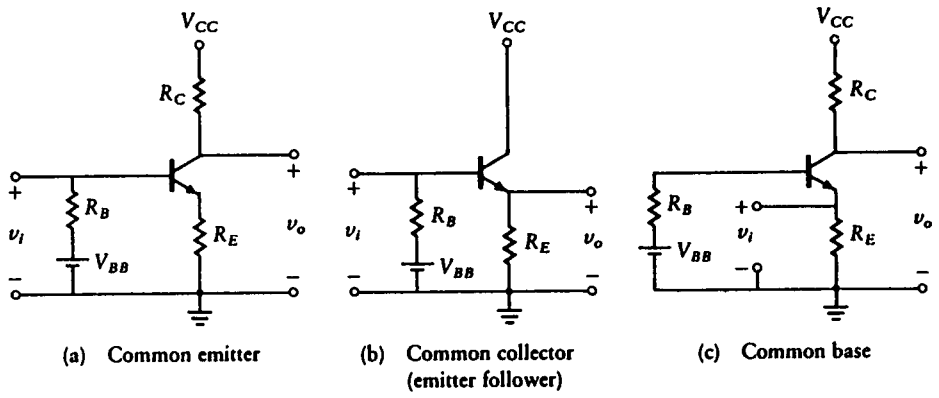


FIGURE 28.1 Amplifier circuits. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 80. With permission.)

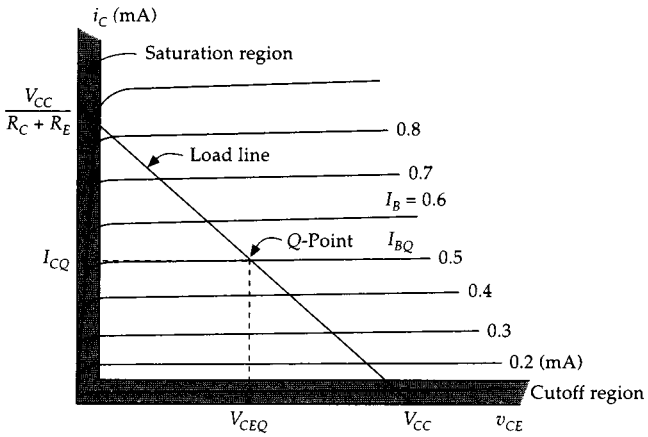


FIGURE 28.2 Transistor characteristic curves. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 82. With permission.)

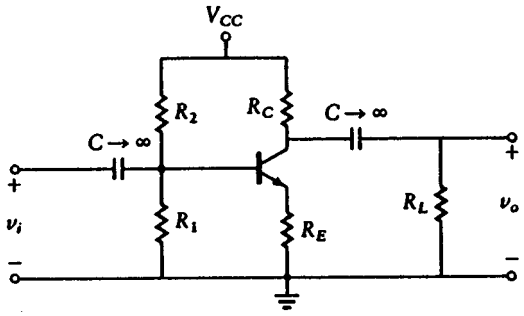


FIGURE 28.3 Amplifier circuit. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 92. With permission.)

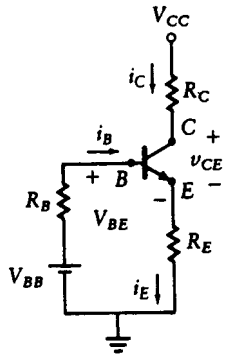


FIGURE 28.4 Amplifier equivalent circuit. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 82. With permission.)

$$V_{BB} = V_{TH} = V_{CC}(R_1)/(R_1 + R_2)$$

$$R_B = R_{TH} = R_1//R_2$$

For this biasing system, the Thévenin equivalent resistance and the Thévenin equivalent voltage can be determined. For design with the biasing system shown in Fig. 28.3, then:

$$R_1 = R_B/(1 - V_{BB}/V_{CC})$$

$$\mathrm{R_2 = R_B\ (V_{CC}/V_{BB})}$$

Graphical Approach

To understand the graphical approach, a clear understanding of the dc and ac load lines is necessary. The dc load line is based on the Kirchhoff's equation from the dc power source to ground (all capacitors open)

$$V_{CC} = v_{CE} + i_C R_{DC}$$

where R_{DC} is the sum of the resistors in the collector-emitter loop.

The ac load line is the loop, assuming the transistor is the ac source and the source voltage is zero, then

$$V'_{CC} = v_{ce} + i_C R_{ac}$$

where R_{ac} is the sum of series resistors in that loop with all the capacitors shorted. The load lines then can be constructed on the characteristic curves as shown in Fig. 28.5. From this it can be seen that to get the maximum output voltage swing, the quiescent point, or Q point, should be located in the middle of the ac load line. To place the Q point in the middle of the ac load line, I_{CQ} can be determined from the equation

$$I_{CQ} = V_{CC}/(R_{DC} + R_{ac})$$

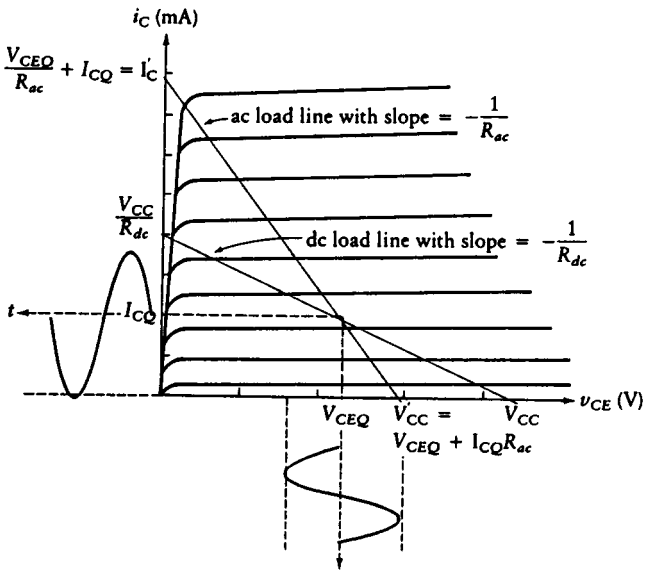


FIGURE 28.5 Load lines. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 94. With permission.)

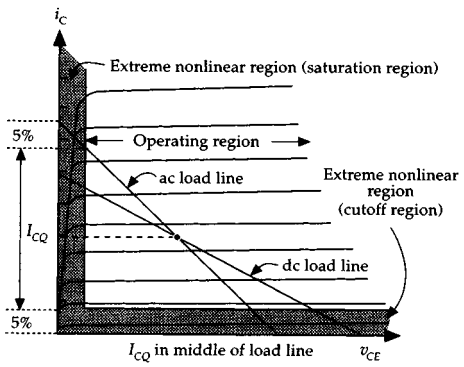


FIGURE 28.6 *Q* point in middle of load line. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 135. With permission.)

To minimize distortion caused by the cutoff and saturation regions, the top 5% and the bottom 5% are discarded. This then results in the equation (Fig. 28.6):

$$V_o \text{ (peak to peak)} = 2\ (0.9)\ I_{CQ}\ (R_C//R_L)$$

If, however, the *Q* point is not in the middle of the ac load line, the output voltage swing will be reduced. Below the middle of the ac load line [Fig. 28.7(a)]:

$$V_o \text{ (peak to peak)} = 2\ (I_{CQ} - 0.05\ I_{C\text{Max}})\ R_C//R_L$$

Above the middle of the ac load line [Fig. 28.7(b)]:

$$V_o \text{ (peak to peak)} = 2\ (0.95\ I_{C\text{Max}} - I_{CQ})\ R_C//R_L$$

These values allow the highest allowable input signal to be used to avoid any distortion by dividing the voltage gain of the amplifier into the maximum output voltage swing. The preceding equations are the same for the CB configuration. For the EF configurations, the R_C is changed to R_E in the equations.

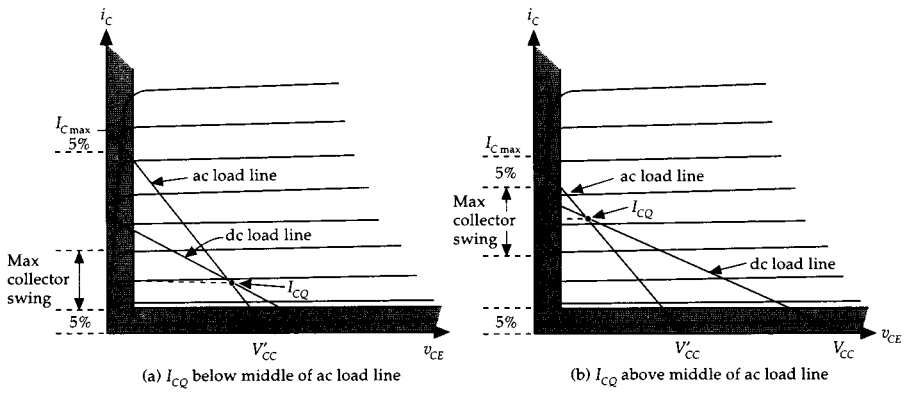


FIGURE 28.7 Reduced output voltage swing. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 136. With permission.)

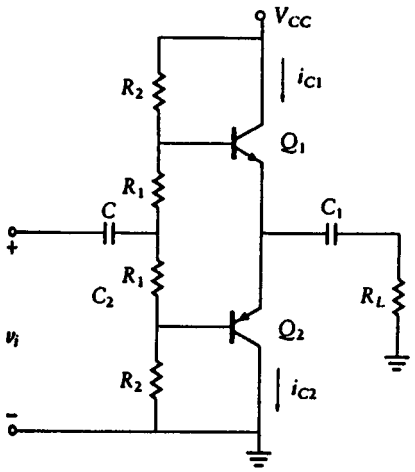


FIGURE 28.8 Complementary symmetry power amplifier. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 248. With permission.)

Power Amplifiers

Emitter followers can be used as power amplifiers. Even though they have less than unity voltage gain they can provide high current gain. Using the standard linear EF amplifier for a maximum output voltage swing provides less than 25% efficiency (ratio of power in to power out). The dc current carrying the ac signal is where the loss of efficiency occurs. To avoid this power loss, the Q point is placed at I_{CQ} equal to zero, thus using the majority of the power for the output signal. This allows the efficiency to increase to as much as 70%. Full signal amplification requires one transistor to amplify the positive portion of the input signal and another transistor to amplify the negative portion of the input signal. In the past, this was referred to as push-pull operation. A better system is to use an NPN transistor for the positive part of the input signal and a PNP transistor for the negative part. This type of operation is referred to as Class B complementary symmetry operation (Fig. 28.8).

In Fig. 28.8, the dc voltage drop across R_1 provides the voltage to bias the transistor at cutoff. Because these are power transistors, the temperature will change based on the amount of power the transistor is absorbing. This means the base-emitter junction voltage will have to change to keep $I_{CQ} = 0$. To compensate for this change in temperature, the R_1 resistors are replaced with diodes or transistors connected as diodes with the same turn-on characteristics as the power transistors. This type of configuration is referred to as the complementary symmetry diode compensated (CSDC) amplifier and is shown in Fig. 28.9. To avoid crossover distortion, small resistors can be placed in series with the diodes so that I_{CQ} can be raised slightly above zero to get increased amplification in the cutoff region. Another problem that needs to be addressed is the possibility of thermal runaway. This can be easily solved by placing small resistors in series with the emitters of the power transistors. For example, if the load is an 8-Ω speaker, the resistors should not be greater than 0.47 Ω to avoid output signal loss.

To design this type of amplifier, the dc current in the bias circuit must be large enough so that the diodes remain on during the entire input signal. This requires the dc diode current to be equal to or larger than the zero to peak current of the input signal, or

$$I_D \geq I_{ac}\,(0 \text{ to peak})$$

$$(V_{CC}/2 - V_{BE})/R_2 = I_B\,(0 \text{ to peak}) + V_L\,(0 \text{ to peak})/R_2$$

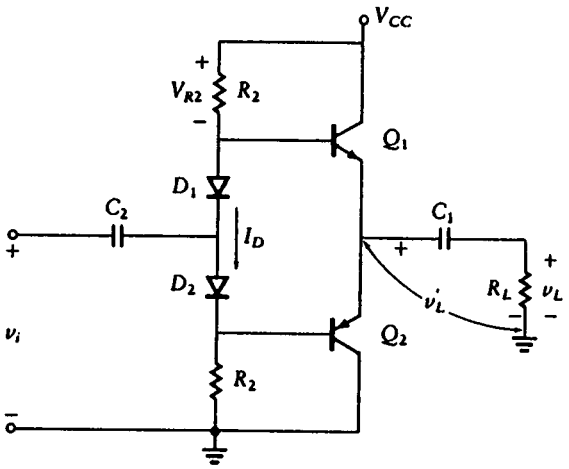


FIGURE 28.9 Complimentary symmetry diode compensated power amplifier. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 251. With permission.)

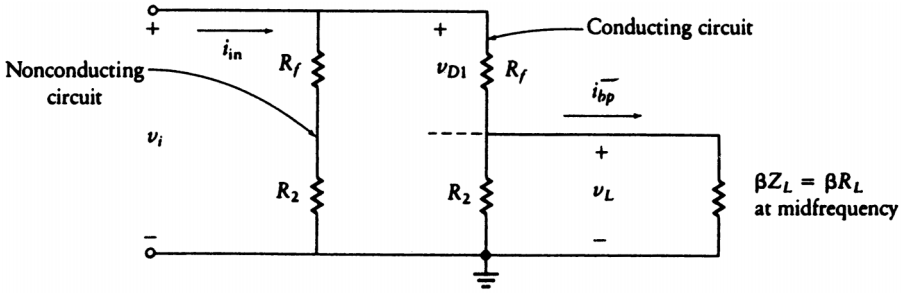


FIGURE 28.10 AC equivalent circuit of the CSDC amplifier. (*Source:* C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991, p. 255. With permission.)

When designing to a specific power, both I_B and V_L can be determined. This allows the selection of the value of R_2 and the equivalent circuit shown in Fig. 28.10 can be developed. Using this equivalent circuit, both the input resistance and the current gain can be shown. R_f is the forward resistance of the diodes.

$$R_{\text{in}} = (R_f + R_2)//[R_f + (R_2//\text{Beta}\, R_L)]$$

$$P_o = I_{\text{Cmax}} R_L/2$$

The power rating of the transistors to be used in this circuit should be greater than

$$P_{\text{rating}} = V_{CC}^2/(4\text{Pi}^2 R_L)$$

$$\text{C}_1 = 1/(2\text{Pi}\, f_{\text{low}} R_L)$$

$$C_2 = 10/[2\text{Pi}\, f_{\text{low}}(R_{\text{in}} + R_i)]$$

where R_i is the output impedance of the previous stage and f_{low} is the desired low frequency cutoff of the amplifier.

Related Topics

24.1 Junction Field-Effect Transistors • 30.1 Power Semiconductor Devices

References

P.R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits,* New York: Wiley, 1984.

J. Millman and A. Grabel, *Microelectronics,* New York: McGraw-Hill, 1987.

P.O. Neudorfer and M. Hassul, *Introduction to Circuit Analysis,* Needham Heights, Mass.: Allyn and Bacon, 1990.

C.J. Savant, M. Roden, and G. Carpenter, *Electronic Design, Circuits and Systems,* 2nd ed., Redwood City, Calif.: Benjamin-Cummings, 1991.

D.L. Schilling and C. Belove, *Electronic Circuits,* New York: McGraw-Hill, 1989.

28.2 Small Signal Analysis

John Choma, Jr.

This section introduces the reader to the analytical methodologies that underlie the design of small signal, analog bipolar junction transistor (BJT) amplifiers. Analog circuit and system design entails complementing basic circuit analysis skills with the art of architecting a circuit topology that produces acceptable input-to-output (I/O) electrical characteristics. Because design is not the inverse of analysis, analytically proficient engineers are not necessarily adept at design. However, circuit and system analyses that conduce an insightful understanding of meaningful topological structures arguably foster design creativity. Accordingly, this section focuses more on the problems of interpreting analytical results in terms of their circuit performance implications than it does on enhancing basic circuit analysis skills. Insightful interpretation breeds engineering understanding. In turn, such an understanding of the electrical properties of circuits promotes topological refinements and innovations that produce reliable and manufacturable, high performance electronic circuits and systems.

Hybrid-Pi Equivalent Circuit

In order for a BJT to function properly in linear amplifier applications, it must operate in the forward active region of its volt–ampere characteristic curves. Two conditions ensure BJT operation in the forward domain. First, the applied emitter-base terminal voltage must forward bias the intrinsic emitter-base junction diode at all times. Second, the instantaneous voltage established across the base-collector terminals of the transistor must preclude a forward biased intrinsic base-collector diode. The simultaneous satisfaction of these two conditions requires appropriate biasing subcircuits, and it imposes restrictions on the amplitudes of applied input signals [Clarke and Hess, 1978].

The most commonly used BJT equivalent circuit for investigating the dynamical responses to small input signals is the **hybrid-pi model** offered in Fig. 28.11 [Sedra and Smith, 1987]. In this model, R_b, R_c, and R_e respectively, represent the *internal base, collector,* and *emitter resistances* of the considered BJT. Although these series resistances vary somewhat with quiescent operating point [de Graaf, 1969], they can be viewed as constants in first-order manual analyses.

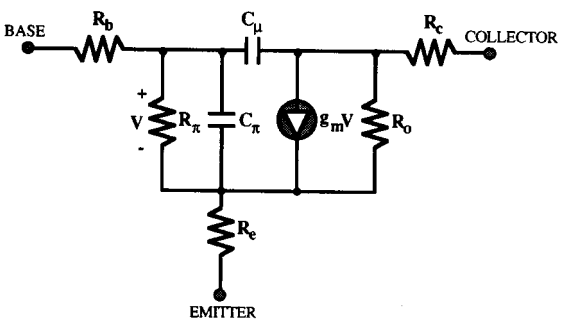


FIGURE 28.11 The small signal equivalent circuit (hybrid-pi model) of a bipolar junction transistor.

The *emitter-base junction diffusion resistance,* R_π, is the small signal resistance of the emitter-base junction diode. It represents the inverse of the slope of the common emitter static input characteristic curves. Analytically, R_π is given by

$$R_\pi = \frac{h_{FE} N_F V_T}{I_{CQ}} \tag{28.1}$$

where h_{FE} is the *static common emitter current gain* of the BJT, N_F is the *emitter-base junction injection coefficient*, V_T is the *Boltzmann voltage corresponding to an absolute junction operating temperature of T*, and I_{CQ} is the *quiescent collector current.*

The expression for the resistance, R_o, which accounts for *conductivity modulation* in the neutral base, is

$$R_o = \frac{V'_{CEQ} + V_{AF}}{I_{CQ}\left(1 - \dfrac{I_{CQ}}{I_{KF}}\right)} \tag{28.2}$$

where V_{AF} is the *forward Early voltage,* V'_{CEQ} is the quiescent voltage developed across the internal collector-emitter terminals, and I_{KF} symbolizes the *forward knee current.* The knee current is a measure of the onset of *high injection effects* [Gummel and Poon, 1970] in the base. In particular, a collector current numerically equal to I_{KF} implies that the forward biasing of the emitter-base junction promotes a net minority carrier charge injected into the base from the emitter that is equal to the background majority charge in the neutral base. The Early voltage is an inverse measure of the slope of the common emitter output characteristic curves.

The final low frequency parameter of the hybrid-pi model is the *forward transconductance*, g_m. This parameter, which is a measure of the forward small signal gain available at a quiescent operating point, is given by

$$g_m = \frac{I_{CQ}}{N_F V_T}\left(\frac{1 - \dfrac{I_{CQ}}{I_{KF}}}{1 + \dfrac{V_{CEQ'}}{V_{AF}}}\right) \tag{28.3}$$

Two capacitances, C_π and C_μ, are incorporated in the small signal model to provide a first-order approximation of steady-state transistor behavior at high signal frequencies. The capacitance, C_π, is the *net capacitance of the emitter-base junction diode* and is given by

$$C_\pi = \frac{C_{JE}}{\left(1 - \dfrac{V_E}{V_{JE} - 2V_T}\right)^{M_{JE}}} + \tau_f g_m \tag{28.4}$$

where the first term on the right-hand side represents the *depletion component* and the second term is the *diffusion component* of C_π. In Eq. (28.4), τ_f is the *average forward transit time of minority carriers in the field-neutral base*, C_{JE} is the *zero bias value of emitter-base junction depletion capacitance,* V_{JE} is the *built-in potential of the junction,* V_E is the *forward biasing voltage developed across the intrinsic emitter-base junction,* and M_{JE} is the *grading coefficient of the junction.* The capacitance, C_μ, has only a depletion component, owing to the reverse

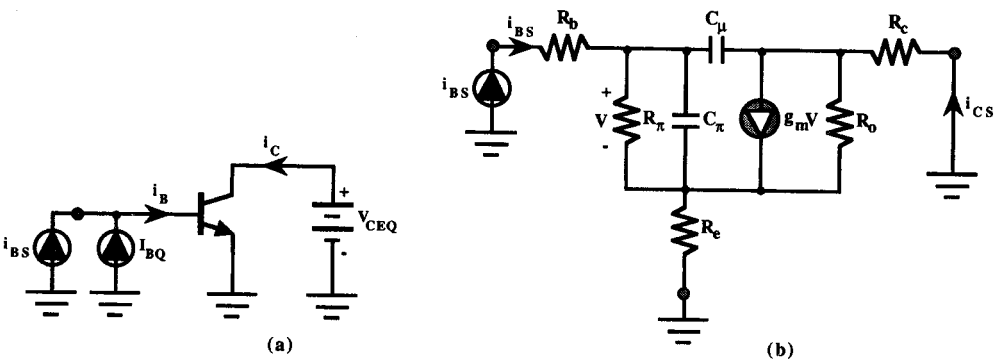


FIGURE 28.12 (a) Schematic diagram pertinent to the evaluation of the short circuit, common emitter, small signal current gain. (b) High frequency small signal model of the circuit in part (a).

(or at most zero) bias impressed across the internal base-collector junction. Accordingly, its analytical form is analogous to the first term on the right-hand side of Eq. (28.4). Specifically,

$$C_{\mu} = \frac{C_{JC}}{\left(1 - \dfrac{V_C}{V_{JC} - 2V_T}\right)^{M_{JC}}} \tag{28.5}$$

where the physical interpretation of C_{JC}, V_{JC}, and M_{JC} is analogous to C_{JE}, V_{JE}, and M_{JE}, respectively.

A commonly invoked figure of merit for assessing the high speed, small signal performance attributes of a BJT is the *common emitter,* **short circuit gain-bandwidth product,** ω_T, which is given by

$$\omega_T = \frac{g_m}{C_{\pi} + C_{\mu}} \tag{28.6}$$

The significance of Eq. (28.6) is best appreciated by studying the simple circuit diagram of Fig. 28.12(a), which depicts the grounded emitter configuration of a BJT biased for linear operation at a quiescent base current of I_{BQ} and a quiescent collector-emitter voltage of V_{CEQ}. Note that the battery supplying V_{CEQ} grounds the collector for small signal conditions. The small signal model of the circuit at hand is resultantly seen to be the topology offered in Fig. 28.12(b), where i_{BS} and i_{CS}, respectively, denote the signal components of the net instantaneous base current, i_B, and the net instantaneous collector current, i_C.

For negligibly small internal collector (R_c) and emitter (R_e) resistances, it can be shown that the *small signal, short circuit, high frequency common emitter current gain,* $\beta_{ac}(j\omega)$, is expressible as

$$\beta_{ac}(j\omega) \triangleq \frac{i_{CS}}{i_{BS}} = \frac{\beta_{ac}\left(1 - \dfrac{j\omega C_{\mu}}{g_m}\right)}{1 + \dfrac{j\omega}{\omega_{\beta}}} \tag{28.7}$$

where β_{ac}, the low frequency value of $\beta_{ac}(j\omega)$, or simply the *low frequency beta,* is

$$\beta_{ac} = \beta_{ac}(0) = g_m R_\pi \tag{28.8}$$

and

$$\omega_\beta = \frac{1}{R_\pi(C_\pi + C_\mu)} \tag{28.9}$$

symbolizes the so-called *beta cutoff frequency* of the BJT. Because the frequency, g_m/C_μ, is typically much larger than ω_β, ω_β is the approximate **3-dB bandwidth** of $\beta_{ac}(j\omega)$; that is,

$$\left|\beta_{ac}(j\omega_\beta)\right| \cong \frac{\beta_{ac}}{\sqrt{2}} \tag{28.10}$$

It follows that the corresponding *gain-bandwidth product*, ω_T, is the product of β_{ac} and ω_β, which, recalling Eq. (28.8), leads directly to the expression in Eq. (28.6). Moreover, in the neighborhood of ω_T,

$$\beta_{ac}(j\omega) \cong \frac{\beta_{ac}\omega_\beta}{j\omega} = \frac{\omega_T}{j\omega} \tag{28.11}$$

which suggests that ω_T is the approximate frequency at which the magnitude of the small signal, short circuit, common emitter current gain degrades to unity.

Hybrid-Pi Equivalent Circuit of a Monolithic BJT

The conventional hyprid-pi model in Fig. 28.11 generally fails to provide sufficiently accurate predictions of the high frequency response of monolithic diffused or implanted BJTs. One reason for this modeling inaccuracy is that the hybrid-pi equivalent circuit does not reflect the fact that monolithic transistors are often fabricated on lightly doped, noninsulating substrates that establish a distributed, large area, *pn* junction with the collector region. Since the substrate-collector *pn* junction is back biased in linear applications of a BJT, negligible static and low frequency signal currents flow from the collector to the substrate. At high frequencies, however, the depletion capacitance associated with the reverse biased substrate-collector junction can cause significant susceptive loading of the collector port. In Fig. 28.13, the lumped capacitance, C_{bb}, whose mathematical definition is similar to that of C_μ in Eq. (28.5), provides a first-order account of this collector loading. Observe that this substrate capacitance appears in series with a substrate resistance, R_{bb}, which reflects the light doping nature of the substrate material. For monolithic transistors fabricated on insulating or semi-insulating substrates, R_{bb} is a very large resistance, thereby rendering C_{bb} unimportant with respect to the problem of predicting steady-state transistor responses at high signal frequencies.

A problem that is even more significant than parasitic substrate dynamics stems from the fact that the hybrid-pi equivalent circuit in Fig. 28.11 is premised on a uniform transistor structure whose emitter-base and base-collector junction areas are identical. In a monolithic device, however, the effective base-collector junction area is much larger than that of the emitter-base junction because the base region is diffused or implanted into the collector [Glaser and Subak-Sharpe, 1977]. The effect of such a geometry is twofold. First, the actual value of C_μ is larger than the value predicated on the physical considerations that surround a simplified uniform structure BJT. Second, C_μ is not a single lumped capacitance that is incident with only the intrinsic base-collector junction. Rather, the effective value of C_μ is distributed between the intrinsic collector and the entire base-collector junction interface. A first-order account of this capacitance distribution entails partitioning C_μ in Fig. 28.11 into two capacitances, say $C_{\mu 1}$ and $C_{\mu 2}$, as indicated in Fig. 28.13. In general, $C_{\mu 2}$ is 3 to 5 times larger than $C_{\mu 1}$. Whereas $C_{\mu 1}$ is proportional to the emitter-base junction area, $C_{\mu 2}$ is proportional to the net base-collector junction area, less the area of the emitter-base junction.

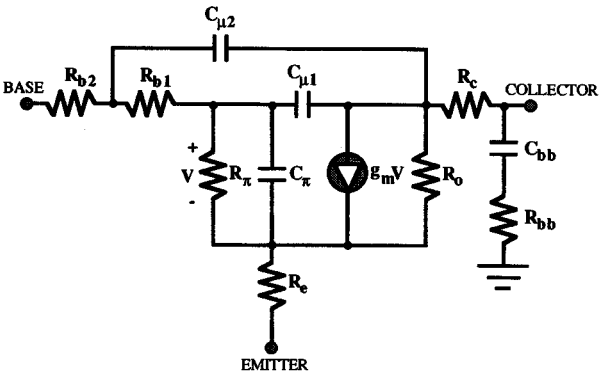


FIGURE 28.13 The hybrid-pi equivalent circuit of a monolithic bipolar junction transistor.

Just as $C_{\mu 1}$ and $C_{\mu 2}$ superimpose to yield the original C_μ in the simplified high frequency model of a BJT, the effective base resistances, R_{b1} and R_{b2}, sum to yield the original base resistance, R_b. The resistance, R_{b1}, is the *contact resistance* associated with the base lead and the inactive BJT base region. It is inversely proportional to the surface area of the base contact. On the other hand, R_{b2}, which is referred to as the *active base resistance*, is nominally an inverse function of emitter finger length. Because of submicron base widths and the relatively light average doping concentrations of active base regions, R_{b2} is significantly larger than R_{b1}.

Common Emitter Amplifier

The most commonly used canonic cell of linear BJT amplifiers is the *common emitter amplifier*, whose basic circuit schematic diagram is depicted in Fig. 28.14(a). In this diagram, R_{ST} is the Thévenin resistance of the applied signal source, V_{ST}, and R_{LT} is the effective, or Thévenin, load resistance driven by the amplifier. The signal source has zero average, or dc, value. Although requisite biasing is not shown in the figure, it is tacitly assumed that the transistor is biased for linear operation. Hence, the diagram at hand is actually the **ac schematic diagram**; that is, it delineates only the signal paths of the circuit. Note that in the common emitter orientation, the input signal is applied to the base of the transistor, while the resultant small signal voltage response, V_{OS}, is extracted at the transistor collector.

The hybrid-pi model of Fig. 28.11 forms the basis for the small signal equivalent circuit of the common emitter cell, which is given in Fig. 28.14(b). In this configuration, the capacitance, C_o, represents an effective output port capacitance that accounts for both substrate loading of the collector port (if the BJT is a monolithic device) and the net effective shunt capacitance associated with the load.

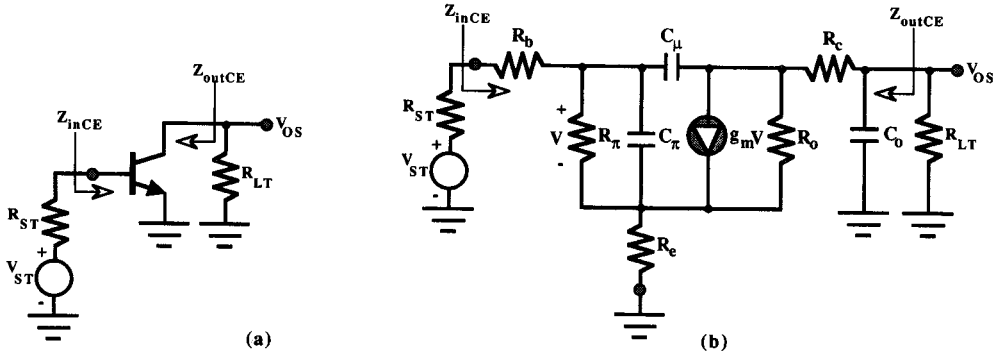


FIGURE 28.14 (a) AC schematic diagram of a common emitter amplifier. (b) Modified small signal, high frequency equivalent circuit of common emitter amplifier.

At low signal frequencies, the capacitors, C_π, C_μ, and C_o in the model of Fig. 28.14(b), can be replaced by open circuits. A straightforward circuit analysis of the resultantly simplified equivalent circuit produces analytical expressions for the low frequency values of the small signal *voltage gain*, $A_{vCE} = V_{OS}/V_{ST}$; the *driving point input impedance*, $Z_{\text{in}CE}$; and the *driving point output impedance*, $Z_{\text{out}CE}$. Because the Early resistance, R_o, is invariably much larger than the resistance sum $(R_c + R_e + R_{LT})$, the low frequency voltage gain of the common emitter cell is expressible as

$$A_{vCE}(0) \cong -\left[\frac{\beta_{ac} R_{LT}}{R_{ST} + R_b + R_\pi + (\beta_{ac} + 1)R_e}\right] \tag{28.12}$$

For large R_o, conventional circuit analyses also produce a low frequency driving point input resistance of

$$R_{\text{in}CE} = Z_{\text{in}CE}(0) \cong R_b + R_\pi + (\beta_{ac} + 1)R_e \tag{28.13}$$

and a low frequency driving point output resistance of

$$R_{\text{out}CE} = Z_{\text{out}CE}(0) \cong \left(\frac{\beta_{ac} R_e}{R_e + R_b + R_\pi + R_{ST}} + 1\right) R_o \tag{28.14}$$

At high signal frequencies, the capacitors in the small signal equivalent circuit of Fig. 28.14(b) produce a third-order voltage gain frequency response whose analytical formulation is algebraically cumbersome [Singhal and Vlach, 1977; Haley, 1988]. However, because the poles produced by these capacitors are real, lie in the left half complex frequency plane, and generally have widely separated frequency values, the dominant pole approximation provides an adequate estimate of high frequency common emitter amplifier response in the usable passband of the amplifier. Accordingly, the high frequency voltage gain, say $A_{vCE}(s)$, of the common emitter amplifier can be approximated as

$$A_{vCE}(s) \cong A_{vCE}(0)\left[\frac{1 + sT_{zCE}}{1 + sT_{pCE}}\right] \tag{28.15}$$

In this expression, T_{pCE} is of the form,

$$T_{pCE} = R_{C\pi} C_\pi + R_{C\mu} C_\mu + R_{Co} C_o \tag{28.16}$$

where $R_{C\pi}$, $R_{C\mu}$, and R_{Co}, respectively, represent the Thévenin resistances seen by the capacitors, C_π, C_μ, and C_o, under the conditions that (1) all capacitors are supplanted by open circuits and (2) the independent signal generator, V_{ST}, is reduced to zero. Analogously, T_{zCE} is of the form

$$T_{zCE} = R_{C\pi o} C_\pi + R_{C\mu o} C_\mu + R_{Coo} C_o \tag{28.17}$$

where $R_{C\pi o}$, $R_{C\mu o}$, and R_{Coo}, respectively, represent the Thévenin resistances seen by the capacitors, C_π, C_μ, and C_o, under the conditions that (1) all capacitors are supplanted by open circuits and (2) the output voltage response, V_{OS}, is constrained to zero while maintaining nonzero input signal source voltage. It can be shown that when R_o is very large and R_c is negligibly small,

$$R_{C\pi} = \frac{R_\pi \| (R_{ST} + R_b + R_e)}{1 + \dfrac{\beta_{ac} R_e}{R_{ST} + R_b + R_\pi + R_e}} \tag{28.18}$$

$$R_{C\mu} = (R_{LT} + R_c) + \{(R_{ST} + R_b) \| [R_\pi + (\beta_{ac} + 1)R_e]\} \left[1 + \frac{\beta_{ac}(R_{LT} + R_c)}{R_\pi + (\beta_{ac} + 1)R_e}\right] \tag{28.19}$$

and

$$R_{Co} = R_{LT} \tag{28.20}$$

Additionally, $R_{C\pi o} = R_{Coo} = 0$, and

$$R_{C\mu o} = -\frac{R_\pi + (\beta_{ac} + 1)R_e}{\beta_{ac}} \tag{28.21}$$

Once T_{pCE} and T_{zCE} are determined, the 3-dB voltage gain bandwidth, B_{CE}, of the common emitter amplifier can be estimated in accordance with

$$B_{CE} \cong \frac{1}{T_{pCE}\sqrt{1 - 2\left(\dfrac{T_{zCE}}{T_{pCE}}\right)^2}} \tag{28.22}$$

The high frequency behavior of both the driving point input and output impedances, $Z_{\text{in}CE}(s)$ and $Z_{\text{out}CE}(s)$, respectively, can be approximated by mathematical functions whose forms are analogous to the gain expression in Eq. (28.15). In particular,

$$Z_{\text{in}CE}(s) \cong R_{\text{in}CE}\left[\frac{1 + sT_{zCE1}}{1 + sT_{pCE1}}\right] \tag{28.23}$$

and

$$Z_{\text{out}CE}(s) \cong R_{\text{out}CE}\left[\frac{1 + sT_{zCE1}}{1 + sT_{pCE2}}\right] \tag{28.24}$$

where $R_{\text{in}CE}$ and $R_{\text{out}CE}$ are defined by Eqs. (28.13) and (28.14). The dominant time constants, T_{pCE1}, T_{zCE1}, T_{pCE2}, and T_{zCE2}, derive directly from Eqs. (28.16) and (28.17) in accordance with [Choma and Witherspoon, 1990]

$$T_{pCE1} = \lim_{R_{ST} \to \infty} [T_{pCE}] \tag{28.25}$$

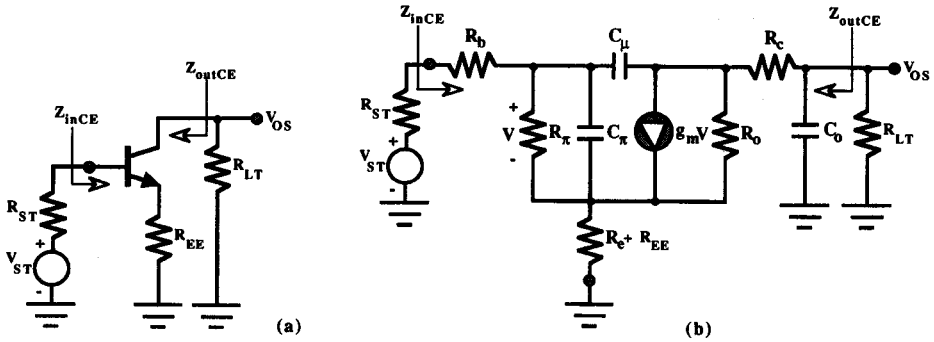


FIGURE 28.15 (a) AC schematic diagram of a common emitter amplifier using an emitter degeneration resistance. (b) Small signal, high frequency equivalent circuit of amplifier in part (a).

$$T_{zCE1} = \lim_{R_{ST} \to 0} [T_{pCE}] \tag{28.26}$$

$$T_{pCE2} = \lim_{R_{LT} \to \infty} [T_{pCE}] \tag{28.27}$$

and

$$T_{zCE2} = \lim_{R_{LT} \to 0} [T_{pCE}] \tag{28.28}$$

For reasonable values of transistor model parameters and terminating resistances, $T_{pCE1} > T_{zCE1}$, and $T_{pCE2} > T_{zCE2}$. It follows that both the input and output ports of a common emitter canonic cell are capacitive at high signal frequencies.

Design Considerations for the Common Emitter Amplifier

Equation (28.12) underscores a serious shortcoming of the canonical common emitter configuration. In particular, since the internal emitter resistance of a BJT is small, the low frequency voltage gain is sensitive to the processing uncertainties that accompany the numerical value of the small signal beta. The problem can be rectified at the price of a diminished voltage gain magnitude by inserting an *emitter degeneration resistance*, R_{EE}, in series with the emitter lead, as shown in Fig. 28.15(a). Since R_{EE} appears in series with the internal emitter resistance, R_e, as suggested in Fig. 28.15(b), the impact of emitter degeneration can be assessed analytically by replacing R_e in Eqs. (28.12) through (28.28) by the resistance sum $(R_e + R_{EE})$. For sufficiently large R_{EE}, such that

$$R_e + R_{EE} \cong R_{EE} >> \frac{R_{ST} + R_b + R_\pi}{\beta_{ac} + 1} \tag{28.29}$$

the low frequency voltage gain becomes

$$A_{vCE}(0) \cong -\frac{\alpha_{ac} R_{LT}}{R_{EE}} \tag{28.30}$$

where α_{ac}, which symbolizes the *small signal, short circuit, common base current gain*, or simply the *ac alpha*, of the transistor is given by

$$\alpha_{ac} = \frac{\beta_{ac}}{\beta_{ac} + 1} \tag{28.31}$$

Despite numerical uncertainties in β_{ac}, minimum values of β_{ac} are much larger than one, thereby rendering the voltage gain in Eq. (28.30) almost completely independent of small signal BJT parameters.

A second effect of emitter degeneration is an increase in both the low frequency driving point input and output resistances. This contention is confirmed by Eq. (28.13), which shows that if R_o remains much larger than $(R_c + R_e + R_{EE} + R_{LT})$, a resistance in the amount of $(\beta_{ac} + 1)R_{EE}$ is added to the input resistance established when the emitter of a common emitter amplifier is returned to signal ground. Likewise, Eq. (28.14) verifies that emitter degeneration increases the low frequency driving point output resistance. In fact, a very large value of R_{EE} produces an output resistance that approaches a limiting value of $(\beta_{ac} + 1)R_o$. It follows that a common emitter amplifier that exploits emitter degeneration behaves as a voltage-to-current converter at low signal frequencies. In particular, its high input resistance does not incur an appreciable load on signal voltage sources that are characterized by even moderately large Thévenin resistances, while its large output resistance comprises an almost ideal current source at its output port.

A third effect of emitter degeneration is a decrease in the effective pole time constant, T_{pCE}, as well as an increase in the effective zero time constant, T_{zCE}, which can be confirmed by reinvestigating Eqs. (28.18) through (28.21) for the case of R_e replaced by the resistance sum $(R_e + R_{EE})$. The use of an emitter degeneration resistance therefore promotes an increased 3-dB circuit bandwidth. Unfortunately, it also yields a diminished circuit gain-bandwidth product; that is, a given emitter degeneration resistance causes a degradation in the low frequency gain magnitude that is larger than the corresponding bandwidth increase promoted by this resistance. This deterioration of circuit gain-bandwidth product is a property of all *negative feedback circuits* [Choma, 1984].

For reasonable values of the emitter degeneration resistance, R_{EE}, the Thévenin time constant, $R_{C\mu}C_\mu$, is likely to be the dominant contribution to the effective first-order time constant, T_{pCE}, attributed to the poles of a common emitter amplifier. Hence, C_μ is the likely device capacitance that dominantly imposes an upper limit to the achievable 3-dB bandwidth of a common emitter cell. The reason for this substantial bandwidth sensitivity to C_μ is the so-called Miller multiplication factor, say M, which appears as the last bracketed term on the right-hand side of Eq. (28.19), namely,

$$M = 1 + \frac{\beta_{ac}(R_{LT} + R_c)}{R_\pi + (\beta_{ac} + 1)R_e} \tag{28.32}$$

The Miller factor, M, which effectively multiplies C_μ in the expression for $R_{C\mu}C_\mu$, increases sharply with the load resistance, R_{LT}, and hence with the gain magnitude of the common emitter amplifier. Note that in the limit of a large emitter degeneration resistance (which adds directly to R_e), Eq. (28.30) reduces Eq. (28.32) to the factor

$$M \cong 1 + \left|A_{vCE}(0)\right| \tag{28.33}$$

Common Base Amplifier

A second canonic cell of linear BJT amplifiers is the *common base amplifier*, whose ac circuit schematic diagram appears in Fig. 28.16(a). In this diagram, R_{ST}, V_{ST}, R_{LT}, and V_{OS} retain the significance they respectively have in the previously considered common emitter configuration. Note that in the common base orientation, the input signal is applied to the base, while the resultant small signal voltage response is extracted at the collector of a transistor.

The relevant small signal model is shown in Fig. 28.16(b). A straightforward application of Kirchhoff's circuit laws gives, for the case of large R_o, a low frequency voltage gain, $A_{vCB}(0) = V_{OS}/V_{ST}$, of

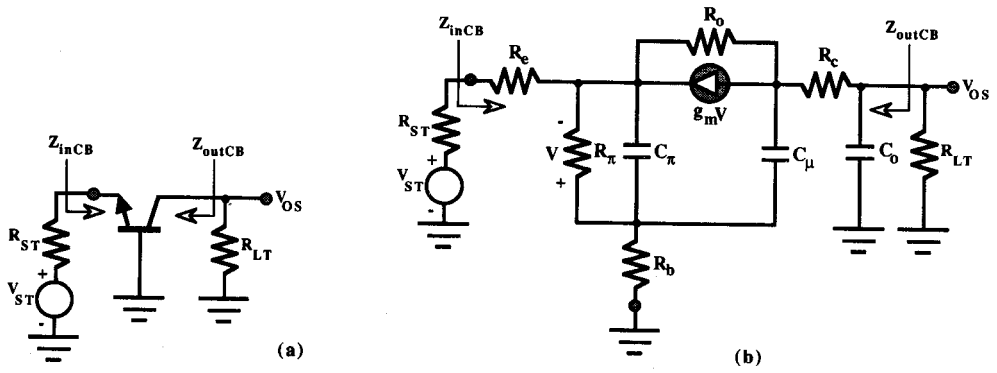


FIGURE 28.16 (a) AC schematic diagram of a common base amplifier. (b) Small signal, high frequency equivalent circuit of amplifier in part (a).

$$A_{vCB}(0) \cong \frac{\alpha_{ac} R_{LT}}{R_{ST} + R_{\text{in}CB}} \tag{28.34}$$

where $R_{\text{in}CB}$ is the low frequency value of the common base driving point input impedance,

$$R_{\text{in}CB} = Z_{\text{in}CB}(0) \cong R_e + \frac{R_b + R_\pi}{\beta_{ac} + 1} \tag{28.35}$$

Moreover, it can be shown that the low frequency driving point output resistance is

$$R_{\text{out}CB} = Z_{\text{out}CB}(0) \cong \left[\frac{\beta_{ac}(R_e + R_{ST})}{R_e + R_b + R_\pi + R_{ST}} + 1 \right] R_o \tag{28.36}$$

The preceding three equations underscore several operating characteristics that distinguish the common base amplifier from its common emitter counterpart. For example, Eq. (28.35) suggests a low frequency input resistance that is significantly smaller than that of a common emitter unit. To underscore this contention, consider the case of two identical transistors, one used in a common emitter amplifier and the other used in a common base configuration, that are biased at identical quiescent operating points. Under this circumstance, Eqs. (28.35) and (28.13) combine to deliver

$$R_{\text{in}CB} \cong \frac{R_{\text{in}CE}}{\beta_{ac} + 1} \tag{28.37}$$

which shows that the common base input resistance is a factor of $(\beta_{ac} + 1)$ times smaller than the input resistance of the common emitter cell. The resistance reflection factor, $(\beta_{ac} + 1)$, in Eq. (28.37) represents the ratio of small signal emitter current to small signal base current. Accordingly, Eq. (28.37) is self-evident when it is noted that the input resistance of a common base stage is referred to an input emitter current, whereas the input resistance of its common emitter counterpart is referred to an input base current.

A second difference between the common emitter and common base amplifiers is that the voltage gain of the latter displays no phase inversion between source and response voltages. Moreover, for the same load and source terminations and for identical transistors biased identically, the voltage gain of the common base cell is likely to be much smaller than that of the common emitter unit. This contention is verified by substituting Eq. (28.37) into Eq. (28.34) and using Eqs. (28.31), (28.13), and (28.12) to write

$$A_{vCB}(0) \cong \frac{|A_{vCE}(0)|}{1 + \dfrac{\beta_{ac} R_{ST}}{R_{ST} + R_{\text{in}CE}}} \tag{28.38}$$

At high signal frequencies, the voltage gain, driving point input impedance, and driving point output impedance can be approximated by functions whose analytical forms mirror those of Eqs. (28.15), (28.23), and (28.24). Let T_{pCB} and T_{zCB} designate the time constants of the effective dominant pole and the effective dominant zero, respectively, of the common base cell. An analysis of the structure of Fig. 28.16(b) resultantly produces, with R_o and R_c ignored,

$$T_{pCB} = R_{C\pi} C_\pi + R_{C\mu} C_\mu + R_{Co} C_o \tag{28.39}$$

where

$$R_{C\pi} = \frac{R_\pi \,||\, (R_{ST} + R_b + R_e)}{1 + \dfrac{\beta_{ac}(R_{ST} + R_e)}{R_{ST} + R_b + R_\pi + R_e}} \tag{28.40}$$

$$R_{C\mu} = R_b \,||\, [R_\pi + (\beta_{ac} + 1)(R_{ST} + R_e)] + R_{LT}\left[1 + \frac{\beta_{ac} R_b}{R_b + R_\pi + (\beta_{ac} + 1)(R_{ST} + R_e)}\right] \tag{28.41}$$

and R_{Co} remains given by Eq. (28.20). Moreover,

$$T_{zCB} = \frac{R_b C_\mu}{\alpha_{ac}} \tag{28.42}$$

Design Considerations for the Common Base Amplifier

An adaptation of Eqs. (28.25) through (28.28) to the common base stage confirms that the driving point input impedance is capacitive at high signal frequencies. On the other hand, $g_m R_b > 1$ renders a common base driving point input impedance that is inductive at high frequencies. This impedance property can be gainfully exploited to realize monolithic shunt peaked amplifiers in which the requisite circuit inductance is synthesized as the driving point input impedance of a common base stage (or the driving point output impedance of a common collector cell) [Grebene, 1984].

The common base stage is often used to broadband the common emitter amplifier by forming the *common emitter–common base cascode*, whose ac schematic diagram is given in Fig. 28.17. The broadbanding afforded by the cascode structure stems from the fact that the effective low frequency load resistance, say R_{Le}, seen by the common emitter transistor, QE, is the small driving point input resistance of the common base amplifier, QB. This effective load resistance, as witnessed by C_μ of the common emitter transistor, is much smaller than the actual load resistance that terminates the output port of the amplifier, thereby decreasing the Miller multiplication of the C_μ in QE. If the time constant savings afforded by decreased Miller multiplication is larger than the sum of the additional time constants presented to the circuit by the common base transistor, an enhancement of common emitter bandwidth occurs. Note that such bandwidth enhancement is realized without compromising the common emitter gain-bandwidth product, since the voltage gain of the common emitter–common base unit is almost identical to that of the common emitter amplifier alone.

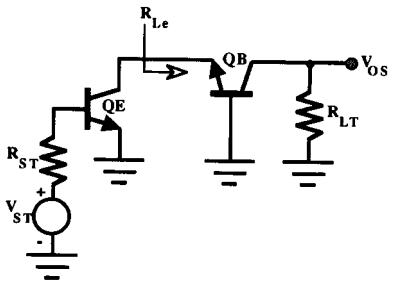


FIGURE 28.17 AC schematic diagram of a common emitter–common base cascode amplifier.

Common Collector Amplifier

The final canonic cell of linear BJT amplifiers is the *common collector amplifier*. The ac schematic diagram of this stage, which is often referred to as an *emitter follower*, is given in Fig. 28.18(a). In emitter followers, the input signal is applied to the base, and the resultant small signal output voltage is extracted at the transistor emitter.

The small signal equivalent circuit corresponding to the amplifier in Fig. 28.18(a) is shown in Fig. 28.18(b). A straightforward circuit analysis gives, for the case of large R_o, a low frequency voltage gain, $A_{vCC}(0) = V_{OS}/V_{ST}$, of

$$A_{vCC}(0) \cong \frac{R_{LT}}{R_{LT} + R_{\text{out}CC}} \tag{28.43}$$

where $R_{\text{out}CC}$ is the low frequency value of the driving point output impedance,

$$R_{\text{out}CC} = Z_{\text{out}CC}(0) \cong R_e + \frac{R_b + R_\pi + R_{ST}}{\beta_{ac} + 1} \tag{28.44}$$

The low frequency driving point output resistance is

$$R_{\text{in}CC} = Z_{\text{in}CC}(0) \cong R_b + R_\pi + (\beta_{ac} + 1)(R_e + R_{LT}) \tag{28.45}$$

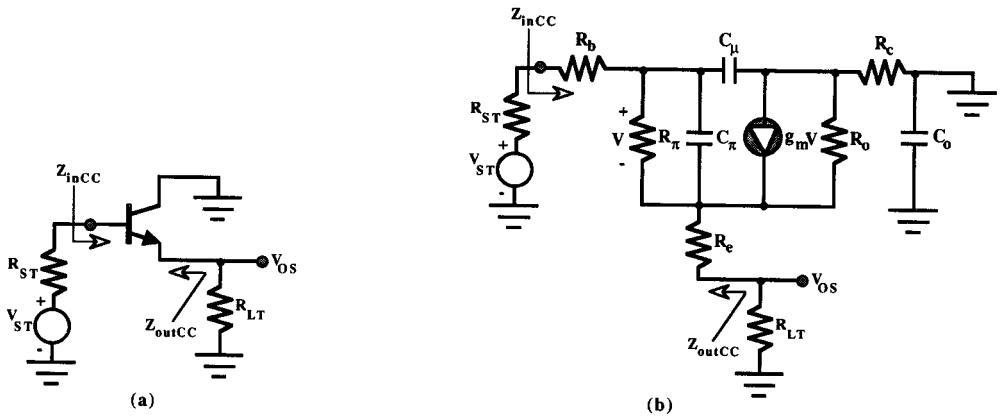


FIGURE 28.18 (a) AC schematic diagram of a common collector (emitter follower) amplifier. (b) Small signal, high frequency equivalent circuit of amplifier in part (a).

The facts that the voltage gain is less than one and is without phase inversion, the output resistance is small, and the input resistance is large make the emitter follower an excellent candidate for impedance buffering applications.

As in the cases of the common emitter and the common base amplifiers, the high frequency voltage gain, driving point input resistance, and driving point output resistance can be approximated by functions having analytical forms that are similar to those of Eqs. (28.15), (28.23), and (28.24). Let T_{pCC} and T_{zCC} designate the time constants of the effective dominant pole and the effective dominant zero, respectively, of the emitter follower. Since the output port capacitance, C_o, appears across a short circuit, T_{pCC} is expressible as

$$T_{pCC} = R_{C\pi}C_{\pi} + R_{C\mu}C_{\mu} \tag{28.46}$$

With R_o ignored,

$$R_{C\pi} = \frac{R_{\pi} \,||\, (R_{ST} + R_b + R_{LT} + R_e)}{1 + \dfrac{\beta_{ac}(R_{LT} + R_e)}{R_{ST} + R_b + R_{\pi} + R_{LT} + R_e}} \tag{28.47}$$

and

$$R_{C\mu} = (R_{ST} + R_b) \,||\, [R_{\pi} + (\beta_{ac} + 1)(R_{LT} + R_e) + \left[1 + \frac{\beta_{ac}(R_{ST} + R_b)}{R_{ST} + R_b + R_{\pi} + (\beta_{ac} + 1)(R_{LT} + R_e)}\right] R_c \tag{28.48}$$

The time constant of the effective dominant zero is

$$T_{zCC} = \frac{R_{\pi}C_{\pi}}{\beta_{ac} + 1} \tag{28.49}$$

Although the emitter follower possesses excellent wideband response characteristics, it should be noted in Eq. (28.48) that the internal collector resistance, R_c, incurs some Miller multiplication of the base-collector junction capacitance, C_{μ}. For this reason, monolithic common collector amplifiers work best in broadband impedance buffering applications when they exploit transistors that have collector sinker diffusions and buried collector layers, which collectively serve to minimize the parasitic internal collector resistance.

Defining Terms

ac schematic diagram: A circuit schematic diagram, divorced of biasing subcircuits, that depicts only the dynamic signal flow paths of an electronic circuit.

Driving point impedance: The effective impedance presented at a port of a circuit under the condition that all other circuit ports are terminated in the resistances actually used in the design realization.

Hybrid-pi model: A two-pole linear circuit used to model the small signal responses of bipolar circuits and circuits fabricated in other device technologies.

Miller effect: The deterioration of the effective input impedance caused by the presence of feedback from the output port to the input port of a phase-inverting voltage amplifier.

Short circuit gain-bandwidth product: A measure of the frequency response capability of an electronic circuit. When applied to bipolar circuits, it is nominally the signal frequency at which the magnitude of the current gain degrades to one.

Three-decibel bandwidth: A measure of the frequency response capability of low-pass and bandpass electronic circuits. It is the range of signal frequencies over which the maximum gain of the circuit is constant to within a factor of the square root of two.

Related Topic

24.2 Bipolar Transistors

References

W.K. Chen, *Circuits and Filters Handbook,* Boca Raton, Fla: CRC Press, 1995.

J. Choma, "A generalized bandwidth estimation theory for feedback amplifiers," *IEEE Transactions on Circuits and Systems*, vol. CAS-31, Oct. 1984.

J. Choma and S. Witherspoon, "Computationally efficient estimation of frequency response and driving point impedances in wide-band analog amplifiers," *IEEE Transactions on Circuits and Systems*, vol. CAS-37, June 1990.

K.K. Clarke and D.T. Hess, *Communication Circuits: Analysis and Design,* Reading, Mass.: Addison-Wesley, 1978.

H.C. de Graaf, "Two New Methods for Determining the Collector Series Resistance in Bipolar Transistors With Lightly Doped Collectors," Phillips Research Report, 24, 1969.

A.B. Glaser and G.E. Subak-Sharpe, *Integrated Circuit Engineering: Design, Fabrication, and Applications,* Reading, Mass.: Addison-Wesley, 1977.

A.B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design,* New York: Wiley Interscience, 1984.

H.K. Gummel and H.C. Poon, "An integral charge-control model of bipolar transistors," *Bell System Technical Journal*, 49, May–June 1970.

S. B. Haley, "The general eigenproblem: pole-zero computation," *Proc. IEEE*, 76, Feb. 1988.

J.D. Irwin, *Industrial Electronics Handbook,* Boca Raton, Fla.: CRC Press, 1997.

A.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., New York: Holt, Rinehart and Winston, 1991.

K. Singhal and J. Vlach, "Symbolic analysis of analog and digital circuits," *IEEE Transactions on Circuits and Systems*, vol. CAS-24, Nov. 1977.

Further Information

The *IEEE Journal of Solid-State Circuits* publishes state-of-the-art articles on all aspects of integrated electronic circuit design. The December issue of this journal focuses on analog electronics.

The *IEEE Transactions on Circuits and Systems* also publishes circuit design articles. Unlike the *IEEE Journal of Solid-State Circuits,* this journal addresses passive and active, discrete component circuits, as well as integrated circuits and systems, and it features theoretic research that underpins circuit design strategies.

The *Journal of Analog Integrated Circuits and Signal Processing* publishes design-oriented papers with emphasis on design methodologies and design results.

Massara, R.E., Steadman, J.W., Wilamowski, B.M., Svoboda, J.A. “Active Filters”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

29

Active Filters

Robert E. Massara
University of Essex

J. W. Steadman
University of Wyoming

B. M. Wilamowski
University of Wyoming

James A. Svoboda
Clarkson University

29.1 Synthesis of Low-Pass Forms
Passive and Active Filters • Active Filter Classification and Sensitivity • Cascaded Second-Order Sections • Passive Ladder Simulation • Active Filters for ICs

29.2 Realization
Transformation from Low-Pass to Other Filter Types • Circuit Realizations

29.3 Generalized Impedance Convertors and Simulated Impedances

29.1 Synthesis of Low-Pass Forms

Robert E. Massara

Passive and Active Filters

There are formal definitions of activity and passivity in electronics, but it is sufficient to observe that passive filters are built from passive components; resistors, capacitors, and inductors are the commonly encountered building blocks although distributed RC components, quartz crystals, and surface acoustic wave devices are used in filters working in the high-megahertz regions. **Active filters** also use resistors and capacitors, but the inductors are replaced by active devices capable of producing power gain. These devices can range from single transistors to integrated circuit (IC) -controlled sources such as the operational amplifier (op amp), and more exotic devices, such as the operational transconductance amplifier (OTA), the generalized impedance converter (GIC), and the frequency-dependent negative resistor (FDNR).

The theory of filter synthesis, whether active or passive, involves the determination of a suitable circuit topology and the computation of the circuit component values within the topology, such that a required network response is obtained. This response is most commonly a voltage transfer function (VTF) specified in the frequency domain. Circuit analysis will allow the performance of a filter to be evaluated, and this can be done by obtaining the VTF, $H(s)$, which is, in general, a rational function of s, the complex frequency variable. The *poles* of a VTF correspond to the roots of its denominator polynomial. It was established early in the history of filter theory that a network capable of yielding complex-conjugate transfer function (TF) pole-pairs is required to achieve high selectivity. A highly selective network is one that gives a rapid transition between passband and stopband regions of the frequency response. Figure 29.1(a) gives an example of a passive low-pass LCR ladder network capable of producing a VTF with the necessary pole pattern.

The network of Fig. 29.1(a) yields a VTF of the form

$$H(s) = \frac{V_{\text{out}}(s)}{V_{\text{in}}(s)} = \frac{1}{a_5 s^5 + a_4 s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0} \tag{29.1}$$

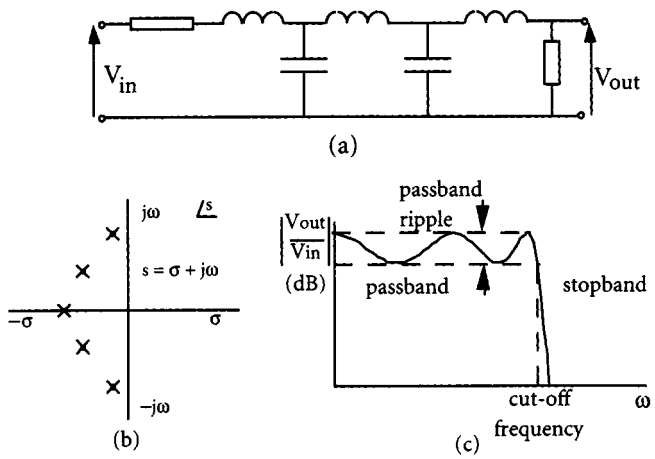


FIGURE 29.1 (a) Passive LCR filter; (b) typical pole plot; (c) typical frequency response.

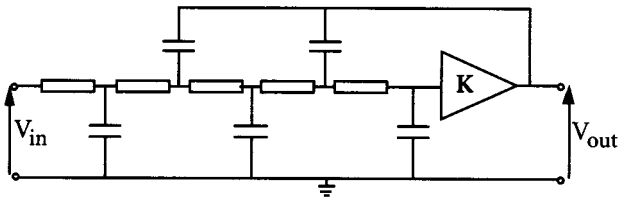


FIGURE 29.2 RC-active filter equivalent to circuit of Fig. 29.1(a).

Figure 29.1(b) shows a typical pole plot for the fifth-order VTF produced by this circuit. Figure 29.1(c) gives a sample sinusoidal steady-state frequency response plot. The frequency response is found by setting $s = j\omega$ in Eq. (29.1) and taking $|H(j\omega)|$. The LCR low-pass ladder structure of Fig. 29.1(a) can be altered to higher or lower order simply by adding or subtracting reactances, preserving the series-inductor/shunt-capacitor pattern. In general terms, the higher the filter order, the greater the selectivity.

This simple circuit structure is associated with a well-established design theory and might appear the perfect solution to the filter synthesis problem. Unfortunately, the problems introduced by the use of the inductor as a circuit component proved a serious difficulty from the outset. Inductors are intrinsically nonideal components, and the lower the frequency range of operation, the greater these problems become. Problems include significant series resistance associated with the physical structure of the inductor as a coil of wire, its ability to couple by electromagnetic induction into fields emanating from external components and sources and from other inductors within the filter, its physical size, and potential mechanical instability. Added to these problems is the fact that the inductor tends not to be an off-the-shelf component but has instead to be fabricated to the required value as a custom device. These serious practical difficulties created an early pressure to develop alternative approaches to electrical filtering. After the emergence of the electronic amplifier based on vacuum rules, it was discovered that networks involving resistors, capacitors, and amplifiers—*RC-active filters*—were capable of producing TFs exactly equivalent to those of LCR ladders. Figure 29.2 shows a single-amplifier multiloop ladder structure that can produce a fifth-order response identical to that of the circuit of Fig. 29.1(a).

The early active filters, based as they were on tube amplifiers, did not constitute any significant advance over their passive counterparts. It required the advent of solid-state active devices to make the RC-active filter a viable alternative. Over the subsequent three decades, active filter theory has developed to an advanced state, and this development continues as new IC technologies create opportunities for novel network structures and applications.

Active Filter Classification and Sensitivity

There are two major approaches to the synthesis of RC-active filters. In the first approach, a TF specification is factored into a product of second-order terms. Each of these terms is realized by a separate RC-active

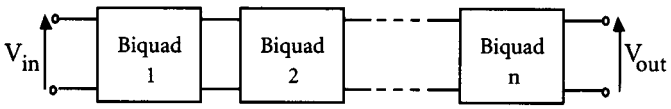


FIGURE 29.3 Biquad cascade realizing high-order filter.

subnetwork designed to allow for non-interactive interconnection. The subnetworks are then connected in cascade to realize the required overall TF, as shown in Fig. 29.3. A first-order section is also required to realize odd-order TF specifications. These second-order sections may, depending on the exact form of the overall TF specification, be required to realize numerator terms of up to second order. An RC-active network capable of realizing a biquadratic TF (that is, one whose numerator *and* denominator polynomials are second-order) is called a **biquad**.

This scheme has the advantage of design ease since simple equations can be derived relating the components of each section to the coefficients of each factor in the VTF. Also, each biquad can be independently adjusted relatively easily to give the correct performance. Because of these important practical merits, a large number of alternative biquad structures have been proposed, and the newcomer may easily find the choice overwhelming.

The second approach to active filter synthesis involves the use of RC-active circuits to simulate passive LCR ladders. This has two important advantages. First, the design process can be very straightforward: the wealth of design data published for passive ladder filters (see Further Information) can be used directly so that the sometimes difficult process of component value synthesis from specification is eliminated. Second, the LCR ladder offers optimal **sensitivity** properties [Orchard, 1966], and RC-active filters designed by ladder simulation share the same low sensitivity features. Chapter 4 of Bowron and Stephenson [1979] gives an excellent introduction to the formal treatment of circuit sensitivity.

Sensitivity plays a vital role in the characterization of RC-active filters. It provides a measure of the extent to which a change in the value of any given component affects the response of the filter. High sensitivity in an RC-active filter should also alert the designer to the possibility of oscillation. A nominally stable design will be unstable in practical realization if sensitivities are such that component value errors cause one or more pairs of poles to migrate into the right half plane. Because any practical filter will be built with components that are not exactly nominal in value, sensitivity information provides a practical and useful indication of how different filter structures will react and provides a basis for comparison.

Cascaded Second-Order Sections

This section will introduce the cascade approach to active filter design. As noted earlier, there are a great many second-order RC-active sections to choose from, and the present treatment aims only to convey some of the main ideas involved in this strategy. The references provided at the end of this section point the reader to several comprehensive treatments of the subject.

Sallen and Key Section

This is an early and simple example of a second-order section building block [Sallen and Key, 1955]. It remains a commonly used filter despite its age, and it will serve to illustrate some key stages in the design of all such RC-active sections. The circuit is shown in Fig. 29.4. A straightforward analysis of this circuit yields a VTF

$$H(s) = \frac{K\dfrac{1}{C_1C_2R_1R_2}}{s^2 + s\left[\dfrac{1}{C_2R_2} + \dfrac{1}{C_2R_1} + \dfrac{1-K}{C_1R_1}\right] + \dfrac{1}{C_1C_2R_1R_2}} \tag{29.2}$$

This is an all-pole low-pass form since the numerator involves only a constant term.

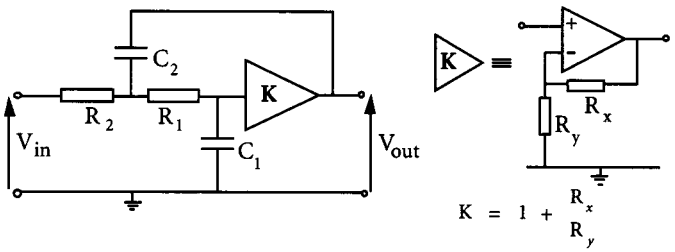


FIGURE 29.4 Sallen and Key second-order filter section.

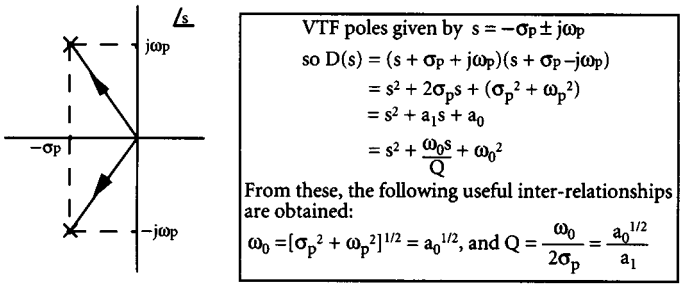


FIGURE 29.5 VTF pole relationships.

Specifications for an all-pole second-order section may arise in coefficient form, where the required *s*-domain VTF is given as

$$H(s) = \frac{k}{s^2 + a_1 s + a_0} \tag{29.3}$$

or in Q-ω_0 standard second-order form

$$H(s) = \frac{k}{s^2 + \dfrac{\omega_0}{Q} s + \omega_0} \tag{29.4}$$

Figure 29.5 shows the relationship between these VTF forms.

As a design example, the VTF for an all-pole fifth-order Chebyshev filter with 0.5-dB passband ripple [see Fig. 29.1(c)] has the factored-form denominator

$$D(s) = (s + 0.36232)(s^2 + 0.22393\,s + 1.0358)(s^2 + 0.58625s + 0.47677) \tag{29.5}$$

Taking the first of the quadratic factors in Eq. (29.5) and comparing like coefficients from Eq. (29.2) gives the following design equations:

$$\frac{1}{C_1 C_2 R_1 R_2} = 1.0358; \qquad \frac{1}{C_2 R_2} + \frac{1}{C_2 R_1} + \frac{1 - K}{C_1 R_1} = 0.22393 \tag{29.6}$$

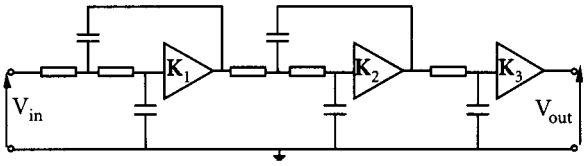


FIGURE 29.6 Form of fifth-order Sallen and Key cascade.

Clearly, the designer has some degrees of freedom here since there are two equations in five unknowns. Choosing to set both (normalized) capacitor values to unity, and fixing the dc stage gain $K = 5$, gives

$$C_1 = C_2 = 1\text{F};\ R_1 = 1.8134\ \Omega;\ R_2 = 1.3705\ \Omega;\ R_x = 4\ \Omega;\ R_y = 1\ \Omega$$

Note that Eq. (29.5) is a normalized specification giving a filter cut-off frequency of 1 rad s^{-1}. These normalized component values can now be denormalized to give a required cut-off frequency and practical component values. Suppose that the filter is, in fact, required to give a cut-off frequency $f_c = 1$ kHz. The necessary shift is produced by multiplying all the capacitors (leaving the resistors fixed) by the factor ω_N/ω_D where ω_N is the normalized cut-off frequency (1 rad s^{-1} here) and ω_D is the required denormalized cut-off frequency ($2\pi \times 1000$ rad s^{-1}). Applying this results in denormalized capacitor values of 159.2 μF. A useful rule of thumb [Waters, 1991] advises that capacitor values should be on the order of magnitude of $(10/f_c)$ μF, which suggests that the capacitors should be further scaled to around 10 nF. This can be achieved without altering of the filter's f_c by means of the impedance scaling property of electrical circuits. Providing all circuit impedances are scaled by the same amount, current and voltage TFs are preserved. In an RC-active circuit, this requires that all resistances are multiplied by some factor while all capacitances are divided by it (since capacitive impedance is proportional to $1/C$). Applying this process yields final values as follows:

$$C_1,\ C_2 = 10\ \text{nF};\ R_1 = 29.86\ \text{k}\Omega;\ R_2 = 21.81\ \text{k}\Omega;\ R_x = 63.66\ \text{k}\Omega;\ R_y = 15.92\ \text{k}\Omega$$

Note also that the dc gain of each stage, $|H(0)|$, is given by K [see Eq. (29.2) and Fig. 29.4] and, when several stages are cascaded, the overall dc gain of the filter will be the product of these individual stage gains. This feature of the Sallen and Key structure gives the designer the ability to combine easy-to-manage amplification with prescribed filtering.

Realization of the complete fifth-order Chebyshev VTF requires the design of another second-order section to deal with the second quadratic term in Eq. (29.5), together with a simple circuit to realize the first-order term arising because this is an odd-order VTF. Figure 29.6 shows the form of the overall cascade. Note that the op amps at the output of each stage provide the necessary interstage isolation. It is finally worth noting that an extended single-amplifier form of the Sallen and Key network exists—the circuit shown in Fig. 29.2 is an example of this—but that the saving in op amps is paid for by higher component spreads, sensitivities, and design complexity.

State-Variable Biquad

The simple Sallen and Key filter provides only an all-pole TF; many commonly encountered filter specifications are of this form—the Butterworth and Chebyshev approximations are notable examples—so this is not a serious limitation. In general, however, it will be necessary to produce sections capable of realizing a second-order denominator together with a numerator polynomial of up to second-order:

$$H(s) = \frac{b_2 s^2 + b_1 s + b_0}{s^2 + a_1 s + a_0} \tag{29.7}$$

The other major filter approximation in common use—the elliptic (or Cauer) function filter—involves quadratic numerator terms in which the b_1 coefficient in Eq. (29.7) is missing. The resulting numerator

polynomial, of the form $b_2\ s^2 + b_0$, gives rise to *s*-plane zeros on the $j\omega$ axis corresponding to points in the stopband of the sinusoidal frequency response where the filter's transmission goes to zero. These notches or *transmission zeros* account for the elliptic's very rapid transition from passband to stopband and, hence, its optimal selectivity.

A filter structure capable of producing a VTF of the form of Eq. (29.7) was introduced as a state-variable realization by its originators [Kerwin et al., 1967]. The structure comprises two op amp integrators and an op amp summer connected in a loop and was based on the integrator-summer analog computer used in control/analog systems analysis, where system state is characterized by some set of so-called state variables. It is also often referred to as a ring-of-three structure. Many subsequent refinements of this design have appeared (Schaumann et al., [1990] gives a useful treatment of some of these developments) and the state-variable biquad has achieved considerable popularity as the basis of many commercial universal packaged active filter building blocks. By selecting appropriate chip/package output terminals, and with the use of external trimming components, a very wide range of filter responses can be obtained.

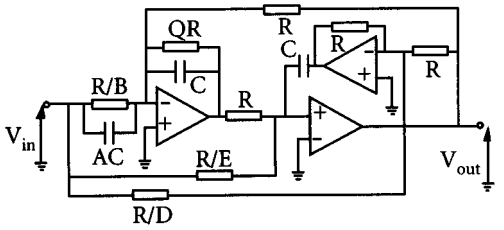


FIGURE 29.7 Circuit schematic for state-variable biquad.

Figure 29.7 shows a circuit developed from this basic state-variable network and described in Schaumann et al. [1990]. The circuit yields a VTF

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = -\frac{As^2 + \omega_0(B - D)s + E\omega_0^2}{s^2 + \dfrac{\omega_0}{Q}s + \omega_0^2}, \text{ with } \omega_0 \triangleq 1/RC \tag{29.8}$$

By an appropriate choice of the circuit component values, a desired VTF of the form of Eq. (29.8) can be realized.

Consider, for example, a specification requirement for a second-order elliptic filter cutting off at 10 kHz. Assume that a suitable normalized (1 rad/s) specification for the VTF is

$$H(s) = -\frac{0.15677(s^2 + 7.464)}{s^2 + 0.9989s + 1.1701} \tag{29.9}$$

From Eq. (29.8) and Eq. (29.9), and referring to Fig. 29.7, normalized values for the components are computed as follows. As the *s* term in the numerator is to be zero, set $B = D = 0$ (which obtains if resistors R/B and R/D are simply removed from the circuit). Setting $C = 1$ F gives the following results:

$$\text{AC} = 0.15677\text{F};\ \text{R} = 1/C\omega_0 = 0.92446\ \Omega;\ \text{QR} = 1.08290\ \Omega;\ \text{R/E} = 0.92446\ \Omega$$

Removing the normalization and setting $C = (10/10\text{ k})\ \mu\text{F} = 1$ nF requires capacitors to be multiplied by 10^{-9} and resistors to be multiplied by 15.9155×10^3. Final denormalized component values for the 10-kHz filter are thus:

$$\text{C} = 1\text{ nF};\ \text{AC} = 0.15677\text{ nF};\ \text{R} = \text{R/E} = 14.713\ \text{k}\Omega;\ \text{QR} = 17.235\ \text{k}\Omega$$

Passive Ladder Simulation

As for the biquad approach, numerous different ladder-based design methods have been proposed. Two representative schemes will be considered here: inductance simulation and ladder transformation.

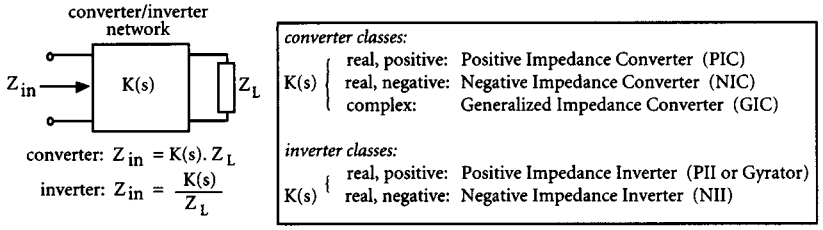


FIGURE 29.8 Generic impedance converter/inverter networks.

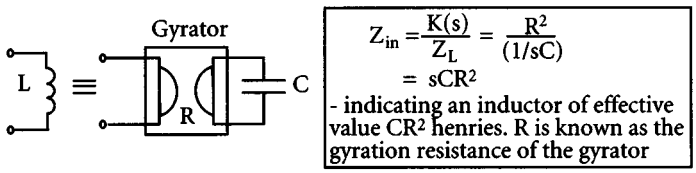


FIGURE 29.9 Gyrator simulation of an inductor.

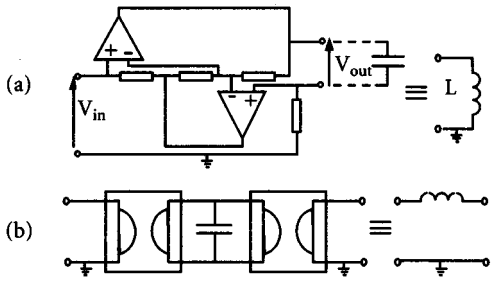


FIGURE 29.10 (a) Practical gyrator and (b) simulation of floating inductor. (Source: A. Antoniou, Proc. IEE, vol. 116, pp. 1838–1850, 1969. With permission.)

Inductance Simulation

In the inductance simulation approach, use is made of impedance converter/inverter networks. Figure 29.8 gives a classification of the various generic forms of device. The NIC enjoyed prominence in the early days of active filters but was found to be prone to instability. Two classes of device that have proved more useful in the longer term are the GIC and the gyrator.

Figure 29.9 introduces the symbolic representation of a gyrator and shows its use in simulating an inductor.

The gyrator can conveniently be realized by the circuit of Fig. 29.10(a), but note that the simulated inductor is grounded at one end. This presents no problem in the case of high-pass filters and other forms requiring a grounded shunt inductor but is not suitable for the low-pass filter. Figure 29.10(b) shows how a pair of back-to-back gyrators can be configured to produce a floating inductance, but this involves four op amps per inductor.

The next section will introduce an alternative approach that avoids the op amp count difficulty associated with simulating the floating inductors directly.

Ladder Transformation

The other main approach to the RC-active simulation of passive ladders involves the transformation of a prototype ladder into a form suitable for active realization. A most effective method of this class is based on the use of the Bruton transformation [Bruton, 1969], which involves the complex impedance scaling of a prototype passive LCR ladder network. All prototype circuit impedances $Z(s)$ are transformed to $Z_T(s)$ with

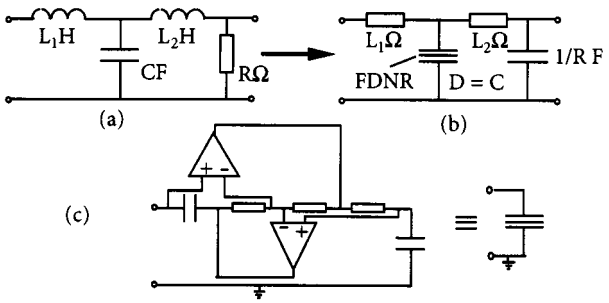


FIGURE 29.11 FDNR active filter.

$$Z_T(s) = \frac{K}{s} \cdot Z(s) \tag{29.10}$$

where K is a constant chosen by the designer and which provides the capacity to scale component values in the final filter. Since impedance transformations do not affect voltage and current transfer ratios, the VTF remains unaltered by this change. The Bruton transformation is applied directly to the elements in the prototype network, and it follows from Eq. (29.10) that a resistance R transforms into a capacitance $C = K/R$, while an inductance L transforms into a resistance $R = KL$. The elimination of inductors in favor of resistors is the key purpose of the Bruton transform method. Applying the Bruton transform to a prototype circuit capacitance C gives

$$Z_T(s) = \frac{K}{s} \cdot \frac{1}{sC} = \frac{K}{s^2C} = \frac{1}{s^2D} \tag{29.11}$$

where $D = C/K$ is the parameter value of a new component produced by the transformation, which is usually referred to as a frequency-dependent negative resistance (FDNR). This name results from the fact that the sinusoidal steady-state impedance $Z_T(j\omega) = -(1/\omega^2 D)$ is frequency-dependent, negative, and real, hence, resistive. In practice, the FDNR elements are realized by RC-active subnetworks using op amps, normally two per FDNR. Figure 29.11(a) and (b) shows the sequence of circuit changes involved in transforming from a third-order LCR prototype ladder to an FDNR circuit. Figure 29.11(c) gives an RC-active realization for the FDNR based on the use of a GIC, introduced in the previous subsection.

Active Filters for ICs

It was noted earlier that the advent of the IC op amp made the RC-active filter a practical reality. A typical state-of-the-art 1960–70s active filter would involve a printed circuit board-mounted circuit comprising discrete passive components together with IC op amps. Also appearing at this time were hybrid implementations, which involve special-purpose discrete components and op amp ICs interconnected on a ceramic or glass substrate. It was recognized, however, that there were considerable benefits to be had from producing an all-IC active filter.

Production of a quality on-chip capacitor involves substantial chip area, so the scaling techniques referred to earlier must be used to keep capacitance values down to the low picofarad range. The consequence of this is that, unfortunately, the circuit resistance values become proportionately large so that, again, there is a chip-area problem. The solution to this dilemma emerged in the late 1970s/early 1980s with the advent of the switched-capacitor (SC) active filter. This device, a development of the active-RC filter that is specifically intended for use in IC form, replaces prototype circuit resistors with arrangements of switches and capacitors that can be shown to simulate resistances, under certain circumstances. The great merit of the scheme is that the values of the capacitors involved in this process of resistor simulation are inversely proportional to the values of the prototype resistors; thus, the final IC structure involves principal and switched capacitors that are

small in magnitude and hence ideal for IC realization. A good account of SC filters is given, for example, in Schaumann et al. [1990] and in Taylor and Huang [1997]. Commonly encountered techniques for SC filter design are based on the two major design styles (biquads and ladder simulation) that have been introduced in this section.

Many commercial IC active filters are based on SC techniques, and it is also becoming usual to find custom and semicustom IC design systems that include active filter modules as components within a macrocell library that the system-level design can simply invoke where analog filtering is required within an all-analog or mixed-signal analog/digital system.

Defining Terms

Active filter: An electronic filter whose design includes one or more active devices.

Biquad: An active filter whose transfer function comprises a ratio of second-order numerator and denominator polynomials in the frequency variable.

Electronic filter: An electronic circuit designed to transmit some range of signal frequencies while rejecting others. Phase and time-domain specifications may also occur.

Sensitivity: A measure of the extent to which a given circuit performance measure is affected by a given component within the circuit.

Related Topic

27.2 Applications

References

A. Antoniou, "Realization of gyrators using operational amplifiers and their use in RC-active network synthesis," *Proc. IEE*, vol. 116, pp. 1838–1850, 1969.

P. Bowron and F.W. Stephenson, *Active Filters for Communications and Instrumentation,* New York: McGraw-Hill, 1979.

L.T. Bruton, "Network transfer functions using the concept of frequency dependent negative resistance," *IEEE Trans.,* vol. CT-18, pp. 406–408, 1969.

W.J. Kerwin, L.P. Huelsman, and R.W. Newcomb, "State-variable synthesis for insensitive integrated circuit transfer functions," *IEEE J.,* vol. SC-2, pp. 87–92, 1967.

H.J. Orchard, "Inductorless filters," *Electron. Letters*, vol. 2, pp. 224–225, 1966.

P.R. Sallen and E.L. Key, "A practical method of designing RC active filters," *IRE Trans.,* vol. CT-2, pp. 74–85, 1955.

R. Schaumann, M.S. Ghausi, and K.R. Laker, *Design of Analog Filters*, Englewood Cliffs, N.J: Prentice-Hall, 1990.

J.T. Taylor and Q. Huang, *CRC Handbook of Electrical Filters,* Boca Raton, Fla.: CRC Press, 1997.

A. Waters, *Active Filter Design*, New York: Macmillan, 1991.

Further Information

Tabulations of representative standard filter specification functions appear in the sources in the References by Schaumann et al. [1990] and Bowron and Stephenson [1979], but more extensive tabulations, including prototype passive filter component values, are given in A. I. Zverev, *Handbook of Filter Synthesis* (New York: John Wiley, 1967). More generally, the Schaumann text provides an admirable, up-to-date coverage of filter design with an extensive list of references as does Taylor and Huang [1997].

The field of active filter design remains active, and new developments appear in *IEEE Transactions on Circuits and Systems* and *IEE Proceedings Part G (Circuits and Systems)*. The IEE publication *Electronic Letters* provides for short contributions. A number of international conferences (whose proceedings can be borrowed through technical libraries) feature active filter and related sessions, notably the *IEEE International Symposium on Circuits and Systems* (ISCAS) and the *European Conference on Circuit Theory and Design* (ECCTD).

29.2 Realization

J. W. Steadman and B. M. Wilamowski

After the appropriate low-pass form of a given **filter** has been synthesized, the designer must address the realization of the filter using **operational amplifiers**. If the required filter is not low-pass but high-pass, bandpass, or bandstop, transformation of the prototype function is also required [Budak, 1974; Van Valkenburg, 1982]. While a detailed treatment of the various transformations is beyond the scope of this work, most of the filter designs encountered in practice can be accomplished using the techniques given here.

When the desired filter function has been determined, the corresponding electronic circuit must be designed. Many different circuits can be used to realize any given transfer function. For purposes of this handbook, we present several of the most popular types of realizations. Much more detailed information on various circuit realizations and the advantages of each may be found in the literature, in particular Van Valkenburg [1982], Huelseman and Allen [1980], and Chen [1986]. Generally the design trade-offs in making the choice of circuit to be used for the realization involve considerations of the number of elements required, the sensitivity of the circuit to changes in component values, and the ease of tuning the circuit to given specifications. Accordingly, limited information is included about these characteristics of the example circuits in this section.

Each of the circuits described here is commonly used in the realization of **active filters**. When implemented as shown and used in the appropriate gain and bandwidth specifications of the amplifiers, they will provide excellent performance. Computer-aided filter design programs are available which simplify the process of obtaining proper element values and simulation of the resulting circuits [Krobe et al., 1989; Wilamowski et al., 1992].

Transformation from Low-Pass to Other Filter Types

To obtain a high-pass, bandpass, or bandstop filter function from a low-pass prototype, one of two general methods can be used. In one of these, the circuit is realized and then individual circuit elements are replaced by other elements or subcircuits. This method is more useful in **passive filter** designs and is not discussed further here. In the other approach, the transfer function of the low-pass prototype is transformed into the required form for the desired filter. Then a circuit is chosen to realize the new filter function. We give a brief description of the transformation in this section, then give examples of circuit realizations in the following sections.

Low-Pass to High-Pass Transformation

Suppose the desired filter is, for example, a high-pass Butterworth. Begin with the low-pass Butterworth transfer function of the desired order and then *transform* each pole of the original function using the formula

$$\frac{1}{S - S_j} \rightarrow \frac{Hs}{s - s_j} \tag{29.12}$$

which results in one complex pole and one zero at the origin for each pole in the original function. Similarly, each zero of the original function is transformed using the formula

$$S - S_j \rightarrow \frac{s - s_j}{Hs} \tag{29.13}$$

which results in one zero on the imaginary axis and one pole at the origin. In both equations, the scaling factors used are

$$H = \frac{1}{S_j} \quad \text{and} \quad s_j = \frac{\omega_0}{S_j} \tag{29.14}$$

where ω_0 is the desired cut-off frequency in radians per second.

Low-Pass to Bandpass Transformation

Begin with the low-pass prototype function in factored, or *pole-zero*, form. Then each pole is transformed using the formula

$$\frac{1}{S - S_j} \rightarrow \frac{Hs}{(s - s_1)(s - s_2)} \tag{29.15}$$

resulting in one zero at the origin and two conjugate poles. Each zero is transformed using the formula

$$S - S_j \rightarrow \frac{(s - s_1)(s - s_2)}{Hs} \tag{29.16}$$

resulting in one pole at origin and two conjugate zeros. In Eqs. (29.15) and (29.16)

$$H = -B; \quad s_{1,2} = \omega_c\left(\alpha \pm \sqrt{\alpha^2 - 1}\right); \text{ and } \alpha = \frac{BS_j}{2\omega_c} \tag{29.17}$$

where ω_c is the center frequency and B is the bandwidth of the bandpass function.

Low-Pass to Bandstop Transformation

Begin with the low-pass prototype function in factored, or pole-zero, form. Then each pole is transformed using the formula

$$\frac{1}{S - S_j} \rightarrow \frac{H(s - s_1)(s - s_2)}{(s - s_3)(s - s_4)} \tag{29.18}$$

transforming each pole into two zeros on the imaginary axis and into two conjugate poles. Similarly, each zero is transformed into two poles on the imaginary axis and into two conjugate zeros using the formula

$$S - S_j \rightarrow \frac{(s - s_3)(s - s_4)}{H(s - s_1)(s - s_2)} \tag{29.19}$$

where

$$H = \frac{1}{S_j}; \; s_{1,2} = \pm j\omega_c; \; s_{3,4} = \omega_c\left(\beta \pm \sqrt{\beta^2 - 1}\right); \text{ and } \beta = \frac{B}{2\omega_c S_j} \tag{29.20}$$

Once the desired transfer function has been obtained through obtaining the appropriate low-pass prototype and transformation, if necessary, to the associated high-pass, bandpass or bandstop function, all that remains is to obtain a circuit and the element values to realize the transfer function.

Circuit Realizations

Various electronic circuits can be found to implement any given transfer function. Cascade filters and ladder filters are two of the basic approaches for obtaining a practical circuit. Cascade realizations are much easier to find and to tune, but ladder filters are less sensitive to element variations. In cascade realizations, the transfer function is simply factored into first- and second-order parts. Circuits are built for the individual parts and then cascaded to produce the overall filter. For simple to moderately complex filter designs, this is the most common method, and the remainder of this section is devoted to several examples of the circuits used to obtain

the first- and second-order filters. For very high-order transfer functions, ladder filters should be considered, and further information can be obtained by consulting the literature.

In order to simplify the circuit synthesis procedure, very often ω_0 is assumed to be equal to one and then after a circuit is found, the values of all capacitances in the circuit are divided by ω_0. In general, the following magnitude and frequency transformations are allowed:

$$R_{\text{new}} = K_M R_{\text{old}} \text{ and } C_{\text{new}} = \frac{1}{K_F K_M} C_{\text{old}} \tag{29.21}$$

where K_M and K_F are magnitude and frequency scaling factors, respectively.

Cascade filter designs require the transfer function to be expressed as a product of first- and second-order terms. For each of these terms a practical circuit can be implemented. Examples of these circuits are presented in Figs. 29.12–29.22. In general the following first- and second-order terms can be distinguished:

(a) First-order low-pass:

$$T(s) = \frac{H\omega_0}{s + \omega_0}$$

Assumption : $r_1 = 1$

$$c_1 = \frac{1}{\omega_0} \qquad r_2 = |H|\,\omega_0$$

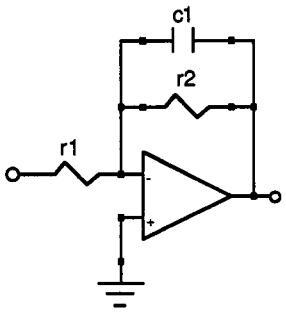


FIGURE 29.12 First-order low-pass filter.

This filter is inverting, i.e., H must be negative, and the scaling factors shown in Eq. (29.21) should be used to obtain reasonable values for the components.

(b) First-order high-pass:

$$T(s) = \frac{Hs}{s + \omega_0}$$

Assumption : $r_1 = 1$

$$c_1 = \frac{1}{\omega_0} \qquad r_2 = |H|$$

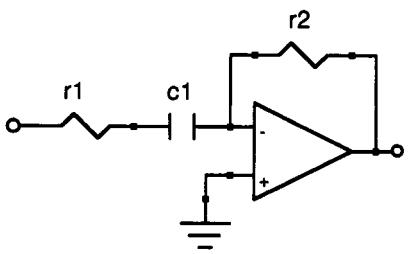


FIGURE 29.13 First-order high-pass filter.

This filter is inverting, i.e., H must be negative, and the scaling factors shown in Eq. (29.21) should be used to obtain reasonable values for the components.

While several passive realizations of first-order filters are possible (low-pass, high-pass, and lead-lag), the active circuits shown here are inexpensive and avoid any loading of the other filter sections when the individual circuits are cascaded. Consequently, these circuits are preferred unless there is some reason to avoid the use of the additional operational amplifier. Note that a second-order filter can be realized using one operational amplifer as shown in the following paragraphs, so it is common practice to choose even-order transfer functions, thus avoiding the use of any first-order filters.

(c) There are several second-order low-pass circuits:

$$T(s) = \frac{H\omega_0^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

$$\text{Assumption}: \quad r_1 = r_2 = 1$$

$$c_1 = \frac{2Q}{\omega_0} \qquad c_2 = \frac{1}{2Q\omega_0}$$

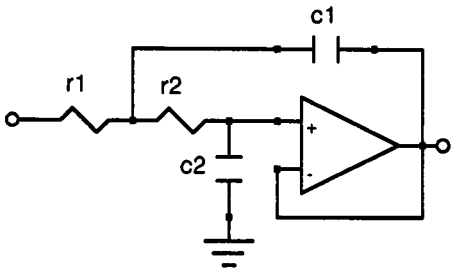


FIGURE 29.14 Second-order low-pass Sallen-Key filter.

This filter is noninverting and unity gain, i.e., H must be one, and the scaling factors shown in Eq. (29.21) should be used to obtain reasonable element values. This is a very popular filter for realizing second-order functions because it uses a minimum number of components and since the operation amplifier is in the unity gain configuration it has very good bandwidth.

Another useful configuration for second-order low-pass filters uses the operational amplifier in its inverting "infinite gain" mode as shown in Fig. 29.15.

$$T(s) = \frac{H\omega_0^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

$$\text{Assumption}: \quad r_1 = r_2 = r_3 = 1$$

$$c_1 = \frac{3Q}{\omega_0} \qquad c_2 = \frac{1}{3Q\omega_0}$$

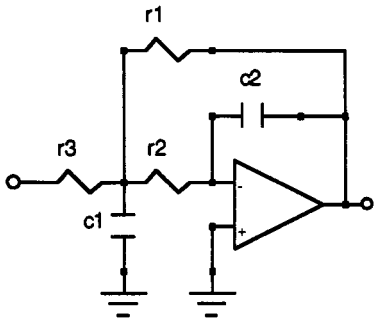


FIGURE 29.15 Second-order low-pass filter using the inverting circuit.

This circuit has the advantage of relatively low sensitivity of ω_0 and Q to variations in component values. In this configuration the operational amplifier's gain-bandwidth product may become a limitation for high-Q and high-frequency applications [Budak, 1974]. There are several other circuit configurations for low-pass filters. The references given at the end of the section will guide the designer to alternatives and the advantages of each.

(d) Second-order high-pass filters may be designed using circuits very much like those shown for the low-pass realizations. For example, the Sallen-Key low-pass filter is shown in Fig. 29.16.

$$T(s) = \frac{Hs^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

$$\text{Assumption}: \quad r_3 = 1$$

$$c_1 = c_2 = 1$$

$$r_1 = r_2 = \frac{1}{\omega_0} \qquad r_4 = 2 - \frac{1}{Q}$$

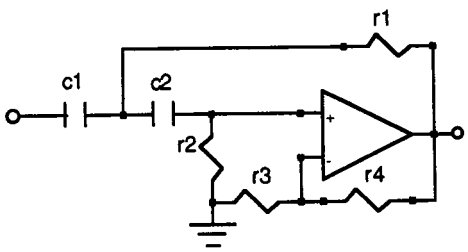


FIGURE 29.16 A second-order high-pass Sallen-Key filter.

As in the case of the low-pass Sallen-Key filter, this circuit is noninverting and requires very little gain from the operational amplifier. For low to moderate values of Q, the **sensitivity functions** are reasonable and the circuit performs well.

The inverting *infinite gain* high-pass circuit is shown in Fig. 29.17 and is similar to the corresponding low-pass circuit.

$$T(s) = \frac{Hs^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

Assumption: $r_1 = 1$

$$r_2 = 9Q^2 \quad c_1 = c_2 = c_3 = \frac{1}{3Q^2}$$

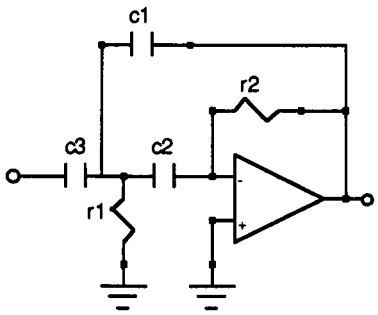


FIGURE 29.17 An inverting second-order high-pass circuit.

This circuit has relatively good sensitivity figures. The principal limitation occurs with high-Q filters since this requires a wide spread of resistor values.

Both low-pass and high-pass frequency response circuits can be achieved using three operational amplifier circuits. Such circuits have some sensitivity function and tuning advantages but require far more components. These circuits are used in the sections describing bandpass and bandstop filters. The designer wanting to use the three-operational-amplifier realization for low-pass or high-pass filters can easily do this using simple modifications of the circuits shown in the following sections.

(e) Second-order bandpass circuits may be realized using only one operational amplifier. The Sallen-Key filter shown in Fig. 29.18 is one such circuit.

$$T(s) = \frac{H\frac{\omega_0}{Q}s}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

Assumption: $c_1 = c_2 = 1; \quad r_5 = 1$

$$r_2 = r_3 = \frac{\sqrt{2}}{\omega_0} \qquad r_1 = \frac{\frac{4Q}{\sqrt{2}} - 1}{H}$$

$$r_4 = \frac{\frac{4Q}{\sqrt{2}} - 1}{\frac{4Q}{\sqrt{2}} - 1 - H} \qquad r_6 = 3 - \frac{\sqrt{2}}{\omega_0}$$

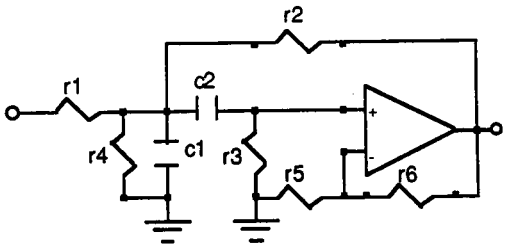


FIGURE 29.18 A Sallen-Key bandpass filter.

This is a noninverting amplifier which works well for low- to moderate-Q filters and is easily tuned [Budak, 1974]. For high-Q filters the sensitivity of Q to element values becomes high, and alternative circuits are recommended. One of these is the bandpass version of the inverting amplifier filter as shown in Fig. 29.19.

$$T(s) = \frac{H\frac{\omega_0}{Q}s}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

$$\text{Assumption}: \quad c_1 = c_2 = \frac{1}{2Q\omega_0}$$

$$r_1 = \frac{2Q^2}{H} \qquad r_2 = 4Q^2 \qquad r_3 = \frac{1}{1 - \frac{H}{2Q^2}}$$

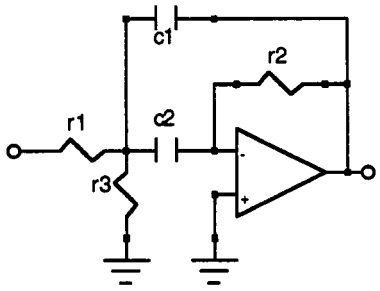
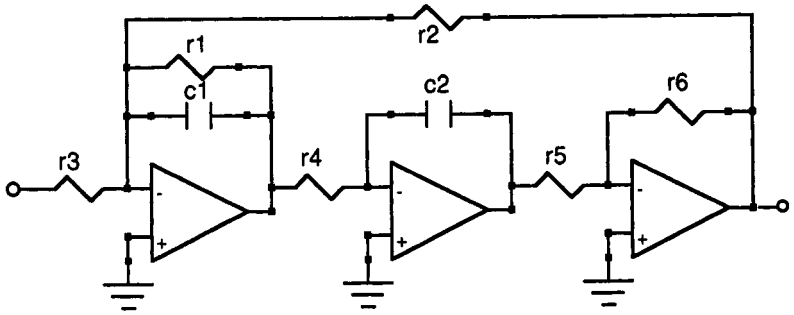


FIGURE 29.19 The inverting amplifier bandpass filter.

This circuit has few components and relatively small sensitivity of ω_0 and Q to variations in element values. For high-Q circuits, the range of resistor values is quite large as r_1 and r_2 are much larger than r_3.

When ease of tuning and small sensitivities are more important than the circuit complexity, the three-operational-amplifier circuit of Fig. 29.20 may be used to implement the bandpass transfer function.



$$T(s) = \frac{H\frac{\omega_0}{Q}s}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2} \qquad c_1 = c_2 = \frac{1}{\omega_0} \qquad r_1 = Q \qquad r_2 = r_4 = r_5 = r_6 = 1 \qquad r_3 = \frac{Q}{|H|}$$

FIGURE 29.20 The three-operational-amplifier bandpass filter.

The filter as shown in Fig. 29.20 is inverting. For a noninverting realization, simply take the output from the middle amplifier rather than the right one. This same configuration can be used for a three-operational-amplifier low-pass filter by putting the input into the summing junction of the middle amplifier and taking the output from the left operational amplifier. Note that Q may be changed in this circuit by varying r_1 and that this will not alter ω_0. Similarly, ω_0 can be adjusted by varying c_1 or c_2 and this will not change Q. If only variable resistors are to be used, the filter can be tuned by setting ω_0 using any of the resistors other than r_1 and then setting Q using r_1.

(f) Second-order bandstop filters are very useful in rejecting unwanted signals such as line noise or carrier frequencies in instrumentation applications. Such filters are implemented with methods very similar to the bandpass filters just discussed. In most cases, the frequency of the zeros is to be the same as the frequency of the poles. For this application, the circuit shown in Fig. 29.21 can be used.

$$T(s) = \frac{H(s^2 + \omega_z^2)}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2}$$

Assumption : $c_1 = c_2 = 1$

$$r_1 = \frac{1}{2Q\omega_0} \qquad r_3 = \frac{1}{Q\omega_0} \qquad r_2 = r_4 = \frac{2Q}{\omega_0}$$

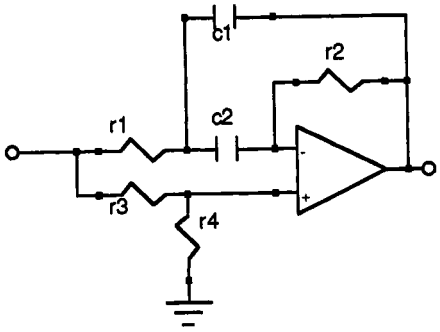
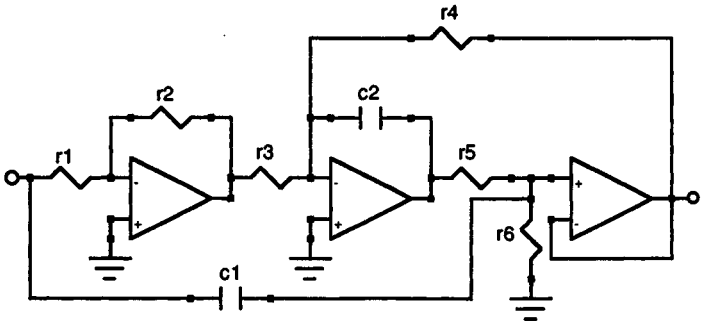


FIGURE 29.21 A single operational-amplifier bandstop filter.

The primary advantage of this circuit is that it requires a minimum number of components. For applications where no tuning is required and the Q is low, this circuit works very well. When the bandstop filter must be tuned, the three-operational-amplifier circuit is preferable.



$$T(s) = \frac{H(s^2 + \omega_z^2)}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2} \quad c_1 = c_2 = \frac{1}{\omega_0} \quad r_1 = 1 \quad r_2 = H \quad r_5 = r_6 = 2Q \quad r_3 = \frac{H\omega_0^2}{2Q\omega_z^2} \quad r_4 = \frac{1}{2Q}$$

FIGURE 29.22 A three-operational-amplifier bandstop filter.

The foregoing circuits provide a variety of useful first- and second-order filters. For higher-order filters, these sections are simply cascaded to realize the overall transfer function desired. Additional detail about these circuits as well as other circuits used for active filters may be found in the references.

Defining Terms

Active filter: A filter circuit which uses active components, usually operational amplifiers.

Filter: A circuit which is designed to be frequency selective. That is, the circuit will emphasize or "pass" certain frequencies and attenuate or "stop" others.

Operational amplifier: A very high-gain differential amplifier used in active filter circuits and many other applications. These monolithic integrated circuits typically have such high gain, high input impedance, and low output impedance that they can be considered "ideal" when used in active filters.

Passive filter: Afilter circuit which uses only passive components, i.e., resistors, inductors, and capacitors. These circuits are useful at higher frequencies and as prototypes for ladder filters that are active.

Sensitivity function: A measure of the fractional change in some circuit characteristic, such as center frequency, to variations in a circuit parameter, such as the value of a resistor. The sensitivity function is normally defined as the partial derivative of the desired circuit characteristic with respect to the element value and is usually evaluated at the nominal value of all elements.

Related Topics

10.3 The Ideal Linear-Phase Low-Pass Filter • 27.1 Ideal and Practical Models

References

A. Budak, *Passive and Active Network Analysis and Synthesis*, Boston: Houghton Mifflin, 1974.

W.K. Chen, *Passive and Active Filters, Theory and Implementations*, New York: Wiley, 1986.

L.P. Huelseman and P.E. Allen, *Introduction to the Theory and Design of Active Filters*, New York: McGraw-Hill, 1980.

M.R. Krobe, J. Ramirez-Angulo, and E. Sanchez-Sinencio, "FIESTA—A filter educational synthesis teaching aid," *IEEE Trans. on Education*, vol. 12, no. 3, pp. 280–286, August 1989.

M.E. Van Valkenburg, *Analog Filter Design*, New York: Holt, Rinehart and Winston, 1982.

B.M. Wilamowski, S.F. Legowski, and J.W. Steadman, "Personal computer support for teaching analog filter analysis and design," *IEEE Trans. on Education*, vol. 35, no. 4, November 1992.

Further Information

The monthly journal *IEEETransactions on Circuits and Systems* is one of the best sources of information on new active filter functions and associated circuits.

The British journal *Electronics Letters* also often publishes articles about active circuits.

The *IEEETransactions on Education* has carried articles on innovative approaches to active filter synthesis as well as computer programs for assisting in the design of active filters.

29.3 Generalized Impedance Convertors and Simulated Impedances

James A. Svoboda

The problem of designing a circuit to have a given transfer function is called filter design. This problem can be solved using passive circuits, that is, circuits consisting entirely of resistors, capacitors, and inductors. Further, these passive filter circuits can be designed to have some attractive properties. In particular, passive filters can be designed so that the transfer function is relatively insensitive to variations in the values of the resistances, capacitances, and inductances. Unfortunately, passive circuits contain inductors. Inductors are frequently large, heavy, expensive, and nonlinear.

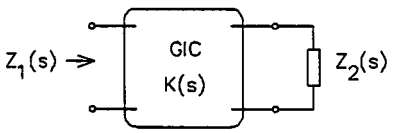


FIGURE 29.23 The GIC converts the impedance $Z_2(s)$ to the impedance $Z_1(s)$.

Generalized impedance convertors (GIC) are electronic circuits used to convert one impedance into another impedance [Bruton, 1981; Van Valkenburg, 1982]. GICs provide a way to get the advantages of passive circuits without the disadvantages of inductors. Figure 29.23 illustrates the application of a GIC. The GIC converts the impedance $Z_2(s)$ to the impedance $Z_1(s)$. The impedances are related by

$$Z_1(s) = K(s)Z_2(s) \tag{29.22}$$

The function $K(s)$ is called the conversion function or, more simply, the gain of the GIC.

Figure 29.24 shows two ways to implement a GIC using operational amplifiers (op amps). The GIC shown in Fig. 29.24a has a gain given by

$$K(s) = -\frac{Z_A(s)}{Z_B(s)} \tag{29.23}$$

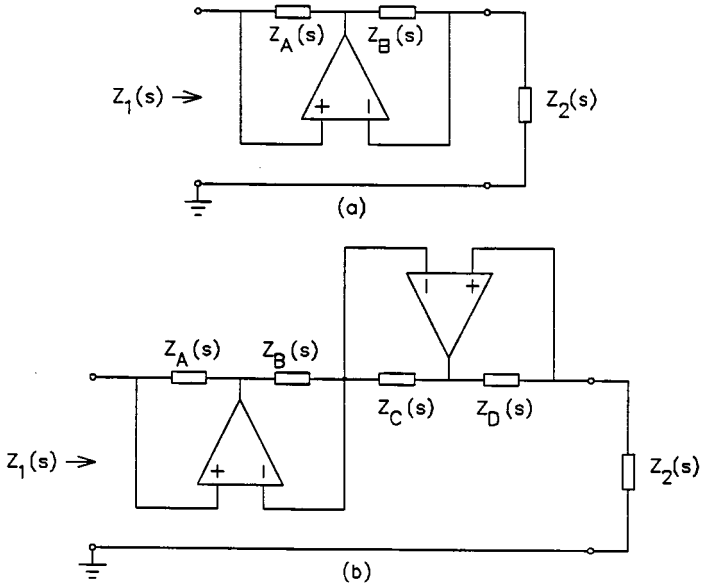


FIGURE 29.24 (a) An inverting GIC and (b) a noninverting GIC.

This GIC is called an inverting GIC because $K(s)$ is negative.

A **negative resistor** is an electronic circuit that acts like a resistor having a negative value of resistance. The inverting GIC can be used to design a negative resistor by taking $Z_A(s) = R_A$, $Z_B(s) = R_B$, and $Z_2(s) = R_2$. Figure 29.25(a) shows the op amp circuit that implements a negative resistor, and Fig. 29.25(b) shows the equivalent circuit. The resistance of the negative resistor is given by

$$R = -\frac{R_A}{R_B} R_2 \tag{29.24}$$

Figure 29.24(b) shows another op amp circuit that implements a GIC. The gain of this GIC is given by

$$K(s) = \frac{Z_A(s)Z_C(s)}{Z_B(s)Z_D(s)} \tag{29.25}$$

This GIC is called a noninverting GIC because $K(s)$ is positive.

A **simulated inductor** is circuit consisting of resistors, capacitors, and amplifiers that acts like an inductor. The noninverting GIC can be used to design a simulated inductor by taking $Z_A(s) = R_A$, $Z_B(s) = R_B$, $Z_C(s) = R_C$, $Z_D(s) = 1/(sC_D)$, and $Z_2(s) = R_2$. Figure 29.25(c) shows the op amp circuit that implements a simulated inductor, and Fig. 29.25(d) shows the equivalent circuit. The inductance of the simulated inductor is given by

$$L = \frac{R_A R_C C_D}{R_B} R_2 \tag{29.26}$$

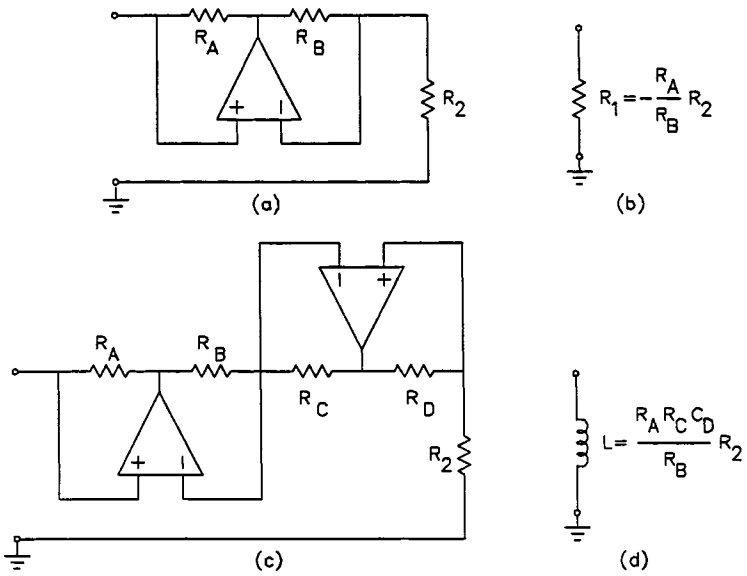


FIGURE 29.25 (a) A grounded negative resistor and (b) its equivalent circuit. (c) A grounded simulated inductor and (d) its equivalent circuit.

Notice that one node of the negative resistor shown in Fig. 29.25(b) and of the simulated inductor shown in Fig. 29.25(d) is grounded. This ground is the ground of the power supplies used to bias the op amp. Op amp circuits implementing floating negative resistors and simulated inductors are more difficult to design [Reddy, 1976]. Floating negative resistors and simulated inductors can be more easily designed using an electronic device called a **current conveyor**. The symbol for the current conveyor is shown in Fig. 29.26. The terminal voltages and currents of the "second-generation" current conveyor [Sedra and Smith, 1971] are represented by

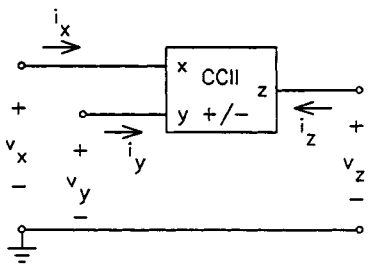


FIGURE 29.26 A CCII current conveyor.

$$\begin{pmatrix} i_y \\ v_x \\ i_z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \pm 1 & 0 \end{pmatrix} \begin{pmatrix} v_y \\ i_x \\ v_z \end{pmatrix} \tag{29.27}$$

There are two kinds of second-generation current conveyor, corresponding to the two possible signs of the ±1 entry in the third row of Eq. (29.27). The + indicates a CCII$^+$ current conveyor while the – indicates a CCII$^-$ current conveyor.

Current conveyors are related to **transimpedance amplifiers** [Svoboda, 1991]. Figure 29.27(a) indicates that a transimpedance amplifier consists of a CCII$^+$ current conveyor and a voltage buffer. Several transimpedance amplifiers, e.g., the AD844, AD846, and AD811, are commercially available. Figure 29.27(b) shows that a CCII$^-$ current conveyor can be constructed from two CCII$^+$ current conveyors.

Figure 29.28(a) presents a current conveyor circuit that implements a floating negative resistor. The resistance of the negative resistor is given simply as

$$R = -R_2 \tag{29.28}$$

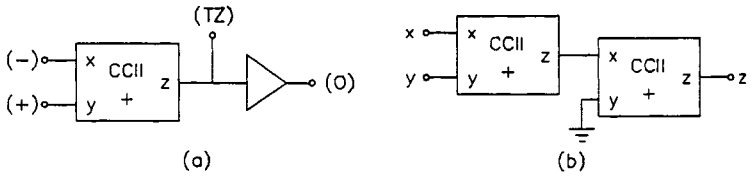


FIGURE 29.27 (a) A transimpedance amplifier consists of a CCII+ current conveyor and a voltage buffer. (b) A CCII− implemented using two CCII+ current conveyors.

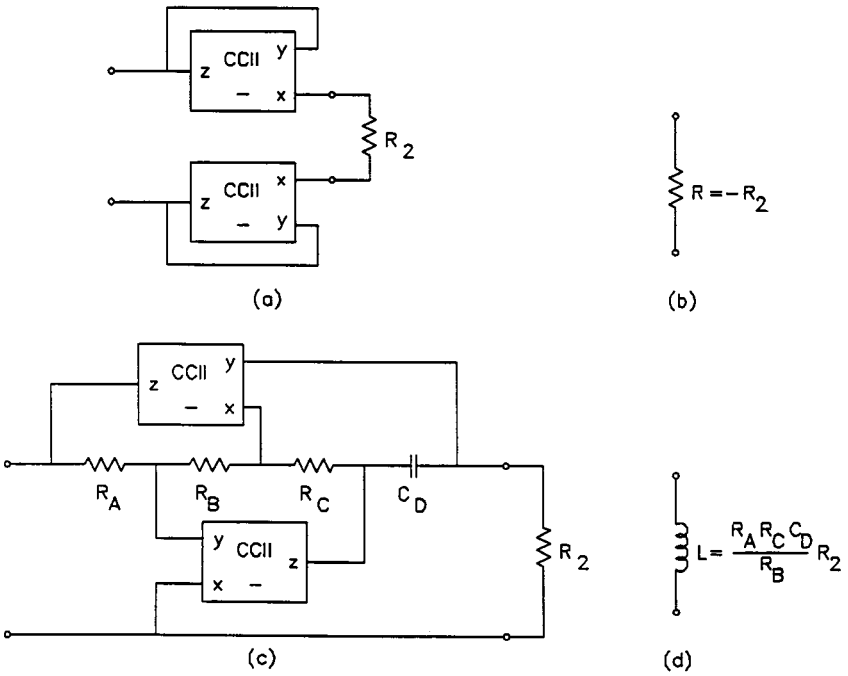


FIGURE 29.28 (a) A floating negative resistor and (b) its equivalent circuit. (c) A floating simulated inductor and (d) its equivalent circuit.

Figure 29.28(b) shows the equivalent circuit of the current conveyor negative resistor. Notice that in Fig. 29.28(b) neither node is required to be ground, in contrast to the equivalent circuit for the op amp negative resistor in Fig. 29.25(b).

Figure 29.28(c) shows a current conveyor circuit that implements a floating simulated inductor. The inductance of this simulated inductor is given by

$$L = \frac{R_A R_C C_D}{R_B} R_2 \tag{29.29}$$

Figure 29.28(d) shows the equivalent circuit of the current conveyor simulated inductor. The current conveyor circuit can simulate a floating inductor, so neither node of the equivalent inductor is required to be grounded.

Figure 29.29 illustrates an application of simulated impedances. The circuit shown in Fig. 29.29(a) implements a voltage-controlled current source (VCCS). This particular VCCS has the advantage of perfect regulation. In other words, the output current, i_o, is completely independent of the load resistance, R_L. The circuit in Fig. 29.29(a) requires a negative resistor, the resistor labeled $-R$. Since one node of this resistor is grounded,

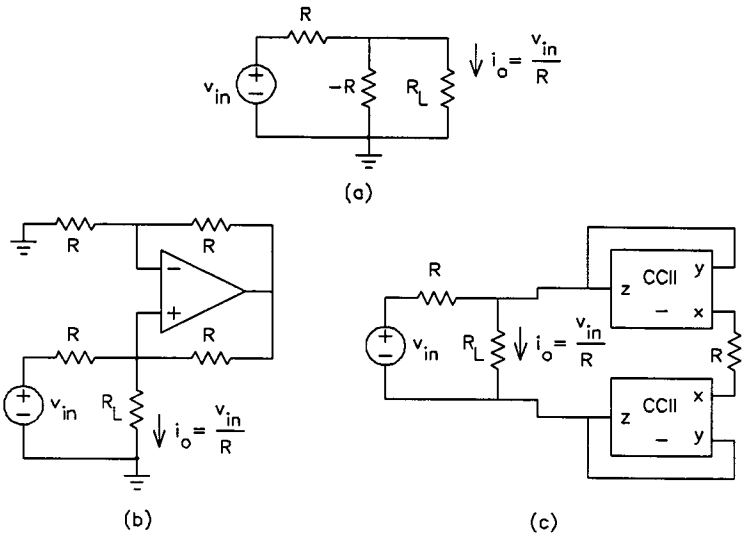


FIGURE 29.29 Three versions of a VCCS: (a) using a negative resistor, (b) using an op amp, and (c) using current conveyors.

this resistor can be implemented using the op amp negative resistor shown in Fig. 29.25(a). The resulting circuit is shown in Fig. 29.29(b).

In Fig. 29.29(a), one node of the load resistor is grounded. As a consequence, one node of the negative resistor was grounded and it was appropriate to use the op amp negative resistor. Sometimes a VCCS is needed to cause a current in an ungrounded load resistance. In this case the negative resistor must also be ungrounded so the current conveyor negative resistor is used. In Fig. 29.29(c) the current conveyor negative resistor is used to implement a VCCS that supplies current to an ungrounded resistor R_L.

Figure 29.30 illustrates the application of a simulated inductor. The circuit shown in Fig. 29.30(a) is a low-pass filter. The transfer function of this filter is

$$\frac{V_o(s)}{V_{\text{in}}(s)} = \frac{\dfrac{1}{LC}}{s^2 + \dfrac{R}{L}s + \dfrac{1}{LC}} \tag{29.30}$$

The filter in Fig. 29.30(a) contains an inductor. This inductor can be implemented as a simulated inductor. Since neither node of the inductor is grounded, it is necessary to use the current conveyor simulated inductor. The resulting circuit is shown in Fig. 29.30(b). The inductance of the simulated inductor is given by Eq. (29.29). Substituting this equation into Eq. (29.30) gives the transfer function of the circuit in Fig. 29.30(b)

$$\frac{V_o(s)}{V_{\text{in}}(s)} = \frac{\dfrac{R_B}{R_A R_C R_2 C_D C}}{s^2 + \dfrac{RR_B}{R_A R_C R_2 C_D}s + \dfrac{R_B}{R_A R_C R_2 C_D C}} \tag{29.31}$$

Similarly, high-pass, bandpass, and notch filters can be designed by rearranging the resistor, capacitor, and inductor in Fig. 29.30(a) to get the desired transfer function and then simulating the inductor. When the inductor is grounded, it can be simulated using the op amp–simulated inductor, but when the inductor is floating, the current conveyor–simulated inductor must be used.

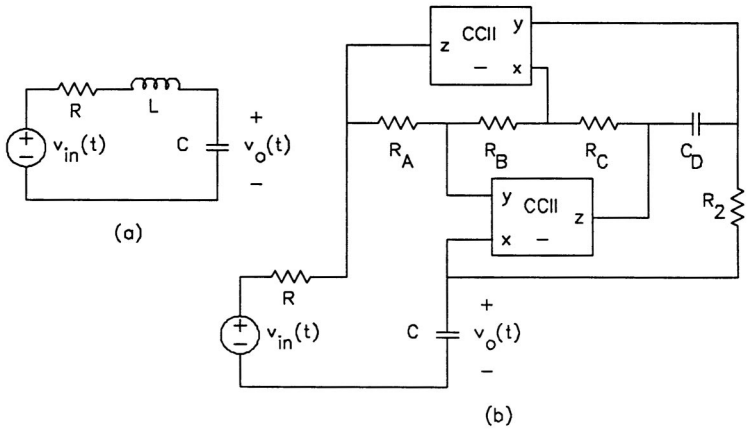


FIGURE 29.30 (a) A low-pass RLC filter and (b) the same low-pass filter implemented using a floating simulated inductor.

Defining Terms

Current conveyor: An electronic device represented by Fig. 29.26 and Eq. (29.27).
Generalized impedance convertors (GIC): Electronic circuits used to convert one impedance into another impedance.
Negative resistor: An electronic circuit that acts like a resistor having a negative value of resistance.
Transimpedance amplifier: An amplifier consisting of a CCII$^+$ current conveyor and a voltage buffer.
Simulated inductor: A circuit consisting of resistors capacitors and amplifiers that acts like an inductor.

Related Topic

27.1 Ideal and Practical Models

References

L. T. Bruton, *RC-Active Circuits,* Englewood Cliffs, N.J.: Prentice-Hall, 1981.
M. A. Reddy, "Some new operational-amplifier circuits for the realization of the lossless floating inductor," *IEEE Transactions on Circuits and Systems,* vol. CAS-23, pp. 171–173, 1976.
A. Sedra and K. C. Smith, "A second generation current conveyor and its application," *IEEE Transactions on Circuit Theory,* vol. CT-17, pp. 132–134, 1970.
J. A. Svoboda, "Applications of a commercially available current conveyor," *International J. of Electronics,* 70, no. 1, pp. 159–164, 1991.
M. E. Van Valkenburg, *Analog Filter Design,* New York: Holt, Rinehart and Winston, 1982.

Further Information

Additional information regarding current conveyors can be found in *Analogue IC Design: The Current Mode Approach* edited by Toumazou, Lidgey, and Haigh. *The Circuits and Filters Handbook* edited by Wai-Kai Chen provides background on circuit design in general and on filters in particular. Several journals, including *IEEE Transactions on Circuits and Systems, The International Journal of Electronics,* and *Electronic Letters,* report on advances in filter design.

Rajashekara, K., Bhat, A.K.S., Bose, B.K. “Power Electronics”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

30

Power Electronics

30.1 Power Semiconductor Devices
Thyristor and Triac • Gate Turn-Off Thyristor (GTO) • Reverse-Conducting Thyristor (RCT) and Asymmetrical Silicon- Controlled Rectifier (ASCR) • Power Transistor • Power MOSFET • Insulated-Gate Bipolar Transistor (IGBT) • MOS Controlled Thyristor (MCT)

30.2 Power Conversion
AC-DC Converters • Cycloconverters • DC-to-AC Converters • DC-DC Converters

30.3 Power Supplies
DC Power Supplies • AC Power Supplies • Special Power Supplies

30.4 Converter Control of Machines
Converter Control of DC Machines • Converter Control of AC Machines

Kaushik Rajashekara
Delphi Energy & Engine Management Systems

Ashoka K. S. Bhat
University of Victoria

Bimal K. Bose
University of Tennessee

30.1 Power Semiconductor Devices

Kaushik Rajashekara

The modern age of power electronics began with the introduction of thyristors in the late 1950s. Now there are several types of power devices available for high-power and high-frequency applications. The most notable power devices are gate turn-off thyristors, power Darlington transistors, power MOSFETs, and insulated-gate bipolar transistors (IGBTs). Power semiconductor devices are the most important functional elements in all power conversion applications. The power devices are mainly used as switches to convert power from one form to another. They are used in motor control systems, uninterrupted power supplies, high-voltage dc transmission, power supplies, induction heating, and in many other power conversion applications. A review of the basic characteristics of these power devices is presented in this section.

Thyristor and Triac

The thyristor, also called a silicon-controlled rectifier (SCR), is basically a four-layer three-junction *pnpn* device. It has three terminals: anode, cathode, and gate. The device is turned on by applying a short pulse across the gate and cathode. Once the device turns on, the gate loses its control to turn off the device. The turn-off is achieved by applying a **reverse voltage** across the anode and cathode. The thyristor symbol and its volt-ampere characteristics are shown in Fig. 30.1. There are basically two classifications of thyristors: converter grade and inverter grade. The difference between a converter-grade and an inverter-grade thyristor is the low turn-off time (on the order of a few microseconds) for the latter. The converter-grade thyristors are slow type and are used in natural commutation (or phase-controlled) applications. Inverter-grade thyristors are used in forced commutation applications such as dc-dc choppers and dc-ac inverters. The inverter-grade thyristors are turned off by forcing the current to zero using an external commutation circuit. This requires additional commutating components, thus resulting in additional losses in the inverter.

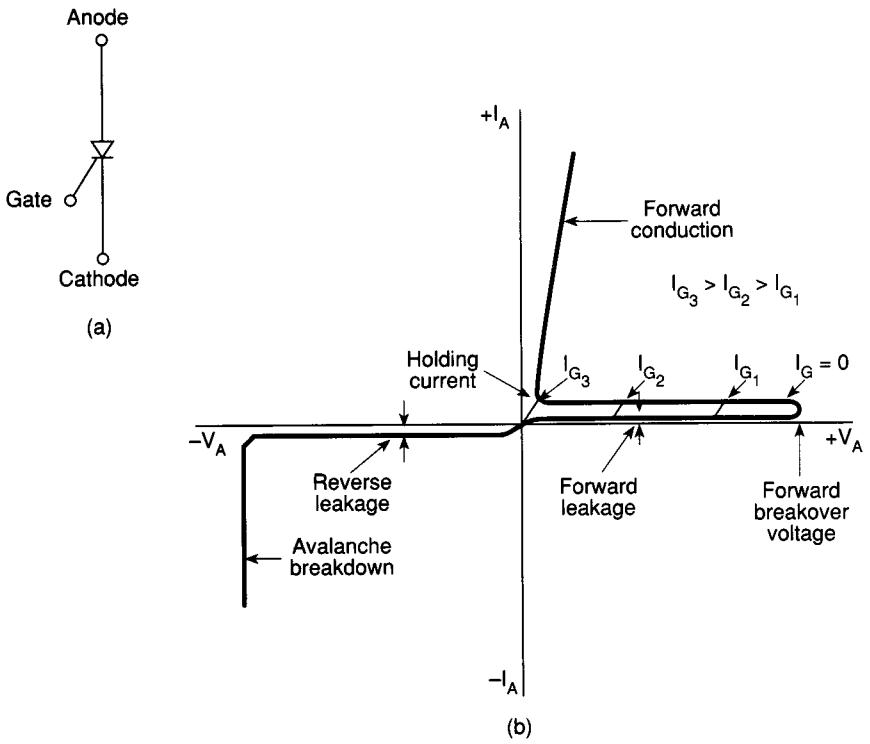


FIGURE 30.1 (a) Thyristor symbol and (b) volt-ampere characteristics. (*Source:* B.K. Bose, *Modern Power Electronics: Evaluation, Technology, and Applications,* p. 5. © 1992 IEEE.)

Thyristors are highly rugged devices in terms of transient currents, ***di/dt,*** and ***dv/dt*** capability. The **forward voltage** drop in thyristors is about 1.5 to 2 V, and even at higher currents of the order of 1000 A, it seldom exceeds 3 V. While the forward voltage determines the on-state power loss of the device at any given current, the switching power loss becomes a dominating factor affecting the device junction temperature at high operating frequencies. Because of this, the maximum switching frequencies possible using thyristors are limited in comparison with other power devices considered in this section.

Thyristors have I^2t withstand capability and can be protected by fuses. The nonrepetitive surge current capability for thyristors is about 10 times their rated root mean square (rms) current. They must be protected by snubber networks for *dv/dt* and *di/dt* effects. If the specified *dv/dt* is exceeded, thyristors may start conducting without applying a gate pulse. In dc-to-ac conversion applications it is necessary to use an antiparallel diode of similar rating across each main thyristor. Thyristors are available up to 6000 V, 3500 A.

A triac is functionally a pair of converter-grade thyristors connected in antiparallel. The triac symbol and volt-ampere characteristics are shown in Fig. 30.2. Because of the integration, the triac has poor reapplied *dv/dt*, poor gate current sensitivity at turn-on, and longer turn-off time. Triacs are mainly used in phase control applications such as in ac regulators for lighting and fan control and in solid-state ac relays.

Gate Turn-Off Thyristor (GTO)

The GTO is a power switching device that can be turned on by a short pulse of gate current and turned off by a reverse gate pulse. This reverse gate current amplitude is dependent on the anode current to be turned off. Hence there is no need for an external commutation circuit to turn it off. Because turn-off is provided by bypassing carriers directly to the gate circuit, its turn-off time is short, thus giving it more capability for high-frequency operation than thyristors. The GTO symbol and turn-off characteristics are shown in Fig. 30.3.

GTOs have the I^2t withstand capability and hence can be protected by semiconductor fuses. For reliable operation of GTOs, the critical aspects are proper design of the gate turn-off circuit and the snubber circuit.

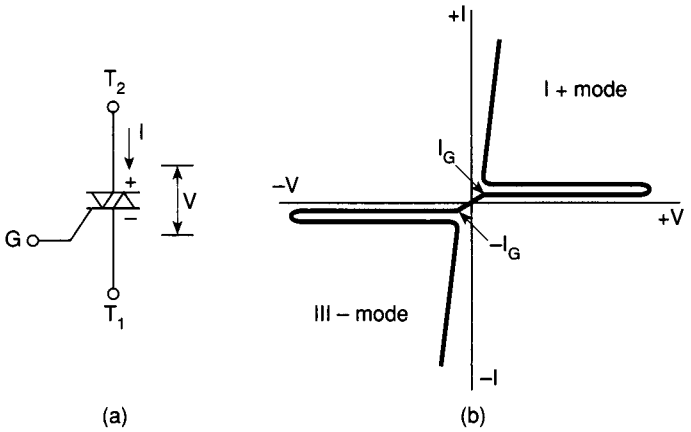


FIGURE 30.2 (a) Triac symbol and (b) volt-ampere characteristics. (*Source:* B.K. Bose, *Modern Power Electronics: Evaluation, Technology, and Applications,* p. 5. © 1992 IEEE.)

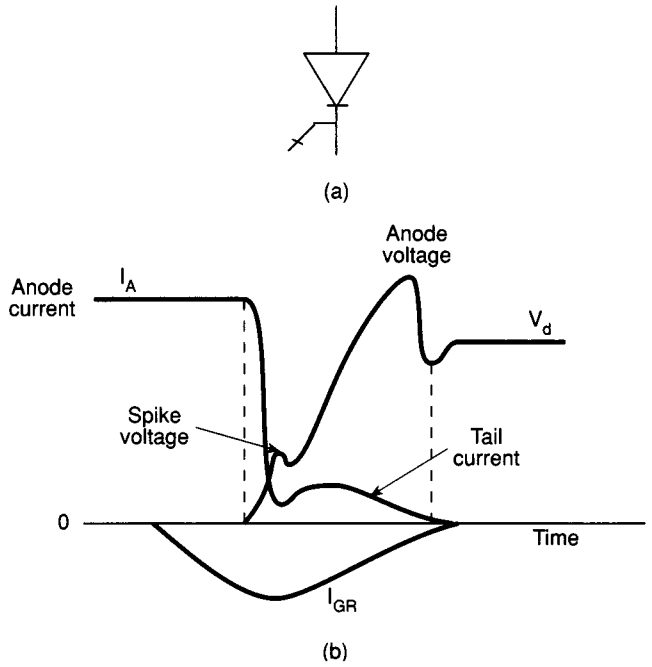


FIGURE 30.3 (a) GTO symbol and (b) turn-off characteristics. (*Source:* B.K. Bose, *Modern Power Electronics: Evaluation, Technology, and Applications,* p. 5. © 1992 IEEE.)

A GTO has a poor turn-off current gain of the order of 4 to 5. For example, a 2000-A peak current GTO may require as high as 500 A of reverse gate current. Also, a GTO has the tendency to latch at temperatures above 125°C. GTOs are available up to about 4500 V, 2500 A.

Reverse-Conducting Thyristor (RCT) and Asymmetrical Silicon-Controlled Rectifier (ASCR)

Normally in inverter applications, a diode in antiparallel is connected to the thyristor for commutation/freewheeling purposes. In RCTs, the diode is integrated with a fast switching thyristor in a single silicon chip. Thus,

the number of power devices could be reduced. This integration brings forth a substantial improvement of the static and dynamic characteristics as well as its overall circuit performance.

The RCTs are designed mainly for specific applications such as traction drives. The antiparallel diode limits the reverse voltage across the thyristor to 1 to 2 V. Also, because of the reverse recovery behavior of the diodes, the thyristor may see very high reapplied *dv/dt* when the diode recovers from its reverse voltage. This necessitates use of large RC snubber networks to suppress voltage transients. As the range of application of thyristors and diodes extends into higher frequencies, their reverse recovery charge becomes increasingly important. High reverse recovery charge results in high power dissipation during switching.

The ASCR has a similar forward blocking capability as an inverter-grade thyristor, but it has a limited reverse blocking (about 20–30 V) capability. It has an on-state voltage drop of about 25% less than an inverter-grade thyristor of a similar rating. The ASCR features a fast turn-off time; thus it can work at a higher frequency than an SCR. Since the turn-off time is down by a factor of nearly 2, the size of the commutating components can be halved. Because of this, the switching losses will also be low.

Gate-assisted turn-off techniques are used to even further reduce the turn-off time of an ASCR. The application of a negative voltage to the gate during turn-off helps to evacuate stored charge in the device and aids the recovery mechanisms. This will in effect reduce the turn-off time by a factor of up to 2 over the conventional device.

Power Transistor

Power transistors are used in applications ranging from a few to several hundred kilowatts and switching frequencies up to about 10 kHz. Power transistors used in power conversion applications are generally *npn* type. The power transistor is turned on by supplying sufficient base current, and this base drive has to be maintained throughout its conduction period. It is turned off by removing the base drive and making the base voltage slightly negative (within $-V_{BE(\max)}$). The saturation voltage of the device is normally 0.5 to 2.5 V and increases as the current increases. Hence the on-state losses increase more than proportionately with current. The transistor off-state losses are much lower than the on-state losses because the leakage current of the device is of the order of a few milliamperes. Because of relatively larger switching times, the switching loss significantly increases with switching frequency. Power transistors can block only forward voltages. The reverse peak voltage rating of these devices is as low as 5 to 10 V.

Power transistors do not have I^2t withstand capability. In other words, they can absorb only very little energy before breakdown. Therefore, they cannot be protected by semiconductor fuses, and thus an electronic protection method has to be used.

To eliminate high base current requirements, Darlington configurations are commonly used. They are available in monolithic or in isolated packages. The basic Darlington configuration is shown schematically in Fig. 30.4. The Darlington configuration presents a specific advantage in that it can considerably increase the current switched by the transistor for a given base drive. The $V_{CE(\text{sat})}$ for the Darlington is generally more than that of a single transistor of similar rating with corresponding increase in on-state power loss. During switching, the reverse-biased collector junction may show hot spot breakdown effects that are specified by reverse-bias safe operating area (RBSOA) and forward bias safe operating area (FBSOA). Modern devices with highly interdigited emitter base geometry force more uniform current distribution and therefore considerably improve second breakdown effects. Normally, a well-designed switching aid network constrains the device operation well within the SOAs.

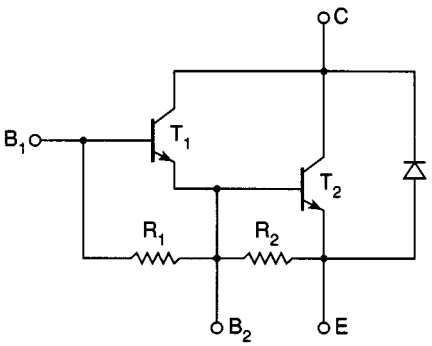


FIGURE 30.4 A two-stage Darlington transistor with bypass diode. (*Source:* B.K. Bose, *Modern Power Electronics: Evaluation, Technology, and Applications,* p. 6. © 1992 IEEE.)

Power MOSFET

Power MOSFETs are marketed by different manufacturers with differences in internal geometry and with different names such as MegaMOS, HEXFET, SIPMOS, and TMOS. They have unique features that make them potentially attractive for switching applications. They are essentially voltage-driven rather than current-driven devices, unlike bipolar transistors.

The gate of a MOSFET is isolated electrically from the source by a layer of silicon oxide. The gate draws only a minute leakage current of the order of nanoamperes. Hence the gate drive circuit is simple and power loss in the gate control circuit is practically negligible. Although in steady state the gate draws virtually no current, this is not so under transient conditions. The gate-to-source and gate-to-drain capacitances have to be charged and discharged appropriately to obtain the desired switching speed, and the drive circuit must have a sufficiently low output impedance to supply the required charging and discharging currents. The circuit symbol of a power MOSFET is shown in Fig. 30.5.

Power MOSFETs are majority carrier devices, and there is no minority carrier storage time. Hence they have exceptionally fast rise and fall times. They are essentially resistive devices when turned on, while bipolar transistors present a more or less constant $V_{CE(\text{sat})}$ over the normal operating range. Power dissipation in MOSFETs is $Id^2R_{DS(\text{on})}$, and in bipolars it is $I_CV_{CE(\text{sat})}$. At low currents, therefore, a power MOSFET may have a lower conduction loss than a comparable bipolar device, but at higher currents, the conduction loss will exceed that of bipolars. Also, the $R_{DS(\text{on})}$ increases with temperature.

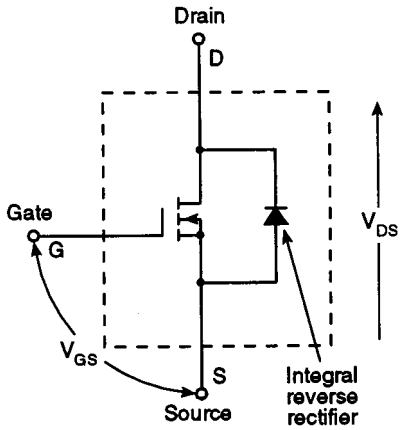


FIGURE 30.5 Power MOSFET circuit symbol. (*Source:* B.K. Bose, *Modern Power Electronics: Evaluation, Technology, and Applications,* p. 7. © 1992 IEEE.)

An important feature of a power MOSFET is the absence of a secondary breakdown effect, which is present in a bipolar transistor, and as a result, it has an extremely rugged switching performance. In MOSFETs, $R_{DS(\text{on})}$ increases with temperature, and thus the current is automatically diverted away from the hot spot. The drain body junction appears as an antiparallel diode between source and drain. Thus power MOSFETs will not support voltage in the reverse direction. Although this inverse diode is relatively fast, it is slow by comparison with the MOSFET. Recent devices have the diode recovery time as low as 100 ns. Since MOSFETs cannot be protected by fuses, an electronic protection technique has to be used.

With the advancement in MOS technology, ruggedized MOSFETs are replacing the conventional MOSFETs. The need to ruggedize power MOSFETs is related to device reliability. If a MOSFET is operating within its specification range at all times, its chances for failing catastrophically are minimal. However, if its absolute maximum rating is exceeded, failure probability increases dramatically. Under actual operating conditions, a MOSFET may be subjected to transients — either externally from the power bus supplying the circuit or from the circuit itself due, for example, to inductive kicks going beyond the absolute maximum ratings. Such conditions are likely in almost every application, and in most cases are beyond a designer's control. Rugged devices are made to be more tolerant for over-voltage transients. Ruggedness is the ability of a MOSFET to operate in an environment of dynamic electrical stresses, without activating any of the parasitic bipolar junction transistors. The rugged device can withstand higher levels of diode recovery *dv/dt* and static *dv/dt.*

Insulated-Gate Bipolar Transistor (IGBT)

The IGBT has the high input impedance and high-speed characteristics of a MOSFET with the conductivity characteristic (low saturation voltage) of a bipolar transistor. The IGBT is turned on by applying a positive voltage between the gate and emitter and, as in the MOSFET, it is turned off by making the gate signal zero or slightly negative. The IGBT has a much lower voltage drop than a MOSFET of similar ratings. The structure of an IGBT is more like a thyristor and MOSFET. For a given IGBT, there is a critical value of collector current

that will cause a large enough voltage drop to activate the thyristor. Hence, the device manufacturer specifies the peak allowable collector current that can flow without latch-up occurring. There is also a corresponding gate source voltage that permits this current to flow that should not be exceeded.

Like the power MOSFET, the IGBT does not exhibit the secondary breakdown phenomenon common to bipolar transistors. However, care should be taken not to exceed the maximum power dissipation and specified maximum junction temperature of the device under all conditions for guaranteed reliable operation. The on-state voltage of the IGBT is heavily dependent on the gate voltage. To obtain a low on-state voltage, a sufficiently high gate voltage must be applied.

In general, IGBTs can be classified as punch-through (PT) and nonpunch-through (NPT) structures, as shown in Fig. 30.6. In the PT IGBT, an N^+ buffer layer is normally introduced between the P^+ substrate and the N^- epitaxial layer, so that the whole N^- drift region is depleted when the device is blocking the off-state voltage, and the electrical field shape inside the N^- drift region is close to a rectangular shape. Because a shorter N^- region can be used in the punch-through IGBT, a better trade-off between the forward voltage drop and turn-off time can be achieved. PT IGBTs are available up to about 1200 V.

High voltage IGBTs are realized through non-punch-through process. The devices are built on a N^- wafer substrate which serves as the N^- base drift region. Experimental NPT IGBTs of up to about 4 KV have been reported in the literature. NPT IGBTs are more robust than PT IGBTs particularly under short circuit conditions. But NPT IGBTs have a higher forward voltage drop than the PT IGBTs.

The PT IGBTs cannot be as easily paralleled as MOSFETs. The factors that inhibit current sharing of parallel-connected IGBTs are (1) on-state current unbalance, caused by $V_{CE}(sat)$ distribution and main circuit wiring resistance distribution, and (2) current unbalance at turn-on and turn-off, caused by the switching time difference of the parallel connected devices and circuit wiring inductance distribution. The NPT IGBTs can be paralleled because of their positive temperature coefficient property.

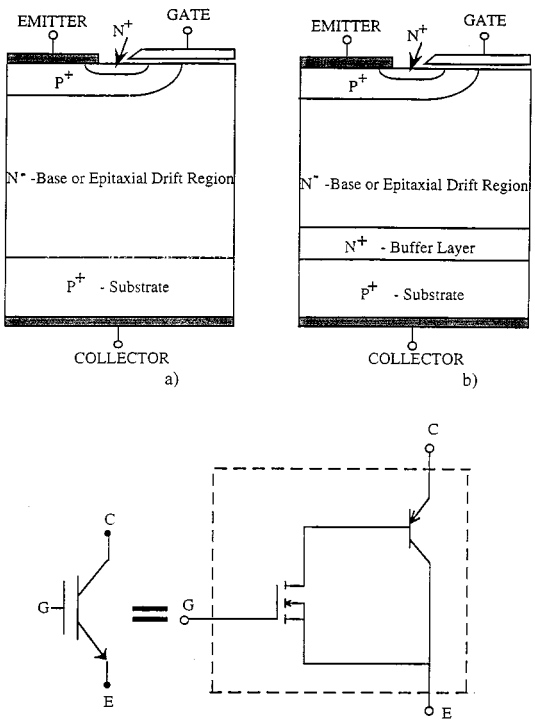


FIGURE 30.6 Nonpunch-through IGBT, (b) Punch-through IGBT, (c) IGBT equivalent circuit.

MOS-Controlled Thyristor (MCT)

The MCT is a new type of power semiconductor device that combines the capabilities of thyristor voltage and current with MOS gated turn-on and turn-off. It is a high power, high frequency, low conduction drop and a rugged device, which is more likely to be used in the future for medium and high power applications. A cross sectional structure of a p-type MCT with its circuit schematic is shown in Fig. 30.7. The MCT has a thyristor type structure with three junctions and PNPN layers between the anode and cathode. In a practical MCT, about 100,000 cells similar to the one shown are paralleled to achieve the desired current rating. MCT is turned on by a negative voltage pulse at the gate with respect to the anode, and is turned off by a positive voltage pulse.

The MCT was announced by the General Electric R & D Center on November 30, 1988. Harris Semiconductor Corporation has developed two generations of p-MCTs. Gen-1 p-MCTs are available at 65 A/1000 V and 75A/600 V with peak controllable current of 120 A. Gen-2 p-MCTs are being developed at similar current and voltage ratings, with much improved turn-on capability and switching speed. The reason for developing p-MCT is the fact that the current density that can be turned off is 2 or 3 times higher than that of an n-MCT; but n-MCTs are the ones needed for many practical applications. Harris Semiconductor Corporation is in the process of developing n-MCTs, which are expected to be commercially available during the next one to two years.

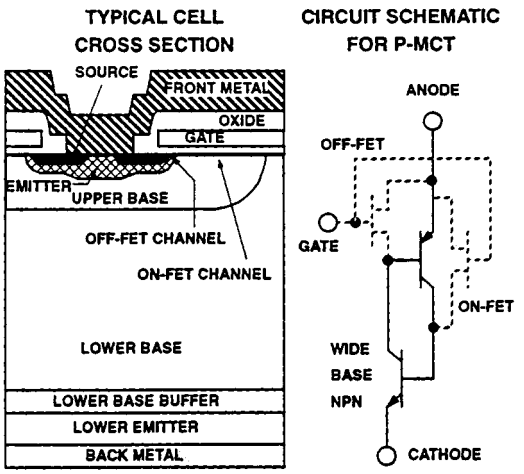


FIGURE 30.7 (*Source:* Harris Semiconductor, *User's Guide of MOS Controlled Thyristor,* With permission.)

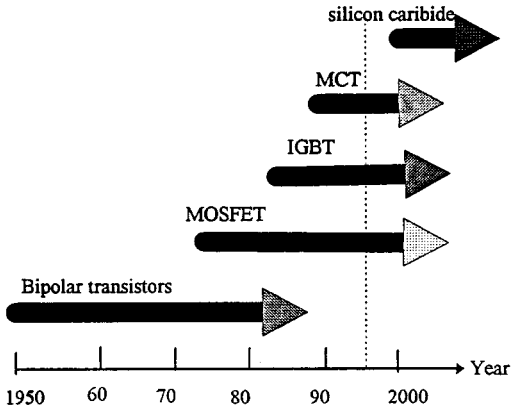


FIGURE 30.8 Current and future pwer semiconductor devices development direction (*Source:* A.Q. Huang, *Recent Developments of Power Semiconductor Devices,* VPEC Seminar Proceedings, pp. 1–9. With permission.)

The advantage of an MCT over-IGBT is its low forward voltage drop. N-type MCTs will be expected to have a similar forward voltage drop, but with an improved reverse bias safe operating area and switching speed. MCTs have relatively low switching times and storage time. The MCT is capable of high current densities and blocking voltages in both directions. Since the power gain of an MCT is extremely high, it could be driven directly from logic gates. An MCT has high *di/dt* (of the order of 2500 A/μs) and high *dv/dt* (of the order of 20,000 V/μs) capability.

The MCT, because of its superior characteristics, shows a tremendous possibility for applications such as motor drives, uninterrupted power supplies, static VAR compensators, and high power active power line conditioners.

The current and future power semiconductor devices developmental direction is shown in Fig. 30.8. High temperature operation capability and low forward voltage drop operation can be obtained if silicon is replaced by silicon carbide material for producing power devices. The silicon carbide has a higher band gap than silicon. Hence higher breakdown voltage devices could be developed. Silicon carbide devices have excellent switching characteristics and stable blocking voltages at higher temperatures. But the silicon carbide devices are still in the very early stages of development.

Defining Terms

di/dt limit: Maximum allowed rate of change of current through a device. If this limit is exceeded, the device may not be guaranteed to work reliably.
dv/dt: Rate of change of voltage withstand capability without spurious turn-on of the device.
Forward voltage: The voltage across the device when the anode is positive with respect to the cathode.
I^2t: Represents available thermal energy resulting from current flow.
Reverse voltage: The voltage across the device when the anode is negative with respect to the cathode.

Related Topic

5.1 Diodes and Rectifiers

References

B.K. Bose, *Modern Power Electronics: Evaluation, Technology, and Applications,* New York: IEEE Press, 1992.
Harris Semiconductor, *User's Guide of MOS Controlled Thyristor.*

A.Q. Huang, *Recent Developments of Power Semiconductor Devices,* VPEC Seminar Proceedings, pp. 1–9, September 1995.
N. Mohan and T. Undeland, *Power Electronics: Converters, Applications, and Design,* New York: John Wiley & Sons, 1995.
J. Wojslawowicz, "Ruggedized transistors emerging as power MOSFET standard-bearers," *Power Technics Magazine,* pp. 29–32, January 1988.

Further Information

B.M. Bird and K.G. King, *An Introduction to Power Electronics,* New York: Wiley-Interscience, 1984.
R. Sittig and P. Roggwiller, *Semiconductor Devices for Power Conditioning,* New York: Plenum, 1982.
V.A.K. Temple, "Advances in MOS controlled thyristor technology and capability," *Power Conversion*, pp. 544–554, Oct. 1989.
B.W. Williams, *Power Electronics, Devices, Drivers and Applications,* New York: John Wiley, 1987.

30.2 Power Conversion

Kaushik Rajashekara

Power conversion deals with the process of converting electric power from one form to another. The power electronic apparatuses performing the power conversion are called *power converters.* Because they contain no moving parts, they are often referred to as *static* power converters. The power conversion is achieved using power semiconductor devices, which are used as switches. The power devices used are SCRs (silicon controlled rectifiers, or thyristors), triacs, power transistors, power MOSFETs, insulated gate bipolar transistors (IGBTs), and MCTs (MOS-controlled thyristors). The power converters are generally classified as:

1. ac-dc converters (phase-controlled converters)
2. direct ac-ac converters (cycloconverters)
3. dc-ac converters (inverters)
4. dc-dc converters (choppers, buck and boost converters)

AC-DC Converters

The basic function of a **phase-controlled converter** is to convert an alternating voltage of variable amplitude and frequency to a variable dc voltage. The power devices used for this application are generally **SCR**s. The average value of the output voltage is controlled by varying the conduction time of the SCRs. The turn-on of the SCR is achieved by providing a gate pulse when it is forward-biased. The turn-off is achieved by the **commutation** of current from one device to another at the instant the incoming ac voltage has a higher instantaneous potential than that of the outgoing wave. Thus there is a natural tendency for current to be commutated from the outgoing to the incoming SCR, without the aid of any external commutation circuitry. This commutation process is often referred to as *natural commutation.*

A single-phase half-wave converter is shown in Fig. 30.9. When the SCR is turned on at an angle α, full supply voltage (neglecting the SCR drop) is applied to the load. For a purely resistive load, during the positive half cycle, the output voltage waveform follows the input ac voltage waveform. During the negative half cycle, the SCR is turned off. In the case of inductive load, the energy stored in the inductance causes the current to flow in the load circuit even after the reversal of the supply voltage, as shown in Fig. 30.9(b). If there is no freewheeling diode D_F, the load current is discontinuous. A freewheeling diode is connected across the load to turn off the SCR as soon as the input voltage polarity reverses, as shown in Fig. 30.9(c). When the SCR is off, the load current will freewheel through the diode. The power flows from the input to the load only when the SCR is conducting. If there is no freewheeling diode, during the negative portion of the supply voltage, SCR returns the energy stored in the load inductance to the supply. The freewheeling diode improves the input power factor.

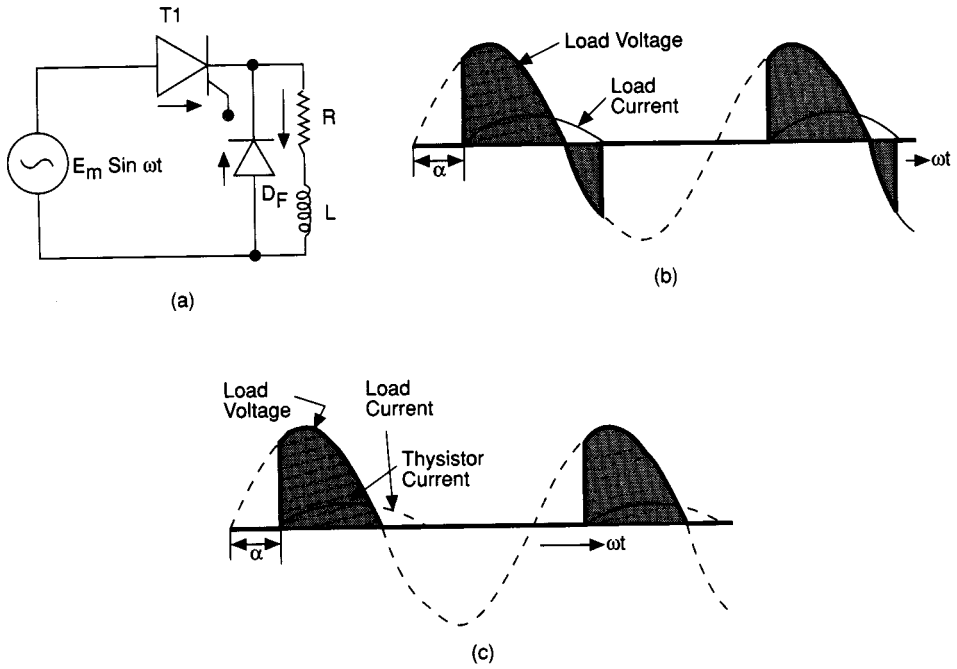


FIGURE 30.9 Single-phase half-wave converter with freewheeling diode. (a) Circuit diagram; (b) waveform for inductive load with no freewheeling diode; (c) waveform with freewheeling diode.

The controlled full-wave dc output may be obtained by using either a center tap transformer (Fig. 30.10) or by bridge configuration (Fig. 30.11). The bridge configuration is often used when a transformer is undesirable and the magnitude of the supply voltage properly meets the load voltage requirements. The average output voltage of a single-phase full-wave converter for continuous current conduction is given by

$$v_{d\alpha} = 2\frac{E_m}{\pi}\cos\alpha$$

where E_m is the peak value of the input voltage and α is the firing angle. The output voltage of a single-phase bridge circuit is the same as that shown in Fig. 30.10. Various configurations of the single-phase bridge circuit can be obtained if, instead of four SCRs, two diodes and two SCRs are used, with or without freewheeling diodes.

A three-phase full-wave converter consisting of six thyristor switches is shown in Fig. 30.12(a). This is the most commonly used three-phase bridge configuration. Thyristors T_1, T_3, and T_5 are turned on during the positive half cycle of the voltages of the phases to which they are connected, and thyristors T_2, T_4, and T_6 are turned on during the negative half cycle of the phase voltages. The reference for the angle in each cycle is at the crossing points of the phase voltages. The ideal output voltage, output current, and input current waveforms are shown in Fig. 30.12(b). The output dc voltage is controlled by varying the firing angle α. The average output voltage under continuous current conduction operation is given by

$$v_o = \frac{3\sqrt{3}}{\pi}E_m\cos\alpha$$

where E_m is the peak value of the phase voltage. At $\alpha = 90°$, the output voltage is zero. For $0 < \alpha < 90°$, v_o is positive and power flows from ac supply to the load. For $90° < \alpha < 180°$, v_o is negative and the converter operates in the inversion mode. If the load is a dc motor, the power can be transferred from the motor to the ac supply, a process known as *regeneration.*

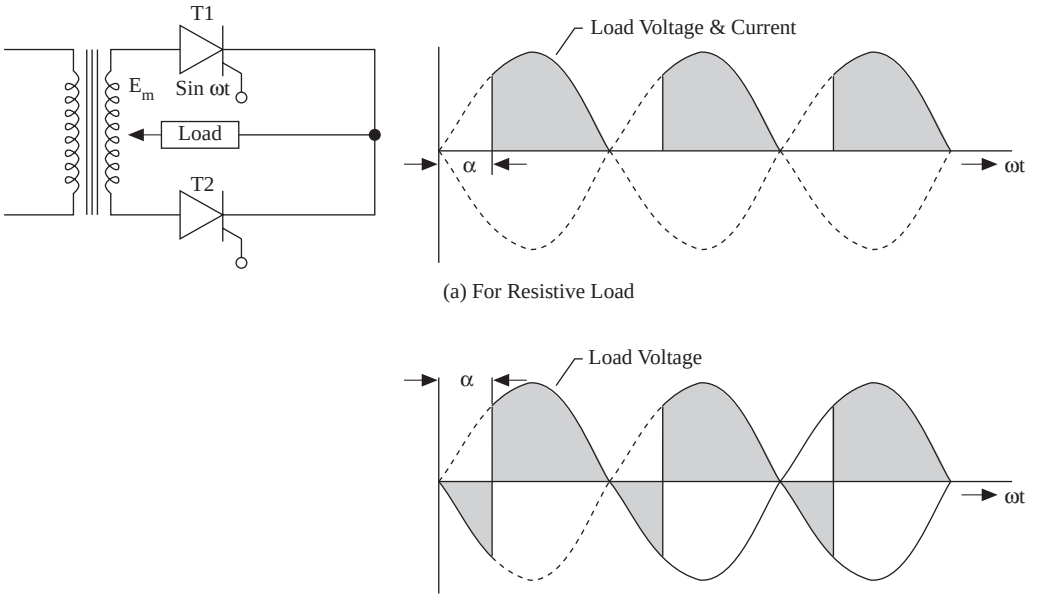


FIGURE 30.10 Single-phase full-wave converter with transformer.

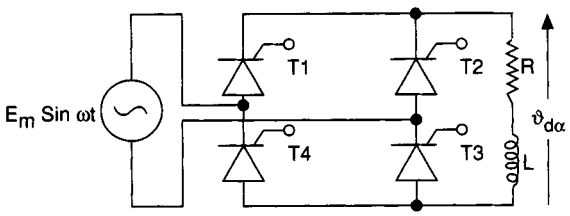


FIGURE 30.11 Single-phase bridge converter.

In Fig. 30.12(a), the top or bottom thyristors could be replaced by diodes. The resulting topology is called a *thyristor semiconverter*. With this configuration, the input power factor is improved, but the regeneration is not possible.

Cycloconverters

Cycloconverters are direct ac-to-ac frequency changers. The term *direct conversion* means that the energy does not appear in any form other than the ac input or ac output. The output frequency is lower than the input frequency and is generally an integral multiple of the input frequency. A cycloconverter permits energy to be fed back into the utility network without any additional measures. Also, the phase sequence of the output voltage can be easily reversed by the control system. Cycloconverters have found applications in aircraft systems and industrial drives. These cycloconverters are suitable for synchronous and induction motor control. The operation of the cycloconverter is illustrated in Section 30.4 of this chapter.

DC-to-AC Converters

The dc-to-ac converters are generally called *inverters*. The ac supply is first converted to dc, which is then converted to a variable-voltage and variable-frequency power supply. This generally consists of a three-phase bridge connected to the ac power source, a dc link with a filter, and the three-phase inverter bridge connected

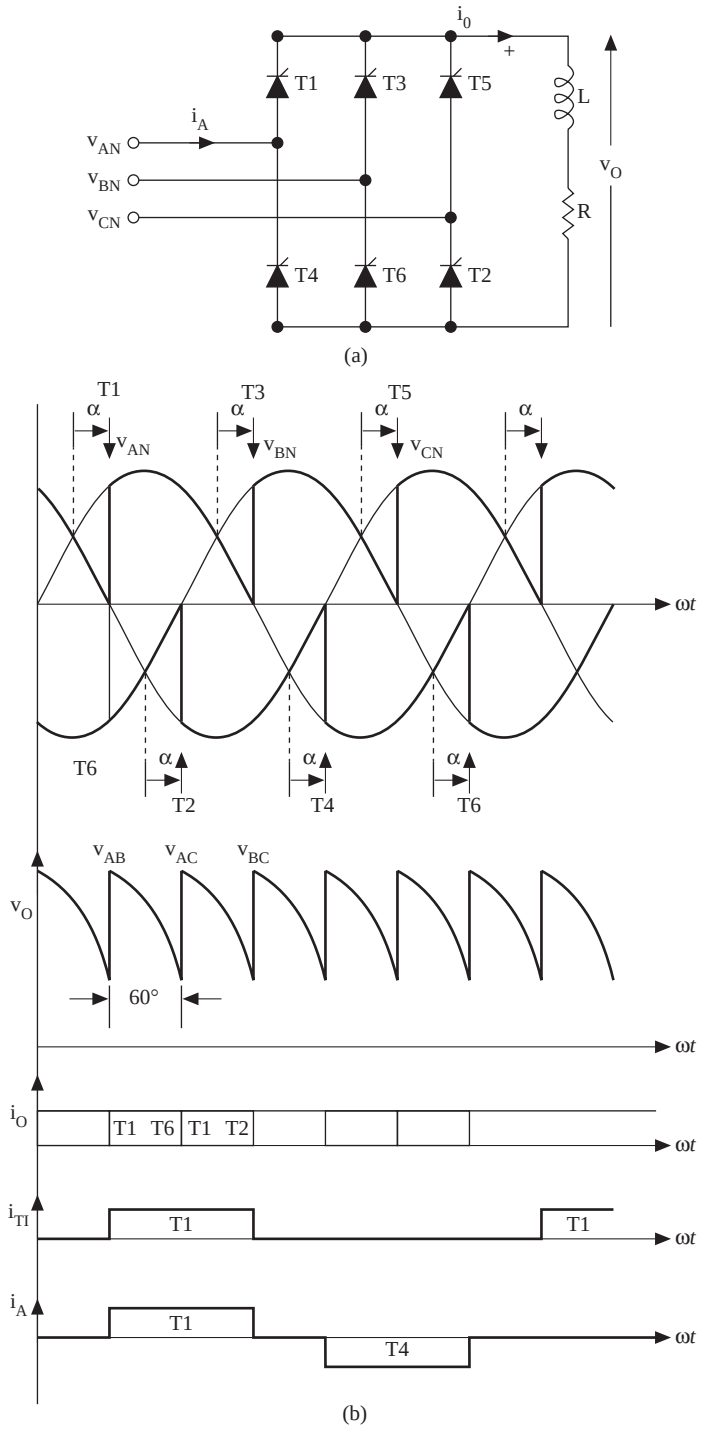


FIGURE 30.12 (a) Three-phase thyristor full bridge configuration; (b) output voltage and current waveforms.

to the load. In the case of battery-operated systems, there is no intermediate dc link. Inverters can be classified as voltage source inverters (VSIs) and current source inverters (CSIs). A voltage source inverter is fed by a stiff dc voltage, whereas a current source inverter is fed by a stiff current source. A voltage source can be converted to a current source by connecting a series inductance and then varying the voltage to obtain the desired current.

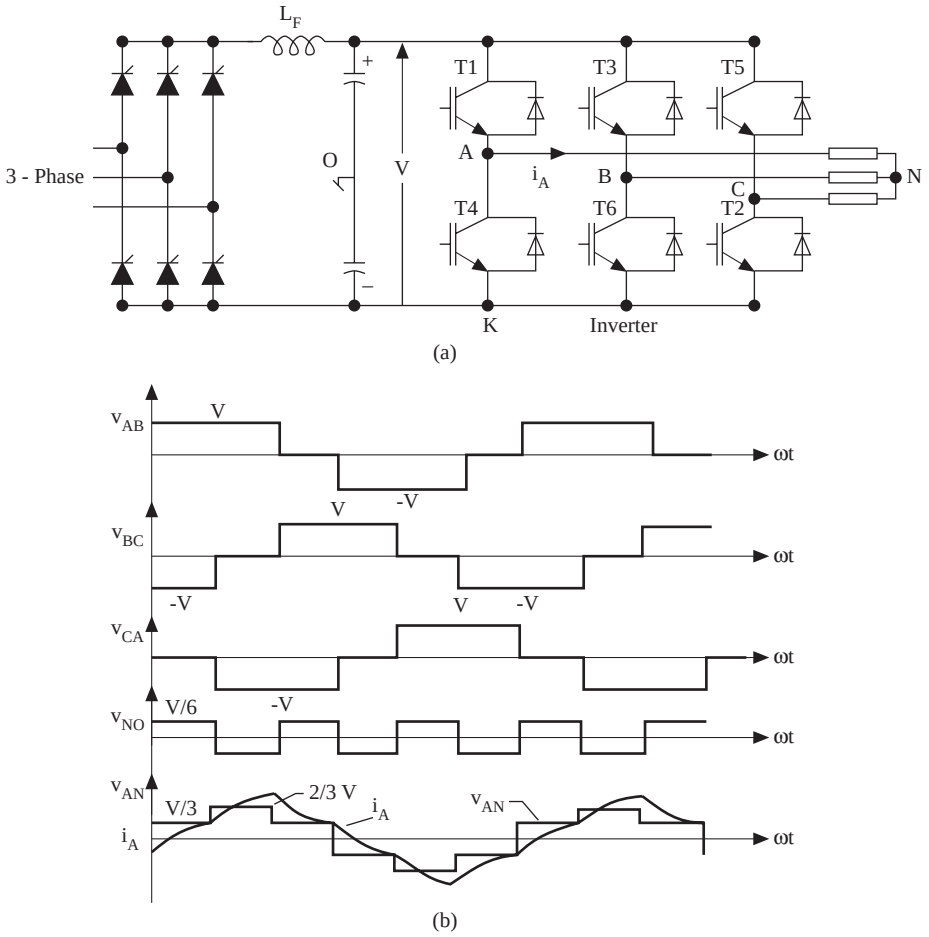


FIGURE 30.13 (a) Three-phase converter and voltage source inverter configuration; (b) three-phase square-wave inverter waveforms.

A VSI can also be operated in current-controlled mode, and similarly a CSI can also be operated in the voltage-control mode. The inverters are used in variable frequency ac motor drives, uninterrupted power supplies, induction heating, static VAR compensators, etc.

Voltage Source Inverter

A three-phase voltage source inverter configuration is shown in Fig. 30.13(a). The VSIs are controlled either in square-wave mode or in pulsewidth-modulated (PWM) mode. In square-wave mode, the frequency of the output voltage is controlled within the inverter, the devices being used to switch the output circuit between the plus and minus bus. Each device conducts for 180 degrees, and each of the outputs is displaced 120 degrees to generate a six-step waveform, as shown in Fig. 30.13(b). The amplitude of the output voltage is controlled by varying the dc link voltage. This is done by varying the firing angle of the thyristors of the three-phase bridge converter at the input. The square-wave-type VSI is not suitable if the dc source is a battery. The six-step output voltage is rich in harmonics and thus needs heavy filtering.

In PWM inverters, the output voltage and frequency are controlled within the inverter by varying the width of the output pulses. Hence at the front end, instead of a phase-controlled thyristor converter, a diode bridge rectifier can be used. A very popular method of controlling the voltage and frequency is by sinusoidal pulsewidth modulation. In this method, a high-frequency triangle carrier wave is compared with a three-phase sinusoidal waveform, as shown in Fig. 30.14. The power devices in each phase are switched on at the intersection of sine

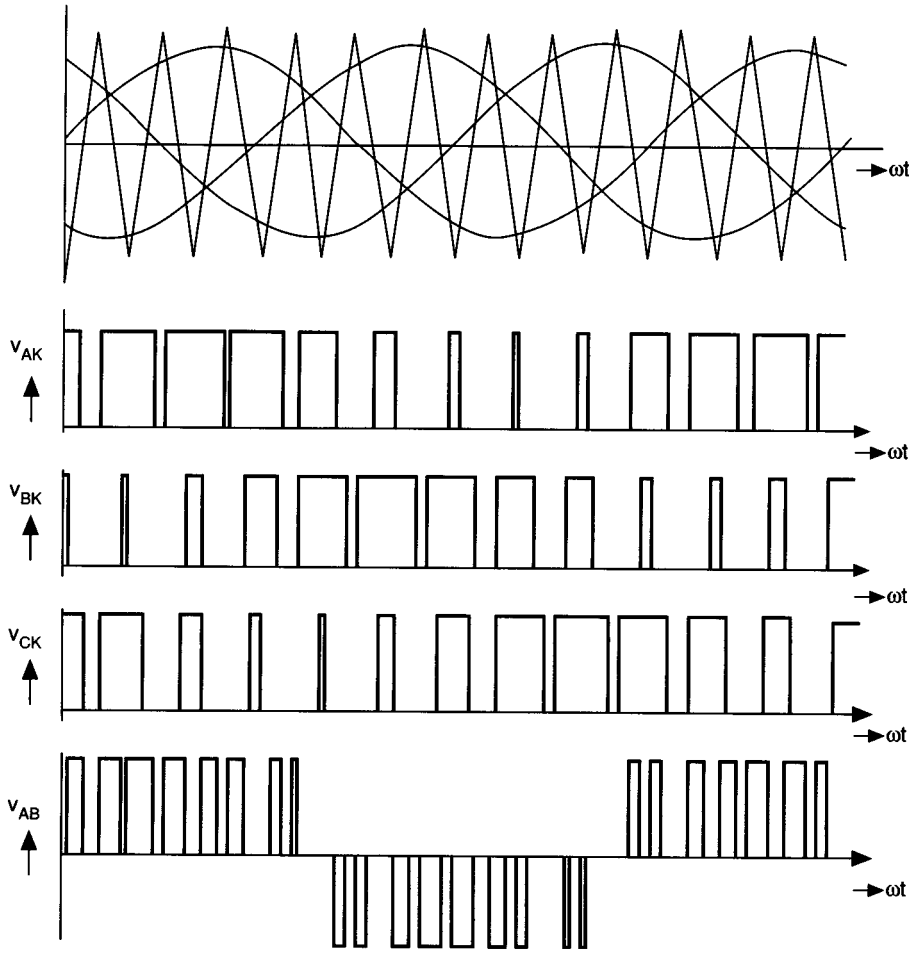


FIGURE 30.14 Three-phase sinusoidal PWM inverter waveforms.

and triangle waves. The amplitude and frequency of the output voltage are varied, respectively, by varying the amplitude and frequency of the reference sine waves. The ratio of the amplitude of the sine wave to the amplitude of the carrier wave is called the *modulation index*.

The harmonic components in a PWM wave are easily filtered because they are shifted to a higher-frequency region. It is desirable to have a high ratio of carrier frequency to fundamental frequency to reduce the harmonics of lower-frequency components. There are several other PWM techniques mentioned in the literature. The most notable ones are selected harmonic elimination, hysteresis controller, and space vector PWM technique.

In inverters, if SCRs are used as power switching devices, an external forced commutation circuit has to be used to turn off the devices. Now, with the availability of **IGBT**s above 1000-A, 1000-V ratings, they are being used in applications up to 300-kW motor drives. Above this power rating, GTOs are generally used. Power Darlington transistors, which are available up to 800 A, 1200 V, could also be used for inverter applications.

Current Source Inverter

Contrary to the voltage source inverter where the voltage of the dc link is imposed on the motor windings, in the current source inverter the current is imposed into the motor. Here the amplitude and phase angle of the motor voltage depend on the load conditions of the motor. The current source inverter is described in detail in Section 30.4.

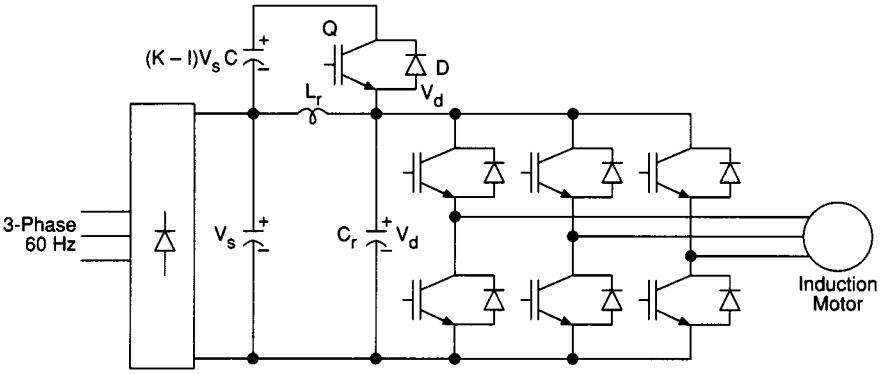


FIGURE 30.15 Resonant dc-link inverter system with active voltage clamping.

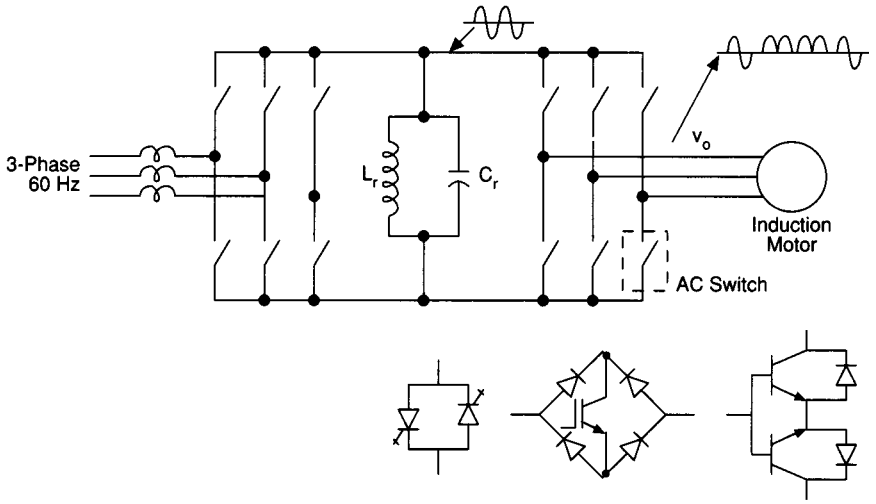


FIGURE 30.16 Resonant ac-link converter system showing configuration of ac switches.

Resonant-Link Inverters

The use of resonant switching techniques can be applied to inverter topologies to reduce the switching losses in the power devices. They also permit high switching frequency operation to reduce the size of the magnetic components in the inverter unit. In the resonant dc-link inverter shown in Fig. 30.15, a resonant circuit is added at the inverter input to convert a fixed dc voltage to a pulsating dc voltage. This resonant circuit enables the devices to be turned on and turned off during the zero voltage interval. Zero voltage or zero current switching is often termed *soft switching*. Under soft switching, the switching losses in the power devices are almost eliminated. The electromagnetic interference (EMI) problem is less severe because resonant voltage pulses have lower *dv/dt* compared to those of hard-switched PWM inverters. Also, the machine insulation is less stretched because of lower *dv/dt* resonant voltage pulses. In Fig. 30.15, all the inverter devices are turned on simultaneously to initiate a resonant cycle. The commutation from one device to another is initiated at the zero dc-link voltage. The inverter output voltage is formed by the integral numbers of quasi-sinusoidal pulses. The circuit consisting of devices *Q*, *D*, and the capacitor *C* acts as an active clamp to limit the dc voltage to about 1.4 times the diode rectifier voltage V_s.

There are several other topologies of resonant link inverters mentioned in the literature. There are also resonant link ac-ac converters based on bidirectional ac switches, as shown in Fig. 30.16. These resonant link converters find applications in ac machine control and uninterrupted power supplies, induction heating, etc. The resonant link inverter technology is still in the development stage for industrial applications.

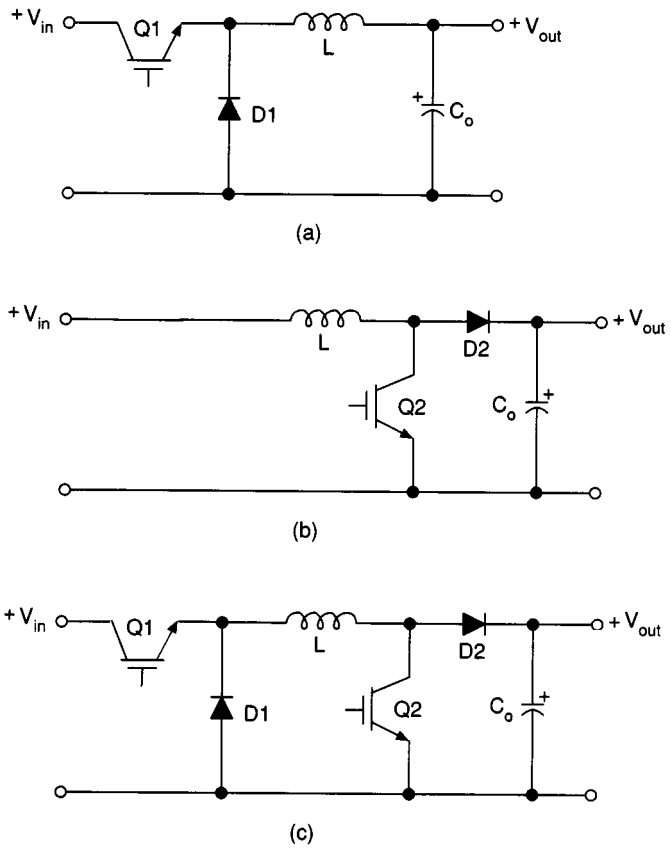


FIGURE 30.17 DC-DC converter configurations: (a) buck converter; (b) boost converter; (c) buck-boost converter.

DC-DC Converters

DC-dc converters are used to convert unregulated dc voltage to regulated or variable dc voltage at the output. They are widely used in switch-mode dc power supplies and in dc motor drive applications. In dc motor control applications, they are called *chopper-controlled drives.* The input voltage source is usually a battery or derived from an ac power supply using a diode bridge rectifier. These converters are generally either hard-switched PWM types or soft-switched resonant-link types. There are several dc-dc converter topologies, the most common ones being buck converter, boost converter, and buck-boost converter, shown in Fig. 30.17.

Buck Converter

A buck converter is also called a *step-down* converter. Its principle of operation is illustrated by referring to Fig. 30.17(a). The IGBT acts as a high-frequency switch. The IGBT is repetitively closed for a time t_{on} and opened for a time t_{off}. During t_{on}, the supply terminals are connected to the load, and power flows from supply to the load. During t_{off}, load current flows through the freewheeling diode D_1, and the load voltage is ideally zero. The average output voltage is given by

$$V_{out} = DV_{in}$$

where D is the **duty cycle** of the switch and is given by $D = t_{on}/T$, where T is the time for one period. $1/T$ is the switching frequency of the power device IGBT.

Boost Converter

A boost converter is also called a *step-up* converter. Its principle of operation is illustrated by referring to Fig. 30.17(b). This converter is used to produce higher voltage at the load than the supply voltage. When the

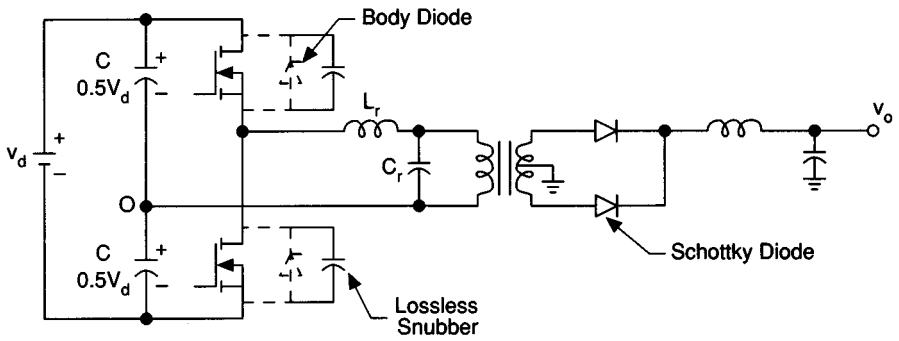


FIGURE 30.18 Resonant-link dc-dc converter.

power switch is on, the inductor is connected to the dc source and the energy from the supply is stored in it. When the device is off, the inductor current is forced to flow through the diode and the load. The induced voltage across the inductor is negative. The inductor adds to the source voltage to force the inductor current into the load. The output voltage is given by

$$V_{\text{out}} = \frac{V_{\text{in}}}{1 - D}$$

Thus for variation of D in the range $0 < D < 1$, the load voltage V_{out} will vary in the range $V_{\text{in}} < V_{\text{out}} < \infty$.

Buck-Boost Converter

A buck-boost converter can be obtained by the cascade connection of the buck and the boost converter. The steady-state output voltage V_{out} is given by

$$V_{\text{out}} = V_{\text{in}} \frac{D}{1 - D}$$

This allows the output voltage to be higher or lower than the input voltage, based on the duty cycle D. A typical buck-boost converter topology is shown in Fig. 30.17(c). When the power device is turned on, the input provides energy to the inductor and the diode is reverse biased. When the device is turned off, the energy stored in the inductor is transferred to the output. No energy is supplied by the input during this interval. In dc power supplies, the output capacitor is assumed to be very large, which results in a constant output voltage. In dc drive systems, the chopper is operated in step-down mode during motoring and in step-up mode during regeneration operation.

Resonant-Link DC-DC Converters

The use of resonant converter topologies would help to reduce the switching losses in dc-dc converters and enable the operation at switching frequencies in the megahertz range. By operating at high frequencies, the size of the power supplies could be reduced. There are several types of resonant converter topologies. The most popular configuration is shown in Fig. 30.18. The dc power is converted to high-frequency alternating power using the MOSFET half-bridge inverter. The resonant capacitor voltage is transformer-coupled, rectified using the two Schottky diodes, and then filtered to get output dc voltage. The output voltage is regulated by control of the inverter switching frequency.

Instead of parallel loading as in Fig. 30.18, the resonant circuit can be series-loaded; that is, the transformer in the output circuit can be placed in series with the tuned circuit. The series resonant circuit provides the short-circuit limiting feature.

There are other forms of resonant converter topologies mentioned in the literature such as quasi-resonant converters and multiresonant converters. These resonant converter topologies find applications in high-density power supplies.

Defining Terms

Commutation: Process of transferring the current from one power device to another.
Duty cycle: Ratio of the on-time of a switch to the switching period.
Full-wave control: Both the positive and negative half cycle of the waveforms are controlled.
IGBT: Insulated-gate bipolar transistor.
Phase-controlled converter: Converter in which the power devices are turned off at the natural crossing of zero voltage in ac to dc conversion applications.
SCR: Silicon-controlled rectifier.

Related Topics

33.2 Heat Transfer Fundamentals • 61.3 High-Voltage Direct-Current Transmission

References

B.K. Bose, *Modern Power Electronics,* New York: IEEE Press, 1992.
Motorola, *Linear/Switchmode Voltage Regulator Handbook,* 1989.
K.S. Rajashekara, H. Le-Huy, et al., "Resonant DC Link Inverter-Fed AC Machines Control," IEEE Power Electronics Specialists Conference, 1987, pp. 491–496.
P.C. Sen, *Thyristor DC Drives,* New York: John Wiley, 1981.
G. Venkataramanan and D. Divan, "Pulse Width Modulation with Resonant DC Link Converters," IEEE IAS Annual Meeting, 1990, pp. 984–990.

Further Information

B.K. Bose, *Power Electronics & AC Drives,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
R. Hoft, *Semiconductor Power Electronics,* New York: Van Nostrand Reinhold, 1986.
B.R. Pelly, *Thyristor Phase Controlled Converters and Cycloconverters,* New York: Wiley-Interscience, 1971.
A.I. Pressman, *Switching and Linear Power Supply, Power Converter Design,* Carmel, Ind.: Hayden Book Company, 1977.
M.H. Rashid, *Power Electronics, Circuits, Devices and Applications,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.

30.3 Power Supplies

Ashoka K. S. Bhat

Power supplies are used in many industrial and aerospace applications and also in consumer products. Some of the requirements of power supplies are small size, light weight, low cost, and high power conversion efficiency. In addition to these, some power supplies require the following: electrical isolation between the source and load, low harmonic distortion for the input and output waveforms, and high power factor (PF) if the source is ac voltage. Some special power supplies require controlled direction of power flow.

Basically two types of power supplies are required: dc power supplies and ac power supplies. The output of dc power supplies is regulated or controllable dc, whereas the output for ac power supplies is ac. The input to these power supplies can be ac or dc.

DC Power Supplies

If an ac source is used, then ac-to-dc **converters** explained in Section 30.2 can be used. In these converters, electrical isolation can only be provided by bulky line frequency transformers. The ac source can be rectified with a diode rectifier to get an uncontrolled dc, and then a dc-to-dc converter can be used to get a controlled dc output. Electrical isolation between the input source and the output load can be provided in the dc-to-dc converter using a high-frequency (HF) transformer. Such HF transformers have small size, light weight, and low cost compared to bulky line frequency transformers. Whether the input source is dc (e.g., battery) or ac, dc-to-dc converters form an important part of dc power supplies, and they are explained in this subsection.

DC power supplies can be broadly classified as linear and switching power supplies.

A *linear power supply* is the oldest and simplest type of power supply. The output voltage is **regulated** by dropping the extra input voltage across a series transistor (therefore, also referred to as a series regulator). They have very small output ripple, theoretically zero noise, large hold-up time (typically 1–2 ms), and fast response. Linear power supplies have the following disadvantages: very low efficiency, electrical isolation can only be on 60-Hz ac side, larger volume and weight, and, in general, only a single output possible. However, they are still used in very small regulated power supplies and in some special applications (e.g., magnet power supplies). Three terminal linear regulator integrated circuits (ICs) are readily available (e.g., μA7815 has +15-V, 1-A output), are easy to use, and have built-in load short-circuit protection.

Switching power supplies use power semiconductor switches in the *on* and *off* switching states resulting in high efficiency, small size, and light weight. With the availability of fast switching devices, HF magnetics and capacitors, and high-speed control ICs, switching power supplies have become very popular. They can be further classified as **pulsewidth-modulated (PWM) converters** and **resonant converters**, and they are explained below.

Pulsewidth-Modulated Converters

These converters employ square-wave pulsewidth modulation to achieve voltage regulation. The average output voltage is varied by varying the duty cycle of the power semiconductor switch. The voltage waveform across the switch and at the output are square wave in nature [refer to Fig. 30.13(b)] and they generally result in higher switching losses when the switching frequency is increased. Also, the switching stresses are high with the generation of large electromagnetic interference (EMI), which is difficult to filter. However, these converters are easy to control, well understood, and have wide load control range.

The methods of control of PWM converters are discussed next.

The Methods of Control. The PWM converters operate with a fixed-frequency, variable duty cycle. Depending on the duty cycle, they can operate in either continuous current mode (CCM) or discontinuous current mode (DCM). If the current through the output inductor never reaches zero (refer to Fig. 30.13), then the converter operates in CCM; otherwise DCM occurs.

The three possible control methods [Severns and Bloom, 1988; Hnatek, 1981; Unitrode Corporation, 1984; Motorola, 1989; Philips Semiconductors, 1991] are briefly explained below.

1. *Direct duty cycle control* is the simplest control method. A fixed-frequency ramp is compared with the control voltage [Fig. 30.19(a)] to obtain a variable duty cycle base drive signal for the transistor. This is the simplest method of control. Disadvantages of this method are (a) provides no voltage feedforward to anticipate the effects of input voltage changes, slow response to sudden input changes, poor audio susceptibility, poor open-loop line regulation, requiring higher loop gain to achieve specifications; (b) poor dynamic response.
2. *Voltage feedforward control.* In this case the ramp amplitude varies in direct proportion to the input voltage [Fig. 30.19(b)]. The open-loop regulation is very good, and the problems in 1(a) above are corrected.
3. *Current mode control.* In this method, a second inner control loop compares the peak inductor current with the control voltage which provides improved open-loop line regulation [Fig. 30.19(c)]. All the problems of the direct duty cycle control method 1 above are corrected with this method. An additional advantage of this method is that the two-pole second-order filter is reduced to a single-pole (the filter capacitor) first-order filter, resulting in simpler compensation networks.

The above control methods can be used in all the PWM converter configurations explained below.

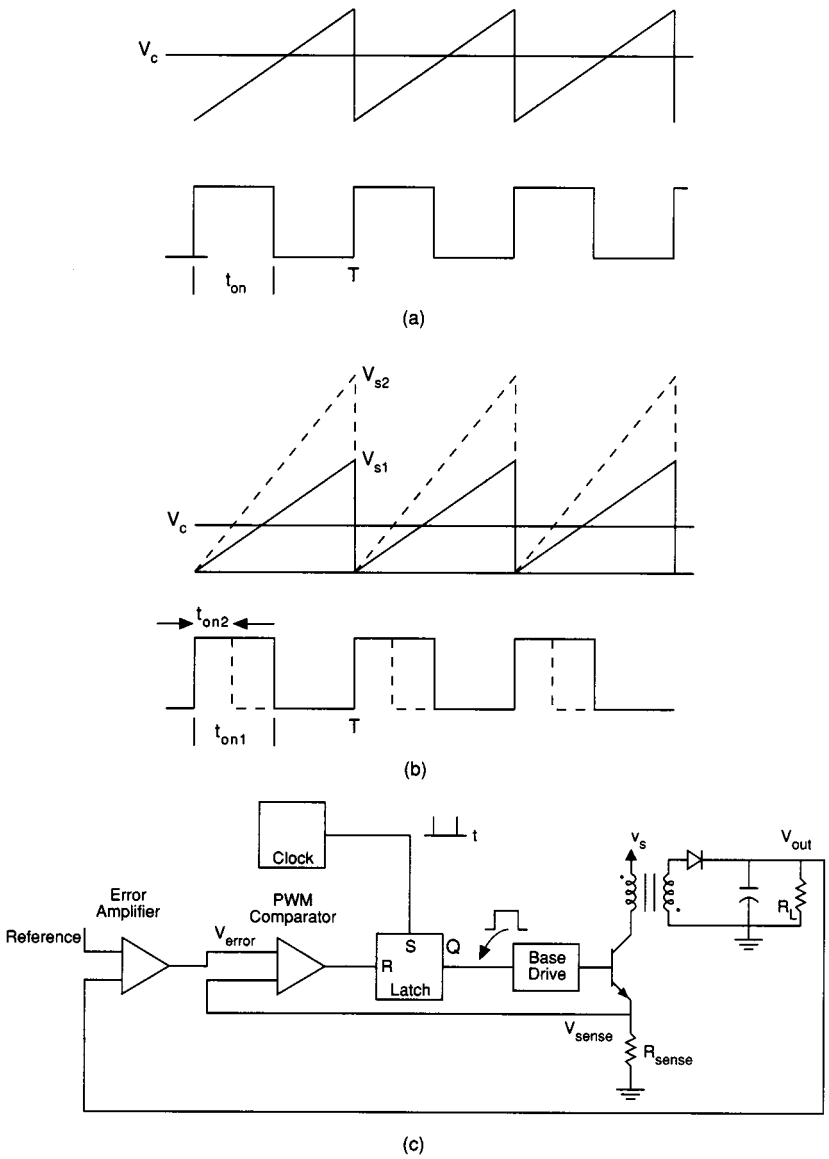


FIGURE 30.19 PWM converter control methods: (a) direct duty cycle control; (b) voltage feedforward control; (c) current mode control (illustrated for flyback converter).

PWM converters can be classified as single-ended and double-ended converters. These converters may or may not have a high-frequency transformer for isolation.

Nonisolated Single-Ended PWM Converters. The basic nonisolated single-ended converters are (a) buck (step-down), (b) boost (step-up), (c) buck-boost (step-up or step-down, also referred to as flyback), and (d) ´Cuk converters (Fig. 30.20). The first three of these converters have been discussed in Section 30.2. The ´Cuk converter provides the advantage of nonpulsating input-output current ripple requiring smaller size external filters. Output voltage expression is the same as the buck-boost converter (refer to Section 30.2) and can be less than or greater than the input voltage. There are many variations of the above basic nonisolated converters, and most of them use a high-frequency transformer for ohmic isolation between the input and the output. Some of them are discussed below.

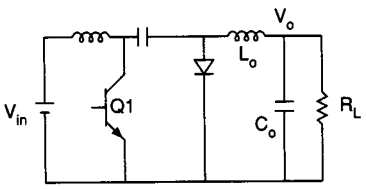


FIGURE 30.20 Nonisolated Ćuk converter.

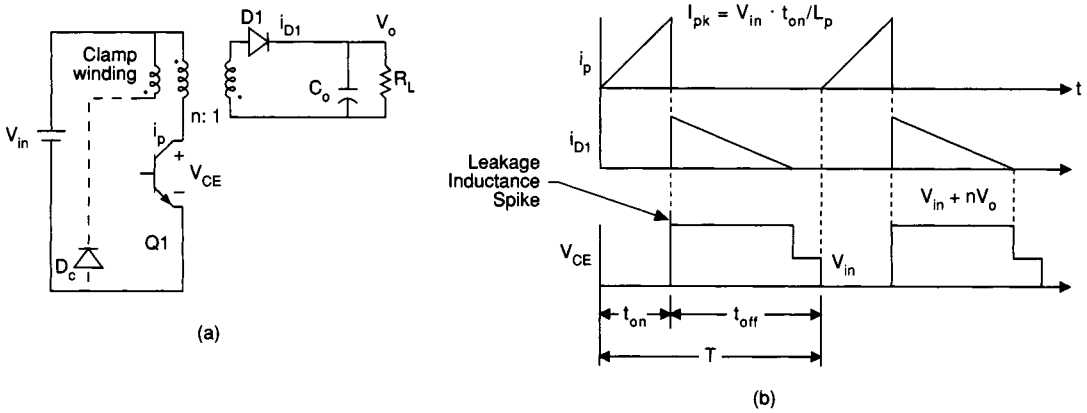


FIGURE 30.21 (a) Flyback converter. The clamp winding shown is optional and is used to clamp the transistor voltage stress to $V_{in} + nV_o$. (b) Flyback converter waveforms without the clamp winding. The leakage inductance spikes vanish with the clamp winding.

Isolated Single-Ended Topologies

1. The flyback converter (Fig. 30.21) is an **isolated** version of the buck-boost converter. In this converter (Fig. 30.21), when the transistor is on, energy is stored in the coupled inductor (not a transformer), and this energy is transferred to the load when the switch is off.

   Some of the advantages of this converter are that the leakage inductance is in series with the output diode when current is delivered to the output, and, therefore, no filter inductor is required; cross regulation for multiple output converters is good; it is ideally suited for high-voltage output applications; and it has the lowest cost.

   Some of the disadvantages are that large output filter capacitors are required to smooth the pulsating output current; inductor size is large since air gaps are to be provided; and due to stability reasons, flyback converters are usually operated in the DCM, which results in increased losses. To avoid the stability problem, flyback converters are operated with current mode control explained earlier. Flyback converters are used in the power range of 20 to 200 W.
2. The forward converter (Fig. 30.22) is based on the buck converter. It is usually operated in the CCM to reduce the peak currents and does not have the stability problem of the flyback converter. The HF transformer transfers energy directly to the output with very small stored energy. The output capacitor size and peak current rating are smaller than they are for the flyback. Reset winding is required to remove the stored energy in the transformer. Maximum duty cycle is about 0.45 and limits the control range. This topology is used for power levels up to about 1 kW.

The flyback and forward converters explained above require the rating of power transistors to be much higher than the supply voltage. The two-transistor flyback and forward converters shown in Fig. 30.23 limit the voltage rating of transistors to the supply voltage.

The Sepic converter shown in Fig. 30.24 is another isolated single-ended PWM converter.

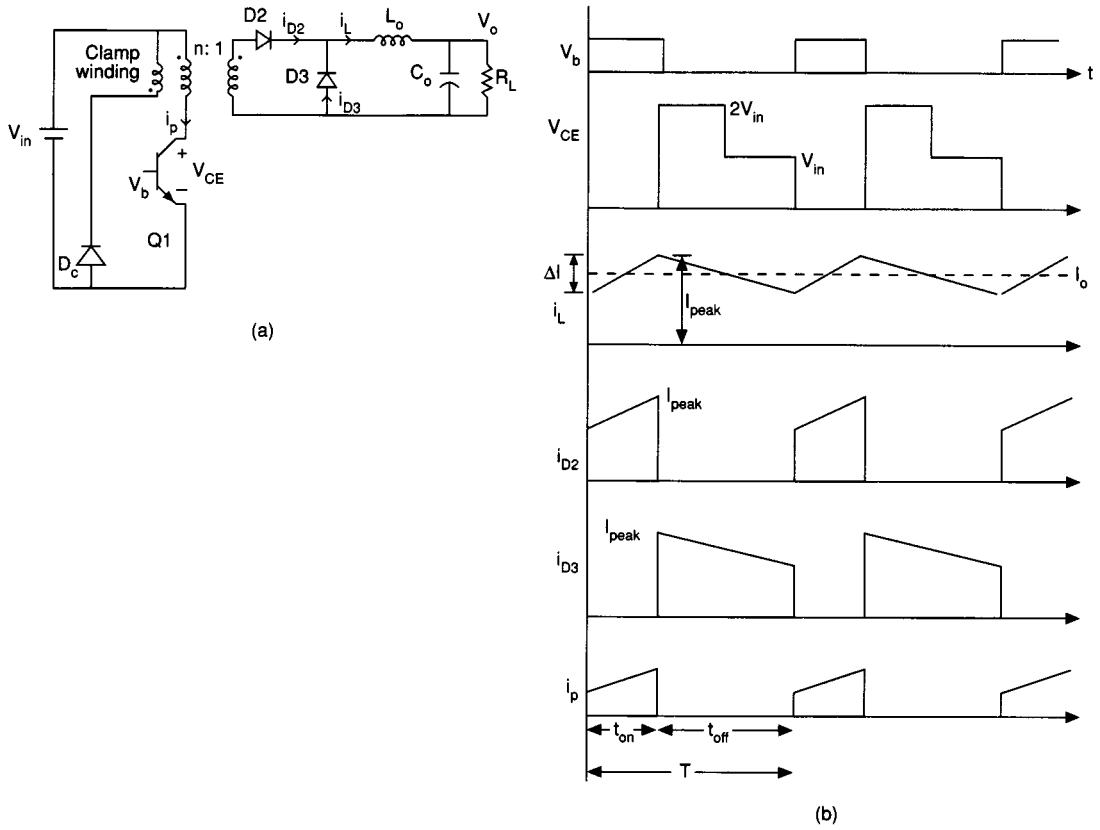


FIGURE 30.22 (a) Forward converter. The clamp winding shown is required for operation. (b) Forward converter waveforms.

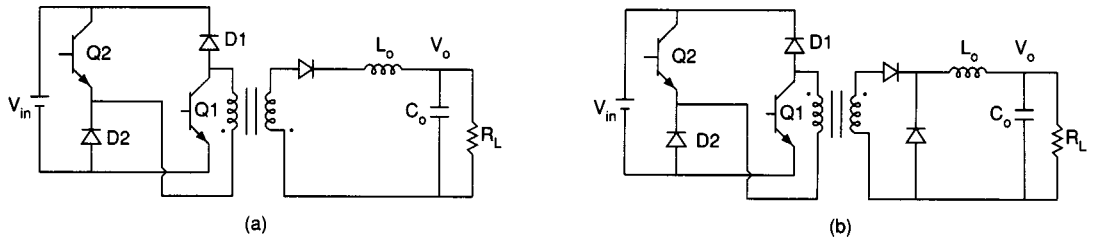


FIGURE 30.23 (a) Two-transistor single-ended flyback converter. (b) Two-transistor single-ended forward converter.

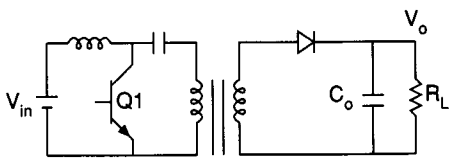


FIGURE 30.24 Sepic converter.

Double-Ended PWM Converters. Usually, for power levels above 300 W, double-ended converters are used. In double-ended converters, full-wave rectifiers are used and the output voltage ripple will have twice the switching frequency. Three important double-ended PWM converter configurations are push-pull (Fig. 30.25), half-bridge (Fig. 30.26), and full-bridge (Fig. 30.27).

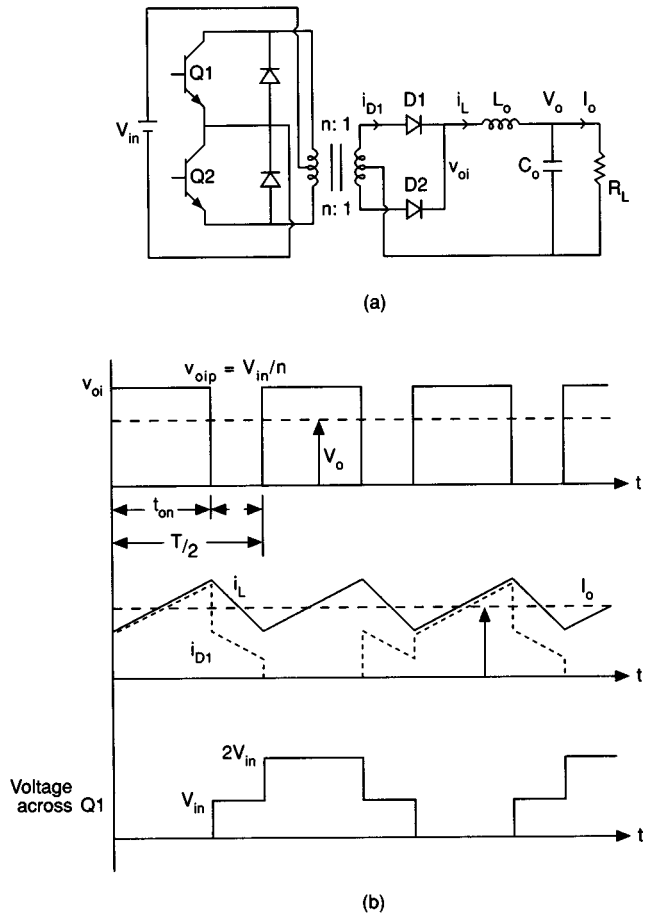


FIGURE 30.25 (a) Push-pull converter and (b) its operating waveforms.

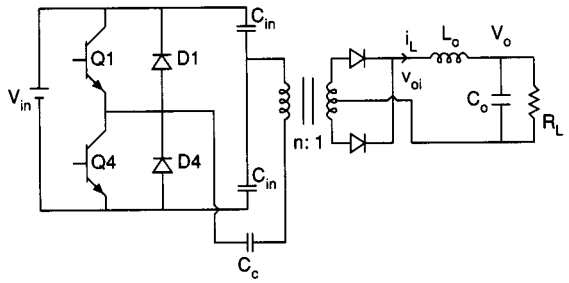


FIGURE 30.26 Half-bridge converter. Coupling capacitor C_c is used to avoid transformer saturation.

1. *The push-pull converter.* The duty ratio of each transistor in a push-pull converter (Fig. 30.25) is less than 0.5. Some of the advantages are that the transformer flux swings fully, thereby the size of the transformer is much smaller (typically half the size) than single-ended converters, and output ripple is twice the switching frequency of transistors, allowing smaller filters.

   Some of the disadvantages of this configuration are that transistors must block twice the supply voltage, flux symmetry imbalance can cause transformer saturation and special control circuitry is required to avoid this problem, and use of center-tap transformer requires extra copper resulting in higher volt-ampere (VA) rating.

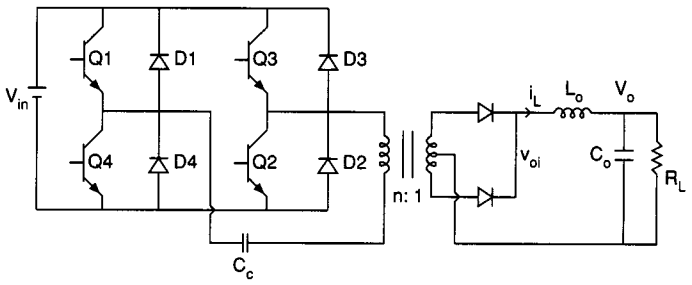


FIGURE 30.27 Full-bridge converter.

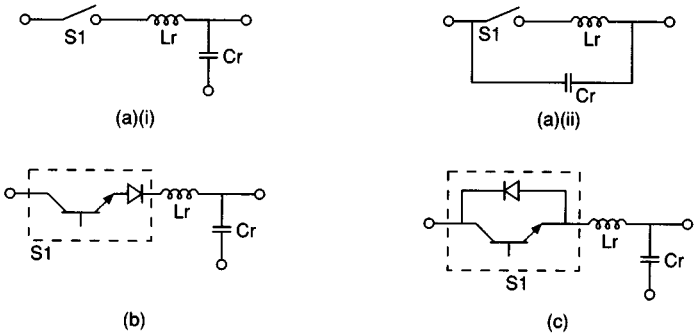


FIGURE 30.28 (a) Zero-current resonant switch: (i) L-type and (ii) M-type. (b) Half-wave configuration using L-type ZC resonant switch. (c) Full-wave configuration using L-type ZC resonant switch.

Current mode control (for the primary current) can be used to overcome the flux imbalance. This configuration is used in 100- to 500-W output range.

2. *The half-bridge.* In the half-bridge configuration (Fig. 30.26) center-tapped dc source is created by two smoothing capacitors (C_{in}), and this configuration utilizes the transformer core efficiently. The voltage across each transistor is equal to the supply voltage (half of push-pull) and, therefore, is suitable for high-voltage inputs. One salient feature of this configuration is that the input filter capacitors can be used to change between 110/220-V mains as selectable inputs to the supply.

   The disadvantage of this configuration is the requirement for large-size input filter capacitors. The half-bridge configuration is used for power levels of the order of 500 to 1000 W.

3. The full-bridge configuration (Fig. 30.27) requires only one smoothing capacitor, and for the same transistor type as that of half-bridge, output power can be doubled. It is usually used for power levels above 1 kW, and the design is more costly due to increased number of components (uses four transistors compared to two in push-pull and half-bridge converters).

   One of the salient features of a full-bridge converter is that by using proper control technique it can be operated in zero-voltage switching (ZVS) mode. This type of operation results in negligible switching losses. However, at reduced load currents, the ZVS property is lost. Recently, there has been a lot of effort to overcome this problem.

Resonant Power Supplies

Similar to the PWM converters, there are two types of resonant converters: single-ended and double-ended. Resonant converter configurations are obtained from the PWM converters explained earlier by adding LC (inductor-capacitor) resonating elements to obtain sinusoidally varying voltage and/or current waveforms. This approach reduces the switching losses and the switch stresses during switching instants, enabling the converter to operate at high switching frequencies, resulting in reduced size, weight, and cost. Some other advantages of resonant converters are that leakage inductances of HF transformers and the junction capacitances of semiconductors can

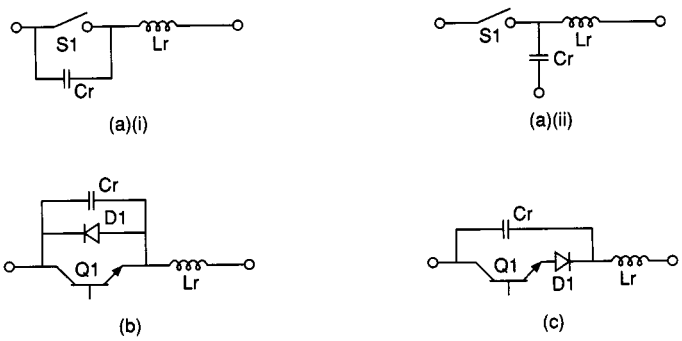


FIGURE 30.29 (a) Zero-voltage resonant switches. (b) Half-wave configuration using ZV resonant switch shown in Fig. (a)(i). (c) Full-wave configuration using ZV resonant switch shown in Fig. (a)(i).

be used profitably in the resonant circuit, and reduced EMI. The major disadvantage of resonant converters is increased peak current (or voltage) stress.

Single-Ended Resonant Converters. They are referred to as quasi-resonant converters (QRCs) since the voltage (or current) waveforms are quasi-sinusoidal in nature. The QRCs can operate with zero-current switching (ZCS) or ZVS or both. All the QRC configurations can be generated by replacing the conventional switches by the resonant switches shown in Figs. 30.28 and 30.29. A number of configurations are realizable. Basic principles of ZCS and ZVS are explained briefly below.

1. *Zero-current switching QRCs* [Sum, 1988; Liu et al., 1985]. Figure 30.30(a) shows an example of a ZCS QR buck converter implemented using a ZC resonant switch. Depending on whether the resonant switch is half-wave type or full-wave type, the resonating current will be only half-wave sinusoidal [Fig. 30.30(b)] or a full sine-wave [Fig. 30.30(c)]. The device currents are shaped sinusoidally, and, therefore, the switching losses are almost negligible with low turn-on and turn-off stresses. ZCS QRCs can operate at frequencies of the order of 2 MHz. The major problems with this type of converter are high peak currents through the switch and capacitive turn-on losses.
2. *Zero-voltage switching QRCs* [Sum, 1988; Liu and Lee, 1986]. ZVS QRCs are duals of ZCS QRCs. The auxiliary LC elements are used to shape the switching device's voltage waveform at off time in order to create a zero-voltage condition for the device to turn on. Fig. 30.31(a) shows an example of ZVS QR boost converter implemented using a ZV resonant switch. The circuit can operate in the half-wave mode [Fig. 30.31(b)] or in the full-wave mode [Fig. 30.31(c)] depending on whether a half-wave or full-wave ZV resonant switch is used, and the name comes from the capacitor voltage waveform. The full-wave mode ZVS circuit suffers from capacitive turn-on losses. The ZVS QRCs suffer from increased voltage stress on the switch. However, they can be operated at much higher frequencies compared to ZCS QRCs.

Double-Ended Resonant Converters. These converters [Sum, 1988; Bhat, 1991; Steigerwald, 1988; Bhat, 1992] use full-wave rectifiers at the output, and they are generally referred to as resonant converters. A number of resonant converter configurations are realizable by using different resonant tank circuits, and the three most popular configurations, namely, the series resonant converter (SRC), the parallel resonant converter (PRC), and the series-parallel resonant converter (SPRC) (also called LCC-type PRC), are shown in Fig. 30.32.

Series resonant converters [Fig. 30.32(a)] have high efficiency from full load to part load. Transformer saturation is avoided due to the series blocking resonating capacitor. The major problems with the SRC are that it requires a very wide change in switching frequency to regulate the load voltage and the output filter capacitor must carry high ripple current (a major problem especially in low output voltage, high output current applications).

Parallel resonant converters [Fig. 30.32(b)] are suitable for low output voltage, high output current applications due to the use of filter inductance at the output with low ripple current requirements for the filter capacitor. The major disadvantage of the PRC is that the device currents do not decrease with the load current, resulting in reduced efficiency at reduced load currents.

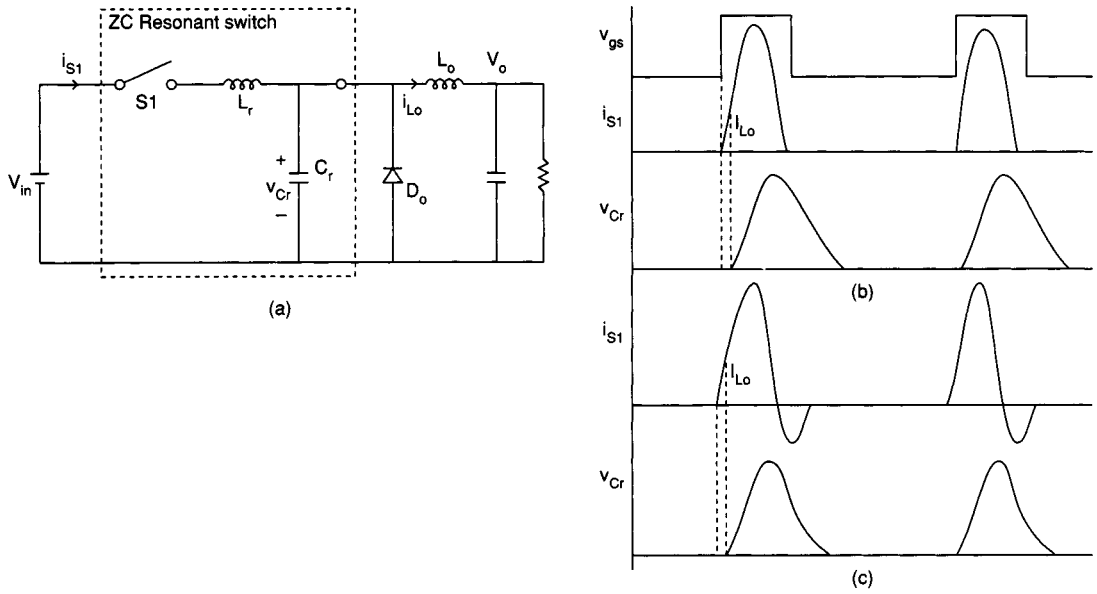


FIGURE 30.30 (a) Implementation of ZCS QR buck converter using L-type resonant switch. (b) Operating waveforms for half-wave mode. (c) Operating waveforms for full-wave mode.

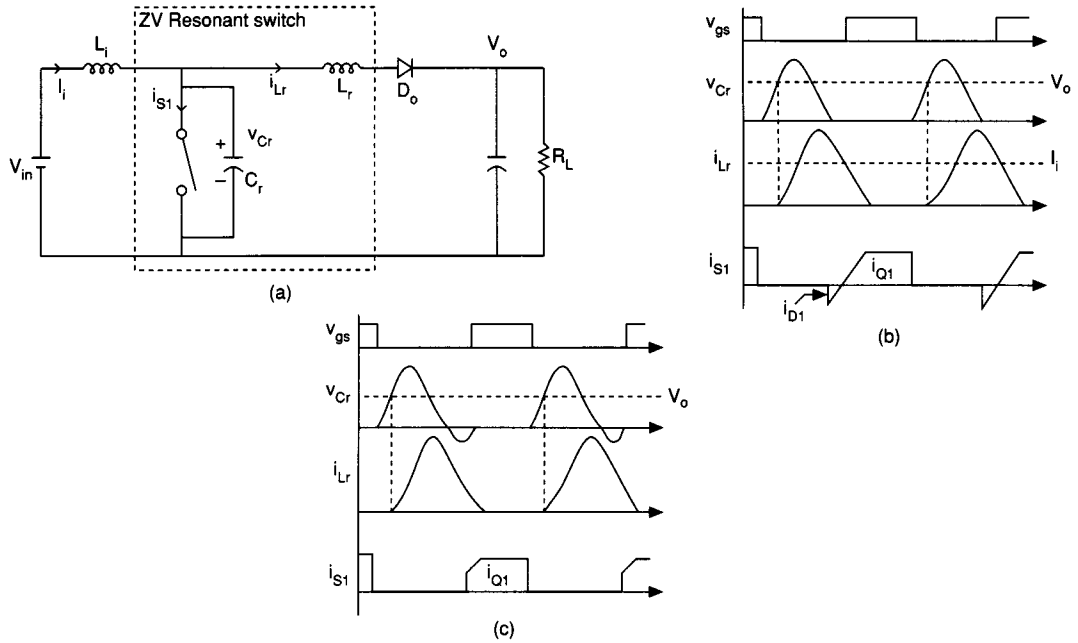


FIGURE 30.31 (a) Implementation of ZVS QR buck converter using resonant switch shown in Fig. 30.28(a)(i). (b) Operating waveforms for half-wave mode. (c) Operating waveforms for full-wave mode.

The SPRC [Fig. 30.32(c)] takes the desirable features of SRC and PRC.

Load voltage regulation in resonant converters for input supply variations and load changes is achieved by either varying the switching frequency or using fixed-frequency (variable pulsewidth) control.

1. *Variable-frequency operation.* Depending on whether the switching frequency is below or above the natural resonance frequency (ω_r), the converter can operate in different operating modes as explained below.

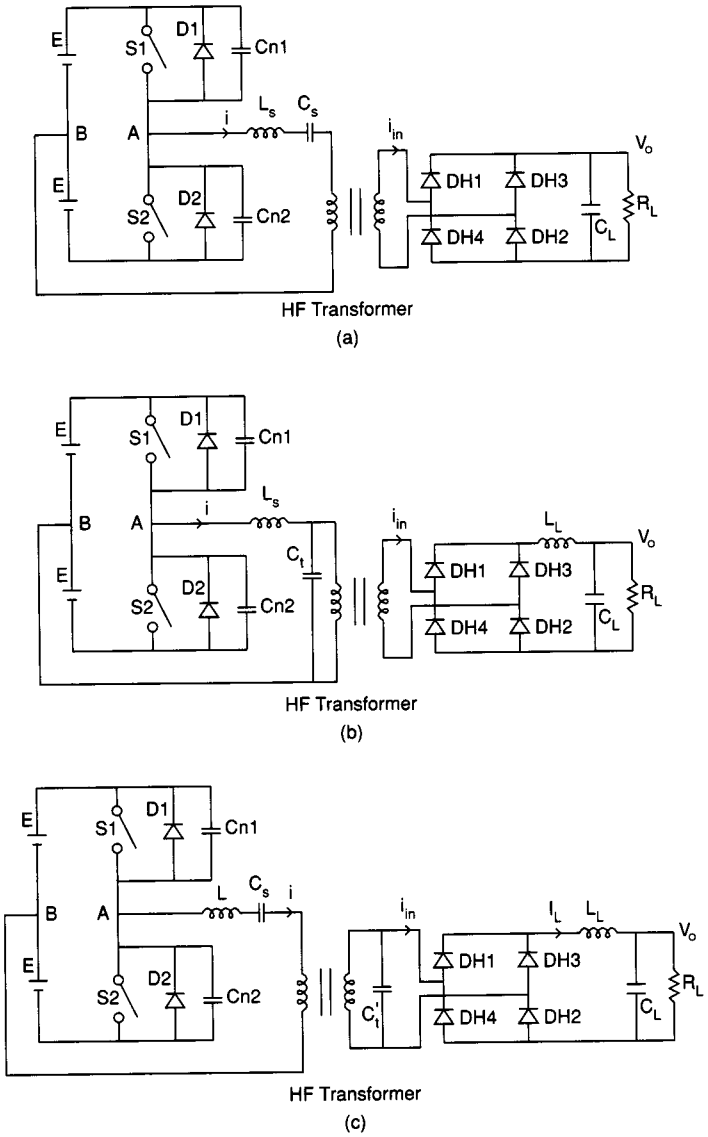


FIGURE 30.32 High-frequency resonant converter (half-bridge version) configurations suitable for operation above resonance. *Cn*1 and *Cn*2 are the snubber capacitors. (*Note:* For operation below resonance, *di/dt* limiting inductors and RC snubbers are required. For operation above resonance, only capacitive snubbers are required as shown.) (a) Series resonant converter. Leakage inductances of the HF transformer can be part of resonant inductance. (b) Parallel resonant converter. (c) Series-parallel (or LCC-type) resonant converter with capacitor C_t placed on the secondary side of the HF transformer.

a. Below-resonance (leading PF) mode. When the switching frequency is below the natural resonance frequency, the converter operates in a below-resonance mode (Fig. 30.33). The equivalent impedance across *AB* presents a leading PF so that natural turn-off of the switches is assured and any type of fast turn-off switch (including asymmetric SCRs) can be used. Depending on the instant of turn-on of switches S_1 and S_2, the converter can enter into two modes of operation, namely, continuous and discontinuous current modes. The steady-state operation in continuous current mode (CCM) [Fig. 30.33(a)] is explained briefly as follows.

Assume that diode D_2 was conducting and switch S_1 is turned on. The current carried by D_2 will be transferred to S_1 almost instantaneously (except for a small time of recovery of D_2 during which input supply is shorted through D_2 and S_1, and the current is limited by the *di/dt* limiting inductors). The

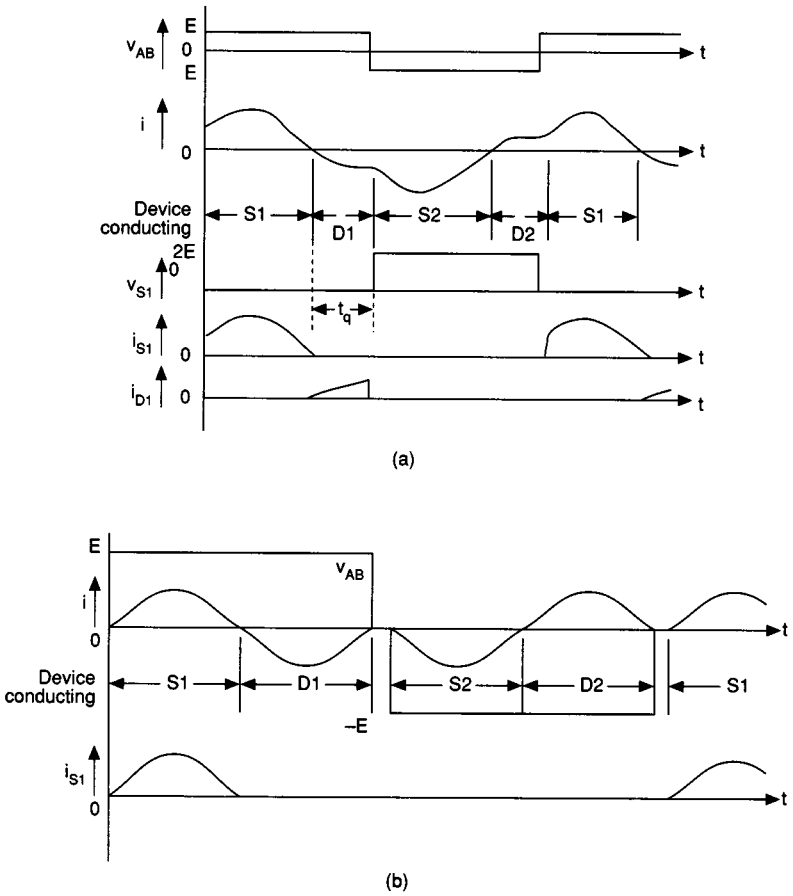


FIGURE 30.33 Typical waveforms at different points of a resonant converter operating below resonance (a) in continuous current mode and (b) in discontinuous current mode.

current i then oscillates sinusoidally and goes to zero in the natural way. The current tries to reverse, and the path for this current is provided by the diode D_1. Conduction of D_1 feeds the reactive energy in the load and the tank circuit back to the supply. The on-state of D_1 also provides a reverse voltage across S_1, allowing it to turn off. After providing a time equal to or greater than the turn-off time of S_1, switch S_2 can be turned on to initiate the second half cycle. The process is similar to the first half cycle, with the voltage across v_{AB} being of opposite polarity, and the functions of D_1, S_1 will be assumed by D_2, S_2. With this type of operation, the converter works in the continuous current mode as the switches are turned on before the currents in the diodes reach zero. If the switching on of S_1 and S_2 is delayed such that the currents through the previously conducting diodes reach zero, then there are zero current intervals and the **inverter** operates in the DCM [Fig. 30.33(b)].

Load voltage regulation is achieved by decreasing the switching frequency below the rated value. Since the inverter output current i leads the inverter output voltage v_{AB}, this type of operation is also called a leading PF mode of operation. If transistors are used as the switching devices, then for operation in DCM, the pulsewidth can be kept constant while decreasing the switching frequency to avoid CCM operation. DCM operation has the advantages of negligible switching losses due to ZCS, lower *di/dt* and *dv/dt* stresses, and simple control circuitry. However, DCM operation results in higher switch peak currents.

From the waveforms shown in Fig. 30.33, the following problems can be identified for operation in the below-resonance mode: requirement of *di/dt* inductors to limit the large turn-on switch currents and a need for lossy RC snubbers and fast recovery diodes. Since the switching frequency is decreased to control the load power, the HF transformer and magnetics must be designed for the lowest switching frequency, resulting in increased size of the converter.

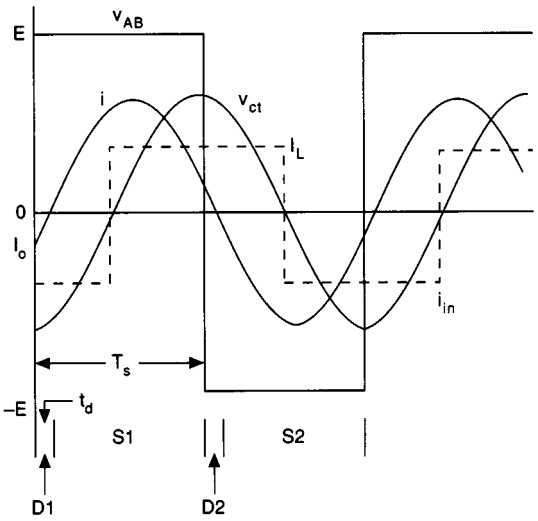


FIGURE 30.34 Typical operating waveforms at different points of an SPRC operating above resonance.

b. Above-resonance (lagging PF) mode. If switches capable of gate or base turn-off (e.g., MOSFETs, bipolar transistors) are used, then the converter can operate in the above-resonance mode (lagging PF mode). Figure 30.34 shows some typical operating waveforms for such type of operation, and it can be noticed that the current i lags the voltage v_{AB}. Since the switch takes current from its own diode across it at zero-current point, there is no need for di/dt limiting inductance, and a simple capacitive snubber can be used. In addition, the internal diodes of MOSFETs can be used due to the large turn-off time available for the diodes. Major problems with the lagging PF mode of operation are that there are switch turn-off losses, and since the voltage regulation is achieved by increasing the switching frequency above the rated value, the magnetic losses increase and the design of a control circuit is difficult.

Exact analysis of resonant converters is complex due to the nonlinear loading on the resonant tanks. The rectifier-filter-load resistor block can be replaced by a square-wave voltage source [for SRC, Fig. 30.32(a)] or a square-wave current source [for PRC and SPRC, Fig. 30.32(b) and (c)]. Using fundamental components of the waveforms, an approximate analysis [Bhat, 1991; Steigerwald, 1988] using a phasor circuit gives a reasonably good design approach. This analysis approach is illustrated next for the SPRC.

2. *Approximate analysis of SPRC.* Figure 30.35 shows the equivalent circuit at the output of the inverter and the phasor circuit used for the analysis. All the equations are normalized using the base quantities

$$\text{Base voltage } V_B = E_{\min}$$

$$\text{Base impedance } Z_B = R'_L = n^2 R_L$$

$$\text{Base current } I_B = V_B/I_B$$

The converter gain [normalized output voltage in per unit (p.u.) referred to the primary-side] can be derived as [Bhat, 1991; Steigerwald, 1988]

$$M = \frac{1}{\left\{\left(\frac{\pi^2}{8}\right)^2\left[1+\left(\frac{C_t}{C_s}\right)\left(1-y_s^2\right)\right]^2 + Q_s^2\left[y_s - \left(\frac{1}{y_s}\right)\right]^2\right\}^{1/2}} \text{ p.u.} \qquad (30.1)$$

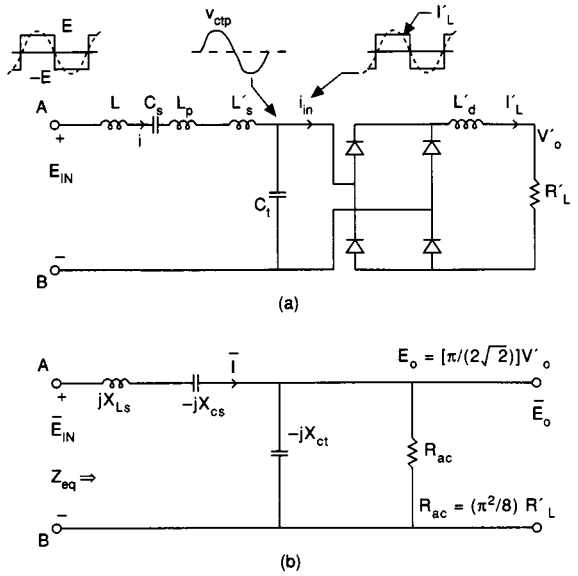


FIGURE 30.35 (a) Equivalent circuit for a SPRC at the output of the inverter terminals (across *AB*) of Fig. 30.31(c), L_p and L'_s are the leakage inductance of the primary and primary referred leakage inductance of the secondary, respectively. (b) Phasor circuit model used for the analysis of the SPRC converter.

where

$$Q_s = \frac{(L_s / C_s)^{1/2}}{R'_L}; L_s = L + L_p + L'_s \tag{30.2}$$

$$y_s = \frac{f_s}{f_r} \tag{30.3}$$

and

$$\begin{aligned} f_s &= \text{switching frequency} \\ f_r &= \text{series resonance frequency} \\ &= \frac{\omega_r}{2\pi} = \frac{1}{2\pi(L_s C_s)^{1/2}} \end{aligned} \tag{30.4}$$

The equivalent impedance looking into the terminals *AB* is given by

$$Z_{eq} = \frac{B_1 + jB_2}{B_3} \quad \text{p.u.} \tag{30.5}$$

where

$$B_1 = \left(\frac{8}{\pi^2}\right)\left(\frac{C_s}{C_t}\right)^2\left(\frac{Q_s}{y_s}\right)^2 \tag{30.6}$$

$$B_2 = Q_s\left(y_s - \frac{1}{y_s}\right)\left[1 + \left(\frac{8}{\pi^2}\right)^2\left(\frac{C_s}{C_t}\right)^2\left(\frac{Q_s}{y_s}\right)^2\right] - \left(\frac{C_s}{C_t}\right)\left(\frac{Q_s}{y_s}\right) \tag{30.7}$$

$$B_3 = 1 + \left(\frac{8}{\pi^2}\right)^2\left(\frac{C_s}{C_t}\right)^2\left(\frac{Q_s}{y_s}\right)^2 \tag{30.8}$$

The peak inverter output (resonant inductor) current can be calculated using

$$I_p = \frac{4}{\pi|Z_{eq}|} \quad \text{p.u.} \tag{30.9}$$

The same current flows through the switching devices.

The value of initial current I_0 is given by

$$I_0 = I_p \sin(-\phi) \text{ p.u.} \tag{30.10}$$

where $\phi = \tan^{-1}(B_2/B_1)$ rad. B_1 and B_2 are given by Eqs. (30.6) and (30.7), respectively.

If I_0 is negative, then forced commutation is necessary and the converter is operating in the lagging PF mode. The peak voltage across the capacitor C'_t (on the secondary side) is

$$V_{\text{ctp}} = \frac{\pi}{2} V_o \quad \text{V} \tag{30.11}$$

The peak voltage across C_s and the peak current through C'_t are given by

$$V_{\text{csp}} = \frac{Q_s}{y_s} I_p \quad \text{p.u.} \tag{30.12}$$

$$I_{\text{ctp}} = \frac{V_{\text{ctp}}}{X_{\text{cptu}} R_L} \quad \text{A} \tag{30.13}$$

$$X_{\text{ctpu}} = \left(\frac{C_s}{C_t}\right)\left(\frac{Q_s}{y_s}\right) \quad \text{p.u.} \tag{30.14}$$

The plot of converter gain versus the switching frequency ratio y_s, obtained using (30.1), is shown for $C_s/C_t = 1$ in Fig. 30.36, for the lagging PF mode of operation. If the ratio C_s/C_t increases, then the converter takes the characteristics of SRC and the load voltage regulation requires a very wide range in the frequency change. Lower values of C_s/C_t take the characteristics of a PRC. Therefore, a compromised value of $C_s/C_t = 1$ is chosen.

It is possible to realize higher-order resonant converters with improved characteristics and many of them are presented in Bhat [1991].

3. *Fixed-frequency operation.* To overcome some of the problems associated with the variable frequency control of resonant converters, they are operated with fixed frequency [Sum, 1988; Bhat, 1992]. A number of configurations and control methods for fixed-frequency operation are available in the literature (Bhat [1992] gives a list of papers). One of the most popular methods of control is the phase-shift control

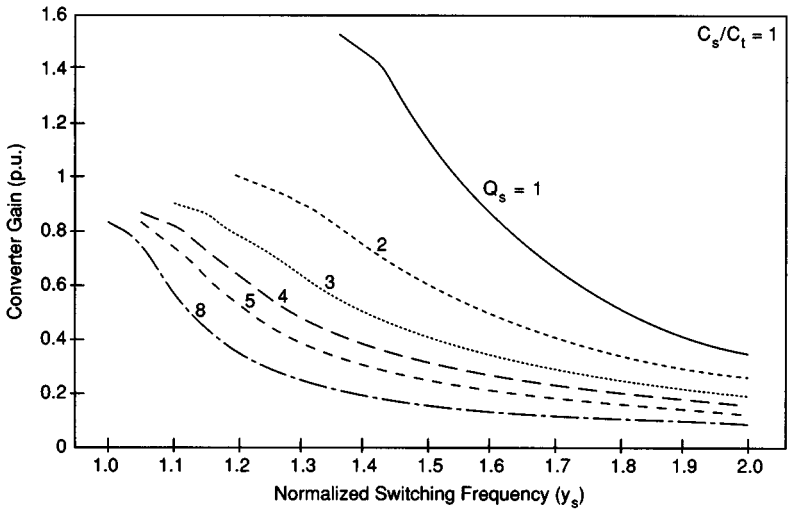


FIGURE 30.36 The converter gain M (p.u.) (normalized output voltage) versus normalized switching frequency y_s of SPRC operating above resonance for $C_s/C_t = 1$.

(also called clamped-mode or PWM operation) method. Figure 30.37 illustrates the clamped-mode fixed-frequency operation of the SPRC. The load power control is achieved by changing the phase-shift angle ϕ between the gating signals to vary the pulsewidth of v_{AB}.

4. *Design example.* Design a 500-W output SPRC (half-bridge version) with secondary-side resonance (operation in lagging PF mode and variable-frequency control) with the following specifications:

$$\text{Minimum input supply voltage} = 2E_{\min} = 230 \text{ V}$$

$$\text{Load voltage, } V_o = 48 \text{ V}$$

$$\text{Switching frequency, } f_s = 100 \text{ kHz}$$

$$\text{Maximum load current} = 10.42 \text{ A}$$

As explained in item 2, $C_s/C_t = 1$ is chosen. Using the constraints (1) minimum kVA rating of tank circuit per kW output power, (2) minimum inverter output peak current, and (3) enough turn-off time for the switches, it can be shown that [Bhat, 1991] $Q_s = 4$ and $y_s = 1.1$ satisfy the design constraints. From Fig. 30.36, $M = 0.8$ p.u.

Average load voltage referred to the primary side of the HF transformer $= 0.8 \times 115\,\text{V} = 92\,\text{V}$. Therefore, the transformer turns ratio required $\simeq 1.84$.

$$R_L' = n^2\left(\frac{V_o^2}{P_o}\right) = 15.6\ \Omega$$

The values of L_s and C_s can be obtained by solving

$$\left(\frac{L_s}{C_s}\right)^{1/2} = 4 \times 15.6\,\Omega \text{ and } \omega_r = \frac{1}{(L_sC_s)^{1/2}} = 2\pi\frac{f_s}{y_s}$$

Solving the above equations gives $L_s = 109\,\mu\text{H}$ and $C_s = 0.0281\,\mu\text{F}$. Leakage inductance $(L_p + L'_s)$ of the HF transformer can be used as part of L_s. Typical value for a 100-kHz practical transformer (using Tokin

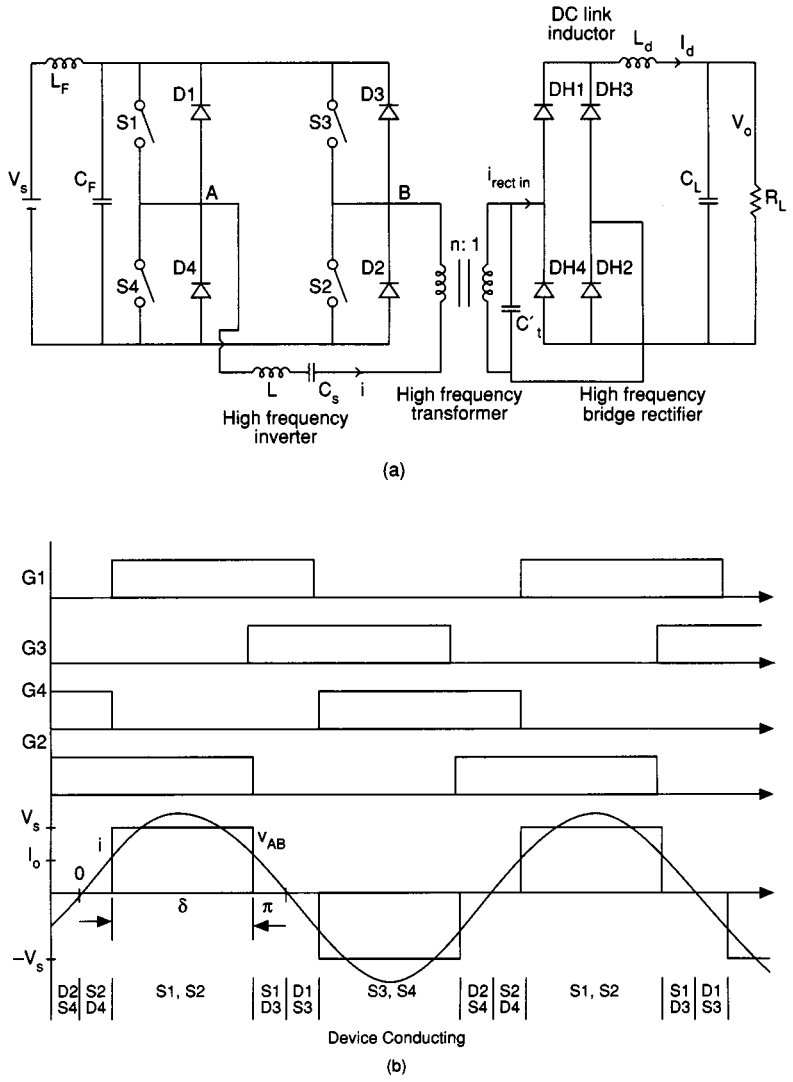


FIGURE 30.37 (a) Basic circuit diagram of series-parallel resonant converter suitable for fixed-frequency operation with PWM (clamped-mode) control. (b) Waveforms to illustrate the operation of fixed-frequency PWM series-parallel resonant converter working with a pulsewidth δ.

Mn-Zn 2500B2 Ferrite, E-I type core) for this application is about 5 μH. Therefore, the external resonant inductance required is $L = 104$ μH.

Since $C_s/C_t = 1$ is chosen, $C_t = 0.0281$ μF. The actual value of C_t used on the secondary side of the HF transformer = $(1.84)^2 \times 0.0281 = 0.09514$ μF. The resonating capacitors must be HF type (e.g., polypropylene) and must be capable of withstanding the voltage and current ratings obtained above (enough safety margin must be provided).

Using Eqs. (30.9) and (30.11) to (30.13):

$$\text{Peak current through switches} = 7.6 \text{ A}$$

$$\text{Peak voltage across } C_s,\ V_{csp} = 430 \text{ V}$$

$$\text{Peak voltage (on secondary side) across } C'_t,\ V_{ctp} = 76 \text{ V}$$

$$\text{Peak current through capacitor } C'_t \text{ (on secondary side)},\ I_{ctp} = 4.54 \text{ A}$$

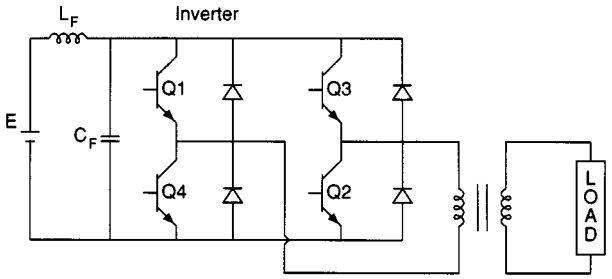


FIGURE 30.38 An inverter circuit to obtain variable-voltage, variable-frequency ac source. Using sinusoidal pulsewidth modulation control scheme, sine-wave ac output voltage can be obtained.

A simple control circuit can be built using PWM IC SG3525 and TSC429 MOSFET driver ICs.

With the development of digital ICs operating on low-voltage (of the order of 3 V) supplies, use of MOSFETs as *synchronous rectifiers* with very low voltage drop (~0.2 V) has become essential [Motorola, 1989] to increase the efficiency of the power supply.

AC Power Supplies

Some applications of ac power supplies are ac motor drives, **uninterruptible power supply (UPS)** used as a standby ac source for critical loads (e.g., in hospitals, computers), and dc source-to-utility interface (either to meet peak power demands or to augment energy by connecting unconventional energy sources like photovoltaic arrays to the utility line). In ac induction motor drives, the ac power main is rectified and filtered to obtain a smooth dc source, and then an inverter (single-phase version is shown in Fig. 30.38) is used to obtain a variable-frequency, variable-voltage ac source. The sinusoidal pulsewidth modulation technique described in Section 30.2 can be used to obtain a sinusoidal output voltage. Some other methods used to get sinusoidal voltage output are [Rashid, 1988] a number of phase-shifted inverter outputs summed in an output transformer to get a stepped waveform that approximates a sine wave and the use of a bang-bang controller in Fig. 30.38. All these methods use line-frequency (60 Hz) transformers for voltage translation and isolation purposes. To reduce the size, weight, and cost of such systems, one can use dc-to-dc converters (discussed earlier) as an intermediate stage. Figure 30.39 shows such a system in block schematic form. One can use an HF inverter circuit (discussed earlier) followed by a **cycloconverter** stage. The major problem with these schemes is the reduction in efficiency due to the extra power stage. Figure 30.40 shows a typical UPS scheme. The battery shown has to be charged by a separate rectifier circuit.

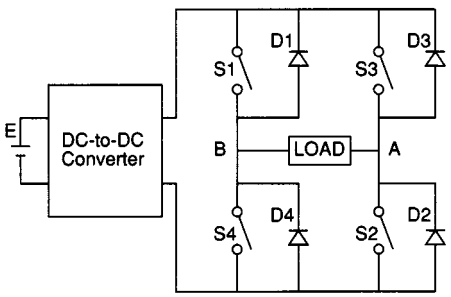


FIGURE 30.39 AC power supplies using HF switching (PWM or resonant) dc-to-dc converter as an input stage. HF transformer isolated dc-to-dc converters can be used to reduce the size and weight of the power supply. Sinusoidal voltage output can be obtained using the modulation in the output inverter stage or in the dc-to-dc converter.

AC-to-ac conversion can also be achieved using cycloconverters [e.g., Rashid, 1988].

Special Power Supplies

Using the inverters and cycloconverters, it is possible to realize bidirectional ac and dc power supplies. In these power supplies [Rashid, 1988], power can flow in both directions, i.e., from input to output or from output to input. It is also possible to control the ac-to-dc converters to obtain sinusoidal line current with unity PF and low harmonic distortion at the ac source.

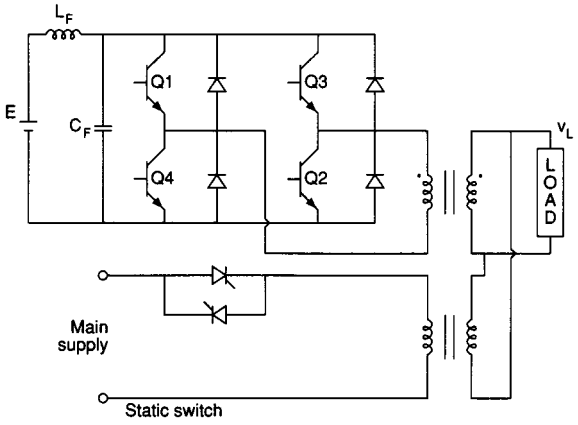


FIGURE 30.40 A typical arrangement of UPS system. The load gets power through the static switch when the ac main supply is present. The inverter supplies power when the main supply fails.

Defining Terms

Converter: A circuit that performs one of the following power conversions — ac to dc, dc to dc, dc to ac, or ac to ac.

Cycloconverter: A power electronic circuit that converts ac input to ac output (generally) of lower frequency than the input source without using any intermediate dc state.

Inverter: A power electronic circuit that converts dc input to ac output.

Isolated: A power electronic circuit that has ohmic isolation between the input source and the load circuit.

Pulsewidth-modulated (PWM) converters: A power electronic converter that employs square-wave switching waveforms with variation of pulsewidth for controlling the load voltage.

Regulated output: Output load voltage is kept at the required value for changes in either the load or the input supply voltage.

Resonant converters: A power electronic converter that employs "LC resonant circuits" to obtain sinusoidal switching waveforms.

Uninterruptible power supply (UPS): A stand-by dc-to-ac inverter used mostly to provide an emergency power to loads at mains frequency (50/60 Hz) in the event of a mains failure.

References

A.K.S. Bhat, "A unified approach for the steady-state analysis of resonant converters," *IEEE Trans. Industrial Electronics*, vol. 38, no. 4, pp. 251–259, Aug. 1991.

A.K.S. Bhat, "Fixed frequency PWM series-parallel resonant converter," *IEEE Trans. Industry Applications*, vol. 28, no. 5, pp. 1002–1009, 1992.

E.R. Hnatek, *Design of Solid-State Power Supplies*, 2nd ed., New York: Van Nostrand Reinhold, 1981.

K.H. Liu and F.C. Lee, "Zero-Voltage Switching Technique In DC/DC Converters," IEEE Power Electronics Specialists Conference Record, 1986, pp. 58–70.

K.H. Liu, R. Oruganti, and F.C. Lee, "Resonant Switches—Topologies and Characteristics," IEEE Power Electronics Specialists Conference Record, 1985, pp. 106–116.

Motorola, *Linear/Switchmode Voltage Regulator Handbook*, 1989.

Philips Semiconductors, *Power Semiconductor Applications*, 1991.

M.H. Rashid, *Power Electronics: Circuits, Devices, and Applications*, Englewood Cliffs, N.J.: Prentice-Hall, 1988.

R. Severns and G. Bloom, *Modern Switching DC-to-DC Converters*, New York: Van Nostrand Reinhold, 1988.

R.L. Steigerwald, "A comparison of half-bridge resonant converter topologies," *IEEE Trans. Power Electron.*, vol. PE-3, no. 2, pp. 174–182, April 1988.

K.K. Sum, *Recent Developments in Resonant Power Conversion*, Calif.: Intertech Communications, 1988.

Unitrode Switching Regulated Power Supply Design Seminar Manual, Lexington, Mass.: Unitrode Corporation, 1984.

Further Information

The following monthly magazines and conference records publish papers on the analysis, design, and experimental aspects of power supply configurations and their applications:

IEEE Transactions on Power Electronics, IEEE Transactions on Industrial Electronics, IEEE Transactions on Industry Applications, and *IEEE Transactions on Aerospace and Electronic Systems.*

IEEE Power Electronics Specialists Conference Records, IEEE Applied Power Electronics Conference Records, IEEE Industry Applications Conference Records, and *IEEE International Telecommunications Energy Conference Records.*

30.4 Converter Control of Machines

Bimal K. Bose

Converter-controlled electrical machine drives are very important in modern industrial applications. Some examples in the high-power range are metal rolling mills, cement mills, and gas line compressors. In the medium-power range are textile mills, paper mills, and subway car propulsion. Machine tools and computer peripherals are examples of converter-controlled electrical machine drive applications in the low-power range. The converter normally provides a variable-voltage dc power source for a dc motor drive and a variable-frequency, variable-voltage ac power source for an ac motor drive. The drive system efficiency is high because the converter operates in switching mode using power semiconductor devices. The primary control variable of the machine may be torque, speed, or position, or the converter can operate as a solid-state starter of the machine. The recent evolution of high-frequency power semiconductor devices and high-density and economical microelectronic chips, coupled with converter and control technology developments, is providing a tremendous boost in the applications of drives.

Converter Control of DC Machines

The speed of a dc motor can be controlled by controlling the dc voltage across its armature terminals. A phase-controlled thyristor converter can provide this dc voltage source. For a low-power drive, a single-phase bridge converter can be used, whereas for a high-power drive, a three-phase bridge circuit is preferred. The machine can be a permanent magnet or wound field type. The wound field type permits variation and reversal of field and is normally preferred in large power machines.

Phase-Controlled Converter DC Drive

Figure 30.41 shows a dc drive using a three-phase thyristor bridge converter. The converter rectifies line ac voltage to variable dc output voltage by controlling the firing angle of the thyristors. With rated field excitation, as the armature voltage is increased, the machine will develop speed in the forward direction until the rated, or base, speed is developed at full voltage when the firing angle is zero. The motor speed can be increased further by weakening the field excitation. Below the base speed, the machine is said to operate in constant

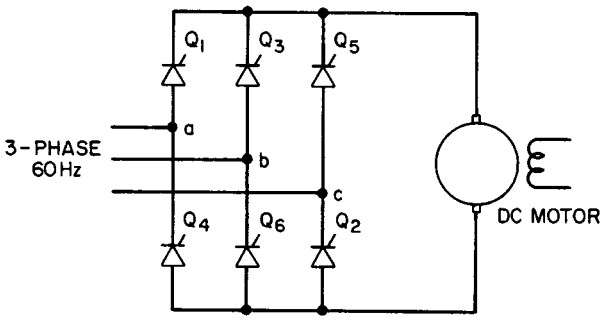


FIGURE 30.41 Three-phase thyristor bridge converter control of a dc machine.

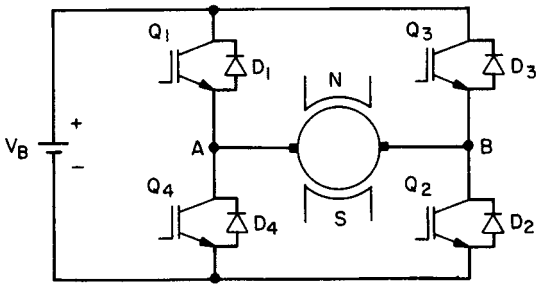


FIGURE 30.42 Four-quadrant dc motor drive using an H-bridge converter.

torque region, whereas the field weakening mode is defined as the constant power region. At any operating speed, the field can be reversed and the converter firing angle can be controlled beyond 90 degrees for **regenerative braking** mode operation of the drive. In this mode, the motor acts as a generator (with negative induced voltage) and the converter acts as an inverter so that the mechanical energy stored in the inertia is converted to electrical energy and pumped back to the source. Such **two-quadrant** operation gives improved efficiency if the drive accelerates and decelerates frequently. The speed of the machine can be controlled with precision by a feedback loop where the command speed is compared with the machine speed measured by a tachometer. The speed loop error generally generates the armature current command through a compensator. The current is then feedback controlled with the firing angle control in the inner loop. Since torque is proportional to armature current (with fixed field), a current loop provides direct torque control, and the drive can accelerate or decelerate with the rated torque. A second bridge converter can be connected in antiparallel so that the dual converter can control the machine speed in all the four quadrants (motoring and regeneration in forward and reverse speeds).

Pulsewidth Modulation Converter DC Machine Drive

Four-quadrant speed control of a dc drive is also possible using an H-bridge pulsewidth modulation (PWM) converter as shown in Fig. 30.42. Such drives (using a permanent magnet dc motor) are popular in low-power applications, such as robotic and instrumentation drives. The dc source can be a battery or may be obtained from ac supply through a diode rectifier and filter. With PWM operation, the drive response is very fast and the armature current ripple is small, giving less harmonic heating and torque pulsation. Four-quadrant operation can be summarized as follows:

Quadrant 1: Forward motoring (buck or step-down converter mode)
- Q_1—on
- Q_3, Q_4—off
- Q_2—chopping
- Current freewheeling through D_3 and Q_1

Quadrant 2: Forward regeneration (boost or step-up converter mode)
- Q_1, Q_2, Q_3—off
- Q_4—chopping
- Current freewheeling through D_1 and D_2

Quadrant 3: Reverse motoring (buck converter mode)
- Q_3—on
- Q_1, Q_2—off
- Q_4—chopping
- Current freewheeling through D_1 and Q_3

Quadrant 4: Reverse regeneration (boost converter mode)
- Q_1, Q_3, Q_4—off
- Q_2—chopping
- Current freewheeling through D_3 and D_4

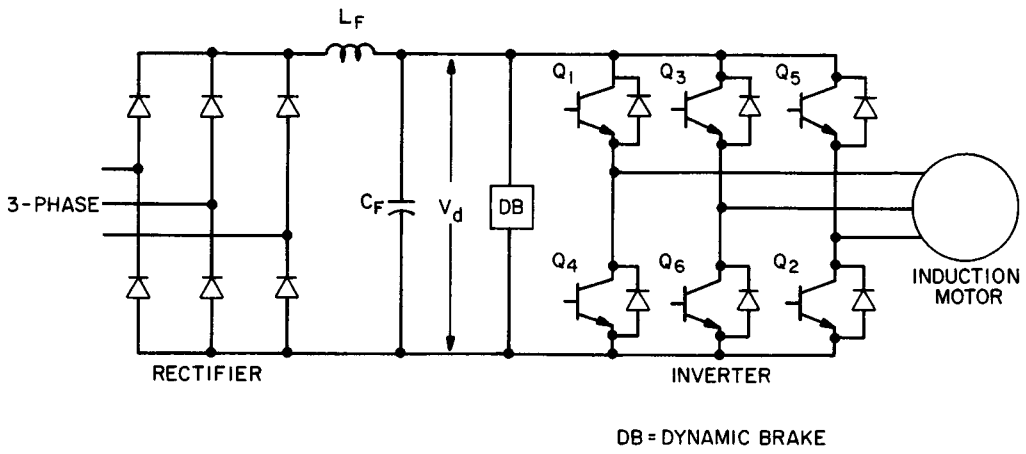


FIGURE 30.43 Diode rectifier PWM inverter control of an induction motor.

Often a drive may need only a one- or two-quadrant mode of operation. In such a case, the converter topology can be simple. For example, in one-quadrant drive, only Q_2 chopping and D_3 freewheeling devices are required, and the terminal A is connected to the supply positive. Similarly, a two-quadrant drive will need only one leg of the bridge, where the upper device can be controlled for motoring mode and the lower device can be controlled for regeneration mode.

Converter Control of AC Machines

Although application of dc drives is quite common, disadvantages are that the machines are bulky and expensive, and the commutators and brushes require frequent maintenance. In fact, commutator sparking prevents machine application in an unclean environment, at high speed, and at high elevation. AC machines, particularly the cage-type induction motor, are favorable when compared with all the features of dc machines. Although converter system, control, and signal processing of ac drives is definitely complex, the evolution of ac drive technology in the past two decades has permitted more economical and higher performance ac drives. Consequently, ac drives are finding expanding applications, pushing dc drives towards obsolescence.

Voltage-Fed Inverter Induction Motor Drive

A simple and popular converter system for speed control of an induction motor is shown in Fig. 30.43. The front-end diode rectifier converts 60 Hz ac to dc, which is then filtered to remove the ripple. The dc voltage is then converted to variable-frequency, variable-voltage output for the machine through a PWM bridge inverter. Among a number of PWM techniques, the sinusoidal PWM is common, and it is illustrated in Fig. 30.44 for one phase only. The stator sinusoidal reference phase voltage signal is compared with a high-frequency carrier wave, and the comparator logic output controls switching of the upper and lower transistors in a phase leg. The phase voltage wave shown refers to the fictitious center tap of the filter capacitor. With the PWM technique, the fundamental voltage and frequency can be easily varied. The stator voltage wave contains high-frequency ripple, which is easily filtered by the machine leakage inductance. The voltage-to-frequency ratio is kept constant to provide constant airgap flux in the machine. The machine voltage-frequency relation, and the corresponding torque, stator current, and slip, are shown in Fig. 30.45. Up to the base or rated frequency ω_b, the machine can develop constant torque. Then, the field flux weakens as the frequency is increased at constant voltage. The speed of the machine can be controlled in a simple open-loop manner by controlling the frequency and maintaining the proportionality between the voltage and frequency. During acceleration, machine-developed torque should be limited so that the inverter current rating is not exceeded. By controlling the frequency, the operation can be extended in the field weakening region. If the supply frequency is controlled to be lower than the machine speed (equivalent frequency), the motor will act as a generator and the inverter will act as a rectifier, and energy from the motor will be pumped back to the dc link. The **dynamic brake** shown is nothing but a buck converter with resistive load that dissipates excess power to maintain the dc bus voltage constant. When

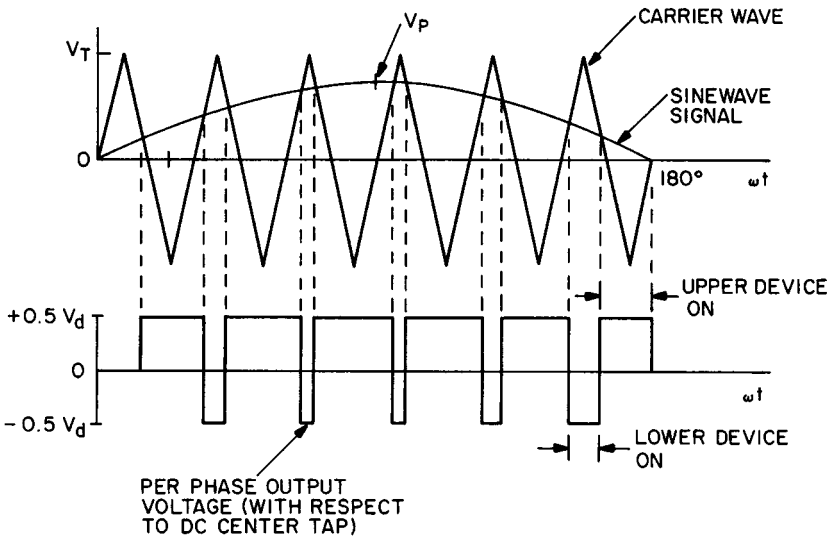


FIGURE 30.44 Sinusoidal pulse width modulation principle.

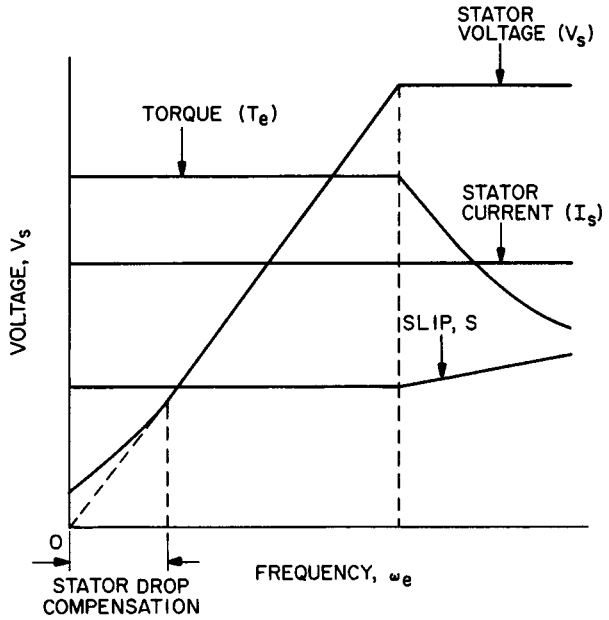


FIGURE 30.45 Voltage-frequency relation of an induction motor.

the motor speed is reduced to zero, the phase sequence of the inverter can be reversed for speed reversal. Therefore, the machine speed can be easily controlled in all four quadrants.

Current-Fed Inverter Induction Motor Drive

The speed of a machine can be controlled by a current-fed inverter as shown in Fig. 30.46. The front-end thyristor rectifier generates a variable dc current source in the dc link inductor. The dc current is then converted to six-step machine current wave through the inverter. The basic mode of operation of the inverter is the same as that of the rectifier, except that it is **force-commutated**, that is, the capacitors and series diodes help commutation of the thyristors. One advantage of the drive is that regenerative braking is easy because the

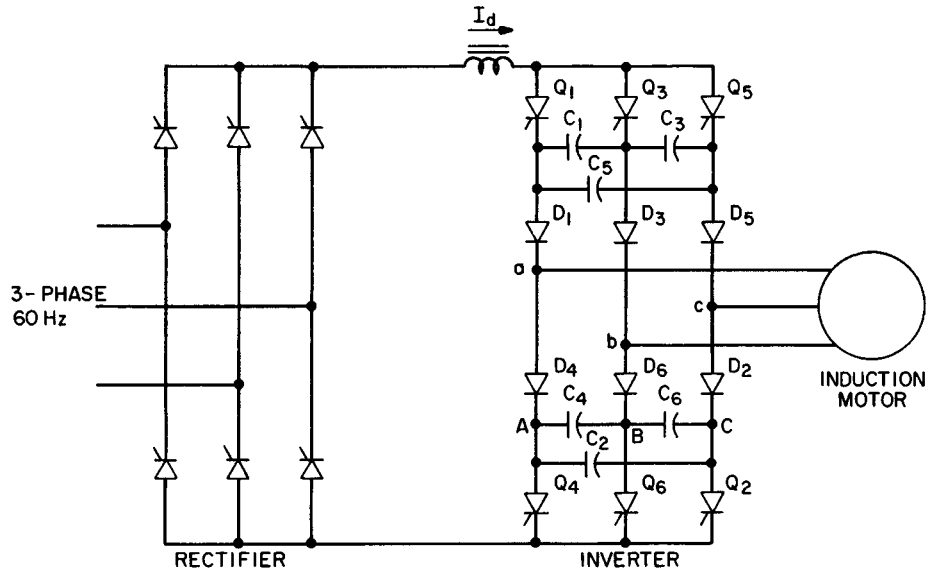


FIGURE 30.46 Force-commutated current-fed inverter control of an induction motor.

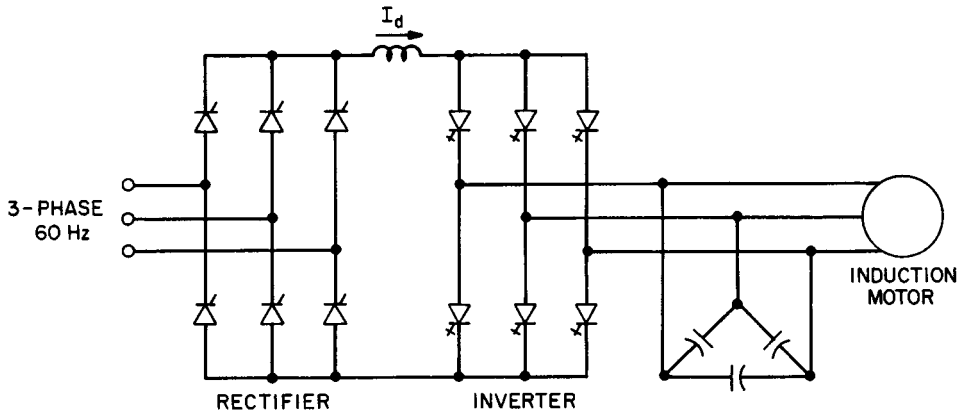


FIGURE 30.47 PWM current-fed inverter control of an induction motor.

rectifier and inverter can reverse their operation modes. Six-step machine current, however, causes large harmonic heating and torque pulsation, which may be quite harmful at low-speed operation. Another disadvantage is that the converter system cannot be controlled in open loop like a voltage-fed inverter.

Current-Fed PWM Inverter Induction Motor Drive

The force-commutated thyristor inverter in Fig. 30.46 can be replaced by a **self-commutating** gate turn-off (GTO) thyristor PWM inverter as shown in Fig. 30.47. The output capacitor bank shown has two functions: (1) it permits PWM switching of the GTO by diverting the load inductive current, and (2) it acts as a low-pass filter causing sinusoidal machine current. The second function improves machine efficiency and attenuates the irritating magnetic noise. Note that the fundamental machine current is controlled by the front-end rectifier, and the fixed PWM pattern is for controlling the harmonics only. The GTO is to be the reverse-blocking type. Such drives are popular in the multimegawatt power range. For lower power, an **insulated gate bipolar transistor (IGBT)** or transistor can be used with a series diode.

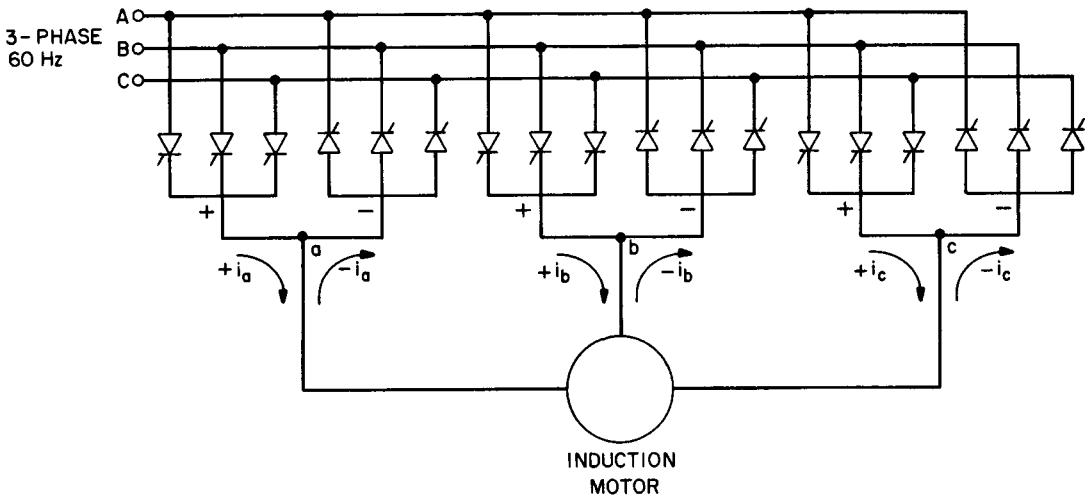


FIGURE 30.48 Cycloconverter control of an induction motor.

Cycloconverter Induction Motor Drive

A phase-controlled cycloconverter can be used for speed control of an ac machine (induction or synchronous type). Figure 30.48 shows a drive using a three-pulse half-wave or 18-thyristor cycloconverter. Each output phase group consists of positive and negative converter components which permit bidirectional current flow. The firing angle of each converter is sinusoidally modulated to generate the variable-frequency, variable-voltage output required for ac machine drive. Speed reversal and regenerative mode operation are easy. The cycloconverter can be operated in blocking or circulating current mode. In blocking mode, the positive or negative converter is enabled, depending on the polarity of the load current. In circulating current mode, the converter components are always enabled to permit circulating current through them. The circulating current reactor between the positive and negative converter prevents short circuits due to ripple voltage. The circulating current mode gives simple control and a higher range of output frequency with lower harmonic distortion.

Slip Power Recovery Drive of Induction Motor

In a cage-type induction motor, the rotor current at slip frequency reacting with the airgap flux develops the torque. The corresponding slip power is dissipated in the rotor resistance. In a wound rotor induction motor, the slip power can be controlled to control the torque and speed of a machine. Figure 30.49 shows a popular slip power-controlled drive, known as a static Kramer drive. The slip power is rectified to dc with a diode rectifier and is then pumped back to an ac line through a thyristor phase-controlled inverter. The method permits speed control in the subsynchronous speed range. It can be shown that the developed machine torque is proportional to the dc link current I_d and the voltage V_d varies directly with speed deviation from the synchronous speed. The current I_d is controlled by the firing angle of the inverter. Since V_d and V_I voltages balance at steady state, at synchronous speed the voltage V_d is zero and the firing angle is 90 degrees. The firing angle increases as the speed falls, and at 50% synchronous speed the firing angle is near 180 degrees. This is practically the lowest speed in static Kramer drive. The transformer steps down the inverter input voltage to get a 180-degree firing angle at lowest speed. The advantage of this drive is that the converter rating is low compared with the machine rating. Disadvantages are that the line power factor is low and the machine is expensive. For limited speed range applications, this drive has been popular.

Wound Field Synchronous Motor Drive

The speed of a wound field synchronous machine can be controlled by a current-fed converter scheme as shown in Fig. 30.46, except that the forced-commutation elements can be removed. The machine is operated at leading power factor by overexcitation so that the inverter can be load commutated. Because of the simplicity of converter topology and control, such a drive is popular in the multimegawatt range.

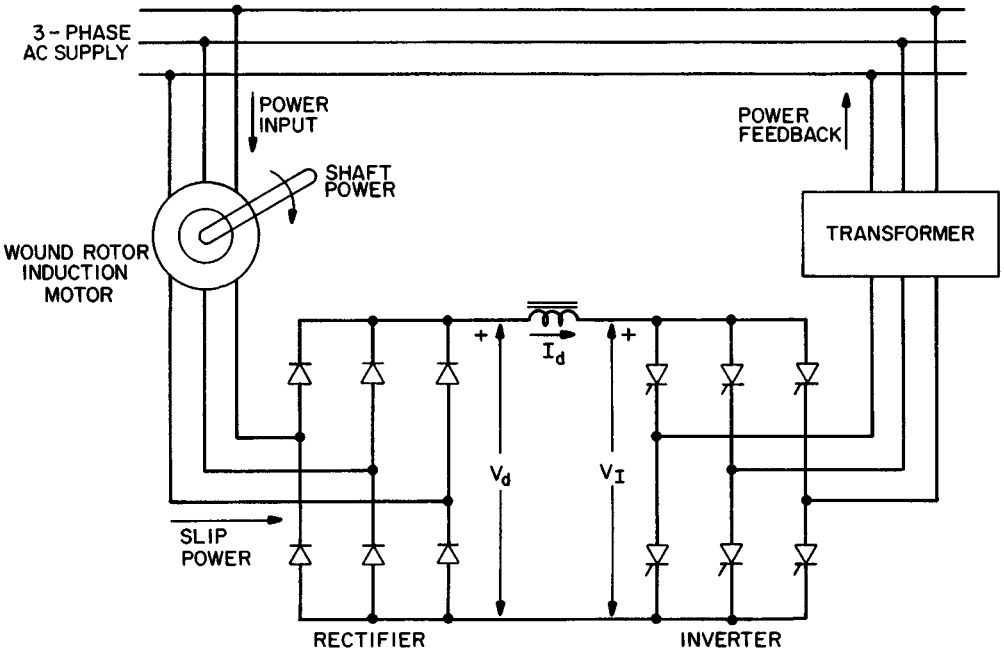


FIGURE 30.49 Slip power recovery control of a wound rotor induction motor.

Permanent Magnet Synchronous Motor Drive

Permanent magnet (PM) machine drives are quite popular in the low-power range. A PM machine can have sinusoidal or concentrated winding, giving the corresponding sinusoidal or trapezoidal induced stator voltage wave. Figure 30.50 shows the speed control system using a trapezoidal machine, and Fig. 30.51 explains the wave forms. The power MOSFET inverter supplies variable-frequency, variable-magnitude six-step current wave to the stator. The inverter is self-controlled, that is, the firing pulses are generated by the machine position sensor through a decoder. It can be shown that such a drive has the features of dc drive and is normally defined as *brushless dc drive*. The speed control loop generates the dc current command, which is then controlled by the **hysteresis-band** method to construct the six-step phase current waves in correct phase relation with the induced voltage waves as shown in Fig. 30.51. The drive can easily operate in four-quadrant mode.

Defining Terms

Dynamic brake: The braking operation of a machine by extracting electrical energy and then dissipating it in a resistor.

Forced-commutation: Switching off a power semiconductor device by external circuit transient.

Four-quadrant: A drive that can operate as a motor as well as a generator in both directions.

Hysteresis-band: A method of controlling current where the instantaneous current can vary within a band.

Insulated gate bipolar transistor (IGBT): A device that combines the features of a power transistor and MOSFET.

Regenerative braking: The braking operation of a machine by converting its mechanical energy into electrical form and then pumping it back to the source.

Self-commutation: Switching off a power semiconductor device by its gate or base drive.

Two-quadrant: A drive that can operate as a motor as well as a generator in one direction.

Related Topics

66.1 Generators • 66.2 Motors

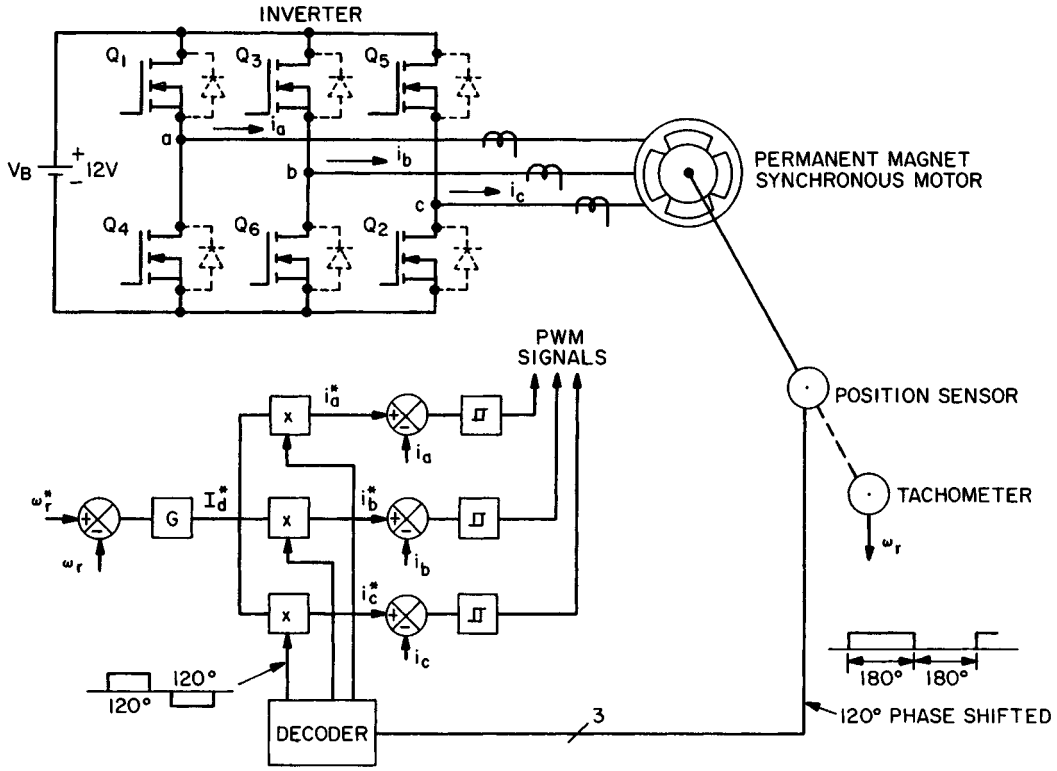


FIGURE 30.50 Permanent magnet synchronous motor control with PWM inverter.

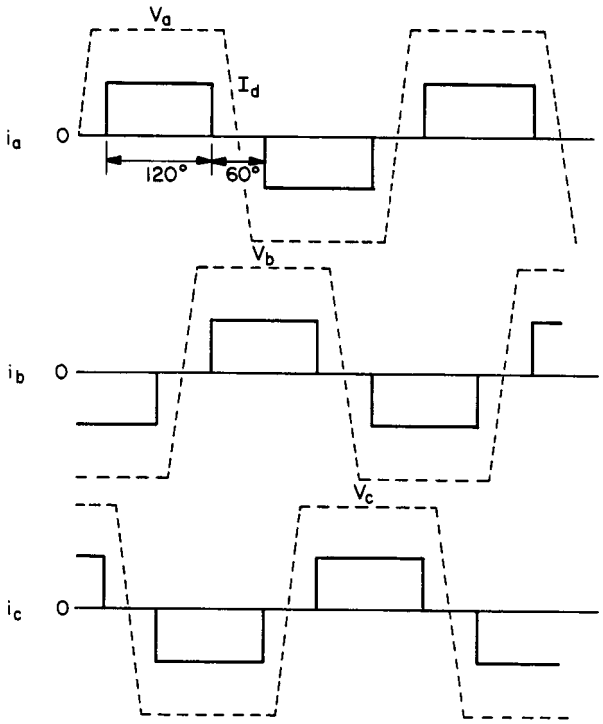


FIGURE 30.51 Phase voltage and current waves in brushless dc drive.

References

B.K. Bose, *Power Electronics and AC Drives*, Englewood Cliffs, N.J.: Prentice-Hall, 1986.
B.K. Bose, "Adjustable speed AC drives—A technology status review," *Proc. IEEE*, vol. 70, pp. 116–135, Feb. 1982.
B.K. Bose, *Modern Power Electronics*, New York: IEEE Press, 1992.
J.M.D. Murphy and F.G. Turnbull, *Power Electronic Control of AC Motors*, New York: Pergamon Press, 1988.
P.C. Sen, *Thyristor DC Drives*, New York: John Wiley, 1981.

Hecht, J., Watkins, L.S., Becker, R.A. “Optoelectronics”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

31

Optoelectronics

Jeff Hecht
Laser Focus World

Laurence S. Watkins
Lucent Technologies

R.A. Becker
Integrated Optical Circuit Consultants

31.1 Lasers
Differences from Other Light Sources • The Laser Industry

31.2 Sources and Detectors
Properties of Light • Absorption • Coherence • Geometric Optics • Incoherent Light • Detectors, Semiconductor • Detectors, Photoemissive • Imaging Detectors • Noise and Detectivity

31.3 Circuits
Integrated Optics • Device Fabrication • Packaging • Applications

31.1 Lasers[1]

Jeff Hecht

The word *laser* is an acronym for "light amplification by the stimulated emission of radiation," a phrase that covers most, though not all, of the key physical processes inside a laser. Unfortunately, that concise definition may not be very enlightening to the nonspecialist who wants to *use* a laser and cares less about its internal physics than its external characteristics. From a practical standpoint, a laser can be considered as a source of a narrow beam of **monochromatic**, coherent light in the visible, infrared, or ultraviolet parts of the spectrum. The power in a continuous beam can range from a fraction of a milliwatt to around 25 kilowatts (kW) in commercial lasers, and up to more than a megawatt in special military lasers. Pulsed lasers can deliver much higher peak powers during a pulse, although the power averaged over intervals while the laser is off and on is comparable to that of continuous lasers.

The range of laser devices is broad. The **laser medium**, or material emitting the laser beam, can be a gas, liquid, glass, crystalline solid, or semiconductor crystal and can range in size from a grain of salt to filling the inside of a moderate-sized building. Not every laser produces a narrow beam of monochromatic, coherent light. Semiconductor diode lasers, for example, produce beams that spread out over an angle of 20 to 40 degrees, hardly a pencil-thin beam. Liquid dye lasers emit at a broad or narrow range of wavelengths, depending on the optics used with them. Other types emit at a number of spectral lines, producing light that is neither truly monochromatic nor coherent. Table 31.1 summarizes important commercial lasers.

Practically speaking, lasers contain three key elements. One is the laser medium itself, which generates the laser light. A second in the power supply, which delivers energy to the laser medium in the form needed to excite it to emit light. The third is the optical cavity or **resonator**, which concentrates the light to stimulate the emission of laser radiation. All three elements can take various forms, and although they are not always immediately evident in all types of lasers, their functions are essential. Figure 31.1 shows these elements in a ruby and a helium-neon laser.

Laser-like devices called optical parametric oscillators have come into increasing use. They are more costly and complex than lasers, but can be tuned across a broad range, with wavelengths from 0.2 to 4 micrometers.

[1]Modified from J. Hecht, *The Laser Guidebook,* 2nd ed., New York: McGraw-Hill, 1991. With permission.

TABLE 31.1 Important Commercial Lasers

Wavelength (μm)	Type	Output Type and Power
0.157	Molecular fluorine (F_2)	Pulsed, avg. to a few watts
0.192	ArF excimer	Pulsed, avg. to tens of watts
0.2–0.35	Doubled dye	Pulsed
0.235–0.3	Tripled Ti-sapphire	Pulsed
0.24–0.27	Tripled alexandrite	Pulsed
0.248	KrF excimer	Pulsed, avg. to over 100 W
0.266	Quadrupled Nd	Pulsed, watts
0.275–0.306	Argon-ion	CW, 1-W range
0.308	XeCl excimer	Pulsed, to tens of watts
0.32–1.0	Pulsed dye	Pulsed, to tens of watts
0.325	He-Cd	CW, to tens of milliwatts
0.337	Nitrogen	Pulsed, under 1 W avg.
0.35–0.47	Doubled Ti-sapphire	Pulsed
0.351	XeF excimer	Pulsed, to tens of watts
0.355	Tripled Nd	Pulsed, to tens of watts
0.36–0.4	Doubled alexandrite	Pulsed, watts
0.37–1.0	CW dye	CW, to a few watts
0.442	He-Cd	CW, to over 0.1 W
0.45–0.53	Ar-ion	CW, to tens of watts
0.51	Copper vapor	Pulsed, tens of watts
0.520–0.569	Kryption ion	CW, >1W
0.523	Doubled Nd-YLF	Pulsed, watts
0.532	Doubled Nd-YAG	Pulsed to 50 W, or CW to watts
0.5435	He-Ne	CW, 1-mW range
0.578	Copper vapor	Pulsed, tens of watts
0.594	He-Ne	CW, to several milliwatts
0.612	He-Ne	CW, to several milliwatts
0.628	Gold vapor	Pulsed
0.6328	He-Ne	CW, to about 50 mW
0.635–0.66	InGaAlP diode	CW, milliwatts
0.647–0.676	Krypton ion	CW, to several watts
0.67	GaInP diode	CW, to 10 mW
0.68–1.13	Ti-sapphire	CW, watts
0.694	Ruby	Pulsed, to a few watts
0.72–0.8	Alexandrite	Pulsed, to tens of watts (CW in lab)
0.75–0.9	GaAlAs diode	CW, to many watts in arrays
0.98	InGaAs diode	CW, to 50 mW
1.047 or 1.053	Nd-YLF	CW or pulsed, to tens of watts
1.061	Nd-glass	Pulsed, to 100 W
1.064	Nd-YAG	CW or pulsed, to kilowatts
1.15	He-Ne	CW, milliwatts
1.2–1.4	InGaAsP diode	CW, to 100 mW
1.313	Nd-YLF	CW or pulsed, to 0.1 W
1.32	Nd-YAG	Pulsed or CW, to a few watts
1.4–1.6	Color center	CW, under 1 W
1.5–1.6	InGaAsP diode	CW, to 100 mW
1.523	He-Ne	CW, milliwatts
1.54	Erbium-glass (bulk)	Pulsed, to 1 W
1.54	Erbium-fiber (amplifier)	CW, milliwatts
1.75–2.5	Cobalt-MgF_2	Pulsed, 1-W range
2.3–3.3	Color center	CW, under 1 W
2.6–3.0	HF chemical	CW or pulsed, to hundreds of watts
3.3–29	Lead-salt diode	CW, milliwatt range
3.39	He-Ne	CW, to tens of milliwatts
3.6–4.0	DF chemical	CW or pulsed, to hundreds of watts
5–6	Carbon monoxide	CW, to tens of watts
9–11	Carbon dioxide	CW or pulsed, to tens of kilowatts
40–100	Far-infrared gas	CW, generally under 1 W

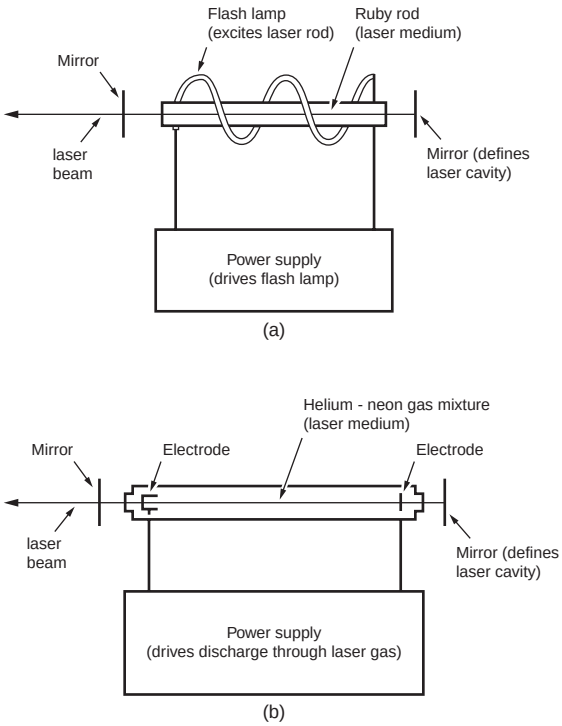


Figure 31.1 Simplified views of two common lasers, (a) ruby and (b) helium-neon, showing the basic components that make a laser.

Several general characteristics are common to most lasers that new users may not expect. Like most other light sources, lasers are inefficient in converting input energy into light. Efficiencies range from less than 0.001 to more than 50%, but except for semiconductor lasers, few types are much above 1% efficient. These low efficiencies can lead to special cooling requirements and duty-cycle limitations, particularly for high-power lasers. In some cases, special equipment may be needed to produce the right conditions for laser operation, such as cryogenic temperatures for lead salt semiconductor lasers. Operating characteristics of individual lasers depend strongly on structural components such as cavity optics, and in many cases a wide range is possible. Packaging can also have a strong impact on laser characteristics and the use of lasers for certain applications. Thus, there are wide ranges of possible characteristics, although single devices will have much more limited ranges of operation.

Differences from Other Light Sources

The basic differences between lasers and other light sources are the characteristics often used to describe a laser: the output beam is narrow, the light is monochromatic, and the emission is coherent. Each of these features is important for certain applications and deserves more explanation.

Most gas or **solid-state lasers** emit beams with divergence angle of about a milliradian, meaning that they spread to about 1 m in diameter after traveling a kilometer. (Semiconductor lasers have much larger beam divergence, but suitable optics can reshape the beam to make it much narrower.) The actual beam divergence depends on the type of laser and the optics used with it. The fact that laser light is contained in a beam serves to concentrate the output power onto a small area. Thus, a modest laser power can produce a high intensity inside the small area of the laser beam; the intensity of light in a 1-mW helium-neon laser beam is comparable to that of sunlight on a clear day, for example. The beams from high-power lasers, delivering tens of watts or more of continuous power or higher peak powers in pulses, can be concentrated to high enough intensities that they can weld, drill, or cut many materials.

The laser beam's concentrated light delivers energy only where it is focused. For example, a tightly focused laser beam can write a spot on a light-sensitive material without exposing the adjacent area, allowing high-resolution printing. Similarly, the beam from a surgical laser can be focused onto a tiny spot for microsurgery, without heating or damaging surrounding tissue. Lenses can focus the parallel rays in a laser beam to a much smaller spot than they can the diverging rays from a point source, a factor that helps compensate for the limited light-production efficiency of lasers.

Most lasers deliver a beam that contains only a narrow range of wavelengths, and thus the beam can be considered monochromatic for all practical purposes. Conventional light sources, in contrast, emit light over much of the visible and infrared spectrum. For most applications, the range of wavelengths emitted by lasers is narrow enough to make life easier for designers by avoiding the need for achromatic optics and simplifying the task of understanding the interactions between laser beam and target. For some applications in spectroscopy and communications, however, that range of wavelengths is not narrow enough, and special line-narrowing options may be required.

One of the beam's unique properties is its **coherence**, the property that the light waves it contains are in phase with one another. Strictly speaking, all light sources have a finite coherence length, or distance over which the light they produce is in phase. However, for conventional light sources that distance is essentially zero. For many common lasers, it is a fraction of a meter or more, allowing their use for applications requiring coherent light. The most important of these applications is probably holography, although coherence is useful in some types of spectroscopy, and there is growing interest in communications using coherent light.

Some types of lasers have two other advantages over other light sources: higher power and longer lifetime. For some high-power semiconductor lasers, lifetime must be traded off against higher power, but for most others the life vs. power trade-off is minimal. The combination of high power and strong directionality makes certain lasers the logical choice to deliver high light intensities to small areas. For some applications, lasers offer longer lifetimes than do other light sources of comparable brightness and cost. In addition, despite their low efficiency, some lasers may be more efficient in converting energy to light than other light sources.

The Laser Industry

Commercial Lasers

There is a big difference between the world of laser research and the world of the commercial laser industry. Unfortunately, many text and reference books fail to differentiate between types of lasers that can be built in the laboratory and those that are readily available commercially. That distinction is a crucial one for laser users.

Laser emission has been obtained from hundreds of materials at many thousands of emission lines in laboratories around the world. Extensive tabulations of these laser lines are available [Weber, 1982], and even today researchers are adding more lines to the list. However, most of these laser lines are of purely academic interest. Many are weak lines close to much stronger lines that dominate the emission in practical lasers. Most of the lasers that have been demonstrated in the laboratory have proved to be cumbersome to operate, low in power, inefficient, and/or simply less practical to use than other types.

Only a couple of dozen types of lasers have proved to be commercially viable on any significant scale; these are summarized in Table 31.1. Some of these types, notably the ruby and helium-neon lasers, have been around since the beginning of the laser era. Others, such as vibronic solid-state, are promising newcomers. The family of commercial lasers is expanding slowly, as new types such as titanium-sapphire come on the market, but with the economics of production a factor to be considered, the number of commercially viable lasers will always be limited.

There are many possible reasons why certain lasers do not find their way onto the market. Some require exotic operating conditions or laser media, such as high temperatures or highly reactive metal vapors. Some emit only feeble powers. Others have only limited applications, particularly lasers emitting low powers in the far-infrared or in parts of the infrared where the atmosphere is opaque. Some simply cannot compete with materials already on the market.

Defining Terms

Coherence: The condition of light waves that stay in the same phase relative to each other; they must have the same wavelength.
Continuous wave (CW): A laser that emits a steady beam rather than pulses.
Laser medium: The material in a laser that emits light; it may be a gas, solid, or liquid.
Monochromatic: Of a single wavelength or frequency.
Resonator: Mirrors that reflect light back and forth through a laser medium, usually on opposite ends of a rod, tube, or semiconductor wafer. One mirror lets some light escape to form the laser beam.
Solid-state laser: A laser in which light is emitted by atoms in a glass or crystalline matrix. Laser specialists do not consider semiconductor lasers to be solid-state types.

Related Topic

42.1 Lightwave Waveguides

References

J. Hecht, *The Laser Guidebook,* 2nd ed., New York: McGraw-Hill, 1991; this section is excerpted from the introduction.
M. J. Weber (ed.), *CRC Handbook of Laser Science and Technology* (2 vols.), Boca Raton, Fla.: CRC Press, 1982.
M. J. Weber (ed.), *CRC Handbook of Laser Science and Technology, Supplement 1,* Boca Raton, Fla.: CRC Press, 1989; other supplements are in preparation.

Further Information

Several excellent introductory college texts are available that concentrate on laser principles. These include: Anthony E. Siegman, *Lasers,* University Science Books, Mill Valley, Calif., 1986, and Orzio Svelto, *Principles of Lasers,* 3rd ed., Plenum, New York, 1989.

Three trade magazines serve the laser field; each publishes an annual directory issue. For further information contact: *Laser Focus World,* PennWell Publishing, Ten Tara Blvd., Nashua, NH 03062; *Lasers & Optronics,* PO Box 650, Morris Plains, N.J. 07950-0650; or *Photonics Spectra,* Laurin Publishing Co., Berkshire Common, PO Box 1146, Pittsfield, Mass. 01202. Write the publishers for information.

31.2 Sources and Detectors

Laurence S. Watkins

Properties of Light

The strict definition of light is electromagnetic radiation to which the eye is sensitive. Optical devices, however, can operate over a larger range of the electromagnetic spectrum, and so the term usually refers to devices which can operate in some part of the spectrum from the near ultraviolet (UV) through the visible range to the near infrared. Figure 31.2 shows the whole spectrum and delineates these ranges.

Optical radiation is electromagnetic radiation and so obeys and can be completely described by Maxwell's equations. We will not discuss this analysis here but just review the important properties of light.

Phase Velocity

In isotropic media light propagates as transverse electromagnetic (TEM) waves. The electric and magnetic field vectors are perpendicular to the propagation direction and orthogonal to each other. The velocity of light propagation in a medium (the velocity of planes of constant phase, i.e., wavefronts) is given by

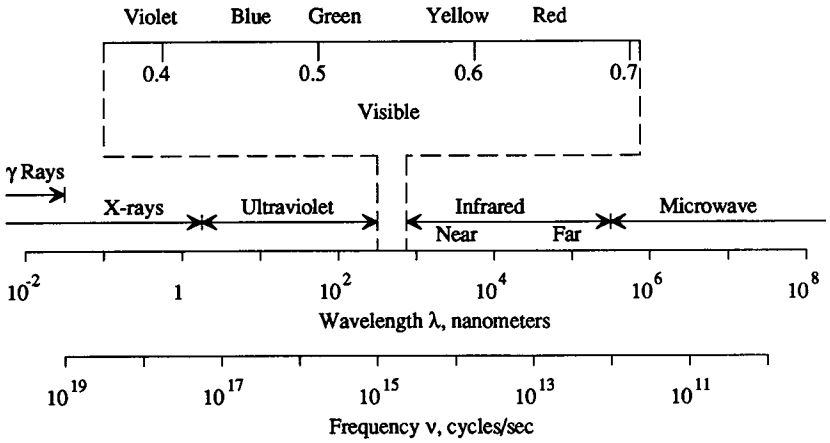


FIGURE 31.2 Electromagnetic spectrum showing visible and optical wavelengths.

$$v = \frac{c}{\sqrt{\varepsilon\mu}} \tag{31.1}$$

where c is the velocity of light in a vacuum (c = 299,796 km/s). The denominator in Eq. (31.1) is a term in optics called the refractive index of the medium

$$n = \sqrt{\varepsilon\mu} \tag{31.2}$$

where ε is the dielectric constant (permittivity) and μ is the magnetic permeability. The wavelength of light, λ, which is the distance between phase fronts is

$$\lambda = \frac{\lambda_0}{n} = \frac{v}{\upsilon} \tag{31.3}$$

where λ_0 is the wavelength in vacuum and υ is the light frequency. The refractive index varies with wavelength, and this is referred to as the dispersive property of a medium.

Another parameter used to describe light frequency is wave number. This is given by

$$\sigma = \frac{1}{\lambda} \tag{31.4}$$

and is usually expressed in cm^{-1}, giving the number of waves in a 1-cm path.

Group Velocity

When traveling in a medium, the velocity of energy transmission (e.g., a light pulse) is less than c and is given by

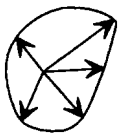
$$u = v - \lambda \frac{dv}{d\lambda} \tag{31.5}$$

In vacuum the phase and group velocities are the same.

Polarization

Light polarization is defined by the direction of the electric field vector. For isotropic media this direction is perpendicular to the propagation direction. It can exist in a number of states, described as follows.

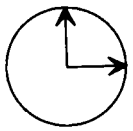
Unpolarized. The electric field vector has a random and constantly changing direction, and when there are multiple frequencies the vector directions are different for each frequency.



Linear. The electric field vector is confined to one direction.

Elliptical. The electric field vector rotates, either left hand or right hand, at the light frequency. The magnitude of the vector (intensity of the light) traces out an ellipse.

Circular. Circular is the special case of the above where the electric field vector traces out a circle.



Absorption

Light in traveling through media can be absorbed. This can be represented in two ways. The light flux propagating through a medium can be written as

$$I = I_0 e^{-\alpha x} \tag{31.6}$$

where x is the distance through the medium with incident light flux I_0. α is the absorption coefficient, usually stated in cm^{-1}. An alternative way of describing absorption is to use the imaginary term in the media refractive index. The complex refractive index is

$$\bar{n} = n(1 + ik) \tag{31.7}$$

where k is the attenuation index. α and k are related as

$$\alpha = \frac{4\pi}{\lambda_0} nk \tag{31.8}$$

Coherence

Light can be partially or fully coherent or incoherent, depending on the source and subsequent filtering operations. Common sources of light are incoherent because they consist of many independent radiators. An example of this is the fluorescent lamp in which each excited atom radiates light independently. There is no fixed phase relationship between the waves from these atoms. In a laser the light is generated in a resonant

cavity using a light amplifier and the resulting coherent light has well-defined phase fronts and frequency characteristics.

Spatial and Temporal Coherence. Spatial coherence describes the phase front properties of light. A beam from a single-mode laser which has one well-defined phase front is fully spatially coherent. A collection of light waves from a number of light emitters is incoherent because the resulting phase front has a randomly indefinable form. Temporal coherence describes the frequency properties of light. A single-frequency laser output is fully temporally coherent. White light, which contains many frequency components, is incoherent, and a narrow band of frequencies is partially coherent.

Laser Beam Focusing

The radial intensity profile of a collimated single-mode TEM_{00} (Gaussian) beam from a laser is given by

$$I(r) = I_0 \exp\left[2\left(\frac{-r^2}{w_0^2}\right)\right] \tag{31.9}$$

where w_0 is the beam radius ($1/e^2$ intensity). This beam will diverge as it propagates out from the laser, and the half angle of the divergence is given by

$$\theta_{1/2} = \frac{\lambda}{\pi w_0} \tag{31.10}$$

When this beam is focused by a lens the resulting light spot radius is given by

$$w_f = \frac{\lambda l}{\pi w_d} \tag{31.11}$$

where l is the distance from the lens to the position of the focused spot and w_d is the beam radius entering the lens. It should be noted that $l \cong f$, the lens focal length, for a collimated beam entering the lens. However, l will be a greater distance than f if the beam is diverging when entering the lens.

Geometric Optics

The wavelength of light can be approximated to zero for many situations. This permits light to be described in terms of light rays which travel in the direction of the wave normal. This branch of optics is referred to geometric optics.

Properties of Light Rays

Refraction. When light travels from one medium into another it changes propagation velocity, Eq. (31.1). This results in refraction (bending) of the light as shown in Fig. 31.3.

The change in propagation direction of the light ray is given by Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{31.12}$$

where n_1 and n_2 are the refractive indices of media 1 and 2, respectively.

Critical Angle. When a light ray traveling in a medium is incident on a surface of a less dense medium, there is an incidence angle θ_2, where $\sin\theta_1 = 1$. This is the critical angle; for light incident at angles greater than θ_2 the light is totally internally reflected as shown in Fig. 31.3(b). The critical angle is given by $\theta_c = \sin^{-1}(n_1/n_2)$.

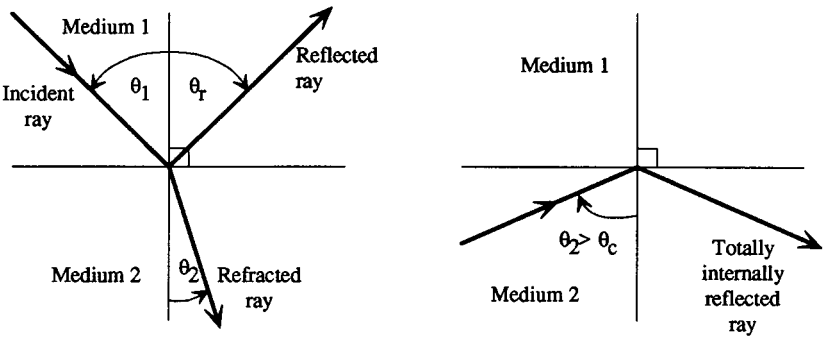


FIGURE 31.3 (a) Diagram of a light ray in medium 1 incident at angle θ_1 on the surface to medium 2. The ray is refracted at angle θ_2. (b) Diagram of the situation when the ray in medium 2 is incident at an angle greater than the critical angle θ_c and totally internally reflected.

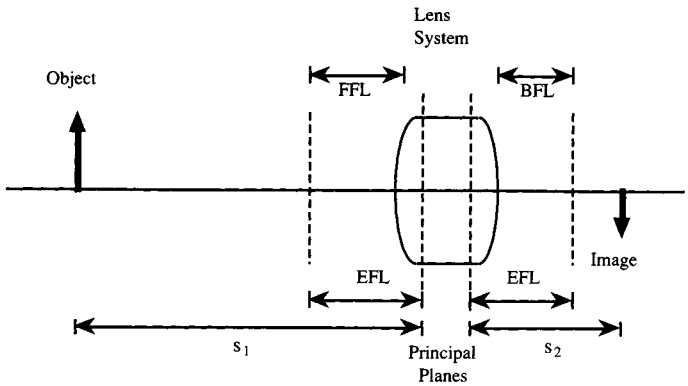


FIGURE 31.4 Schematic of an optical system forming an image of an object. Light rays from the object are captured by the lens which focuses them to form the image. EFL, effective focal length, f, of the lens; FFL and BFL, distances from the focal points to the outer lens surface. Principal planes are the positions to which the focal points, object distance, and image distance are measured; in a simple lens they are coincident.

Image Formation with a Lens

Many applications require a lens to focus light or to form an image onto a detector. A well-corrected lens usually consists of a number of lens elements in a mount, and this can be treated as a black box system. The characteristics of this lens are known as the cardinal points. Figure 31.4 shows how a lens is used to form an image from an illuminated object.

The equation which relates the object, image, and lens system is

$$\frac{1}{f} = \frac{1}{s_1} + \frac{1}{s_2} \tag{31.13}$$

The image magnification is given by $M = s_2/s_1$. When the object is very far away s_1 is infinite and the image is formed at the back focal plane.

Incoherent Light

When two or more incoherent light beams are combined, the resulting light flux is the sum of their energies. For coherent light this is not necessarily true and the resulting light intensity depends on the phase relationships between the electric fields of the two beams, as well as the degree of coherence.

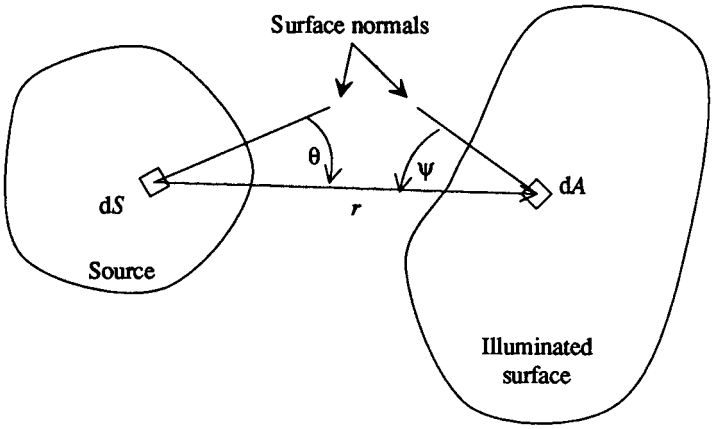


FIGURE 31.5 Surface being illuminated by an extended source. Illumination of surface element *dA* is calculated by summing the effects of elements *dS*.

Brightness and Illumination

The flux density of a light beam emitted from a **point source** decreases with the square of distance from it. Light sources are typically extended sources (being larger than point sources). The illumination of a surface from light emitted from an **extended source** can be calculated using Fig. 31.5.

The flux incident on a surface element *dA* from a source element *dS* is given by

$$dE = \frac{B\,dA\cos\theta\,dS\cos\psi}{r^2} \tag{31.14}$$

The constant *B* is called the luminance or photometric brightness of the source. Its units are candles per square meter (1 stilb = π lamberts) and *dE* is the luminous flux in lumens. The total illumination *E* of the surface element is calculated by integrating over the source. The illuminance or flux density on the surface is thus

$$I = \frac{E}{dA}\ (\text{lumens/cm}^2) \tag{31.15}$$

Two methods are commonly used for quantifying light energy, namely, the radiometric unit of watts and the photometric unit of candelas. The candela is an energy unit which is derived from light emission from a blackbody source. The two can be related using the relative visibility curve $V(\lambda)$, which describes the eye's sensitivity to the visible light spectrum, it being maximum near a wavelength of 550 nm. The constant which relates lumens to watts at this wavelength is 685 lm/W. The luminous flux emitted by a source can therefore be written as

$$F = 685\int V(\lambda)\,P(\lambda)\,d\lambda\ \ (\text{lumens}) \tag{31.16}$$

where *V* is the spectral response of the eye and *P* is the source radiant intensity in watts.

The source radiance is normally stated as luminance in candle per square centimeter (1 lumen per steradian per square centimeter) or radiance in watts per square centimeter per steradian per nanometer. The lumen is defined as the luminous flux emitted into a solid angle of a steradian by a point source of intensity 1/60th that of a 1-cm² blackbody source held at 2042 K temperature (molten platinum).

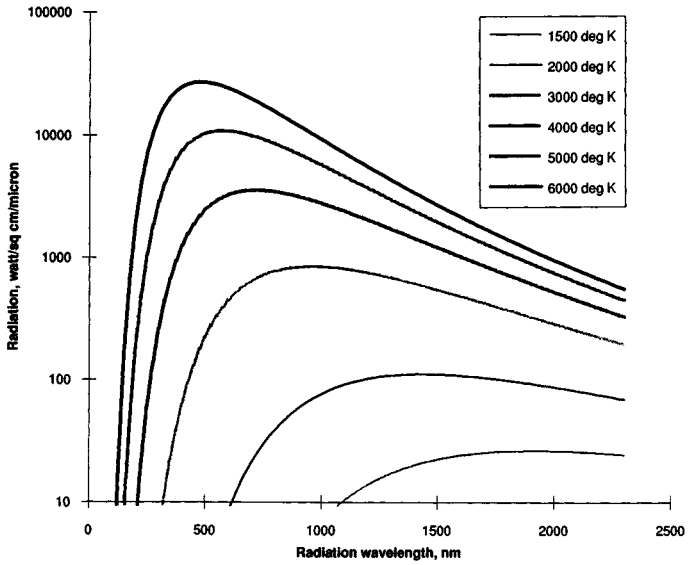


FIGURE 31.6 Plot of blackbody radiation for a series of temperatures. Radiation is in watts into a hemisphere direction from a 1-cm^2 of surface in a 1-μm wavelength band.

Thermal Sources

Objects emit and absorb radiation, and as their temperature is increased the amount of radiation emitted increases. In addition, the spectral distribution changes, with proportionally more radiation emitted at shorter wavelengths. A blackbody is defined as a surface which absorbs all radiation incident upon it, and Kirchhoff's law of radiation is given by

$$\frac{W}{a} = \text{constant} = WB \tag{31.17}$$

stating that the ratio of emitted to absorbed radiation is a constant a at a given temperature.

The energy or wavelength distribution for a blackbody is given by Planck's law

$$W = \frac{c_1}{\lambda^5}\left[\exp\left(\frac{c_2}{\lambda T}\right) - 1\right]^{-1} \quad (\text{watts/cm}^2 \text{ area per } \mu\text{m wavelength})$$
$$c_1 = 3.7413 \times 10^4 \tag{31.18}$$
$$c_2 = 1.4380 \times 10^4$$

T is in degrees Kelvin, λ is in micrometers, and W is the power emitted into a hemisphere direction. Blackbody radiation is incoherent, with atoms or molecules emitting radiation independently. Figure 31.6 is a plot of the blackbody radiation spectrum for a series of temperatures.

Very few materials are true blackbodies; carbon lampblack is one. For this reason a surface emissivity is used which describes the ratio of actual radiation emitted to that from a perfect blackbody. Table 31.2 is a listing of emissivities for some common materials.

Tungsten Filament Lamp

In the standard incandescent lamp a tungsten filament is heated to greater than 2000°C, and it is protected from oxidation and vaporization by an inert gas. In a quartz halogen lamp the envelope is quartz, which allows

TABLE 31.2 Emissivities of Some Common Materials

Material	Temperature (°C)	Emissivity
Tungsten	2000	0.28
Nickel-chromium (80-20)	600	0.87
Lampblack	20–400	0.96
Polished silver	200	0.02
Glass	1000	0.72
Platinum	600	0.1
Graphite	3600	0.8
Aluminum (oxidized)	600	0.16
Carbon filament	1400	0.53

the filament to run at a higher temperature. This increases the light output and gives a whiter wavelength spectrum with proportionally more visible radiation to infrared.

Standard Light Source—Equivalent Black Blackbody

Because the emissivity of incandescent materials is less than 1, an equivalent source is needed for measurement and calibration purposes. This is formed by using an enclosed space which has a small opening in it. Provided the opening is much smaller than the enclosed area, the radiation from the opening will be nearly equal to that from a blackbody at the same temperature, as long as the interior surface emissivity is > 0.5. Blackbody radiation from such a source at the melting point of platinum is defined as 1/60 cd/cm^2.

Arc Lamp

A gas can be heated to temperatures of 6000 K or more by generating an electric arc between two electrodes. The actual resulting temperature is dependent on the current flowing through the arc, the gas pressure and its composition, and other factors. This does provide a light source which is close to the temperature of the sun. Using an inert xenon gas results in essentially a white light spectrum. The use of a gas such as mercury gives more light in the UV as well as a number of strong peak light intensities at certain wavelengths. This is due to excitation and fluorescence of the mercury atoms.

Fluorescent Lamp

A fluorescent source is a container (transparent envelope) in which a gas is excited by either a dc discharge or an RF excitation. The excitation causes the electrons of the gas to move to higher energy orbits, raising the atoms to a higher excited state. When the atoms relax to lower states they give off energy, and some of this energy can be light. The wavelength of the light is characteristically related to the energy levels of the excited states of the gas involved. Typically a number of different wavelengths are associated with a particular gas.

Low-pressure lamps have relatively low luminance but provide light with narrow linewidths and stable spectral wavelengths. If only one wavelength is required, then optical filters can be used to isolate it by blocking the unwanted wavelengths.

Higher luminance is achieved by using higher gas pressures. The fluorescent lamp is very efficient since a high proportion of the input electrical energy is converted to light. White light is achieved by coating the inside of the container with various types of phosphor. The gas, for example a mercury–argon mixture, provides UV and violet radiation which excites the phosphor. Since the light is produced by fluorescence and phosphorescence, the spectral content of the light does not follow Planck's radiation law but is characteristic of the coating (e.g., soft white, cool white).

Light-Emitting Diodes (LED)

Light can be emitted from a semiconductor material when an electron and hole pair recombine. This is most efficient in a direct gap semiconductor like GaAs and the emitted photons have energy close to the bandgap energy E_g. The wavelength is then given by

$$\lambda \cong \frac{hc}{E_g} \tag{31.19}$$

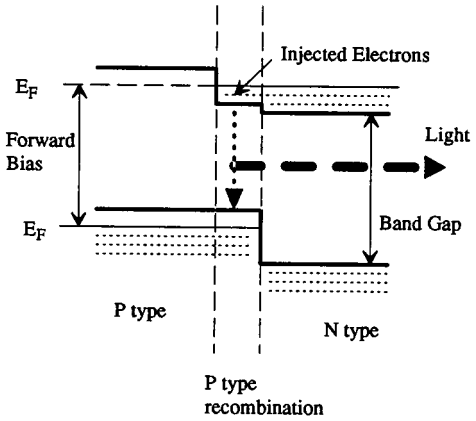


FIGURE 31.7 Band structure of a double heterostructure LED. Forward bias injects holes and electrons into the junction region where they recombine and emit light.

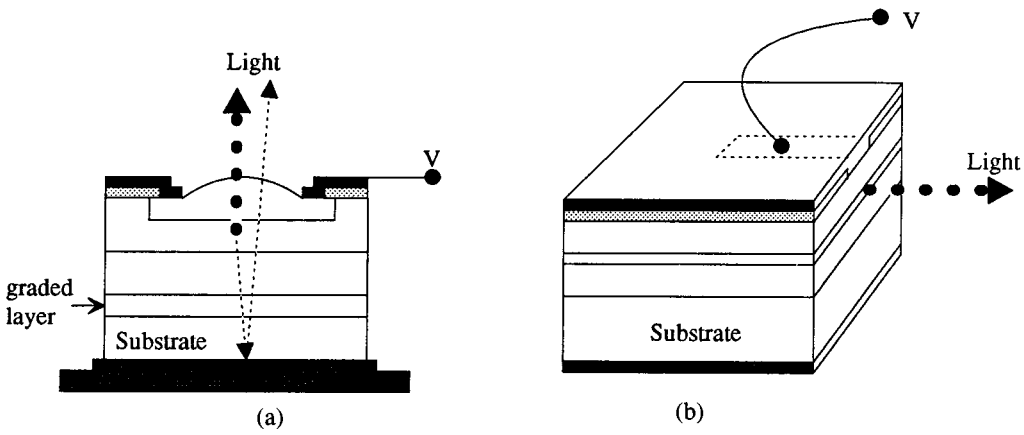


FIGURE 31.8 Cross-sectional diagrams of (a) surface emitting LED and (b) edge emitting LED. The light output from the edge emitter is more directional because of confinement by the junction guide region.

where h is Planck's constant (6.626×10^{-34} J-s) and c the velocity of light in vacuum. The spectral width of the emission is quite broad, a few hundred nanometers, and is a function of the density of states, transition probabilities, and temperature.

For **light emission** to occur, the conduction band must be populated with many electrons. This is achieved by forward biasing a *pn* junction to inject electrons and holes into the junction region as shown in Fig. 31.7.

Figure 31.8(a) shows the cross section of a surface emitting LED with an integral lens fabricated into the surface. The light from the LED is incoherent and emitted in all directions. The lens and the bottom reflecting surface increase the amount of light transmitted out of the front of the device. The output from the LED is approximately linear with current but does decrease with increasing junction temperature.

Figure 31.8(b) shows an edge emitting LED. Here the light is generated in a waveguide region which confines the light, giving a more directional output beam.

Various wavelengths are available and are obtained by using different bandgap semiconductors. This is done by choosing different binary, ternary, and quaternary compositions. Table 31.3 is a listing of the more common ones.

The output power is usually specified in milliwatts per milliamp current obtained in a given measurement situation, e.g., into a fiber or with a 0.5 numerical aperture large area detector. Other parameters are peak wavelength, wavelength band (usually full width half max), and temperature characteristics.

TABLE 31.3 Common Light-Emitting Diode Compounds and Wavelengths

Compound	Wavelength (nm)	Color
GaP	565	Green
GaAsP	590	Yellow
GaAsP	632	Orange
GaAsP	649	Red
GaAlAs	850	Near IR
GaAs	940	Near IR
InGaAs	1060	Near IR
InGaAsP	1300	Near IR
InGaAsP	1550	Near IR

LEDs for Fiber Optic Communications

GaAs and InGaAsP LEDs are commonly used as sources for fiber optic communications. Since they are an incoherent source, it is only practical to use them with multimode fiber. Only multimode fiber has a large enough core diameter and numerical aperture (NA) to couple in enough light to be able to propagate any useful distance. Applications for LEDs in fiber optics are for short distance links using glass or plastic fiber at relatively low bandwidths, typically in the Mb/s rather than Gb/s. Primary applications of these are for low cost datalinks.

The detector can be packaged two ways: first with a fiber pigtail directly attached to the detector package; or a more common package is to have a fiber connector molded in as part of the package so that a connectorized fiber can be plugged in to it. Many LEDs for fiber optics are now packaged with electronic drive circuits to form a transmitter module ready to receive standard format data signals.

Detectors, Semiconductor

When light interacts electronically with a medium, by changing the energy of electrons or creating carriers, for example, it interacts in a quantized manner. The light energy can be quantized according to Planck's theory

$$E = h\upsilon \tag{31.20}$$

where υ is the light frequency and h is Planck's constant. The energy of each photon is very small; however, it does increase with shorter wavelengths.

Photoconductors

Semiconductors can act as photoconductors, where incident light increases the carrier density, thus increasing the conductivity. There are two basic types, intrinsic and extrinsic. Figure 31.9 shows a simple energy diagram containing conduction and valence bands. Also indicated are the levels which occur with the introduction of donor and acceptor impurities.

Intrinsic photoconduction effect is when a photon with energy $h\upsilon$, which is greater than the bandgap energy, excites an electron from the valence band into the conduction band, creating a hole–electron pair. This increases the conductivity of the material. The spectral response of this type of detector is governed by the bandgap of the semiconductor.

In an extrinsic photoconductor (see Fig. 31.9), the photon excites an electron from the valence band into the acceptor level corresponding to the hole of the acceptor atom. The resulting energy $h\upsilon$ is much smaller than the bandgap and is the reason why these detectors have applications for long wavelength infrared sensors. Table 31.4 is a list of commercial photoconductors and their peak wavelength sensitivities.

The doping material in the semiconductor determines the acceptor energy level, and so both the host material and the dopant are named. Since the energy level is quite small it can be populated by a considerable amount by thermal excitation. Thus, for useful detection sensitivity the devices are normally operated at liquid nitrogen and sometimes liquid helium temperatures. The current response, i, of a photoconductor can be written as

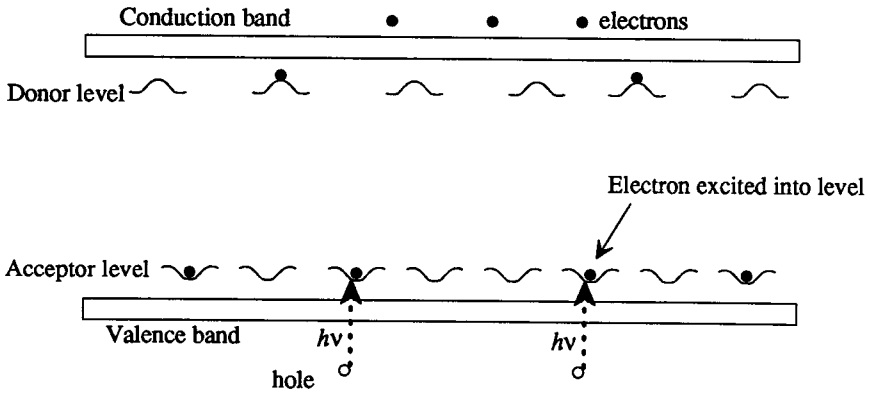


FIGURE 31.9 A simplified energy diagram for a photoconductive semiconductor, showing extrinsic effect of electrons into the acceptor level.

TABLE 31.4 Photoconductor Materials and Their Peak Wavelength Sensitivity

Photoconductor	Peak Wavelength (μm)
PbS	3
PbSe	5
HgCd	4
HgCaTe	10
HgCdTe	11
Si:Ga (4.2 K)	11
Si:As (4.2 K)	20
Si:Sb (4.2 K)	28

$$i = \frac{P\eta\tau_0 ev}{h\upsilon d} \tag{31.21}$$

where P is the optical power at frequency υ; h is Planck's constant; v is drift velocity = μE, where μ is mobility and E is electric field; η is quantum efficiency (at frequency υ); τ_0 is lifetime of carriers; and e is charge on electron.

Charge Amplification. For semiconductor photoconductors like CdS there can be traps. These are holes, which under the influence of a bias field will be captured for a period of time. This allows electrons to move to the anode instead of recombining with a hole, resulting in a longer period for the conduction increase. This provides a photoconductive gain which is equal to the mean time the hole is trapped divided by the electron transit time in the photoconductor. Gains of 10^4 are typical.

The charge amplification can be written as

$$\frac{\tau_0}{\tau_d} \tag{31.22}$$

where $\tau_d = d/v$, the drift time for a carrier to go across the semiconductor. The response time of this type of sensor is consequently slow, ~10 ms, and the output in quite nonlinear.

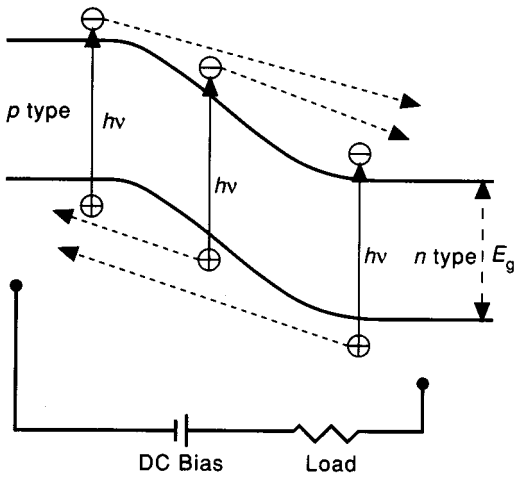


FIGURE 31.10 Energy diagram of a pn junction photodiode showing the three ways electron–hole pairs are created by absorbing photons and the contribution to current flow in the circuit.

Junction Photodiodes

In a simple junction photodiode a *pn* junction is fabricated in a semiconductor material. Figure 31.10 shows the energy diagram of such a device with a reverse voltage bias applied. Incident light with energy greater than the bandgap creates electrons in the *p* region and holes in the *n* region. Those which are within the diffusion length of the junction are swept across by the field. The light also creates electron–hole pairs in the junction region, and these are separated by the field. In both cases an electron charge is contributed to the external circuit. In the case of no bias the carrier movement creates a voltage with *p* region being positive. The maximum voltage is equal to the difference in the Fermi levels in the *p* and *n* regions and approaches the bandgap energy E_g.

PIN Photodiodes. The carriers which are generated in the junction region experience the highest field and so, being separated rapidly, give the fastest time response. The PIN diode has an extra intrinsic high field layer between the p and n regions, designed to absorb the light. This minimizes the generation of slow carriers and results in a fast response detector.

The signal current generated by incident light power *P* is

$$i = \frac{Pe\eta}{h\upsilon} + \text{dark current} \tag{31.23}$$

The output current is linear with incident power plus a constant dark current due to thermal generation of carriers; η is the quantum efficiency.

Avalanche Photodiodes

When the reverse bias of a photodiode is increased to near the breakdown voltage, carriers in the depletion region can be accelerated to the point where they will excite electrons from the valence band into the conduction band, creating more carriers. This current multiplication is called avalanche gain, and typical gains of 50 are available. Avalanche diodes are specially designed to have uniform junction regions to handle the high applied fields.

Detectors for Fiber Optic Communications

A major application for junction photodioldes is detectors for fiber optic communications. Silicon detectors are typically used for short wavelength light such as with GaAs sources. InP detectors are used for the 1.3 and 1.5 μm wavelength bands. The specific type and design of a detector is tailored to the fiber optics application, depending on whether it is low cost lower frequency datalinks or higher cost high frequency bit-rates in the

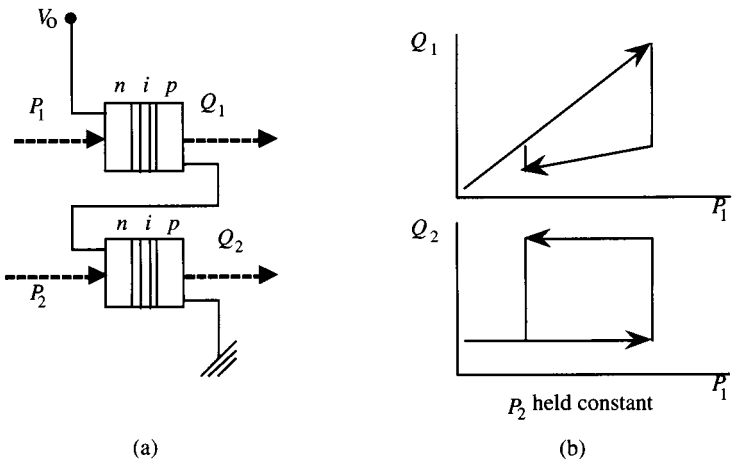


FIGURE 31.11 (a) S-SEED with voltage bias applied; (b) bistable outputs Q as a result of varying the input light power P_1 holding input power P_2 constant.

Gb/s. The detector is packaged either with a fiber pigtail or with a fiber connector receptacle molded as part of the package body.

Fiber optics detectors can also be packaged with pre-amplifier electronics or complete receiver and communications electronics into a module. For very high frequency response it is important to minimize the capacitance of the detector and the attached preamplifier circuit.

Solar Cells

Solar cells are large-area *pn* junction photodiodes, usually in silicon, which are optimized to convert light to electrical power. They are normally operated in the photovoltaic mode without a reverse voltage bias being applied.

Linear Position Sensors

Large-area photodiodes can be made into single axis and two axis position sensors. The single axis device is a long strip detector, and the two axis is normally square. In the single axis device the common terminal is in the middle and there are two signal terminals, one at each end. When a light beam is directed onto the detector, the relative output current from each signal terminal depends on how close the beam is to the terminal. The sum of the output currents from both terminals is proportional to the light intensity.

Phototransistors

For bipolar devices the light generates carriers which inject current into the base of the transistor. This modulates the collector base current, providing a higher output signal. For a field effect device the light generates carriers which create a gate voltage. PhotoFETs can have very high sensitivities.

SEEDs

A self-electro-optic effect device (SEED) is a multiple quantum well semiconductor optical pin device and forms the combination of a photodiode and a modulator. It can operate as a photodetector where incident light will generate a photocurrent in a circuit. It can also act as a modulator where the light transmitted through the device is varied by an applied voltage.

Devices are normally connected in pairs to form symmetric SEEDs as demonstrated in Fig. 31.11(a). These can then be operated as optical logic flip-flop devices. They can be set in one of two bistable states by application of incident light beams. The bistable state can be read out by similar light beams which measure the transmitted intensity. The hysteresis curve is shown in Fig. 31.11(b). These and similar devices are the emerging building blocks for optical logic and are sometimes referred to as smart pixels.

Detectors, Photoemissive

In the photoemissive effect, light falls onto a surface (photocathode) and the light energy causes electrons to be emitted. These electrons are then collected at a positively biased anode. There is a threshold energy required for the electron to be emitted from the surface. This energy is called the work function, f, and is a property of the surface material. The photon energy $h\upsilon$ must be greater than f, and this determines the longest wavelength sensitivity of the photocathode.

Vacuum Photodiodes

A vacuum photodiode comprises a negatively biased photocathode and a positive anode in a vacuum envelope. Light falling on the cathode causes electrons to be emitted, and these electrons are collected at the anode. Not all photons cause photoelectrons to be emitted, and quantum efficiencies, η, typically run 0.5–20%. These devices are not very sensitive; however, they have very good linearity of current to incident light power, P. They are also high-speed devices, with rise time being limited by the transit time fluctuations of electrons arriving at the anode. The photocurrent is given by

$$i = \frac{Pe\eta}{h\upsilon} + \text{dark current} \tag{31.24}$$

This kind of detector exhibits excellent short-term stability. The emissive surface can fatigue with exposure to light but will recover if the illumination is not excessive. Because of these properties, these devices have been used for accurate light measurement, although in many cases semiconductor devices are now supplanting them.

Gas-Filled Tubes

The light sensitivity of vacuum phototubes can be increased by adding 0.1 mm pressure of argon. The photoelectrons under the influence of the anode voltage accelerate and ionize the gas, creating more electrons. Gains of 5–10 can be realized. These devices are both low frequency, in the 10-kHz range, and nonlinear and are suitable only for simple light sensors. Semiconductor devices again are displacing these devices for most applications.

Photomultiplier Tubes

Photomultiplier tubes are the most sensitive light sensors, especially for visible radiation. Figure 31.12 is a schematic showing the electrical circuit used to bias it and form the output voltage signal. Light is incident on the photocathode, and the resulting photoelectrons are accelerated to a series of dynodes to generate secondary electrons and through this **electron multiplication** amplify the signal. Gains of 10^8 can be achieved with only minor degradation of the linearity and speed of vacuum photodiodes. The spectral response is governed by the emission properties of the photocathode.

There are various types of photomultipliers with different physical arrangements to optimize for a specific application. The high voltage supply ranges from 700 to 3000 V, and the electron multiplication gain is normally adjusted by varying the supply voltage. The linearity of a photomultiplier is very good, typically 3% over 3 decades of light level. Saturation is normally encountered at high anode currents caused by space charge effects at the last dynode where most of the current is generated. The decoupling capacitors, C_1, on the last few dynodes are used for high-frequency response and to prevent saturation from the dynode resistors.

Photon Counting

For the detection of very low light levels and for measuring the statistical properties of light, photon counting can be done using photomultipliers. A pulse of up to 10^8 electrons can be generated for each photoelectron emitted from the cathode, and so the arrival of individual photons can be detected. There is a considerable field of study into the statistical properties of light fields as measured by photon counting statistics.

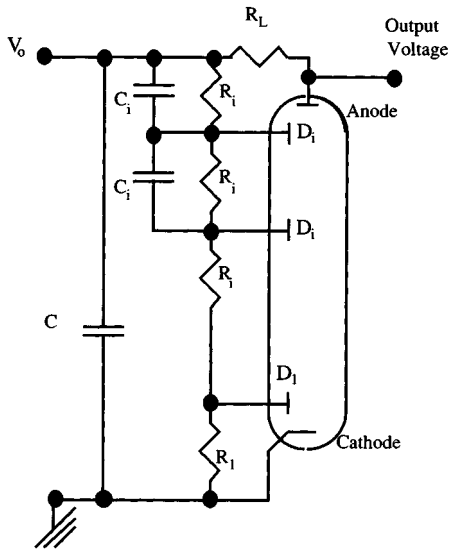


FIGURE 31.12 The basic layout of a photomultiplier tube showing the dynodes and the electrical circuit to bias them.

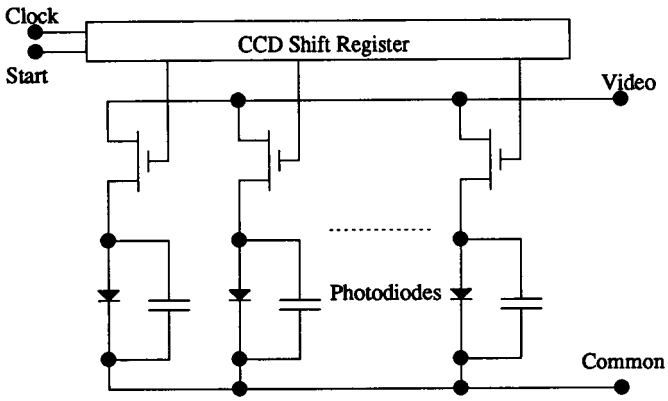


FIGURE 31.13 Schematic diagram of a linear CCD diode array sensor. CCD shift register sequentially clocks out charge from each photodiode to the video line.

Imaging Detectors

A natural extension to single photodetectors is to arrange them in arrays, both linear single dimension and two dimensions. Imaging detectors are made from both semiconductors and vacuum phototubes.

Semiconductor Detector Arrays

Detector arrays have been made using either photodiodes or photoconductors. The applications are for visible and infrared imaging devices. For small-sized arrays each detector is individually connected to an electrical lead on the package. This becomes impossible for large arrays, however, and these contain additional electronic switching circuits to provide sequential access to each diode.

Figure 31.13 show an example of a **charge-coupled device** (CCD) linear photodiode array. The device consists of a linear array of *pn* junction photodiodes. Each diode has capacitance associated with it, and when light falls on the detector the resulting photocurrent charges this capacitance. The charge is thus the time integral of the light intensity falling on the diode. The CCD periodically and sequentially switches the charge to the video line, resulting in a series of pulses. These pulses can be converted to a voltage signal which represents the light pattern incident on the array.

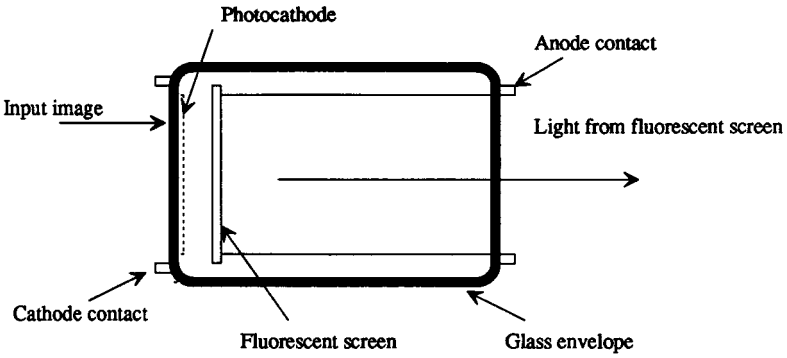


FIGURE 31.14 Diagram of a simple image-intensifier tube. More complex ones use improved electron optics.

The location of the diodes is accurately defined by the lithographic fabrication process and, being solid state, is also a rugged detector. These devices are thus very suitable for linear or two-dimensional optical image measurement. The devices can be quite sensitive and can have variable sensitivity by adjusting the CCD scan speed since the diode integrates the current until accessed by the CCD switch. The spectral sensitivity is that of the semiconductor photodiode, and the majority of devices now available are silicon. Smaller arrays are becoming more available in many types of semiconductors, however.

Image-Intensifier Tubes

An image-intensifier tube is a vacuum device which consists of a photoemissive surface onto which a light image is projected, an electron accelerator, and a phosphor material to view the image. This device, shown in Fig. 31.14, can have a number of applications, for example, brightening a very weak image for night vision or converting an infrared image into a visible one.

Light falling on the cathode causes electrons to be emitted in proportion to the light intensity. These electrons are accelerated and focused by the applied electric field onto the fluorescent screen to form a visible image. Luminance gains of 50–100 times can be achieved, and a sequence of devices can be used to magnify the gain even more.

Image Orthicon Tube (TV Camera)

There are two basic types of **television** (TV) camera tubes, the orthicon and the vidicon. The orthicon uses the photoemissive effect. A light image is focused onto the photocathode, and the electrons emitted are attracted toward a positively based target (see Fig. 31.15). The target is a wire mesh, and the electrons pass through it to be collected on a glass electron target screen. This also causes secondary electrons to be emitted, and they also collect on the screen. This results in a positive charge image which replicates the light image on the photocathode.

A low-velocity electron beam is raster scanned across the target to neutralize the charge. The surplus electrons return to the electron multiplier and generate a current for the signal output. The output current is thus inversely proportional to the light level at the scanning position of the beam. The orthicon tube is very sensitive because there is both charge accumulation between scans and gain from the electron multiplier.

Vidicon Camera Tube

A simple TV camera tube is the vidicon. This is the type used in camcorders and for many video applications where a rugged, simple, and inexpensive camera is required. Figure 31.16 is a schematic of a vidicon tube; the optical image is formed on the surface of a large-area photoconductor, causing corresponding variations in the conductivity. This causes the rear surface to charge toward the bias voltage V_b in relation to the conductivity image. The scanning electron beam periodically recharges the rear side to 0 V, resulting in a recharging current flow in the output. The output signal is a current signal proportional to the light incident at the position of the scanning electron beam.

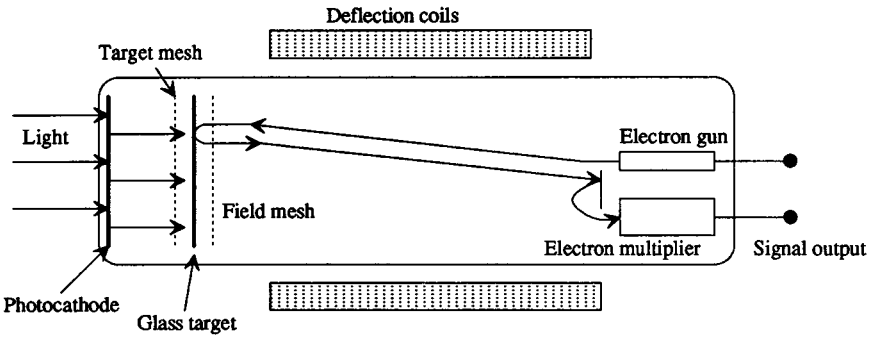


FIGURE 31.15 Schematic diagram of an image orthicon TV camera tube.

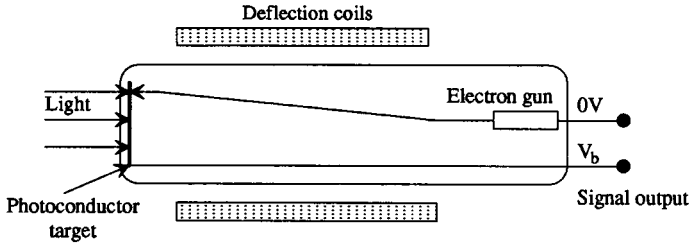


FIGURE 31.16 Schematic of a vidicon TV camera tube.

The primary disadvantages of the vidicon are its longer response time and smaller dynamic range. The recent availability of longer wavelength photoconducting films has resulted in new infrared cameras becoming available.

A recent advance in these types of image sensor is to replace the photoconductor with a dense array of very small semiconductor photodiodes. Photocurrent in the diode charges a capacitor connected to it. The raster scanned electron beam discharges this capacitor in the same way.

Image Dissector Tube

The image dissector tube is a photosensitive device which uses an electron deflection lens to image the electron from the cathode onto a pinhole in front of an electron multiplier. The image can be deflected around in front of the pinhole in a random access manner. The primary application of this kind of device is for tracking purposes.

Noise and Detectivity

Noise

There are two primary sources of noise in photodetectors: Johnson noise due to thermal effects in the resistive components of the device and its circuits, and shot noise or its equivalent, which is due to the quantized nature of electro-optic interactions.

In semiconductor devices noise is usually given in terms of noise current,

$$\delta\bar{i}^2 = 2eiM^{2+x}\Delta f + \frac{4kT\Delta f}{R} \tag{31.25}$$

where i includes signal and dark currents, e is electron charge, M is avalanche gain (x depends on avalanche photodetector characteristics), Δf is frequency bandwidth, k is Boltzmann's constant, T is in degrees Kelvin, and R is the total circuit resistance at temperature, T.

For photoconductor devices (including effects of charge amplification) the noise current is given by

$$\delta\bar{i}^2 = \frac{4ei(\tau_0/\tau_d)\Delta f}{1 + 4\pi^2\upsilon^2\tau_0^2} + \frac{4kT\Delta f}{R} \tag{31.26}$$

The first term is analogous to shot noise but includes the effects of carrier creation and recombination. τ_0 is the carrier lifetime, τ_d is the drift time for a carrier to go across the photoconductor, and υ is the light frequency.

The noise for photoemissive devices is usually written as a noise voltage and is given by

$$\delta_v^{-2} = 2ei\,G^2\Delta f R^2 + 4kT\Delta f R \tag{31.27}$$

where G is the current gain for the photomultiplier.

Detectivity

The performance of a detector is often described using the term D^*, detectivity. This term is useful for comparison purposes by normalizing with respect to detector size and/or noise bandwidth. This is written as

$$D^* = \frac{\sqrt{A\Delta f}}{\text{NEP}} \tag{31.28}$$

where NEP is the noise equivalent power (for signal-to-noise ratio equal to 1) and A is detector area. The term $D^*(\lambda)$ is used for quoting the result using a single-wavelength light source and $D^*(\mathrm{T})$ is used for the unfiltered blackbody radiation source.

Defining Terms

Charge-coupled device (CCD): A series of electronic logic cells in a device in which a signal is represented and stored as an electronic charge on a capacitor. The signal is moved from one cell (memory position or register) to an adjacent cell by electronically switching the charge between the capacitors.

Electron multiplication: The phenomenon where a high-energy electron strikes a surface and causes additional electrons to be emitted from the surface. Energy from the incident electron transfers to the other electrons to cause this. The result is electron gain which is proportional to the incident electron energy.

Extended source: A light source with finite size where the source size and shape can be determined from the emitted light characteristics. The light is spatially incoherent.

Light detection: The conversion of light energy into an electrical signal, either current or voltage.

Light emission: The creation or emission of light from a surface or device.

Point source: A light source which is so small that its size and shape cannot be determined from the characteristics of the light emanating from it. The light emitted has a spherical wave front and is spatially coherent.

Television (TV): The process of detecting an image and converting it to a serial electronic representation. A detector raster scans the image, producing a voltage proportional to the light intensity. The time axis represents the distance along the raster scan. Several hundred horizontal scans make up the image starting at the top. The raster scan is repeated to provide a continuing sequence of images.

Related Topic

42.2 Optical Fibers and Cables

References

B. Crosignani, P. DiPorto, and M. Bartolotti, *Statistical Properties of Scattered Light,* New York: Academic Press, 1975.

A.L. Lentine et al., "A 2 kbit array of symmetric self-electrooptic effect devices," *IEEE Photonics Technol. Lett.*, vol. 2, no. 1, 1990.

Reticon Corp., subsidiary of EG&G, Inc., Application notes #101.

Further Information

W.J. Smith, *Modern Optical Engineering,* New York: McGraw Hill, 1966.

M.J. Howes and D.V. Morgan, *Gallium Arsenide Materials, Devices and Circuits,* New York: John Wiley, 1985.

M.K. Baroski, *Fundamentals of Optical Fiber Communications,* New York: Academic Press, 1981.

C.Y. Wyatt, *Electro-Optic System Design for Information Processes,* New York: McGraw-Hill, 1991.

S. Ungar, *Fibre Optics—Theory and Applications,* New York: John Wiley, 1990.

31.3 Circuits

R.A. Becker

In 1969, Stewart Miller of AT&T Bell Laboratories published his landmark article on **integrated optics**. This article laid the foundation for what has now developed into optoelectronic circuits. In it he described the concepts of planar **optical guided-wave devices** formed as thin films on various substrates using fabrication techniques similar to those used in the semiconductor integrated circuit (IC) industry. The attributes of these new circuits included small size, weight, power consumption, and mechanical robustness because all components were integrated on a single substrate. The field of optoelectronic circuits began as a hybrid implementation where optical sources (laser diodes) and detectors have historically been fabricated on separate semiconductor substrates, and waveguide devices, such as modulators and switches, have been fabricated on electro-optic single-crystal oxides such as **lithium niobate** ($LiNbO_3$). Often, the two dissimilar substrates have been connected using single-mode polarization preserving optical fiber. Now, although the hybrid concept is finding commercial applications, most active research is performed on monolithic implementations, where all devices are fabricated on a common semiconductor substrate. After a brief summary discussion of semiconductor, glass, and polymer material systems, we will deal exclusively with the most mature hybrid implementation of optoelectronic circuits based on $LiNbO_3$.

Because sources and detectors have been covered in previous sections, in this section the devices that are utilized in between, i.e., modulators and switches, will be discussed.

Integrated Optics

Integrated optics can be defined as the monolithic integration of one or more optical guided-wave structures on a common substrate. These structures can be passive, such as a fixed optical power splitter, or active, such as an optical switch. Active devices are realized by placing metal electrodes in close proximity to the optical waveguides. Applying a voltage to the electrodes changes the velocity of the light within the waveguide. Depending on the waveguide geometry and the electrode placement, a wide variety of technologically useful devices and circuits can be realized.

The technological significance of integrated optics stems from its natural compatibility with two other rapidly expanding technologies: fiber optics and semiconductor laser diodes. These technologies have moved in the past 10 years from laboratory curiosities to large-scale commercial ventures. Integrated optic devices typically use laser diode optical sources, diode-pumped yttrium, aluminum, garnet (YAG) lasers, and transmit the modified optical output on a single-mode optical fiber. Integrated optic devices are typically very high speed, compact, and require only moderate control voltages compared to their bulk-optical counterparts.

In integrated optic devices, the optical channel waveguides are formed on a thin, planar, optically polished substrate using photolithographic techniques similar to those used in the semiconductor IC industry. Waveguide routing is accomplished by the mask used in the photolithographic process, similar to the way electrically conductive paths are defined in semiconductor ICs. The photolithographic nature of device fabrication offers the potential of readily scaling the technology to large volumes, as is done in the semiconductor IC industry. For example, the typical device is 0.75 in. × 0.078 in. in size. Dividing the substrate size by the typical device size and assuming a 50% area usage indicates that one can achieve 50 devices per 3-in. wafer.

Substrate materials for integrated optics include semiconductors, such as GaAs and InP, glass, polymer coated glass or Si, and $LiNbO_3$. Recently, primarily passive glass-based devices have been commercially introduced as replacements for passive all-fiber devices such as splitters and combiners. In addition, there are slow-speed switches (millisecond) now available that utilize the thermooptic effect in glass. Glass-based devices are fabricated by either depositing glass waveguiding layers on Si, or through the indiffusion of dopants into glass which results in a waveguiding layer. Both fabrication approaches are used in commercially available devices.

Very recently, low-speed polymer-on-Si switches have been commercially introduced. These also operate via the thermooptic effect. However, since polymers can be engineered with electrooptic properties, high-speed devices may also be available in the future. The primary impediment to market penetration of polymer-based devices has been their relatively poor stability, especially at temperatures above 100°C. However, if polymers can be produced with both strong electrooptic properties and enhanced stability with temperature, they could be the material system of choice for many applications because of their low-cost potential.

The area of semiconductor-based integrated optics has attracted much attention worldwide because if offers the potential of integrating electronic circuitry, optical sources and detectors, and optical waveguides on a single substrate. While being quite promising, the technology is still 5 years away from commercialization. Technical problems in semiconductor-based integrated optics include low electrooptic coefficients, higher optical waveguide attenuation, and an incompatibility of the processing steps needed to fabricate the various types of devices on a single substrate. However, considerable attention is being paid to these problems, and improvements are continually occurring.

The primary substrate material in integrated optics is the widely available synthetic crystal, lithium niobate ($LiNbO_3$), which has been commercially produced in volume for more than 20 years. This material is transparent to optical wavelengths between 400 and 4500 nm, has a hardness similar to glass, and is nontoxic.

$LiNbO_3$-based devices have been commercially available since 1985 and have been incorporated in a large number of experimental systems. The basic $LiNbO_3$ waveguide fabrication technique was developed in 1974 and has been continually refined and improved during subsequent years. The material itself finds wide application in a number of electrical and optical devices because of its excellent optical, electrical, acoustic, and electro- and acousto-optic properties. For example, almost all color television sets manufactured today incorporate a surface-acoustic-wave (SAW) electrical filter based on $LiNbO_3$.

In $LiNbO_3$-based integrated optics, optical waveguides are formed in one of two ways. The first uses photolithographically patterned lines of titanium (Ti), several hundred angstroms thick, on the substrate surface. The titanium is then diffused into the substrate surface at a temperature of about 1000°C for several hours. This process locally raises the refractive index in the regions where titanium has been diffused, forming high-refractive index stripes that will confine and guide light. Because the diffusion is done at exceedingly high temperatures, the waveguide stability is excellent. The waveguide mechanism used is similar to that used in fiber optics, where the higher-index, doped cores guide the light. The exact titanium stripe width, the titanium thickness, and diffusion process are critical parameters in implementing a low-loss single-mode waveguide. Different fabrication recipes are required to optimize the waveguides for operation at the three standard diode laser wavelengths: 800 nm, 1300 nm, and 1500 nm. The second approach uses a technique known as proton exchange. In this approach, a mask is used to define regions of the substrate where hydrogen will be exchanged for lithium, resulting in an increase in the refractive index. This reaction takes place at lower temperatures (200–250°C) but has been found to produce stable waveguides if an anneal at 350–400°C is performed. Waveguides formed using the proton exchange method support only one polarized mode of propagation, whereas those formed using Ti indiffusion support two. Proton exchange waveguides are also capable of handling much higher optical power densities, especially at the shorter wavelengths, than are those formed by Ti indiffusion. More fabrication detail will be provided later.

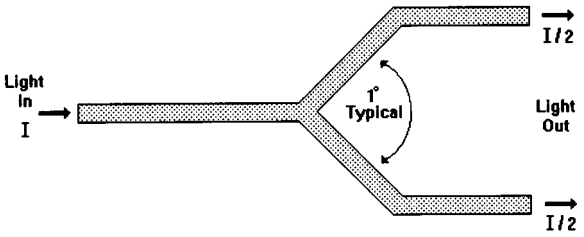


FIGURE 31.17 Passive Y-splitter.

Light modulation is realized via the electro-optic effect, i.e., inducing a small change in the waveguide refractive index by applying an electric field within the waveguide. On an atomic scale the applied electric field causes slight changes in the basic crystal unit cell dimensions, which changes the crystal's refractive index. The magnitude of this change depends on the orientation of the applied electric field and the optical polarization. As a result, only certain crystallographic orientations are useful for device fabrication and devices are typically polarization dependent. The electro-optic coefficients of $LiNbO_3$ are among the highest (30.8 pm/V) of any inorganic material, making the material very attractive for integrated optic applications.

Combining the concepts of optical waveguides and electro-optic modulation with the geometric freedom of photolithographic techniques leads to an extremely diverse array of passive and active devices.

Passive components do not require any electric fields and are used for power splitting and combining functions. Two types of passive power division structures have been fabricated: Y-junctions and directional couplers. A single waveguide can be split into two by fabricating a shallow-angle Y-junction as shown in Fig. 31.17.

An optical signal entering from the single-waveguide side of the junction is split into two optical signals with the same relative phase but one-half the original intensity. Conversely, light incident on the two-waveguide side of the junction will be combined into the single waveguide with a phase and intensity dependent on the original inputs. Directional couplers consist of two or more waveguides fabricated in close proximity to each other so that the optical fields overlap as shown in Fig. 31.18. As a result, optical power is transferred between the waveguides. The extent of the power transfer is dependent on the waveguide characteristics, the waveguide spacing, and the interaction length.

A different type of passive component is an optical polarizer, which can be made using several different techniques. One such method is the metal-clad, dielectric-buffered waveguide shown in Fig. 31.19. In this passive device, the TM polarization state is coupled into the absorbing metal and is thus attenuated, while the TE polarization is virtually unaffected. Measurements of a 2-mm-long polarizer of this type have demonstrated TM attenuations exceeding 50 dB (100,000:1). Polarizers can also be fabricated in others ways. One interesting technique involves the diffusion of hydrogen ions into the $LiNbO_3$. This results in a waveguide which, as discussed earlier, will only support the TE-polarized mode and, thus, is a natural polarizer.

Active components are realized by placing electrodes in close proximity to the waveguide structures. Depending on the substrate crystallographic orientation, the waveguide geometry, and the electrode geometry, a wide variety of components can be demonstrated. The simplest active device is the **phase modulator**, which is a single waveguide with electrodes on either side as shown in Fig. 31.20. Applying a voltage across the electrodes induces an electric field across the waveguide, which changes its refractive index via the electro-optic effect. For 800-nm wavelength operation, a typical phase modulator would be 6 mm long and would induce a π-phase shift for an applied voltage of 4 V. The transfer function (light out versus voltage in) can be expressed as

$$I_0(V) = I_i \exp(j\omega t + \pi V/V_\pi) \tag{31.29}$$

where V_π is the voltage required to cause a 180-degree phase shift. Note that there is no change in the intensity of the light. Coherent techniques are used to measure the amount of phase change.

Optical **intensity modulators** can be fabricated by combining two passive Y-junctions with a phase modulator situated between them. The result, which is shown in Fig. 31.21, is a guided-wave implementation of the classic Mach–Zehnder interferometer. In this device the incoming light is split into two equal components by the first

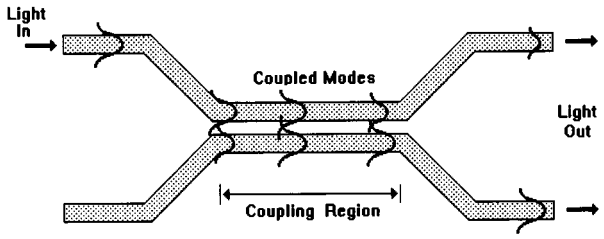


FIGURE 31.18 Directional coupler power splitter.

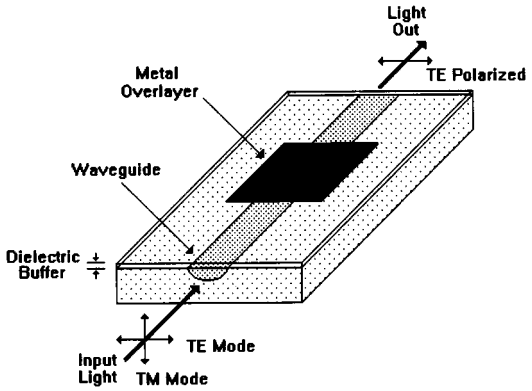


FIGURE 31.19 Thin-film optical polarizer.

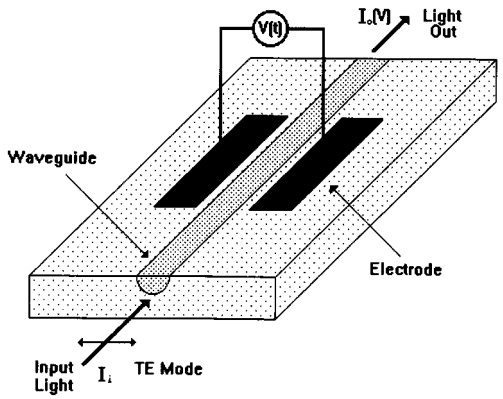


FIGURE 31.20 Electro-optic integrated optic phase modulator.

Y-junction. An electrically controlled differential phase shift is then introduced by the phase modulator, and the two optical signals are recombined in the second Y-junction. If the two signals are exactly in phase, then they recombine to excite the lowest-order mode of the output waveguide and the intensity modulator is turned fully on. If instead there exists a π-phase shift between the two signals, then they recombine to form the second mode, which is radiated into the substrate and the modulator is turned fully off. Contrast ratios greater than 25 dB (300:1) are routinely achieved in commercial devices. The transfer function for the Mach–Zehnder modulator can be expressed as

$$I_0(V) = I_i \cos^2(\pi V/2V_\pi + \phi) \tag{31.30}$$

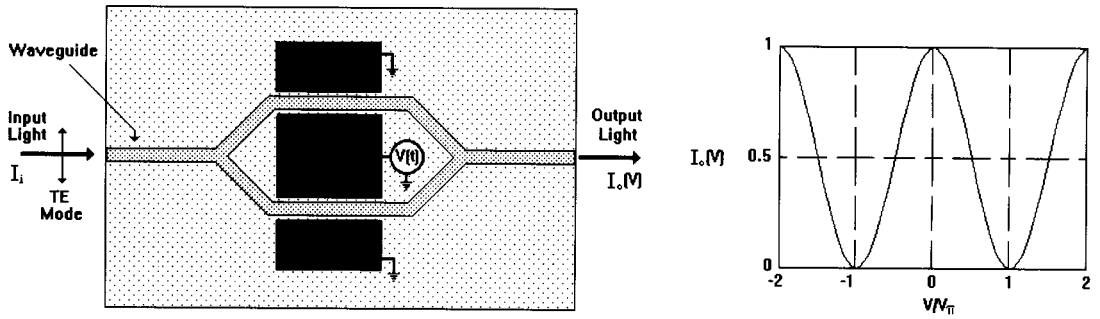


FIGURE 31.21 Mach–Zehnder intensity modulator and transfer function.

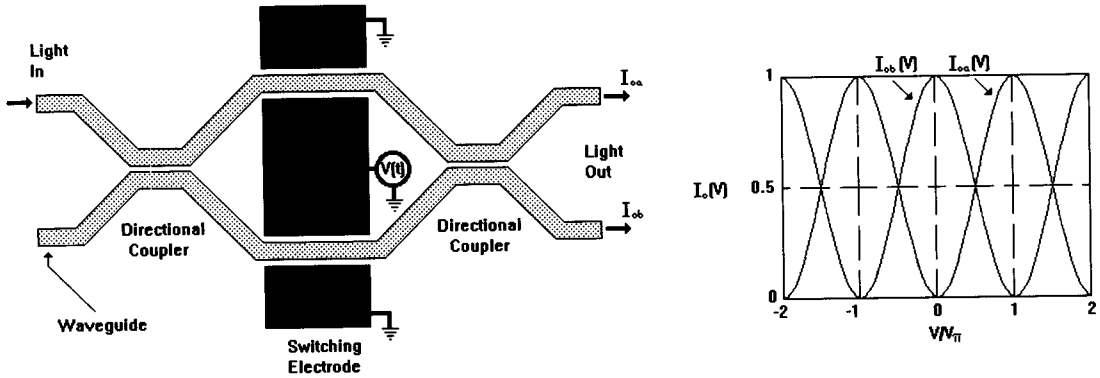


FIGURE 31.22 Balanced-bridge modulator/switch and transfer function.

where V_π is the voltage required to turn the modulator from on to off, and ϕ is any static phase imbalance between the interferometer arms. This transfer function is shown graphically in Fig. 31.21. Note that the modulator shown in Fig. 31.21 has push-pull electrodes. This means that when a voltage is applied, the refractive index is changed in opposite directions in the two arms, yielding a twice-as-efficient modulation.

Optical switches can be realized using a number of different waveguide, electrode, and substrate orientations. Two different designs are used in commercially available optical switches: the balanced-bridge and the $\Delta\beta$ directional coupler. The balanced-bridge design is similar to that of the Mach–Zehnder interferometer, except that the Y-junctions have been replaced by 3-dB directional couplers as shown in Fig. 31.22.

Similar to the Mach–Zehnder, the first 3-dB coupler splits the incident signal into two signals, ideally of equal intensity. Once again, if a differential phase shift is electro-optically induced between these signals, then when they recombine in the second 3-dB coupler, the ratio of power in the two outputs will be altered. Contrast ratios greater than 20 dB (100:1) are routinely achieved in commercial devices. The transfer function for this switch can be expressed as

$$I_{0a} = I_i \cos^2(\pi V/2V_\pi + \pi/2) \tag{31.31}$$

$$I_{0b} = I_i \sin^2(\pi V/2V_\pi + \pi/2) \tag{31.32}$$

and is graphically depicted in Fig. 31.22. In the other type of switch, the $\Delta\beta$ directional coupler, the electrodes are placed directly over the directional coupler as shown in Fig. 31.23. The applied electric field alters the power transfer between the two adjacent waveguides. Research versions of this switch have demonstrated contrast ratios greater than 40 dB (10,000:1); however, commercial versions typically achieve 20 dB, which is competitive

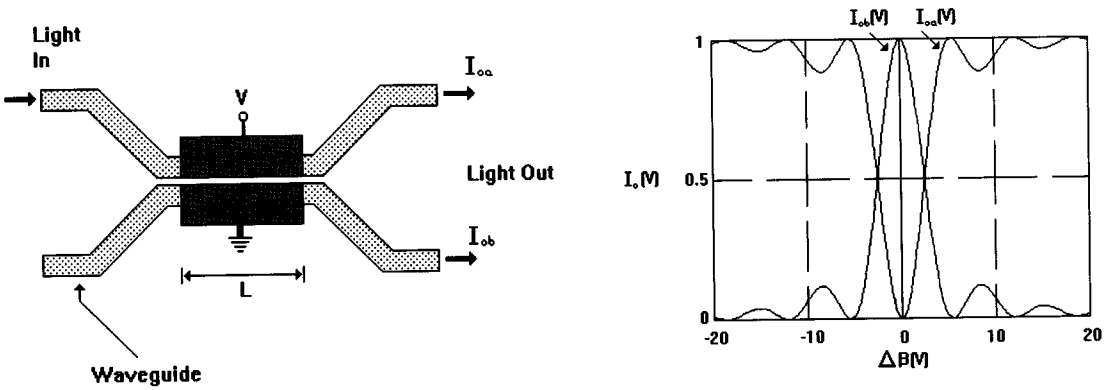


FIGURE 31.23 Directional coupler switch and transfer function.

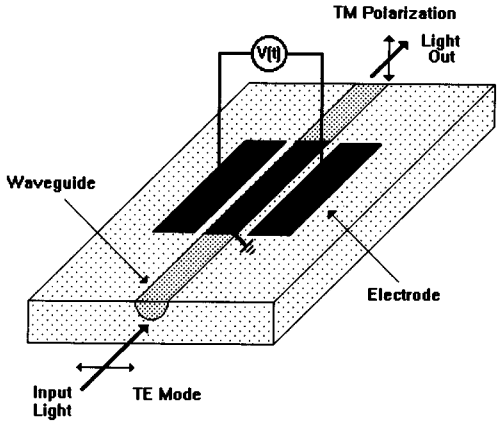


FIGURE 31.24 Guided-wave polarization controller.

with that achieved with the balanced-bridge switch. The transfer function for the Δβ directional coupler switch can be expressed as

$$I_{0a} = \sin^2 \kappa L^* \text{sqrt}(1 + (\Delta\beta/2\kappa)^2)/(1 + (\Delta\beta/2\kappa)^2) \tag{31.33}$$

$$I_{0b} = 1 - I_{0a} \tag{31.34}$$

where κ is the coupling constant and Δβ is the voltage-induced change in the propagation constant. This transfer function is depicted in Fig. 31.23.

Another type of active component that has recently become available commercially is the **polarization controller.** This component allows the incoming optical polarization to be continuously adjustable. The device functions as an electrically variable optical waveplate, where both the birefringence and axes orientation can be controlled. The controller is realized by using a three-electrode configuration as shown in Fig. 31.24 on a substrate orientation where the TE and TM optical polarizations have almost equal velocities. Typical performance values are TE/TM conversion of greater than 99% with less than 50 V.

One of the great strengths of integrated optic technology is the possibility of integrating different types or multiple copies of the same type of device on a single substrate. While this concept is routinely used in the semiconductor IC industry, its application in the optical domain is novel. The scale of integration in integrated optics is quite modest by semiconductor standards. To date the most complex component demonstrated is an

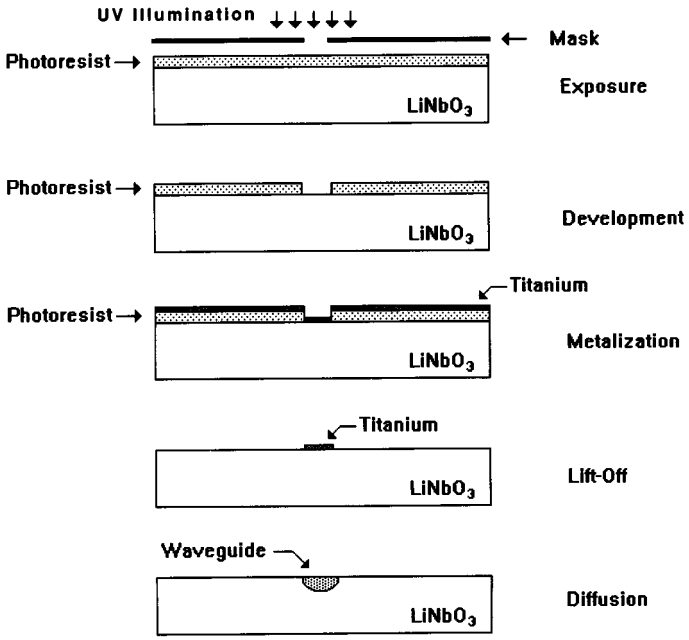


FIGURE 31.25 Ti-indiffused $LiNbO_3$ waveguide fabrication.

8×8 optical switch matrix that uses 64 identical 2×2 optical switches. The most device diversity on a given substrate is found in fiber gyro applications. Here, components incorporating six phase modulators, two electrically tunable directional couplers, and two passive directional couplers have been demonstrated.

Device Fabrication

The fabrication of an integrated optic device uses the same techniques as used in the semiconductor IC industry. Device designs are first entered into a computer-aided design (CAD) system for accurate feature placement and dimensional control. This design is then output as a digitized tape that will control a pattern generation system for fabrication of the chrome masks that are used in device fabrication. A variety of equipment such as step-and-repeat and E-beam systems has been developed for the semiconductor IC industry for the generation of chrome masks. These same systems are used today for generation of masks for integrated optic devices.

The waveguides can be fabricated by using either the Ti indiffusion method or the proton exchange method. The first step in fabricating a waveguide device using Ti indiffusion is the patterning in titanium. The bare $LiNbO_3$ surface is first cleaned and then coated with photoresist. Next, the coated substrate is exposed using the waveguide-layer chrome mask. The photoresist is then developed. The areas that have been exposed are removed in the development cycle. The patterned substrates are then coated with titanium in a vacuum evaporator. The titanium covers the exposed regions of the substrate as well as the surface of the remaining photoresist. The substrate is next soaked in a photoresist solvent. This causes all the residual photoresist (with titanium on top) to be removed, leaving only the titanium that coated the bare regions of the substrate. This process is known as *lift-off*. Finally, the substrate, which is now patterned with titanium, is placed in a diffusion system. At temperatures above 1000°C the titanium diffuses into the substrate, slightly raising the refractive index of these regions. This process typically takes less than 10 hours. This sequence of steps is depicted in Fig. 31.25. The proton exchange method is depicted in Fig. 31.26. Here a chrome masking layer is first deposited on the $LiNbO_3$ substrate. It is patterned using photoresist and etching. Next, the substrate is submerged in hot benzoic acid. Finally, the chrome mask is removed and the substrate is annealed. The regions that have been exposed to the benzoic acid will have an increased refractive index and will guide light.

If the devices being fabricated are to be active (i.e., voltage controlled), then an electrode fabrication step is also required. This sequence of steps parallels the waveguide fabrication sequence. The only differences are that

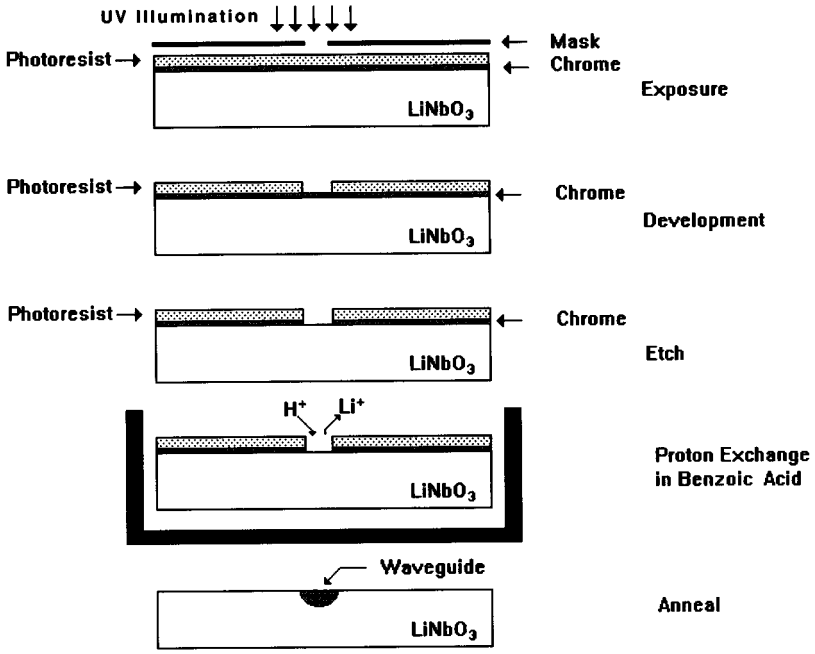


FIGURE 31.26 Proton exchange $LiNbO_3$ waveguide fabrication.

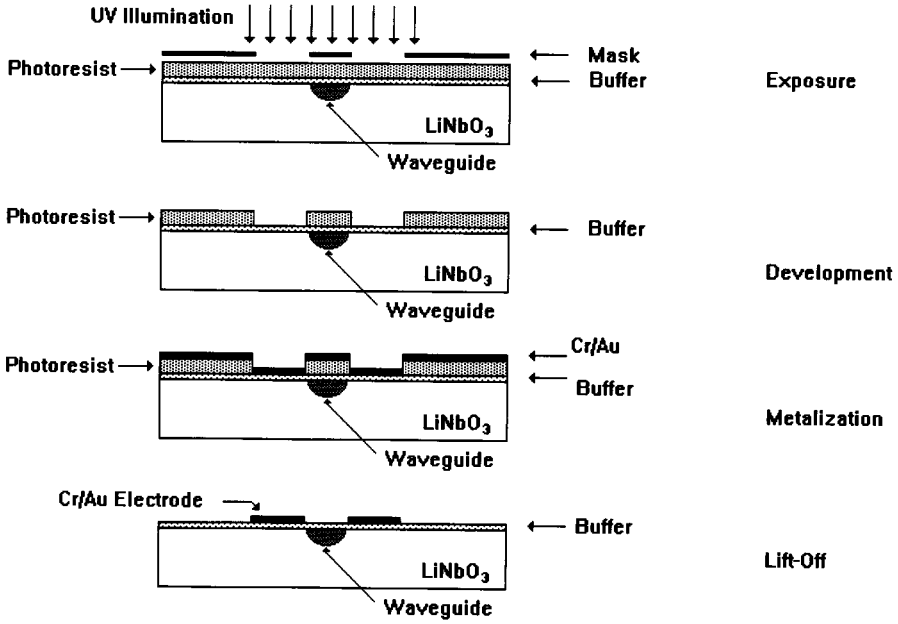


FIGURE 31.27 Electrode fabrication via lift-off.

an electrode mask is used and the vacuum-deposited metal used is chrome/gold or chrome/aluminum. This sequence of steps is shown in Fig. 31.27.

In order to get the light in and out of the waveguide, the endfaces have to be lapped and polished flat with chip-free knife edges. This is currently accomplished using standard lapping and polishing techniques. After

this step, the substrate can be diced into as many devices as were included on the substrate. Finally, the diced parts need to be electrically and optically packaged.

Packaging

To get the light in and out of an integrated optic waveguide requires a tiny optical window to be polished onto the waveguide's end. Currently, the entire endface of the substrate is polished to a sharp, nearly perfect corner, making the whole endface into an optical window. An optical fiber can then be aligned to the waveguide end and attached. Typically, centration of the fiber axis to the waveguide axis must be better than 0.2 μm. Some devices require multiple inputs and outputs. In this case the fibers are prealigned in silicon V-grooves. These V-grooves are fabricated by anisotropic etching of a photolithographically defined pattern on the silicon. The center-to-center spacing of the fiber V-groove array can be made to closely match that of the multiple waveguide inputs and outputs.

Integrated optic devices built on $LiNbO_3$ are inherently single-mode devices. This means that the light is confined in a cross-sectional area of approximately 30 μm^2. The optical mode has a near-field pattern that is 5 to 10 μm across and 3 to 6 μm deep, depending on the wavelength. These mode spot sizes set limits on how light can be coupled in and out. There are a number of methods that can be used to couple the light into $LiNbO_3$ waveguides. These include prism coupling, grating coupling, end-fire coupling with lenses, and end-fire coupling with single-mode optical fibers. In general, most of these techniques are only useful for laboratory purposes. The most practical real-world technique is end-fire coupling with an optical fiber. In this case the optical fiber is aligned to the waveguide end. This is an excellent practical method since integrated optic devices are most often used in fiber optic systems. Therefore, the coupling problem is one of aligning and fixing a single-mode fiber to the single-mode $LiNbO_3$ waveguide. The size of the single-mode radiation pattern and its angular divergence set the alignment tolerances. A low-loss connection between a fiber and a $LiNbO_3$ waveguide requires <1/20 of a mode spot diameter (0.25–0.5 μm) in transverse offset, angular tilt of <2 degrees, and a longitudinal offset of <1 mode spot diameter (5–10 μm). These are very stringent alignment requirements, especially if they have to be maintained over a wide temperature range.

Another aspect of the problem is that many integrated optic devices require a single, well-defined linearly polarized input. Ordinary single-mode fiber is not suitable in this case. The solution is to use polarization preserving fiber. This fiber is made such that it will maintain a single linearly polarized input over long distances. The use of polarization preserving fiber, however, adds another requirement to the coupling problem. This requirement is that the fiber must be rotationally aligned about its cylindrical axis so that the linearly polarized light coincides with the desired rotational axis of the $LiNbO_3$ waveguide. The rotational precision needed is <0.5 degrees.

Many $LiNbO_3$ devices, such as fiber gyro components, require multiple input and/or output optical connections. Thus, the packaging must be able to accommodate multiple inputs/outputs and maintain strict alignment for all connections.

The method of end-fire coupling optical fibers to the $LiNbO_3$ waveguide is commonly called *pigtailing*. This is the only practical packaging method now used for integrated optic devices that operate in a real system and outside the laboratory. The reasons for this are quite logical. The end user installs the device in his system by connecting to the fiber pigtails. The connection can be made with single-mode connectors or by splicing. Flexibility is one of the big advantages of using fiber pigtails.

The typical $LiNbO_3$ device is packaged in a metallic case with optical fiber pigtails connected at both ends. Electrical connections are provided by RF connectors or pins, which are common in the electronics industry. If hermetically sealed packages are desired, then the optical fiber pigtails must be hermetically sealed to the metallic package.

Applications

Many useful systems have been demonstrated using $LiNbO_3$-based integrated optic devices. These system applications can be grouped into four broad categories: telecommunications, instrumentation, signal processing, and sensors. In some cases only a single integrated optic device is used, while in other applications a multifunction component is required.

Optical switches have been shown to be quite useful in the telecommunications area. High-speed 2×2 switches for time-domain mux/demux as well as lower-speed 4×4 switch arrays have been successfully demonstrated. In both cases a major advantage of optical switching is that the switch data transmission rate is not limited by the switch itself as is the case for electronic switches. Thus, it is possible to route optical signals at data rates exceeding terabits/per second (1,000,000,000,000 bits/s).

Aside from the switching application in telecommunications, there also is the high-speed laser modulation application. Using an external $LiNbO_3$-based optic intensity modulator, both analog and digital data transmission systems have been demonstrated. Analog transmission systems using integrated optic devices are particularly attractive as remote antenna links because of their high speed and ability to be driven directly by the received signal without amplification. Recently, the use of high-power diode-pumped YAG lasers operating at 1300 nm and external intensity modulators based on $LiNbO_3$ have found wide application in the cable TV industry. In addition, the ability of the Mach–Zehnder intensity modulator to control intensity with a controlled wavelength change (i.e., chirp) has allowed its use in long-haul telecommunications systems.

Another demonstrated application of integrated optics in the telecommunications area is in coherent communication systems. These systems require both phase modulators and polarization controllers. Current optical fiber transmission systems rely on intensity modulated data transmission schemes. Coherent communication systems are attractive because of the promise of higher bit rates, wavelength division multiplexing capability, and greater noise immunity. In coherent communication systems the information is coded by varying either the phase or frequency of the optical carrier with a phase modulator. At the receiver, a polarization controller is used to ensure a good signal-to-noise ratio in the heterodyne detection system.

One promising application of integrated optic devices in instrumentation is a high-speed, polarization-independent optical switch for use in optical-time-domain reflectometers (OTDR). OTDRs are used to locate breaks or poor splices in fiber optic networks. The instruments work by sending out an optical pulse and measuring the backscattered radiation returning to the instrument as a function of time. The next generation of OTDRs will possibly employ an optical switch, which will be used to rapidly switch the optical fiber under test from the pulsed light source to the OTDR receiver. Such an instrument could detect faults closer to the OTDR than currently possible, which is important in the short-haul systems now being installed. This feature is necessary for local area network (LAN) installations.

Several types of sensors using integrated optic devices have been demonstrated. Two of the most promising are electric/magnetic field or voltage sensors and rotation sensors (fiber optic gyro). Electric field sensors typically consist of either a Mach–Zehnder intensity modulator or an optical switch that is biased midway between the on and off states. For small modulation depths about this midpoint the induced optical modulation is linear with respect to applied voltage. Linear dynamic ranges in excess of 80 dB have been accomplished. This is larger than that obtained using any other known technology.

Perhaps the most promising near-term application of integrated optic devices in the field of rotation sensing is as a key component in optical fiber gyroscopes. A typical fiber optic gyro component is shown in Fig. 31.28. The device consists of a polarizer, a Y-junction, and two phase modulators, all integrated on a single substrate. In fiber gyro systems, the integrated optic component replaces individual, fiber-based components that perform the same function. The integrated optic component offers a greatly improved performance, at a significant reduction in cost, compared to the fiber-based components. Most fiber optic gyro development teams have done away with the fiber components and have adopted $LiNbO_3$-based components as the technology of choice.

Defining Terms

Integrated optics: The monolithic integration of one or more optical guided-wave devices on a common substrate.

Intensity modulator: A modulator that alters only the intensity of the incident light.

Lithium niobate ($LiNbO_3$): A single-crystal oxide that displays electro-optic, acousto-optic, piezoelectric, and pyroelectric properties that is often the substrate of choice for surface acoustic wave devices and integrate optical devices.

Optical guided-wave device: An optical device that transmits or modifies light while it is confined in a thin-film optical waveguide.

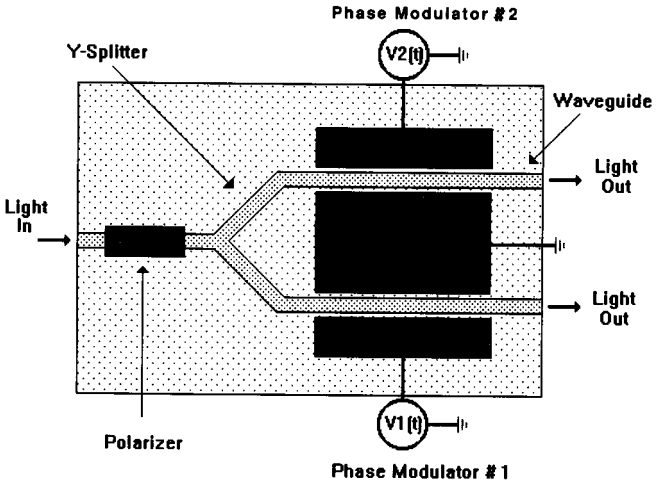


FIGURE 31.28 Fiber-optic gyro chip.

Phase modulator: A modulator that alters only the phase of the incident light.
Polarization controller: A device that alters only the polarization state of the incident light.

Related Topic

42.2 Optical Fibers and Cables

References

R. Alferness, "Waveguide electrooptic modulators," *IEEE Trans. Microwave. Theory Tech.,* vol. MTT-30, p. 1121, 1982.

R.A. Becker, "Commercially available integrated optics products and services," *SPIE,* vol. 993, p. 246, 1988.

R. Childs and V. O'Byrne, "Predistortion Linearization of Directly Modulated DFB Lasers and External Modulators for AM Video Transmission," *OFC'90 Tech. Dig.,* Paper WHG, 1990, p. 79.

C. Cox, G. Betts, and L. Johnson, "An analytic and experimental comparison of direct and external modulation in analog fiber-optic links," *IEEE Trans. Microwave Theory Tech.,* vol. 38, p. 501, 1990.

T. Findakly and M. Bramson, "High-performance integrated-optical chip for a broad range of fiber-optic gyro applications," *Opt. Lett.,* vol. 15, p. 673, 1990.

P. Granestrand, B. Stoltz, L. Thylen, K. Bergvall, W. Doldisen, H. Heinrich, and D. Hoffmann, "Strictly nonblocking 8×8 integrated optical switch matrix," *Electron. Lett.,* vol. 22, p. 816, 1986.

M. Howerton, C. Bulmer, and W. Burns, "Effect of intrinsic phase mismatch on linear modulator performance of the 1×2 directional coupler and Mach–Zehnder interferometer," *J. Lightw. Tech.,* vol. 8, p. 1177, 1990.

S.E. Miller, "Integrated optics: An introduction," *Bell Syst. Tech. J.,* vol. 48, p. 2059, 1969.

Further Information

Integrated Optical Circuits and Components, Design and Applications, edited by Lynn D. Hutcheson (Marcel Dekker, Inc., New York, 1987) and *Optical Integrated Circuits,* by H. Nishihara, M. Haruna, and T. Suhara (McGraw-Hill Book Company, New York, 1989) provide excellent overviews of the field of integrated and guided-wave optics.

Integrated Optics: Devices and Applications, edited by J. T. Boyd (IEEE Press, New York, 1991) provides an excellent cross section of recent publications in the field.

In addition, the monthly magazine *IEEE Journal of Lightwave Technology* provides many publications on current research and development on integrated and guide-wave optic devices and systems.

Garrod, S.A.R. “D/A and A/D Converters”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

32

D/A and A/D Converters

32.1 D/A and A/D Circuits
D/A and A/D Converter Performance Criteria • D/A Conversion Processes • D/A Converter ICs • A/D Conversion Processes • A/D Converter ICs • Grounding and Bypassing on D/A and A/D ICs • Selection Criteria for D/A and A/D Converter ICs

Susan A.R. Garrod
Purdue University

Digital-to-analog (D/A) conversion is the process of converting digital codes into a continuous range of analog signals. *Analog-to-digital (A/D) conversion* is the complementary process of converting a continuous range of analog signals into digital codes. Such conversion processes are necessary to interface real-world systems, which typically monitor continuously varying analog signals, with digital systems that process, store, interpret, and manipulate the analog values.

D/A and A/D applications have evolved from predominately military-driven applications to consumer-oriented applications. Up to the mid-1980s, the military applications determined the design of many D/A and A/D devices. The military applications required very high performance coupled with hermetic packaging, radiation hardening, shock and vibration testing, and military specification and record keeping. Cost was of little concern, and "low power" applications required approximately 2.8 W. The major applications up the mid-1980s included military radar warning and guidance systems, digital oscilloscopes, medical imaging, infrared systems, and professional video.

The applications requiring D/A and A/D circuits in the 1990s have different performance criteria from those of earlier years. In particular, low power and high speed applications are driving the development of D/A and A/D circuits, as the devices are used extensively in battery-operated consumer products. The predominant applications include cellular telephones, hand-held camcorders, portable computers, and set-top cable TV boxes. These applications generally have low power and long battery life requirements, and they may have high speed and high resolution requirements, as is the case with the set-top cable TV boxes.

32.1 D/A and A/D Circuits

D/A and A/D conversion circuits are available as integrated circuits (ICs) from many manufacturers. A huge array of ICs exists, consisting of not only the D/A or A/D conversion circuits, but also closely related circuits such as sample-and-hold amplifiers, analog multiplexers, voltage-to-frequency and frequency-to-voltage converters, voltage references, calibrators, operation amplifiers, isolation amplifiers, instrumentation amplifiers, active filters, dc-to-dc converters, analog interfaces to digital signal processing systems, and data acquisition subsystems. Data books from the IC manufacturers contain an enormous amount of information about these devices and their applications to assist the design engineer.

The ICs discussed in this chapter will be strictly the D/A and A/D conversion circuits. Table 32.1 lists a small sample of the variety of the D/A and A/D converters currently available. The ICs usually perform either D/A or A/D conversion. There are serial interface ICs, however, typically for high-performance audio and digital signal processing applications, that perform both A/D and D/A processes.

TABLE 32.1 D/A and A/D Integrated Circuits

D/A Converter ICs	Resolution, b	Multiplying vs. Fixed Reference	Settling Time, μs	Input Data Format
Analog Devices AD558	8	Fixed reference	3	Parallel
Analog Devices AD7524	8	Multiplying	0.400	Parallel
Analog Devices AD390	Quad, 12	Fixed reference	8	Parallel
Analog Devices AD1856	16	Fixed reference	1.5	Serial
Burr-Brown DAC729	18	Fixed reference	8	Parallel
DATEL DAC-HF8	8	Multiplying	0.025	Parallel
National DAC0800	8	Multiplying	0.1	Parallel

A/D Converter ICs	Resolution, b	Signal Inputs	Conversion Speed, μs	Output Data Format
Analog Devices AD572	12	1	25	Serial & parallel
Burr-Brown ADC803	12	1	1.5	Parallel
Burr-Brown ADC701	16	1	1.5	Parallel
National ADC1005B	10	1	50	Parallel
TI, National ADC0808	8	8	100	Parallel
TI, National ADC0834	8	4	32	Serial
TI TLC0820	8	1	1	Parallel
TI TLC1540	10	11	21	Serial

A/D and D/A Interface ICs	Resolution, b	On-Board Filters	Sampling Rate, kHz	Data Format
TI TLC32040	14	Yes	19.2 (programmable)	Serial
TI 2914 PCM codec & filter	8	Yes	8	Serial

D/A and A/D Converter Performance Criteria

The major factors that determine the quality of performance of D/A and A/D converters are *resolution, sampling rate, speed,* and *linearity.*

The *resolution* of a D/A circuit is the smallest change in the output analog signal. In an A/D system, the resolution is the smallest change in voltage that can be detected by the system and that can produce a change in the digital code. The resolution determines the total number of digital codes, or *quantization levels*, that will be recognized or produced by the circuit.

The *resolution* of a D/A or A/D IC is usually specified in terms of the bits in the digital code or in terms of the least significant bit (LSB) of the system. An *n*-bit code allows for 2^n quantization levels, or $2^n - 1$ steps between quantization levels. As the number of bits increases, the step size between quantization levels decreases, therefore increasing the accuracy of the system when a conversion is made between an analog and digital signal. The system resolution can be specified also as the voltage step size between quantization levels. For A/D circuits, the resolution is the smallest input voltage that is detected by the system.

The *speed* of a D/A or A/D converter is determined by the time it takes to perform the conversion process. For D/A converters, the speed is specified as the *settling time.* For A/D converters, the speed is specified as the *conversion time.* The settling time for D/A converters will vary with supply voltage and transition in the digital code; thus, it is specified in the data sheet with the appropriate conditions stated.

A/D converters have a maximum *sampling rate* that limits the speed at which they can perform continuous conversions. The sampling rate is the number of times per second that the analog signal can be sampled and converted into a digital code. For proper A/D conversion, the minimum sampling rate must be at least two times the highest frequency of the analog signal being sampled to satisfy the Nyquist sampling criterion. The conversion speed and other timing factors must be taken into consideration to determine the maximum sampling rate of an A/D converter. **Nyquist A/D converters** use a sampling rate that is slightly more than twice

the highest frequency in the analog signal. **Oversampling A/D converters** use sampling rates of *N* times this rate, where *N* typically ranges from 2 to 64.

Both D/A and A/D converters require a voltage reference in order to achieve absolute conversion accuracy. Some conversion ICs have internal voltage references, while others accept external voltage references. For high-performance systems, an external precision reference is needed to ensure long-term stability, load regulation, and control over temperature fluctuations. External precision voltage reference ICs can be found in manufacturers' data books.

Measurement accuracy is specified by the converter's *linearity. Integral linearity* is a measure of linearity over the entire conversion range. It is often defined as the deviation from a straight line drawn between the endpoints and through zero (or the offset value) of the conversion range. Integral linearity is also referred to as *relative accuracy.* The *offset* value is the reference level required to establish the zero or midpoint of the conversion range. *Differential linearity* is the linearity between code transitions. Differential linearity is a measure of the *monotonicity* of the converter. A converter is said to be monotonic if increasing input values result in increasing output values.

The accuracy and linearity values of a converter are specified in the data sheet in units of the LSB of the code. The linearity can vary with temperature, so the values are often specified at +25°C as well as over the entire temperature range of the device.

D/A Conversion Processes

Digital codes are typically converted to analog voltages by assigning a voltage weight to each bit in the digital code and then summing the voltage weights of the entire code. A general D/A converter consists of a network of precision resistors, input switches, and level shifters to activate the switches to convert a digital code to an analog current or voltage. D/A ICs that produce an analog current output usually have a faster settling time and better linearity than those that produce a voltage output. When the output current is available, the designer can convert this to a voltage through the selection of an appropriate output amplifier to achieve the necessary response speed for the given application.

D/A converters commonly have a fixed or variable reference level. The reference level determines the switching threshold of the precision switches that form a controlled impedance network, which in turn controls the value of the output signal. **Fixed reference D/A converters** produce an output signal that is proportional to the digital input. **Multiplying D/A converters** produce an output signal that is proportional to the product of a varying reference level times a digital code.

D/A converters can produce bipolar, positive, or negative polarity signals. A four-quadrant multiplying D/A converter allows both the reference signal and the value of the binary code to have a positive or negative polarity. The four-quadrant multiplying D/A converter produces bipolar output signals.

D/A Converter ICs

Most D/A converters are designed for general-purpose control applications. Some D/A converters, however, are designed for special applications, such as video or graphic outputs, high-definition video displays, ultra high-speed signal processing, digital video tape recording, digital attenuators, or high-speed function generators.

D/A converter ICs often include special features that enable them to be interfaced easily to microprocessors or other systems. Microprocessor control inputs, input latches, buffers, input registers, and compatibility to standard logic families are features that are readily available in D/A ICs. In addition, the ICs usually have laser-trimmed precision resistors to eliminate the need for user trimming to achieve full-scale performance.

A/D Conversion Processes

Analog signals can be converted to digital codes by many methods, including integration, **succesive approximation,** parallel (flash) conversion, **delta modulation, pulse code modulation,** and **sigma-delta conversion.** Two of the most common A/D conversion processes are successive approximation A/D conversion and parallel or flash A/D conversion. Very high-resolution digital audio or video systems require specialized A/D techniques that often incorporate one of these general techniques as well as specialized A/D conversion processes. Examples

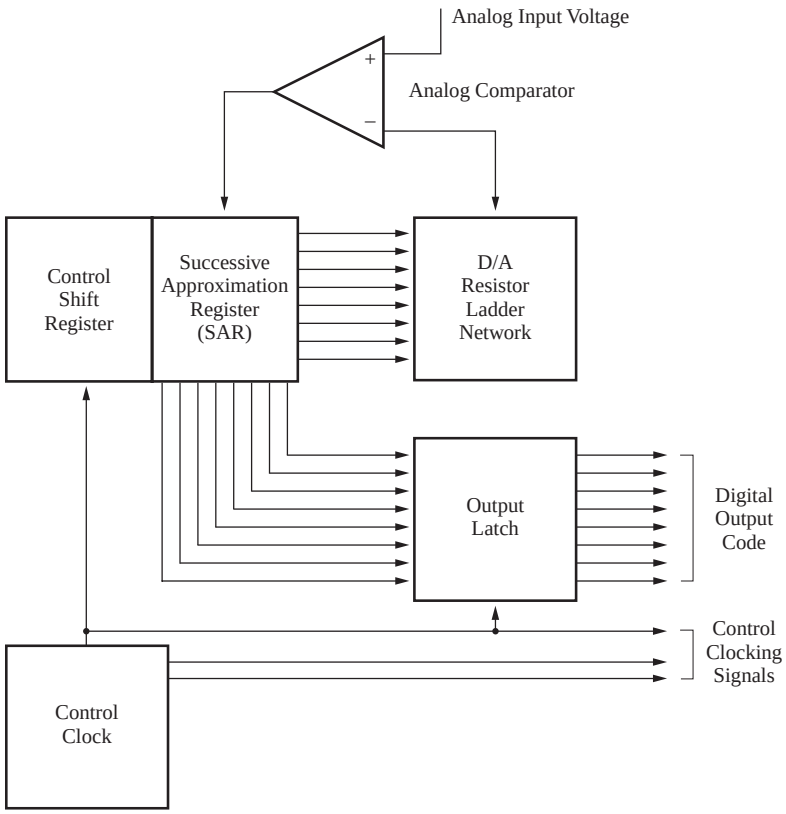


FIGURE 32.1 Successive approximation A/D converter block diagram.

of specialized A/D conversion techniques are pulse code modulation (PCM), and sigma-delta conversion. PCM is a common voice encoding scheme used not only by the audio industry in digital audio recordings but also by the telecommunications industry for voice encoding and multiplexing. Sigma-delta conversion is an oversampling A/D conversion where signals are sampled at very high frequencies. It has very high resolution and low distortion and is being used in the digital audio recording industry.

Successive approximation A/D conversion is a technique that is commonly used in medium- to high-speed data acquisition applications. It is one of the fastest A/D conversion techniques that requires a minimum amount of circuitry. The conversion times for successive approximation A/D conversion typically range from 10 to 300 μs for 8-bit systems.

The successive approximation A/D converter can approximate the analog signal to form an n-bit digital code in n steps. The successive approximation register (SAR) individually compares an analog input voltage to the midpoint of one of n ranges to determine the value of one bit. This process is repeated a total of n times, using n ranges, to determine the n bits in the code. The comparison is accomplished as follows: The SAR determines if the analog input is above or below the midpoint and sets the bit of the digital code accordingly. The SAR assigns the bits beginning with the most significant bit. The bit is set to a 1 if the analog input is greater than the midpoint voltage, or it is set to a 0 if it is less than the midpoint voltage. The SAR then moves to the next bit and sets it to a 1 or a 0 based on the results of comparing the analog input with the midpoint of the next allowed range. Because the SAR must perform one approximation for each bit in the digital code, an n-bit code requires n approximations.

A successive approximation A/D converter consists of four functional blocks, as shown in Fig. 32.1: the SAR, the analog comparator, a D/A converter, and a clock.

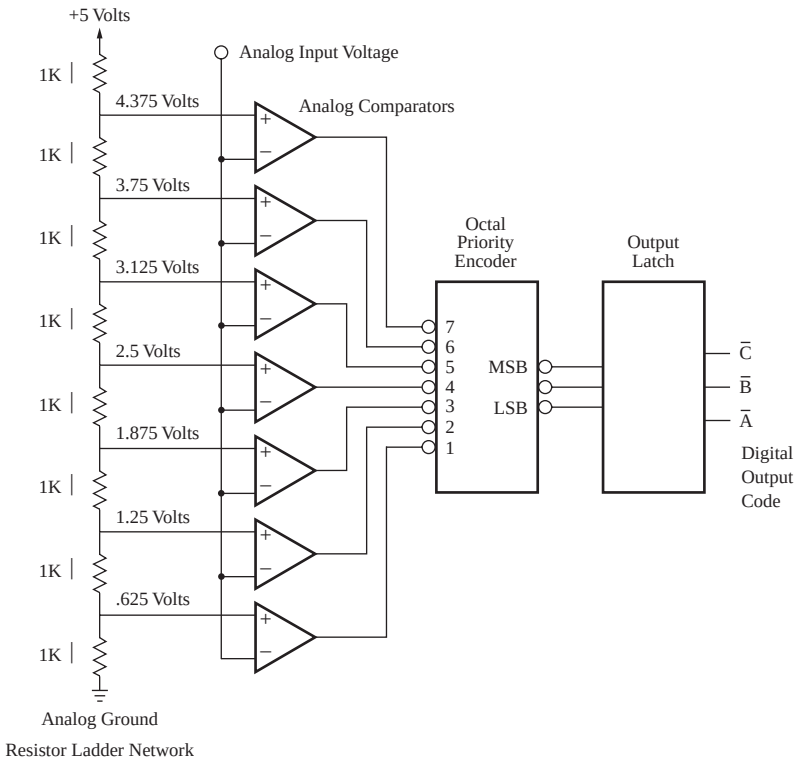


FIGURE 32.2 Flash A/D converter block diagram.

Parallel or flash A/D conversion is used in high-speed applications such as video signal processing, medical imaging, and radar detection systems. A flash A/D converter simultaneously compares the input analog voltage to $2^n - 1$ threshold voltages to produce an n-bit digital code representing the analog voltage. Typical flash A/D converters with 8-bit resolution operate at 20 to 100 MHz.

The functional blocks of a flash A/D converter are shown in Fig. 32.2. The circuitry consists of a precision resistor ladder network, $2^n - 1$ analog comparators, and a digital priority encoder. The resistor network establishes threshold voltages for each allowed quantization level. The analog comparators indicate whether or not the input analog voltage is above or below the threshold at each level. The output of the analog comparators is input to the digital priority encoder. The priority encoder produces the final digital output code that is stored in an output latch.

An 8-bit flash A/D converter requires 255 comparators. The cost of high-resolution A/D comparators escalates as the circuit complexity increases and as the number of analog converters rises by $2^n - 1$. As a low-cost alternative, some manufacturers produce modified flash A/D converters that perform the A/D conversion in two steps to reduce the amount of circuitry required. These modified flash A/D converters are also referred to as *half-flash* A/D converters, since they perform only half of the conversion simultaneously.

A/D Converter ICs

A/D converter ICs can be classified as general-purpose, high-speed, flash, and sampling A/D converters. The *general-purpose A/D converters* are typically low speed and low cost, with conversion times ranging from 2 μs to 33 ms. A/D conversion techniques used by these devices typically include successive approximation, tracking, and integrating. The general-purpose A/D converters often have control signals for simplified microprocessor interfacing. These ICs are appropriate for many process control, industrial, and instrumentation applications, as well as for environmental monitoring such as seismology, oceanography, meteorology, and pollution monitoring.

High-speed A/D converters have conversion times typically ranging from 400 ns to 3 μs. The higher speed performance of these devices is achieved by using the successive approximation technique, modified flash techniques, and statistically derived A/D conversion techniques. Applications appropriate for these A/D ICs include fast Fourier transform (FFT) analysis, radar digitization, medical instrumentation, and multiplexed data acquisition. Some ICs have been manufactured with an extremely high degree of linearity, to be appropriate for specialized applications in digital spectrum analysis, vibration analysis, geological research, sonar digitizing, and medical imaging.

Flash A/D converters have conversion times ranging typically from 10 to 50 ns. Flash A/D conversion techniques enable these ICs to be used in many specialized high-speed data acquisition applications such as TV video digitizing (encoding), radar analysis, transient analysis, high-speed digital oscilloscopes, medical ultrasound imaging, high-energy physics, and robotic vision applications.

Sampling A/D converters have a sample-and-hold amplifier circuit built into the IC. This eliminates the need for an external sample-and-hold circuit. The throughput of these A/D converter ICs ranges typically from 35 kHz to 100 MHz. The speed of the system is dependent on the A/D technique used by the sampling A/D converter.

A/D converter ICs produce digital codes in a serial or parallel format, and some ICs offer the designer both formats. The digital outputs are compatible with standard logic families to facilitate interfacing to other digital systems. In addition, some A/D converter ICs have a built-in analog multiplexer and therefore can accept more than one analog input signal.

Pulse code modulation (PCM) ICs are high-precision A/D converters. The PCM IC is often refered to as a PCM *codec* with both encoder and decoder functions. The encoder portion of the codec performs the A/D conversion, and the decoder portion of the codec performs the D/A conversion. The digital code is usually formatted as a serial data stream for ease of interfacing to digital transmission and multiplexing systems.

PCM is a technique where an analog signal is sampled, quantized, and then encoded as a digital word. The PCM IC can include successive approximation techniques or other techniques to accomplish the PCM encoding. In addition, the PCM codec may employ nonlinear data compression techniques, such as **companding**, if it is necessary to minimize the number of bits in the output digital code. Companding is a logarithmic technique used to compress a code to fewer bits before transmission. The inverse logarithmic function is then used to expand the code to its original number of bits before converting it to the analog signal. Companding is typically used in telecommunications transmission systems to minimize data transmission rates without degrading the resolution of low-amplitude signals. Two standardized companding techniques are used extensively: A-law and μ-law. The A-law companding is used in Europe, whereas the μ-law is used predominantly in the U.S. and Japan. Linear PCM conversion is used in high-fidelity audio systems to preserve the integrity of the audio signal throughout the entire analog range.

Digital signal processing (DSP) techniques provide another type of A/D conversion ICs. Specialized A/D conversion such as *adaptive differential pulse code modulation* (ADPCM), *sigma-delta modulation, speech sub-band encoding, adaptive predictive speech encoding,* and *speech recognition* can be accomplished through the use of DSP systems. Some DSP systems require analog front ends that employ traditional PCM codec ICs or DSP interface ICs. These ICs can interface to a digital signal processor for advanced A/D applications. Some manufacturers have incorporated DSP techniques on board the single-chip A/D IC, as in the case of the DSP56ACD16 sigma-delta modulation IC by Motorola.

Integrating A/D converters are used for conversions that must take place over a long period of time, such as digital voltmeter applications or sensor applications such as thermocouples. The integrating A/D converter produces a digital code that represents the average of the signal over time. Noise is reduced by means of the signal averaging, or integration. Dual-slope integration is accomplished by a counter that advances while an input voltage charges a capacitor in a specified time interval, *T.* This is compared to another count sequence that advances while a reference voltage discharges across the same capacitor in a time interval, δ *T.* The ratio of the charging count value to the discharging count value is proportional to the ration of the input voltage to the reference voltage. Hence, the integrating converter provides a digital code that is a measure of the input voltage averaged over time. The conversion accuracy is independent of the capacitor and the clock frequency since they affect both the charging and discharging operations. The charging period, T, is selected to be the

period of the fundamental frequency to be rejected. The maximum conversion rate is slightly less than 1/(2T) conversions per second. While this limits the conversion rate to be too slow for high-speed data acquisition applications, it is appropriate for long-duration applications of slowly varying input signals.

Grounding and Bypassing on D/A and A/D ICs

D/A and A/D converter ICs require correct grounding and capacitive bypassing in order to operate according to performance specifications. The digital signals can severely impair analog signals. To combat the electromagnetic interference induced by the digital signals, the analog and digital grounds should be kept separate and should have only one common point on the circuit board. If possible, this common point should be the connection to the power supply.

Bypass capacitors are required at the power connections to the IC, the reference signal inputs, and the analog inputs to minimize noise that is induced by the digital signals. Each manufacturer specifies the recommended bypass capacitor locations and values in the data sheet. The 1-μF tantalum capacitors are commonly recommended, with additional high-frequency power supply decoupling sometimes being recommended through the use of ceramic disc shunt capacitors. The manufacturers' recommendations should be followed to ensure proper performance.

Selection Criteria for D/A and A/D Converter ICs

Hundreds of D/A and A/D converter ICs are available, with prices ranging from a few dollars to several hundred dollars each. The selection of the appropriate type of converter is based on the application requirements of the system, the performance requirements, and cost. The following issues should be considered in order to select the appropriate converter.

1. What are the input and output requirements of the system? Specify all signal current and voltage ranges, logic levels, input and output impedances, digital codes, data rates, and data formats.
2. What level of accuracy is required? Determine the resolution needed throughout the analog voltage range, the dynamic response, the degree of linearity, and the number of bits encoding.
3. What speed is required? Determine the maximum analog input frequency for sampling in an A/D system, the number of bits for encoding each analog signal, and the rate of change of input digital codes in a D/A system.
4. What is the operating environment of the system? Obtain information on the temperature range and power supply to select a converter that is accurate over the operating range.

Final selection of D/A and A/D converter ICs should be made by consulting manufacturers to obtain their technical specifications of the devices. Major manufacturers of D/A and A/D converters include Analog Devices, Burr-Brown, DATEL, Maxim, National, Phillips Components, Precision Monolithics, Signetics, Sony, Texas Instruments, Ultra Analog, and Yamaha. Information on contacting these manufacturers and others can be found in an *IC Master Catalog*.

Defining Terms

Companding: A process designed to minimize the transmission bit rate of a signal by compressing it prior to transmission and expanding it upon reception. It is a rudimentary "data compression" technique that requires minimal processing.

Delta modulation: An A/D conversion process where the digital output code represents the change, or slope, of the analog input signal, rather than the absolute value of the analog input signal. A 1 indicates a rising slope of the input signal. A 0 indicates a falling slope of the input signal. The sampling rate is dependent on the derivative of the signal, since a rapidly changing signal would require a rapid sampling rate for acceptable performance.

Fixed reference D/A converter: The analog output is proportional to a fixed (nonvarying) reference signal.

Flash A/D: The fastest A/D conversion process available to date, also referred to as parallel A/D conversion. The analog signal is simultaneously evaluated by $2^n - 1$ comparators to produce an n-bit digital code in one step. Because of the large number of comparators required, the circuitry for flash A/D converters can be very expensive. This technique is commonly used in digital video systems.

Integrating A/D: The analog input signal is integrated over time to produce a digital signal that represents the area under the curve, or the integral.

Multiplying D/A: A D/A conversion process where the output signal is the product of a digital code multiplied times an analog input reference signal. This allows the analog reference signal to be scaled by a digital code.

Nyquist A/D converters: A/D converters that sample analog signals that have a maximum frequency that is less than the Nyquist frequency. The Nyquist frequency is defined as one-half of the sampling frequency. If a signal has frequencies above the Nyquist frequency, a distortion called *aliasing* occurs. To prevent aliasing, an *antialiasing filter* with a flat passband and very sharp roll-off is required.

Oversampling converters: A/D converters that sample frequencies at a rate much higher than the Nyquist frequency. Typical oversampling rates are 32 and 64 times the sampling rate that would be required with the Nyquist converters.

Pulse code modulation (PCM): An A/D conversion process requiring three steps: the analog signal is sampled, quantized, and encoded into a fixed length digital code. This technique is used in many digital voice and audio systems. The reverse process reconstructs an analog signal from the PCM code. The operation is very similar to other A/D techniques, but specific PCM circuits are optimized for the particular voice or audio application.

Sigma-delta A/D conversion: An *oversampling* A/D conversion process where the analog signal is sampled at rates much higher (typically 64 times) than the sampling rates that would be required with a Nyquist converter. Sigma-delta modulators integrate the analog signal before performing the delta modulation. The integral of the analog signal is encoded rather than the change in the analog signal, as is the case for traditional delta modulation. A digital sample rate reduction filter (also called a digital decimation filter) is used to provide an output sampling rate at twice the Nyquist frequency of the signal. The overall result of oversampling and digital sample rate reduction is greater resolution and less distortion compared to a Nyquist converter process.

Successive approximation: An A/D conversion process that systematically evaluates the analog signal in n steps to produce an n-bit digital code. The analog signal is successively compared to determine the digital code, beginning with the determination of the most significant bit of the code.

Related Topic

15.1 Coding, Transmission, and Storage

References

Analog Devices, *Analog Devices Data Conversion Products Data Book*, Norwood, Mass.: Analog Devices, Inc., 1989.

Burr–Brown, *Burr-Brown Integrated Circuits Data Book*, Tucson, Ariz.: Burr-Brown, 1989.

DATEL, *DATEL Data Conversion Catalog*, Mansfield, Mass.: DATEL, Inc., 1988.

W. Drachler, and M. Bill, "New high-speed, low-power data-acquistion ICs," *Analog Dialogue,* vol. 29, no. 2, pp. 3–6, Norwood, Mass.: Analog Devices, Inc., 1995.

S. Garrod and R. Borns, *Digital Logic: Analysis, Application and Design*, Philadelphia, Pa.: Saunders College Publishing, 1991, Chap. 16.

J.M. Jacob, *Industrial Control Electronics*, Englewood Cliffs, N.J.: Prentice–Hall, 1989, Chap. 6.

B. Keiser and E. Strange, *Digital Telaphony and Network Integration*, 2nd ed., New York: Van Nostrand Reinhold, 1995.

Motorola, *Motorola Telecommunications Data Book*, Phoenix, Ariz.: Motorola, Inc., 1989.

National Semiconductor, *National Semiconductor Data Acquisition Linear Devices Data Book*, Santa Clara, Calif.: National Semiconductor Corp., 1989.

S. Park, *Principles of Sigma-Delta Modulation for Analog-to-Digital Converters*, Phoenix, Ariz.: Motorola, Inc., 1990.

Texas Instruments, *Texas Instruments Digital Signal Processing Applications with the TMS320 Family*, Dallas, Tex.: Texas Instruments, 1986.

Texas Instruments, 1989. *Texas Instruments Linear Circuits Data Acquisition and Conversion Data Book*, Dallas, Tex.: Texas Instruments, 1989.

Further Information

Analog Devices, Inc. has edited or published several technical handbooks to assist design engineers with their data acquisition system requirements. These references should be consulted for extensive technical information and depth. The publications include *Analog-Digital Conversion Handbook*, by the engineering staff of Analog Devices, published by Prentice–Hall, Englewood Cliffs, N.J., 1986; *Nonlinear Circuits Handbook, Transducer Interfacing Handbook,* and *Synchro and Resolver Conversion,* all published by Analog Devices Inc., Norwood, Mass.

Engineering trade journals and design publications often have articles describing recent A/D and D/A circuits and their applications. These publications include *EDN Magazine, EE Times,* and *IEEE Spectrum.* Research-related topics are covered in *IEEE Transactions on Circuits and Systems,* and also the *IEEE Transactions on Instrumentation and Measurement.*

Bar-Cohen, A. “Thermal Management of Electronics”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

33

Thermal Management of Electronics

33.1 Introduction
Motivation • Requirements
33.2 Heat Transfer Fundamentals
33.3 Chip Module Thermal Resistance
Definition • Internal Resistance • External Resistance • Total Resistance • Multichip Modules

Avram Bar-Cohen
University of Minnesota

33.1 Introduction

Motivation

In the thermal control of microelectronic components, it is necessary to provide an acceptable *microclimate* for a diversity of devices and packages, which vary widely in size, power dissipation, and sensitivity to temperature. Although the **thermal management** of all electronic components is motivated by a common set of concerns, this diversity often leads to the design and development of distinct thermal control systems for different types of electronic equipment. Moreover, due to substantial variations in the performance, cost, and environmental specifications across product categories, the thermal control of similar components may require widely differing thermal management strategies.

The prevention of **catastrophic thermal failure,** defined as an immediate, thermally induced, total loss of electronic function, must be viewed as the primary and foremost aim of electronics thermal control. Catastrophic failure may result from a significant deterioration in the performance of the component/system or from a loss of structural integrity at one of the relevant packaging levels. In early microelectronic systems, catastrophic failure was primarily *functional* and thought to result from changes in the bias voltage, *thermal runaway* produced by regenerative heating, and dopant migration, all occurring at elevated transistor junction temperatures. While these failure modes may still occur during the device development process, improved silicon simulation tools and thermally compensated integrated circuits have largely quieted these concerns and substantially broadened the operating temperature range of today's silicon-based logic and memory devices. Similar concerns do still exist in the use of CMOS devices for high-performance systems. Because of the dependence of CMOS circuit speed on temperature, it may be necessary to limit the maximum chip temperature to achieve a desired cycle time and/or to maintain timing margins in the system.

More generally, however, thermal design in the 1990s is aimed at preventing thermally induced physical failures, through reduction of the temperature rise above ambient and minimization of temperature variations within the packaging structure(s). The use of many low-temperature materials and the structural complexity of chip packages and printed circuit boards has increased the risk of catastrophic failures associated with the vaporization of organic materials, the melting of solders, and thermal-stress fracture of leads, joints, and seals, as well as the fatigue-induced delamination and fracture or creep-induced deformation of encapsulants and

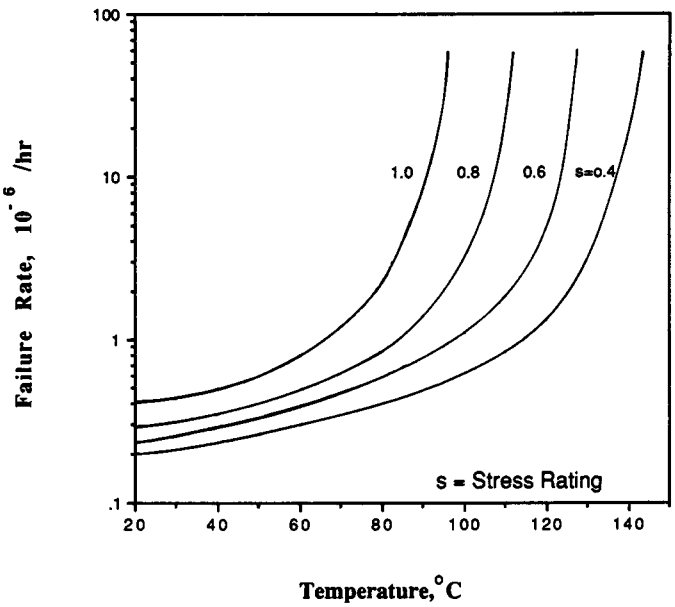


FIGURE 33.1 Exponential dependence of failure rate on component temperature.

laminates. To prevent catastrophic thermal failure, the designer must know the maximum allowable temperatures, acceptable internal temperature differences, and the power consumption/dissipation of the various components. This information can be used to select the appropriate fluid, heat transfer mode, and inlet temperature for the coolant and to thus establish the thermal control strategy early in the design process.

After the selection of an appropriate thermal control strategy, attention can be turned to meeting the desired system-level reliability and the target failure rates of each component and subassembly. Individual solid-state electronic devices are inherently reliable and can typically be expected to operate, at room temperature, for some 100,000 years, i.e., with a base failure rate of 1 FIT (failures in 10^9 h). However, since the number of devices in a typical logic component is rapidly approaching 1 million and since an electronic system may consist of many tens to several hundreds of such components, achieving a system Mean Time Between Failures of several thousand hours in military equipment and 40,000–60,000 hours in commercial electronics is a most formidable task.

Many of the failure mechanisms, which are activated by prolonged operation of electronic components, are related to the local temperature and/or temperature gradients, as well as the thermal history of the package [Pecht et al., 1992]. Device-related functional failures often exhibit a strong relationship between failure rate and operating temperature. This dependence, illustrated in Fig. 33.1, is exponential in nature and commonly represented in the form of an Arrhenius relation, with unique, empirically determined coefficients for each component type. In the normal operating range of microelectronic components, a 10–20°C increase in chip temperature is thought to double the component failure rate, and even a 1°C decrease may lower the predicted failure rate associated with such mechanisms by 2–4% [Morrison et al., 1982].

Unfortunately, it is not generally possible to characterize thermally induced structural failures, which develop as a result of differential thermal expansion among the materials constituting an electronic package, in the form of an Arrhenius relation. Although these mechanical stresses may well increase as the temperature of the component is elevated, thermal stress failures are, by their nature, dependent on the details of the local temperature fields, as well as the assembly, attachment, and local operating history of the component. Furthermore, thermal stress generation in packaging materials and structures is exacerbated by power transients, as well as by the periodically varying environmental temperatures, experienced by most electronic systems, during both qualification tests and actual operation. However, stress variations in the elastic domain or in the range below the fatigue limit may have little effect on the component failure rate. Consequently, the minimization or elimination of thermally induced failures often requires careful attention to both the temperature and stress

fields in the electronic components and necessitate the empirical validation of any proposed thermostructural design criteria.

To initiate the development of a thermal design for a specified electronic product, it is first necessary to define the relevant packaging level. The commonly accepted categorization places the chip package, which houses and protects the chip, at the bottom of the packaging hierarchy (Level 1), the printed circuit board, which provides for chip-to-chip interconnect, as Level 2, the backplane, or "motherboard," which interconnects the printed circuit boards, as Level 3, and defines the box, rack, or cabinet, which houses the entire system, as Level 4.

The primary thermal transport mechanisms and commonly used heat removal techniques vary substantially from one packaging level to the next. While Level 1 thermal packaging is primarily concerned with conducting heat from the chip to the package, at Level 2 attention must be devoted to heat spreading by conduction in the printed circuit board and convection of the heat tot he ambient air, and/or transport of the heat to the board edge. Many of today's electronic systems, as might be surmised from the frequently cited "computer-on-a-chip" or "computer-on-a-board" terminology, can be adequately packaged at Level 1 or 2. Heat sinks, or finned surfaces protruding into the air stream, are often used at Level 1 and 2 to aid in the transfer of heat into the ambient air. When Level 3 and/or 4 are present, thermal packaging generally involves the use of active thermal control measures, such as air handling systems, refrigeration systems, or water channels, heat exchangers, and pumps.

Requirements

Consideration of the Arrhenius relationship has resulted in peak allowable temperatures of 110–120°C for most military equipment [Morrison et al., 1982] and has led designers of commercial equipment to specify average chip operating temperatures in the 65–85°C range [Bar-Cohen, 1987, 1988]. Theoretical predictions of dramatic reductions in component failure rates have been used to justify the use of refrigerated avionics [Morrison, 1982] and cryogenic electronics [Jaeger, 1986; Vacca et al., 1987]. To accomodate rising power dissipation, commercial chip operating temperatures are expected to increase past 100°C in the coming decade.

The stabilization of component temperature and minimization of the temperature differences between adjacent devices, components, and various packaging levels have long been known to reduce failure rates in electronic systems [Hilbert and Kube, 1969]. In layered structures, such as chip packages and printed circuit boards, and in the joints of surface mounted components, temperature nonuniformities, in all but the most clever designs, accentuate the differences in the thermal expansion coefficients among the various materials and can frequently result in thermal stresses that threaten the integrity of these components and joints [Englemier, 1984; Suhir, 1988]. The growing integration on a single chip of functionally distinct and thermally diverse devices, as in the power-integrated chips of the late 1980s and in the microsensor, RF, and in optoelectronic chips under development today, can be expected to focus renewed attention on the minimization of transient temperature and stress fields produced by localized heat sources.

Despite the precipitous drop in transistor switching energy from more than 10^{-9} J in 1960 to nearly 10^{-13} J in devices used during the late 1980s and down to 10^{-14} J in the early 1990s, the cooling requirements of microelectronic packages have not diminished. Because of increased device densities and higher operating speeds, chip heat removal requirements have actually risen from 0.1 to 0.3 W, typical of the SSI devices used in the early 1960s, to 1 to 5 W in the LSI ECL components and VLSI CMOS devices of the mid-1980s, and to values in the range of 15 to 30 W for commercial equipment in the early 1990s. Projections of current trends suggest that by the year 2000, the thermal designer will have to contend with chip power dissipations in excess of 150 W, producing surface heat fluxes of nearly 80 W/cm^2 for the smaller chips and approximately 40 W/cm^2 for the larger (2×2 cm) chips likely to be available in that time period. It may be anticipated that, by the turn of the century, substrate heat fluxes of more than 25 W/cm^2 will be encountered in both large ($30 \times 30 \times 5$ cm) and small ($5 \times 5 \times 2$ cm) multichip modules.

The successful removal of these heat fluxes, in the presence of severe electrical, manufacturing cost, and reliability constraints, poses a formidable challenge to the packaging community. Nevertheless, it must be noted that the heat fluxes encountered in today's "cutting edge technology" chips already pose a significant challenge to the thermal packaging engineer. Chip heat fluxes in the mid-1980s typically ranged from 5 W/cm^2 to nearly

30 W/cm^2, for both single-chip packages and multichip modules [Bar-Cohen, 1987]. Recently released commercial computers often include chips dissipating 15 to 30 W/cm^2 [e.g., Kaneko et al., 1990; Pei et al., 1990], and laboratory prototypes have extended the chip heat flux range to nearly 65 W/cm^2. These heat fluxes are comparable, at the upper end, to the thermal loading experienced by reentry vehicles and even, at the lower end, to heat fluxes imposed on rocket motor cases. The anticipated peak heat flux in the year 2000 of approximately 100 W/cm^2 is in the range of thermal loadings associated with nuclear blasts.

Design Procedure

Generation of an appropriate thermal design begins with tabulation of the specified critical temperatures (source and sink) and the heat generation rate. These parameters can be used to define the target thermal resistance, as

$$R_{\text{target}} = (T_{\text{source}} - T_{\text{sink}})/q_{\text{gen}} \ [\text{K/W}]$$

The electronics thermal control literature, as well as subsequent sections of this chapter, present much of the relevant thermal packaging information in the form of thermal resistances. At its most fundamental level, the thermal packaging task involves selecting a combination of heat removal mechanisms which yield an overall thermal resistance that is not greater than the target value. The implementation of such a system will assure that the heat source, typically a microprocessor or memory chip, will operate at an acceptable temperature.

As will be discussed later in this chapter, in nearly all modes of heat transfer the geometry of the heat flow path, i.e., length and area, play an important role in determining the heat source temperature. Consequently, it is desirable to obtain the relevant geometric details (lengths, thicknesses, areas, volumes) at this early stage in the design process. Combining the geometric information with the target thermal resistance, it is often convenient to define an area-specific or volume-specific target thermal resistance in units of K/(W/cm^2) and K/(W/cm^3), respectively.

When the target thermal resistance is known, first-order (or "back-of-the-envelope") calculations are performed to evaluate the severity of the thermal management problem. Although some designs can be completed at this stage, often more precise calculations are needed to verify that the proposed approach does indeed meet the target thermal resistance value. Such calculations can be performed analytically by drawing on the wealth of knowledge in the thermal sciences, numerically using commercial general purpose software, or with commercial software specifically tailored to the thermofluid and thermostructural configurations encountered in packaging. Due to the difficulty in predicting thermal contact resistances at lightly and variably loaded mechanical interfaces and in determining the convective resistances associated with irregular package and printed circuit board geometries, some experimental data is generally needed to establish key parameters or verify system performance.

The search for an adequate thermal packaging strategy generally begins with consideration of passive transfer modes—conduction, natural convection, and radiation, which require no external motive power and no moving parts. Due to its near-universal availability, air is the most common and preferred cooling fluid. Many electronic systems can be successfully cooled by passive means and especially by natural convection in air. When such passive means are incapable of properly controlling the heat source temperature, the designer must examine the use of active thermal control techniques, including blown air, pumped water, and circulated refrigerants. Immersion of the electronic components directly in dielectric liquids, which can then be pumped or allowed to circulate naturally, provides an additional, though less common, alternative.

33.2 Heat Transfer Fundamentals

To determine the temperature differences encountered in the flow of heat within electronic systems, it is necessary to recognize several different heat transfer mechanisms and their governing relations. In a typical system, heat removal from the active regions of the chip(s) requires the use of several mechanisms, some operating in series and others in parallel, to transport the generated heat to the coolant.

Thermal transport through solids is governed by the Fourier equation, which, in one-dimensional form, is expressible as

$$q = kAdT/dx \ [\text{W}] \tag{33.1}$$

TABLE 33.1 Thermal Conductivities of Typical Packaging Materials at Room Temperature

Materials	Thermal Conductivity (W/m K)
Air	0.024
Mylar	0.19
Silicone rubber	0.19
Solder mask	0.21
Epoxy (dielectric)	0.23
Ablefilm 550 dielectric	0.24
Nylon	0.24
Polytetrafluorethylene	0.24
RTV	0.31
Polyimide	0.33
Epoxy (conductive)	0.35
Water	0.59
Mica	0.71
Ablefilm 550 K	0.78
Thermal greases/pastes	1.10
Borosilicate glass	1.67
Glass epoxy	1.70
Stainless steel	15
Kovar	16.60
Solder (Pb-In)	22
Alumina	25
Solder 80-20 Au-Sn	52
Silicon	118
Molybdenum	138
Aluminum	156
Beryllia	242
Gold	298
Copper	395
Silver	419
Diamond	2000

where q is the heat flow, k is the thermal conductivity of the medium, A is the cross-sectional area for heat flow, and dT/dx the temperature gradient.

The temperature difference resulting from the conduction of heat is thus related to the thermal conductivity of the material, the cross-sectional area, and the path length, Δx, or

$$(T_1 - T_2)_{\text{cond}} = q(\Delta x/kA) \ [\text{K}] \tag{33.2}$$

The form of this equation suggests that, by analogy to electrical current flow in a conductor, it is possible to define a conduction thermal resistance as [Kraus, 1958]

$$R_{\text{cond}} = (T_1 - T_2)/q = \Delta x/kA \ [\text{K/W}] \tag{33.3}$$

Using the thermal conductivities tabulated in Table 33.1, conduction resistance values for packaging materials with typical dimensions can be found by use of Eq. 33.3 or by inspection of Fig. 33.2. Values are seen to range from 2K/W for a 100 mm^2 by 1 mm thick layer of epoxy encapsulant to 0.0006 K/W for a 100 mm^2 by 25 micron (1 mil) thick layer of copper. Similarly, the values of the conduction resistance for typical "soft" bonding materials are found to lie in the range of 0.1 K/W for solder and 1–3K/W for epoxies and thermal pastes, for Δx/A ratios of 0.25 to 1 m^{-1}.

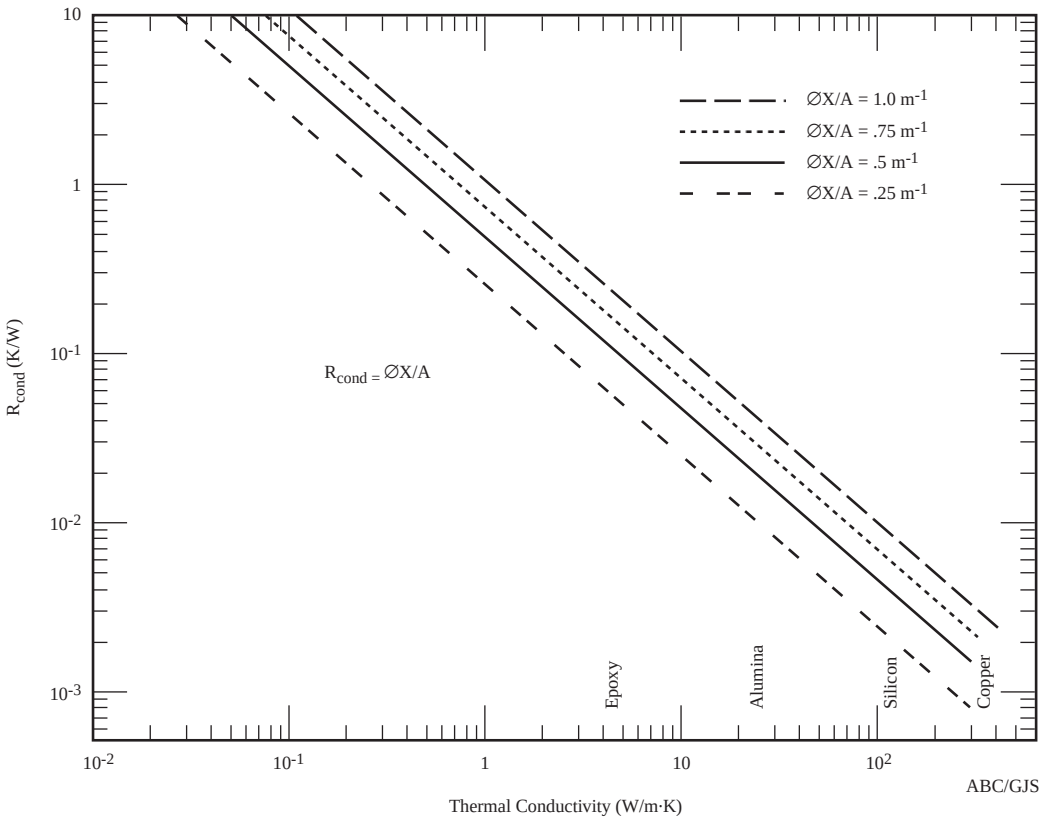


FIGURE 33.2 Conductive thermal resistances for packaging materials.

Thermal transport from a surface to a fluid in motion is called **convective heat transfer** and can be related to the **heat transfer coefficient**, h, the surface-to-fluid temperature difference, and the "wetted" area, in the form

$$q = hA(T_{surf} - T_{fluid}) \text{ [W]} \tag{33.4}$$

The differences among convection to a fast-moving fluid, a slowly flowing fluid, and a stagnant fluid, as well as variations in the convective heat transfer rate among various fluids, are reflected in the value of h. Some theoretical and many empirical correlations are available for determining this convective heat transfer coefficient (e.g., Kraus and Bar-Cohen, 1983). Using Eq. (33.4), it is possible to define the convective **thermal resistance,** as

$$R_{conv} = (hA)^{-1} \text{ [K/W]} \tag{33.5}$$

Values of this convective resistance, for a variety of coolants and heat transfer mechanisms, are shown in Fig. 33.3 for a typical heat source area of 10 cm^2 and a velocity range of 2–8 m/s. These resistances are seen to vary from 100 K/W for natural convection in air to 33 K/W for forced convection in air, to 1 K/W in fluorocarbon liquid in forced convection to less than 0.5 K/W for boiling in fluorocarbon liquids.

Unlike conduction and convection, **radiative heat transfer** between two surfaces or between a surface and its surroundings is not linearly dependent on the temperature difference and is expressed instead as

$$q = \sigma AF(T_1^4 - T_2^4) \text{ [W]} \tag{33.6}$$

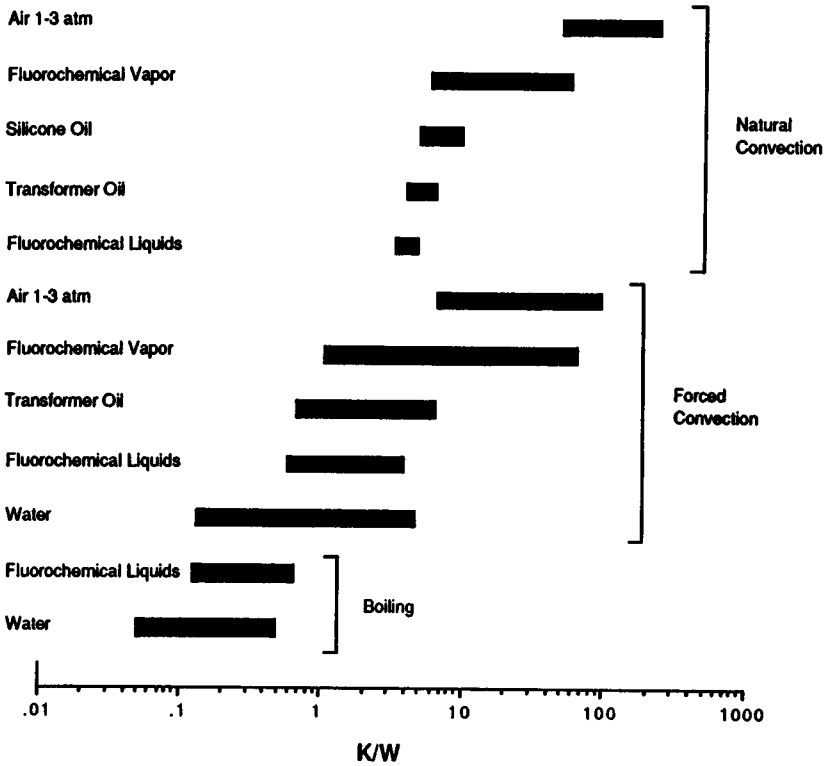


FIGURE 33.3 External thermal resistances for various fluids and cooling modes.

where F includes the effect of surface properties and geometry and σ is the Stefan-Boltzmann constant, which equals 5.67×10^{-8} W/m^{2}K^4. For an ideal, or black, radiating surface in a perfectly absorbing environment, F equals unity.

For modest temperature differences, this equation can be linearized to the form

$$q_r = h_r A(T_1 - T_2) \text{ [W]} \tag{33.7}$$

where h_r is the effective "radiation" heat transfer coefficient and is approximately equal to $4\sigma F (T_1 T_2)^{1.5}$. It is of interest to note that for temperature differences on the order of 10 K, the radiative heat transfer coefficient, h_r, for a radiationally ideal surface in an absorbing environment, is approximately equal to the heat transfer coefficient in natural convection of air. Noting the form of Eq. (33.7), the radiational thermal resistance, analogous to the convective resistance, is seen to equal $(h_r A)^{-1}$.

Ebullient thermal transport displays a complex dependence on the temperature difference between the heated surface and the saturation temperature (boiling point) of the liquid. In nucleate boiling, the primary region of interest, the **ebullient heat transfer** rate can be approximated by a relation of the form

$$q_b = C'_{sf} A(T_{\text{surf}} - T_{\text{sat}})^3 \text{ [W]} \tag{33.8}$$

where C'_{sf} is a function of the surface/fluid combination and T_{sat} is the boiling point of the liquid. For comparison purposes, it is possible to define a boiling heat transfer coefficient, h_b, equal to $C'_{sf}(T_1 - T_{\text{sat}})^2$, which, however, will vary strongly with surface temperature.

In the thermal design of electronic equipment, frequent use is made of heat sinks, involving finned or *extended* surfaces (Kraus and Bar-Cohen, 1995). While such finning can substantially increase the surface area in contact with the coolant, conduction in the **thermal fin** reduces the average temperature of the exposed surface relative

to the fin base. In the analysis of such finned surfaces, it is thus common to define a **fin efficiency**, η, equal to the ratio of the average temperature rise of the fin (above the coolant) to the temperature rise of the fin base. Using this approach, heat transfer by a fin or fin structure can be expressed in the form

$$q_f = hA\eta(T_o - T_f) \text{ [W]} \tag{33.9}$$

where T_o is the temperature of the fin base. The thermal resistance of a finned surface is given by $(\eta hA)^{-1}$ and for a properly designed surface, the fin efficiency can be expected to lie between 0.5 and 0.8.

The transfer of heat to a flowing gas or liquid, not undergoing a phase change, results in an increase in the coolant temperature, according to

$$q = \dot{m}c_p(T_{\text{out}} - T_{\text{in}}) = \rho\Theta c_p(T_{\text{out}} - T_{\text{in}}) \text{ [W]} \tag{33.10}$$

where $\dot{m}$ is the mass flow rate of the coolant, ρ is the density, and Θ is the volumetric flow rate. Based on this relation, it is possible to define an effective flow resistance, R_f, as

$$R_f = (\dot{m}c_p)^{-1} \text{ [K/W]} \tag{33.11}$$

In a first-order thermal model, it is generally appropriate to relate the heat source temperature to the average (rather than the outlet) coolant temperature. In such calculations the flow resistance should be taken to equal one-half of the value given by Eq. (33.11). These average flow resistances for the three common coolants in electronics thermal management, i.e., air, water, and FC-72 (3M Trade Name), are shown in Figure 33.4.

The expression of the governing heat transfer relations in the form of thermal resistances greatly simplifies the first-order thermal analysis of electronic systems. Following the established rules for resistance networks, thermal resistances that occur sequentially along a thermal path can be simply summed to establish the overall thermal resistance for that path. Similarly, the reciprocal of the effective overall resistance of several parallel heat transfer paths can be found by summing the reciprocals of the individual resistances. In refining the thermal design of an electronic system, prime attention should, then, be devoted to reducing the largest resistances along a specified thermal path and/or providing parallel paths for heat removal from a critical area.

While the thermal resistances associated with various paths and thermal transport mechanisms constitute the building blocks in performing a detailed thermal analysis, they have also found widespread application as figures-of-merit in evaluating and comparing the thermal efficacy of various packaging techniques and thermal management strategies. The determination of the relevant thermal resistances is, thus, the key task in the thermal design of an electronic system.

33.3 Chip Module Thermal Resistance

Definition

The thermal performance of chip packaging techniques is commonly compared on the basis of the overall (junction-to-coolant) thermal resistance, R_T. This packaging figure-of-merit is generally defined in a purely empirical fashion to equal

$$R_T = (T_j - T_f)/q_c \text{ [K/W]} \tag{33.12}$$

where T_j and T_f are the junction and coolant (fluid) temperatures, respectively, and q_c is the chip heat dissipation.

Unfortunately, however, most measurement techniques are incapable of detecting the actual junction temperature, i.e., the temperature of the small volume at the interface of *p*-type and *n*-type semiconductors, and, hence, this term generally refers to the average temperature or a representative temperature on the chip. Because

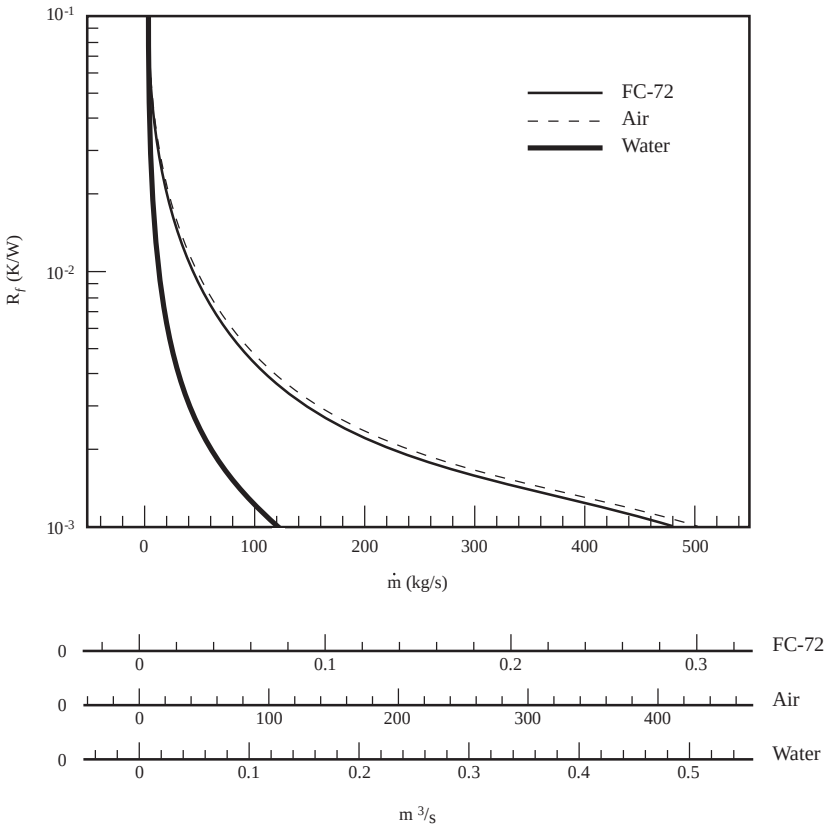


FIGURE 33.4 Flow thermal resistances for typical electronic coolants.

many of the failure mechanisms of integrated circuits are accelerated by an increase in the average chip temperature, low thermal resistances are to be preferred in nearly all categories of electronic packaging.

Single-chip packages can be characterized by their internal, or so-called junction-to-case, resistance. The convective heat removal techniques applied to the external surfaces of the package, including the effect of finned heat sinks and other thermal enhancements, can be compared on the basis of the external thermal resistance. The complexity of heat flow and coolant flow paths in a multichip module generally requires that the thermal capability of these packaging configurations be examined on the basis of overall, or chip-to-coolant, thermal resistance.

Examination of various packaging techniques reveals that the junction-to-coolant thermal resistance is, in fact, composed of an internal, largely conductive, resistance and an external, primarily convective, resistance. As shown in Fig. 33.5, the internal resistance, R_{jc}, is encountered in the flow of dissipated heat from the active chip surface, through the materials used to support and bond the chip, and on to the case of the integrated circuit package. The flow of heat from the case directly to the coolant, or indirectly through a fin structure and then to the coolant, must overcome the external resistance, R_{ex}.

Internal Resistance

As previously noted, conductive thermal transport is governed by the Fourier equation (Eq. 33.1). For composite, rectilinear structures, as encountered in many chip modules, the Fourier equation (with temperature and time invariant properties), takes the form

$$q = (T_i - T_e) \Big/ \sum_p (\Delta x/kA) \ [\mathrm{W}] \tag{33.13}$$

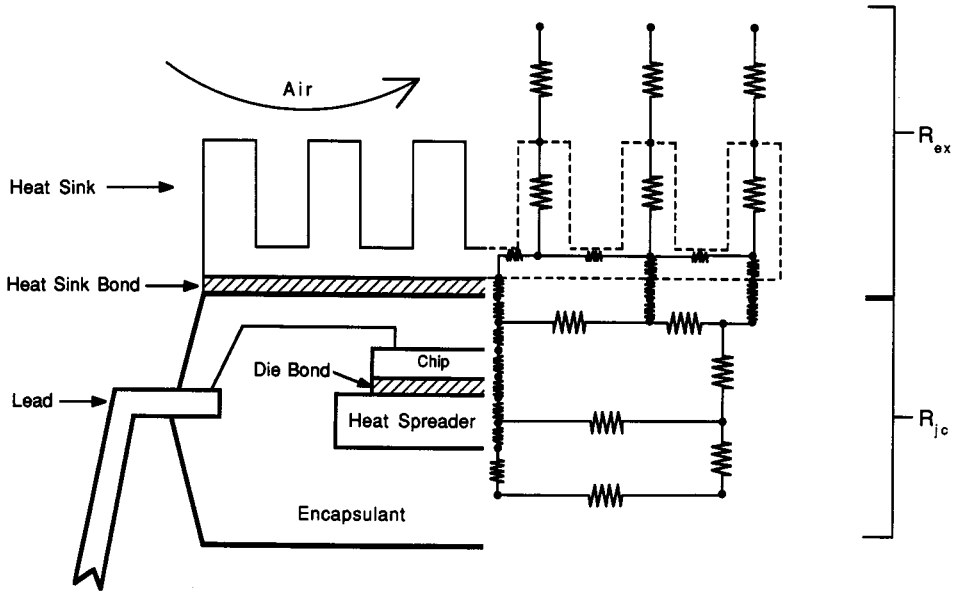


FIGURE 33.5 Thermal resistances in a single-chip package.

where T_i and T_e are the temperatures internal and external to the composite structure, respectively, Δx is the thickness of each material in the direction of heat flow, and the summation sign pertains to p distinct layers of material. The thermal conductivities of typical packaging materials are tabulated in Table 33.1.

Assuming that power is dissipated uniformly across the chip surface and that heat flow is largely one-dimensional, Eq. (33.13) can be used to provide a first-order approximation for the internal chip module resistance, as

$$R_{jc} = (T_j - T_c)/q_c = \sum_p (\Delta x/kA) \ [\text{K/W}] \tag{33.14}$$

where the summed terms represent the thermal resistances of the individual layers of silicon, solder, copper, alumina, etc. It is to be noted that the contact resistances that occur at the interfaces between pairs of materials can be added, as appropriate, to this summation. As may be seen in Fig. 32.2, decreasing the thickness of each layer and/or increasing the thermal conductivity and cross-sectional area, reduce the resistance of the individual layers.

In chip packages that provide for lateral spreading of the heat generated in the chips, the increasing cross-sectional area for heat flow at successive layers reduces the internal thermal resistance. Unfortunately, however, there is an additional resistance associated with the lateral flow of the heat, which must be taken into account in determining the chip-to-case temperature difference.

Following Yovanovich and Antonetti [1988], the spreading resistance for a small heat source on a thick substrate (typically 3–5 times thicker than the square root of the heat source area) can be expressed as

$$R_c = (0.475 - 0.62\varepsilon + 0.13\varepsilon^3)/k\,(A_c)^{0.5} \ [\text{K/W}] \tag{33.15}$$

where ε is the square root of the ratio of the heat source area to the substrate area, k the thermal conductivity of the substrate, and A_c the area of the heat source. For relatively thin layers on thicker substrates, Eq. (33.15) cannot provide an acceptable prediction of R_c. Instead, use can be made of the numerical results plotted in Fig. 33.6 to obtain the requisite value of the spreading resistance.

The internal thermal resistance of a chip package can be expected to vary from approximately 50 K/W for a plastic package with no heat spreader, to 10 to 15 K/W for a plastic package with heat spreader, and to 3 to

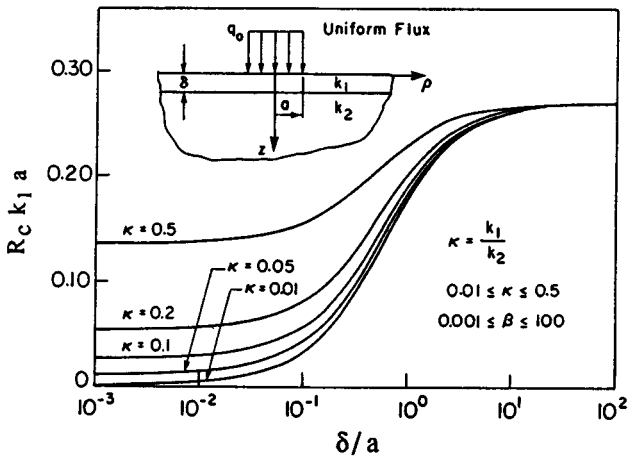


FIGURE 33.6 The thermal resistance for a circular heat source on a two-layer substrate. (*Source:* M.M. Yovanovich and V.W. Antonetti, "Application of Thermal Contact Resistance Theory toElectronic Packages," in *Advances in Thermal Modeling of Electronic Components and Systems,* vol. 1, A. Bar-Cohen and A.D. Kraus, Eds., New York: Hemisphere, 1988, pp. 79–128. With permission.)

8 K/W for a ceramic package or a specially designed plastic chip package. Carefully designed chip packages can attain even lower values of R_{jc}, but the conductive thermal resistances at the interfaces between materials, and especially along the chip surfaces where the heat fluxes are highest, can typically reach 1 K/W for epoxied interfaces and impose a lower bound on the internal package resistance.

External Resistance

To precisely determine the resistance to thermal transport from a surface to a fluid in motion, i.e., the convective resistance, it is necessary to quantify the heat transfer coefficient, *h*. For a particular geometry and flow regime, *h* may be found from available empirical correlations and/or theoretical relations. For flow along plates and in the inlet zones of parallel-plate channels, as may well be encountered in electronic cooling applications, the low velocity, or laminar flow, average convective heat transfer coefficient is given by [Kraus and Bar-Cohen, 1983]

$$h = 0.664(k/l)(\mathrm{Re})^{0.5}(\mathrm{Pr})^{0.333}\ [\mathrm{W/m^2K}] \qquad (33.16)$$

$$\text{for } \mathrm{Re} < 2 \times 10^5$$

where *k* is the fluid thermal conductivity, *l* the length (in the flow direction) of the surface, Re the **Reynolds number** (equal to the product of velocity, density, and length divided by the fluid viscosity), and Pr the **Prandtl number** (equal to the product of specific heat and viscosity divided by the thermal conductivity of the fluid).

Inserting the various parameters associated with the Re and Pr in Eq. (33.16), the laminar heat transfer coefficient is found to be directly proportional to the square root of fluid velocity and inversely proportional to the square root of the length. Furthermore, increases in the thermal conductivity of the fluid and in the Pr, as are encountered in replacing air with a liquid coolant, can be expected to result in higher heat transfer coefficients. In studies of low-velocity convective air cooling of simulated integrated circuit packages, *h* has been found to depend somewhat more strongly on Re than suggested in Eq. (33.16), and to display an Re exponent of 0.54-0.72 [Buller and Kilburn, 1981; Sparrow et al., 1982; Wirtz and Dykshoorn, 1984].

In higher velocity turbulent flow, the dependence of the convective heat transfer coefficient on the Re increases and is typically given by [Kraus and Bar-Cohen, 1983]:

$$h = 0.036(k/l)\ (\mathrm{Re})^{0.8}\ (\mathrm{Pr})^{0.333},\ \mathrm{W/m^2K} \qquad (33.17)$$

$$\text{for } \mathrm{Re} > 3 \times 10^5$$

In this flow regime, the convective heat transfer coefficient is, thus, found to vary directly with the velocity to the 0.8 power and inversely with the characteristic dimension to the 0.2 power. The dependence on fluid conductivity and Pr remains unchanged.

Applying Eq. (33.16) or (33.17) to the transfer of heat from the case of a chip module to the coolant, the external resistance, $R_{ex} = 1/hA$, is found to be inversely proportional to the wetted area and to the coolant velocity to the 0.5 to 0.8 power and directly proportional to the length in the flow direction to the 0.5 to 0.2 power. It may, thus, be observed that the external resistance can be strongly influenced by the fluid velocity and package dimensions and that these factors must be addressed in any meaningful evaluation of the external thermal resistances offered by various packaging technologies.

As previously noted, values of the external resistance, for a variety of coolants and heat transfer mechanisms, are shown in Fig. 33.3 for a typical component wetted area of 10 cm^2 and a velocity range of 2 m/s to 8 m/s. Clearly, larger chip packages will experience proportionately lower external resistances than the tabulated values, and conduction of heat through the leads and package base into the printed circuit board or substrate will serve to further reduce the effective thermal resistance.

When the direct cooling of the package surface is inadequate to maintain the desired chip temperature, it is common to attach finned heat sinks, or compact heat exchangers, to the chip package. These heat sinks can considerably increase the wetted area but may act to reduce the convective heat transfer coefficient and most definitely introduce additional conductive resistances in the adhesive used to bond the heat sink to the package and in the body of the heat sink. Typical air-cooled heat sinks for a single chip package with a base area of 3.3 × 3.3 cm on one side can reduce the external resistance to approximately 15 K/W in natural convection and as low as 5 K/W for moderate forced convection velocities.

When a heat sink or compact heat exchanger is attached to the package, the external resistance can be modified to account for the bond-layer conduction and fin efficiency as

$$R_{ex} = (T_c - T_f)/q_c = (\Delta x/kA)_b + (\eta hA)^{-1} \tag{33.18}$$

In properly designed fin structure, η can be expected to fall in the range of 0.5 to 0.8 [Kraus and Bar-Cohen, 1983]. Relatively thick fins in a low velocity flow of gas are likely to yield fin efficiencies approaching unity. This same unity value would be appropriate, as well, for an unfinned surface and, thus, serve to generalize the use of Eq. (33.18) to all package configurations.

Total Resistance

Based on the accuracy of the assumptions used in the preceding development, the overall chip module resistance, relating the chip temperature to the inlet temperature of the coolant, can be found by summing the internal, external, and flow resistances to yield:

$$R_T = R_{jc} + R_{ex} + R_f = \sum_p [R_c + (\Delta x/kA)] + (\eta hA)^{-1} + \dot{m}c_p/2 \tag{33.19}$$

In evaluating the thermal resistance by this relation, care must be taken to determine the effective cross-sectional area for heat flow at each layer in the module. For single-chip modules, the requisite areas can be readily obtained, although care must be taken to consider possible voidage in solder and adhesive layers.

As previously noted in the development of the relations for external and internal resistances, Eq. (33.19) shows R_T to be a strong function of the convective heat transfer coefficient, the flowing heat capacity of the coolant, and geometric parameters (thickness and cross-sectional area of each layer). Thus, the introduction of a superior coolant, use of thermal enhancement techniques that increase the local heat transfer coefficient, and selection of a heat transfer mode with inherently high heat transfer coefficients (e.g., boiling) will all be reflected in appropriately lower external and total thermal resistances. Similarly, improvements in the thermal conductivity of and reduction in the thickness of the relatively low conductivity bonding materials (e.g., soft solder, epoxy, silicone) would act to reduce the internal and total thermal resistances.

Frequently, however, even more dramatic reductions in the total resistance can be achieved simply by increasing the cross-sectional area for heat flow, within the chip module (e.g., chip, substrate, heat spreader) as well as along the wetted exterior surface. The implementation of such a scale-up generally results in a larger module footprint and/or lower volumetric packaging density, both of which are highly undesirable, and yet is rewarded with a lower thermal resistance. In evaluating packaging approaches, it must, therefore, be understood that the thermal resistance is a somewhat flawed figure-of-merit and that a better reflection of the efficacy of a thermal management technique can be obtained by normalizing R_T with respect to the packaging density, using the number of chips on a substrate or number of chips/packages on a printed circuit board.

Multichip Modules

The thermostructural complexity of the multichip modules in current use hampers effective thermal characterization and introduces significant uncertainty in any attempt to compare the thermal performance of these packaging configurations. Variations in heat generation patterns across the active chips (e.g., devices versus drivers), as well as nonuniformities in heat dissipation among the chips assembled in a single module, further complicate this task. To establish a common, though approximate, basis for comparison of multichip modules, it is possible to neglect these variations and consider that the heat generated by each chip flows through a unit cell of the module structure to the external coolant [Bar-Cohen, 1987, 1988]. For a given structure, increasing the area of the unit cell allows heat to spread from the chip to a larger cross section, reducing the heat flux at some of the thermally critical interfaces and at the convectively cooled surfaces. Consequently, the thermal performance of a multichip module can be best represented by the area-specific thermal resistance, i.e., the temperature difference between the chip and the coolant divided by the substrate heat flux. This figure-of-merit is equivalent to the inverse of the overall heat transfer coefficient, U, commonly used in the compact heat exchanger literature. Despite significant variation in design and fabrication, all the late-1980s water-cooled modules and one air-cooled module provide a specific thermal resistance of approximately 20°C for every watt per square centimeter at the substrate. In cutting-edge multichip modules in use in the 1990s, this value was reduced to 5–10 K/(W/cm^2).

Nomenclature

A	area, m^2
c_p	specific heat, W/kgK
C'_{sf}	boiling surface parameter, W/m^{2}K^3
F	radiational factor
h	heat transfer coefficient, W/m^{2}K
k	thermal conductivity, W/mK
l	path length, m
m	mass flow, kg/s
Pr	Prandtl Number, $\equiv c_p\mu/k$
q	heat flow, W
R	thermal resistance, K/W
Re	Reynolds Number, $\equiv \rho vl/\mu$
T	temperature, K
v	velocity, m/s
x	length, m
σ	Stefan-Boltzmann constant, $5.67 \times 10{-}8$ W/m^{2}K^4
η	fin efficiency
ρ	density, kg/m^3
$\mathcal{Q}$	volumetric flow rate, m^3/s
ε	ratio of heater to substrate size
μ	viscosity, kg/ms

Subscripts

b	bond layer
b	boiling
c	constriction or spreading
cond	conduction
conv	convection
f, fluid	fluid
i	internal
in	inlet
j	junction
jc	junction to case
o	base, external
out	outlet
r	radiation
sat	thermodynamic saturation
surf	surface
T	total

Defining Terms

Catastrophic thermal failure: An immediate, thermally induced total loss of electronic function by a component or system.

Conductive heat transfer: The process by which heat diffuses through a solid or stationary fluid.

Convective heat transfer: The process by which a moving fluid transfers heat to or from a wetted surface.

Ebullient heat transfer: The heat transfer process associated with the formation and release of vapor bubbles on a heated surface.

Fin efficiency: A thermal characteristic of an extended surface that relates the heat transfer ability of the additional area to that of the base area.

Heat transfer coefficient: A characteristic parameter of convective heat transfer that determines the heat flux that can be transported from a wetted surface with a specified temperature difference.

Prandtl number: A nondimensional characteristic of fluids, relating the rate of momentum diffusion to heat diffusion.

Radiative heat transfer: The process by which long-wave electromagnetic radiation transports heat from a surface to its surroundings.

Reynolds number: A nondimensional parameter used to determine the transition to turbulence in a fluid flowing in pipes or past surfaces.

Thermal fin: An extension of the surface area in contact with a heat transfer fluid, usually in the form of a cylinder or rectangular prism protruding from the base surface.

Thermal management or control: The process or processes by which the temperature of a specified component or system is maintained at the desired level.

Thermal resistance: A thermal characteristic of a heat flow path, establishing the temperature drop required to transport heat across the specified segment or surface; analogous to electrical resistance.

Related Topics

1.1 Resistors • 30.2 Power Conversion

References

A. Bar-Cohen, "Thermal management of air- and liquid-cooled multichip modules," *IEEE CHMT Transactions,* vol. CHMT-10, no. 2, pp. 159–175, 1987.

A. Bar-Cohen, "Thermal Design and Control," in *Physical Architecture of VLSI Systems,* R.J. Hannemann, A.D. Kraus, and M. Pecht, Eds., New York: John Wiley & Sons, 1994, Chapter 9, pp. 541–605.

A. Bar-Cohen, "Addendum and correction to thermal management of air- and liquid-cooled multichip modules," *IEEE CHMT Transactions,* vol. CHMT-11, no. 3, pp. 333–334, 1988.

M.L. Buller and R.F. Kilburn, "Evaluation of surface heat transfer coefficients for electronic module packages," *Heat Transfer in Electronic Equipment,* vol. HTD-20, ASME, New York, 1981.

W. Englemier, "Functional cycles and surface mounting attachment reliability," in *Thermal Management Concepts in Microelectronic Packaging*, R.T. Howard, et al., Eds., ISHM Technical Monograph Series, 6984-003, ISHM, New York, 1984, pp. 83–109.

C.A. Harper, Ed., *Electronic Packaging and Interconnection Handbook,* New York: McGraw-Hill, 1991, p. 2.7.

W.F. Hilbert and F.H. Kube, "Effects on electronic equipment reliability of temperature cycling in equipment," Final Report, Grumman Aircraft Engineering Co., Report No. EC-69-400, Bethpage, N.Y., 1969.

R.C. Jaeger, "Development of low temperature CMOS for high performance computer systems," *IEEE International Conference on Computer Design: VLSI in Computers*, 1986, pp. 128–130.

A. Kaneko, K. Seyama, and M. Suzuki, "LSI packaging and cooling technologies for Fujitsu VP-2000 Series," *Fujitsu*, vol. 41, no. 1, pp. 12–19, 1990.

A.D. Kraus, "The use of steady state electrical network analysis in solving heat flow problems," 2nd National Heat Transfer Conference, Chicago, Ill., 1958.

A.D. Kraus and A. Bar-Cohen, *Thermal Analysis and Control of Electronic Equipment,* New York: Hemisphere Publishing Corporation, 1983.

A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks,* New York: John Wiley & Sons, 1995.

G.N. Morrison, J.M. Kallis, L.A. Strattan, I.R. Jones, and A.L. Lena, "RADC thermal guide for reliability engineers," Report Number RADC-TR-82-172, Rome Air Development Center, Air Force Systems Command, Griffis Air Force Base, New York, 1982.

R.A. Morrison, "Improved avionics reliability through phase change conductive cooling," *Proceedings, IEEE National Telesystems Conference,* pp. B5.6.1–B5.6.5, 1982.

W. Nakayama, "Thermal management of electronic equipment: A review of technology and research topics," in *Advances in Thermal Modeling of Electronic Components and Systems,* vol. 1, A. Bar-Cohen and A.D. Kraus, Eds., New York: Hemisphere Publishing Corporation, 1988, pp. 1–78.

M. Pecht, P. Lall, and E. Hakim, "The influence of temperature on integrated circuit failure mechanisms," *Advances in Thermal Modeling of Electronic Components and Systems,* vol. 3, A. Bar-Cohen and A.D. Kraus, Eds., New York: ASME Press, 1992.

J. Pei, S. Heng, R. Charlantini, and P. Gildea, "Cooling components used in the Vax 9000 family of computers," *Proceedings, 1990 International Electronic Packaging Society Conference*, 1990, pp. 587–601.

E.M. Sparrow, J.E. Niethhammer, and A. Chaboki, "Heat transfer and pressure drop characteristics of arrays of rectangular modules encountered in electronic equipment," *Int. J. Heat & Mass Transfer,* vol. 25, no. 7, pp. 961–973, 1982.

E. Suhir, "Thermal stress in electronic components," in *Advances in Thermal Modeling of Electronic Components and Systems,* vol. 1, A. Bar-Cohen and A.D. Kraus, Eds., New York: Hemisphere Publishing Corporation, 1988, pp. 337–412.

R.R. Tummala and E.J. Rymaszewski, *Microelectronics Packaging Handbook,* New York: Van Nostrand Reinhold, 1989, p. 174.

A. Vacca, D. Resnick, D. Frankel, R. Bach, J. Kreilich, and D. Carlson, "A cryogenically cooled CMOS VLSI supercomputer," *VLSI Systems Design,* pp. 80–88, 1987.

R.A. Wirtz and P. Dykshoorn, "Heat transfer from arrays of flat packs in channel flow," *Proceedings, 4th International Electronic Packaging Society Conference*, New York, 1984, pp. 318–326.

M.M. Yovanovich and V.W. Antonetti, "Application of thermal contact resistance theory to electronic packages," in *Advances in Thermal Modeling of Electronic Components and Systems,* vol. 1, A. Bar-Cohen and A.D. Kraus, Eds., New York: Hemisphere Publishing Corporation, 1988, pp. 79–128.

Dewey, A. “Digital and Analog Electronic Design Automation”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

34

Digital and Analog Electronic Design Automation

34.1 Introduction
34.2 Design Entry
34.3 Synthesis
34.4 Verification
Timing Analysis • Simulation • Analog Simulation • Emulation
34.5 Physical Design
34.6 Test
Fault Modeling • Fault Testing
34.7 Summary

Allen Dewey
Duke University

34.1 Introduction

The field of **design automation** (DA) technology, also commonly called *computer-aided design* (CAD) or *computer-aided engineering* (CAE), involves developing computer programs to conduct portions of product design and manufacturing on behalf of the designer. Competitive pressures to produce more efficiently new generations of products having improved function and performance are motivating the growing importance of DA. The increasing complexities of microelectronic technology, shown in Fig. 34.1, illustrate the importance of relegating portions of product development to computer automation [Barbe, 1980].

Advances in microelectronic technology enable over 1 million devices to be manufactured on an **integrated circuit** that is smaller than a postage stamp; yet the ability to exploit this capability remains a challenge. Manual design techniques are unable to keep pace with product design cycle demands and are being replaced by automated design techniques [Saprio, 1986; Dillinger, 1988].

Figure 34.2 summarizes the historical development of DA technology. DA computer programs are often simply called *applications* or *tools*. DA efforts started in the early 1960s as academic research projects and captive industrial programs; these efforts focused on tools for physical and logical design. Follow-on developments extended logic simulation to more-detailed *circuit* and *device* simulation and more-abstract *functional* simulation. Starting in the mid to late 1970s, new areas of test and synthesis emerged and vendors started offering commercial DA products. Today, the electronic design automation (EDA) industry is an international business with a well-established and expanding technical base [Trimberger, 1990]. EDA will be examined by presenting an overview of the following areas:

- Design entry,
- Synthesis,
- Verification,
- Physical design, and
- Test.

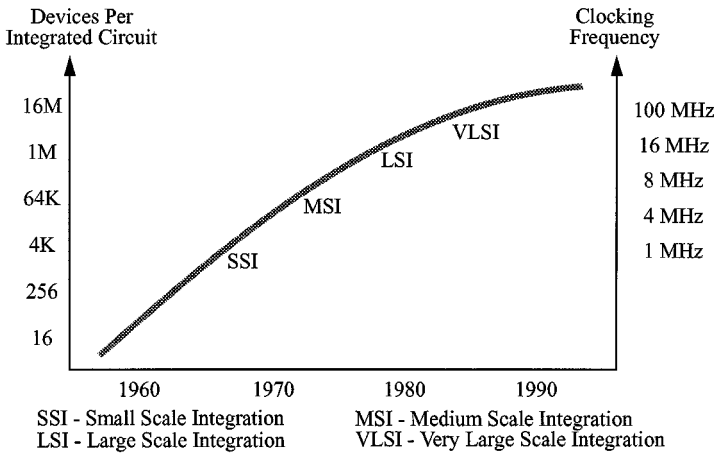


FIGURE 34.1 Microelectronic technology complexity.

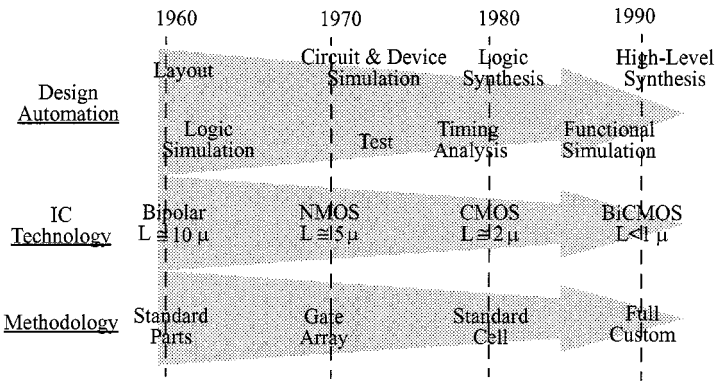


FIGURE 34.2 DA technology development.

34.2 Design Entry

Design entry, also called *design capture*, is the process of communicating with a DA system. In short, design entry is how an engineer "talks" to a DA application and/or system.

Any sort of communication is composed of two elements: language and mechanism. Language provides common semantics; mechanism provides a means by which to convey the common semantics. For example, people communicate via a language, such as English or German, and a mechanism, such as a telephone or electronic mail. For design, a digital system can be described in many ways, involving different perspectives or *abstractions*. An abstraction defines at a particular level of detail the behavior or semantics of a digital system, i.e., how the outputs respond to the inputs. Fig. 34.3 illustrates several popular levels of abstractions. Moving from the lower left to the upper right, the level of abstraction generally increases, meaning that physical models are the most detailed and specification models are the least detailed. The trend toward higher levels of design entry abstraction supports the need to address greater levels of complexity [Peterson, 1981].

The physical level of abstraction involves geometric information that defines electrical devices and their interconnection. Geometric information includes the shape of objects and how objects are placed relative to each other. For example, Fig. 34.4 shows the geometric shapes defining a simple complementary metal-oxide semiconductor (**CMOS**) inverter. The shapes denote different materials, such as aluminum and polysilicon, and connections, called *contacts* or **vias**.

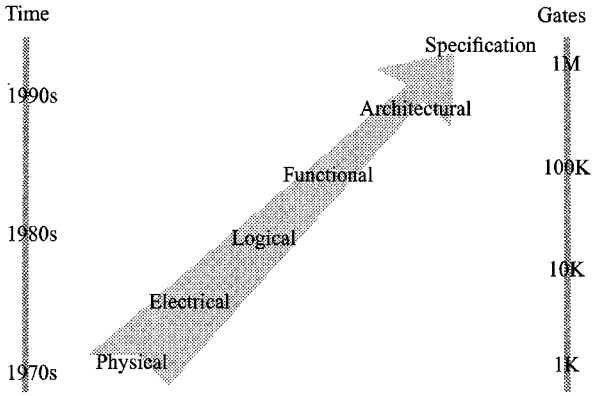


FIGURE 34.3 DA abstractions.

Design entry mechanisms for physical information involve textual and graphical techniques. With textual techniques, geometric shape and placement are described via an artwork description language, such as Caltech Intermediate Form (CIF) or Electronic Design Intermediate Form (EDIF). With graphical techniques, geometric shape and placement are described by rendering the objects on a display terminal.

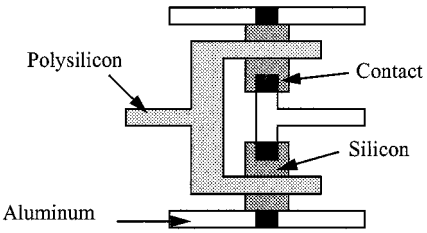


FIGURE 34.4 Physical abstraction.

The electrical level abstracts physical information into corresponding electrical devices, such as **capacitors**, **transistors**, and **resistors**. Electrical information includes device behavior in terms of terminal current and voltage relationships. Device behavior may also be defined in terms of manufacturing parameters. Fig. 34.5 shows the electrical symbols denoting a CMOS inverter.

The logical level abstracts electrical information into corresponding logical elements, such as **and** gates, **or** gates, and inverters. Logical information includes truth table and/or characteristic-switching algebra equations and active-level designations. Fig. 34.6 shows the logical symbol for a CMOS inverter. Notice how the amount of information decreases as the level of abstraction increases.

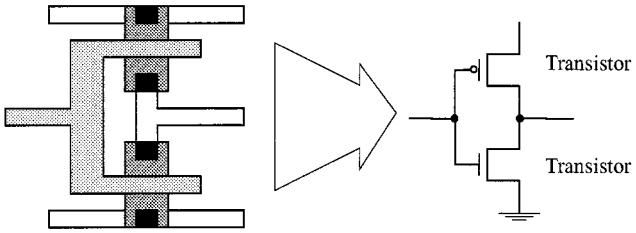


FIGURE 34.5 Electrical abstraction.

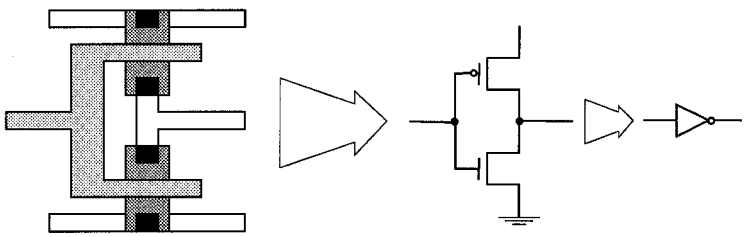


FIGURE 34.6 Logical abstraction.

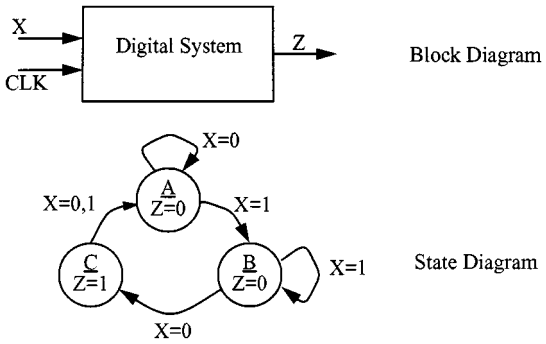


FIGURE 34.7 State diagram.

Design entry mechanisms for electrical and logical abstractions are collectively called *schematic capture* techniques. Schematic capture defines hierarchical structures, commonly called **netlists**, of components. A designer creates instances of components supplied from a library of predefined components and connects component pins or ports via wires [Douglas-Young, 1988; Pechet, 1991].

The functional level abstracts logical elements into corresponding computational units, such as registers, multiplexers, and arithmetic logic units (ALUs). The architectural level abstracts functional information into computational algorithms or paradigms. Examples of common computational paradigms are listed below:

- State diagrams,
- Petri nets,
- Control/data flow graphs,
- Function tables,
- Spreadsheets, and
- Binary decision diagrams.

These higher levels of abstraction support a more expressive, "higher-bandwidth" communication interface between engineers and DA programs. Engineers can focus their creative, cognitive skills on concept and behavior, rather than on the complexities of detailed implementation. Associated design entry mechanisms typically use hardware description languages with a combination of textual and graphic techniques [Birtwistle and Subrahmanyan, 1988].

Figure 34.7 shows an example of a simple state diagram. The state diagram defines three states, denoted by circles. State-to-state transitions are denoted by labeled arcs; state transitions depend on the present state and the input X. The output, Z, per state is given within each state. Since the output is dependent on only the present state, the digital system is classified as a Moore **finite state machine**. If the output is dependent on the present state and input, then the digital system is classified as a Mealy finite state machine.

A hardware description language model written in **VHDL** of the Moore finite state machine is given in Fig. 34.8. The VHDL model, called a *design entity*, uses a "**data flow**" description style to describe the state machine [Dewey, 1983, 1992, 1997]. The entity statement defines the interface, i.e., the ports. The ports include two input signals, X and CLK, and an output signal Z. The ports are of type BIT, which specifies that the signals may only carry the values 0 or 1. The architecture statement defines the input/output transform via two concurrent signal assignment statements. The internal signal STATE holds the finite state information and is driven by a guarded, conditional concurrent signal assignment statement that executes when the associated block expression

(CLK='1' **and not** CLK'STABLE)

is true, which is only on the rising edge of the signal CLK. STABLE is a predefined attribute of the signal CLK; CLK'STABLE is true if CLK has *not* changed value. Thus, if "**not** CLK'STABLE" is true, meaning that CLK has

```
-- entity statement
entity MOORE_MACHINE is
 port (X, CLK : in BIT;   Z : out BIT);
end MOORE_MACHINE;

-- architecture statement
architecture FSM of MOORE_MACHINE is
  type STATE_TYPE is (A, B, C);
  signal STATE : STATE_TYPE := A;
begin
 NEXT_STATE:
 block (CLK='1' and not CLK'STABLE)
 begin
    -- guarded conditional concurrent signal assignment statement
    STATE <= guarded B when (STATE=A and X='1') else
                          C when (STATE=B and X='0') else
                          A when (STATE=C) else
                          STATE;
 end block NEXT_STATE;
 -- unguarded selected concurrent signal assignment statement
 with STATE select
   Z <= '0' when A,
        '0' when B,
        '1' when C;
end FSM;
```

FIGURE 34.8 VHDL model.

just changed value, and "CLK='1'," then a rising transition has occurred on CLK. The output signal Z is driven by a nonguarded, selected concurrent signal assignment statement that executes any time STATE changes value.

34.3 Synthesis

Figure 34.9 shows that the **synthesis** task generally follows the design entry task. After describing the desired system via design entry, synthesis DA programs are invoked to assist generating the required detailed design.

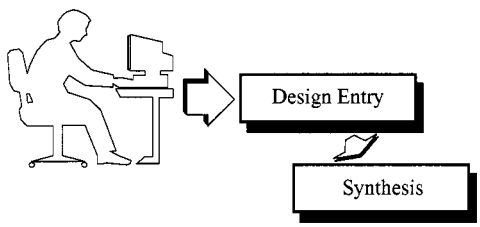


FIGURE 34.9 Design process — synthesis.

Synthesis translates or transforms a design from one level of abstraction to another, more-detailed level of abstraction. The more-detailed level of abstraction may be only an intermediate step in the entire design process, or it may be the final implementation. Synthesis programs that yield a final implementation are sometimes called **silicon compilers** because the programs generate sufficient detail to proceed directly to silicon fabrication [Ayres, 1983; Gajski, 1988].

Like design abstractions, synthesis techniques can be hierarchically categorized, as shown in Fig. 34.10. The higher levels of synthesis offer the advantage of less complexity, but also the disadvantage of less control over the final design.

Algorithmic synthesis, also called *behavioral* synthesis, addresses "multicycle" behavior, which means behavior that spans more than one *control step*. A control step equates to a clock cycle of a synchronous, sequential digital system, i.e., a state in a finite-state machine controller or a microprogram step in a microprogrammed controller.

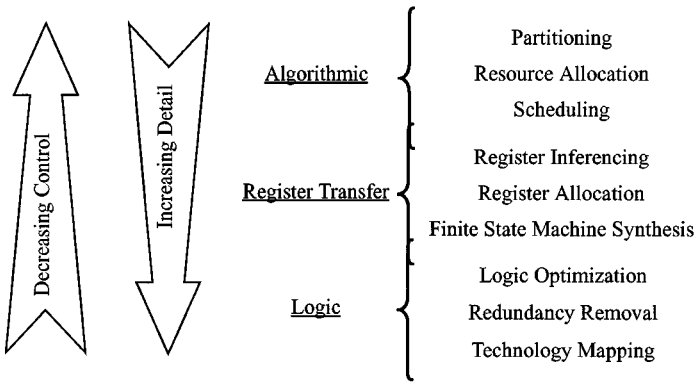


FIGURE 34.10 Taxonomy of synthesis techniques.

Algorithmic synthesis typically accepts sequential design descriptions that define an input/output transform, but provide little information about the parallelism of the final design [Camposano and Wolfe, 1991; Gajski et al., 1992].

Partitioning decomposes the design description into smaller behaviors. Partitioning is an example of a high-level transformation. High-level transformations include common software programming compiler optimizations, such as loop unrolling, subprogram in-line expansion, constant propagation, and common subexpression elimination.

Resource allocation associates behaviors with hardware computational units, and scheduling determines the order in which behaviors execute. Behaviors that are mutually exclusive can potentially share computational resources. Allocation is performed using a variety of graph clique covering or node coloring algorithms. Allocation and scheduling are interdependent, and different synthesis strategies perform allocation and scheduling different ways. Sometimes scheduling is performed first, followed by allocation; sometimes allocation is performed first, followed by scheduling; and sometimes allocation and scheduling are interleaved.

Scheduling assigns computational units to control steps, thereby determining which behaviors execute in which clock cycles. At one extreme, all computational units can be assigned to a single control step, exploiting maximum concurrency. At the other extreme, computational units can be assigned to individual control steps, exploiting maximum sequentiality. Several popular scheduling algorithms are listed below:

- As-soon-as-possible (ASAP),
- As-late-as-possible (ALAP),
- List scheduling,
- Force-directed scheduling, and
- Control step splitting/merging.

ASAP and ALAP scheduling algorithms order computational units based on data dependencies. List scheduling is based on ASAP and ALAP scheduling, but considers additional, more-global constraints, such as maximum number of control steps. Force-directed scheduling computes the probabilities of computational units being assigned to control steps and attempts to evenly distribute computation activity among all control steps. Control step splitting starts with all computational units assigned to one control step and generates a schedule by splitting the computational units into multiple control steps. Control step merging starts with all computational units assigned to individual control steps and generates a schedule by merging or combining units and steps [Paulin and Knight, 1989; Camposano and Wolfe, 1991].

Register transfer synthesis takes as input the results of algorithmic synthesis and addresses "per-cycle" behavior, which means the behavior during one clock cycle. Register transfer synthesis selects logic to realize the hardware computational units generated during algorithmic synthesis, such as realizing an addition operation with a carry–save adder or realizing addition and subtraction operations with an arithmetic logic unit.

Data that must be retained across multiple clock cycles are identified, and registers are allocated to hold the data. Finally, finite-state machine synthesis involves state minimization and state assignment. State minimization seeks to eliminate redundant or equivalent states, and state assignment assigns binary encodings for states to minimize combinational logic [Brayton et al., 1992; Sasao, 1993].

Logic synthesis optimizes the logic generated by register transfer synthesis and maps the optimized logic operations onto physical gates supported by the target fabrication technology. Technology mapping considers the foundry cell library and associated electrical restrictions, such as **fan-in/fan-out** limitations.

34.4 Verification

Figure 34.11 shows that the **verification** task generally follows the synthesis task. The verification task checks the correctness of the function and performance of a design to ensure that an intermediate or final design faithfully realizes the initial, desired specification. Three major types of verification are listed below:

- Timing analysis,
- Simulation, and
- Emulation.

Timing Analysis

Timing analysis checks that the overall design satisfies operating speed requirements and that individual signals within a design satisfy transition requirements. Common signal transition requirements, also called *timing hazards*, include *rise* and *fall times*, *propagation delays*, *clock periods*, *race conditions*, *glitch detection*, and *setup* and *hold times*. For instance, setup and hold times specify relationships between data and control signals to ensure that memory devices (level-sensitive latches or edge-sensitive flip-flops) correctly and reliably store desired data. The data signal carrying the information to be stored in the memory device must be stable for a period equal to the setup time prior to the control signal transition to ensure that the correct value is sensed by the memory device. Also, the data signal must be stable for a period equal to the hold time after the control signal transition to ensure that the memory device has enough time to store the sensed value.

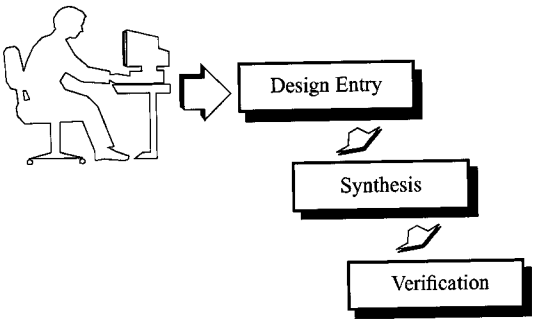


FIGURE 34.11 Design process — verification.

Another class of timing transition requirements, commonly called signal integrity checks, include *reflections*, *crosstalk*, **ground bounce**, and *electromagnetic interference*. Signal integrity checks are typically required for high-speed designs operating at clock frequencies above 75 MHz. At such high frequencies, the transmission line behavior of wires must be analyzed. A wire should be properly terminated, i.e., connected, to a port having an impedance matching the wire characteristic impedance to prevent signal reflections. Signal reflections are portions of an emanating signal that "bounce back" from the destination to the source. Signal reflections reduce the power of the emanating signal and can damage the source. Crosstalk refers to unwanted reactive coupling between physically adjacent signals, providing a connection between signals that are supposed to be electrically isolated. Ground bounce is another signal integrity problem. Since all conductive material has a finite impedance, a ground signal network does not in practice offer the exact same electrical potential throughout an entire design. These potential differences are usually negligible because the distributive impedance of the ground signal network is small compared with other finite-component impedances. However, when many signals switch value simultaneously, a substantial current can flow through the ground signal network. High intermittent currents yield proportionately high intermittent potential drops, i.e., ground bounces, which can

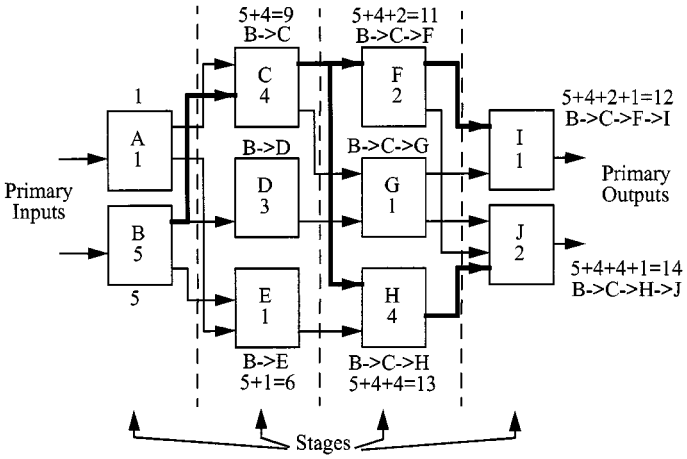


FIGURE 34.12 Block-oriented static timing analysis.

cause unwanted circuit behavior. Finally, electromagnetic interference refers to signal harmonics radiating from design components and interconnects. This harmonic radiation may interfere with other electronic equipment or may exceed applicable environmental safety regulatory limits [McHaney, 1991].

Timing analysis can be performed dynamically or statically. Dynamic timing analysis exercises the design via simulation or emulation for a period of time with a set of input stimuli and records the timing behavior. Static timing analysis does not exercise the design via simulation or emulation. Rather, static analysis records timing behavior based on the timing behavior, e.g., propagation delay, of the design components and their interconnection.

Static timing analysis techniques are primarily *block oriented* or *path oriented*. Block-oriented timing analysis generates design input (also called primary input) to design output (also called primary output), and propagation delays by analyzing the design "stage-by-stage" and by summing up the individual stage delays. All devices driven by primary inputs constitute stage 1, all devices driven by the outputs of stage 1 constitute stage 2, and so on. Starting with the first stage, all devices associated with a stage are annotated with worst-case delays. A worst-case delay is the propagation delay of the device plus the delay of the last input to arrive at the device, i.e., the signal path with the longest delay leading up to the device inputs. For example, the device labeled "H" in stage 3 in Fig. 34.12 is annotated with the worst-case delay of 13, representing the device propagation delay of 4 and the delay of the last input to arrive through devices "B" and "C" of 9 [McWilliams and Widdoes, 1978]. When the devices associated with the last stage, i.e., the devices driving the primary outputs, are processed, the accumulated worst-case delays record the longest delay from primary inputs to primary outputs, also call the critical paths. The critical path for each primary output is highlighted in Fig. 34.12.

Path-oriented timing analysis generates primary input to primary output propagation delays by traversing all possible signal paths one at a time. Thus, finding the critical path via path-oriented timing analysis is equivalent to finding the longest path through a directed acyclic graph, where devices are graph vertices and interconnections are graph edges [Sasiki et al., 1978].

To account for realistic variances in component timing due to manufacturing tolerances, aging, or environmental effects, timing analysis often provides stochastic or statistical checking capabilities. Statistical timing analysis uses random-number generators based on empirically observed probabilistic distributions to determine component timing behavior. Thus, statistical timing analysis describes design performance and the likelihood of the design performance.

Simulation

Simulation exercises a design over a period of time by applying a series of input stimuli and generating the associated output responses. The general event-driven, also called schedule-driven, simulation algorithm is diagrammed in Fig. 34.13. An event is a change in signal value. Simulation starts by initializing the design;

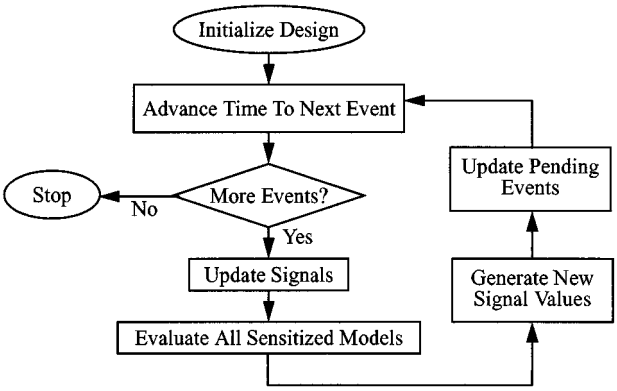


FIGURE 34.13 General event-driven simulation algorithm.

initial values are assigned to all signals. Initial values include starting values and pending values that constitute future events. Simulation time is advanced to the next pending event(s), signals are updated, and sensitized models are evaluated [Pooch, 1993]. The process of evaluating the sensitized models yields new, potentially different, values for signals, i.e., a new set of pending events. These new events are added to the list of pending events, time is advanced to the next pending event(s), and the simulation algorithm repeats. Each pass through the loop in Fig. 34.13 of evaluating sensitized models at a particular time step is called a **simulation cycle**. Simulation ends when the design yields no further activity, i.e., when there are no more pending events to process.

Logic simulation is computationally intensive for large, complex designs. As an example, consider simulating 1 s of a 200K-gate, 20-MHz processor design. By assuming that, on average, only 10% of the total 200K gates are active or sensitized on each processor clock cycle, Eq. 34.1 shows that simulating 1 s of actual processor time equates to 400 billion events.

$$\begin{aligned} 400 \text{ billion events} &= (20 \text{ million clock cycles})(200\text{K gates})(10\% \text{ activity}) \\ 140 \text{ h} &= (400 \text{ billion events})\left(\frac{50 \text{ instructions}}{\text{event}}\right)\left(\frac{50 \text{ million instructions}}{\text{s}}\right) \end{aligned} \tag{34.1}$$

Assuming that, on average, a simulation program executes 50 computer instructions per event on a computer capable of processing 50 million instructions per second (MIP), Eq. 34.1 also shows that processing 400 billion events requires 140 h or just short of 6 days. Fig. 34.14 shows how simulation computation generally scales with design complexity.

To address the growing computational demands of simulation, several simulation acceleration techniques have been introduced. Schedule-driven simulation, explained above, can be accelerated by removing layers of interpretation and running a simulation as a native executable image; such an approach is called complied, scheduled-driven simulation.

As an alternative to schedule-driven simulation, *cycle-driven* simulation avoids the overhead of event queue processing by evaluating all devices at regular intervals of time. Cycle-driven simulation is efficient when a design exhibits a high degree of concurrency, i.e., when a large percentage of the devices are active per simulation cycle. Based on the staging of devices, devices are *rank-ordered* to determine the order in which they are evaluated at each time step to ensure the correct causal behavior yielding the proper ordering of events. For functional verification, logic devices are often assigned zero-delay and memory devices are assigned unit-delay. Thus, any number of stages of logic devices may execute between system clock periods.

In another simulation acceleration technique, *message-driven* simulation, also called *parallel* or *distributed* simulation, device execution is divided among several processors and the device simulations communicate

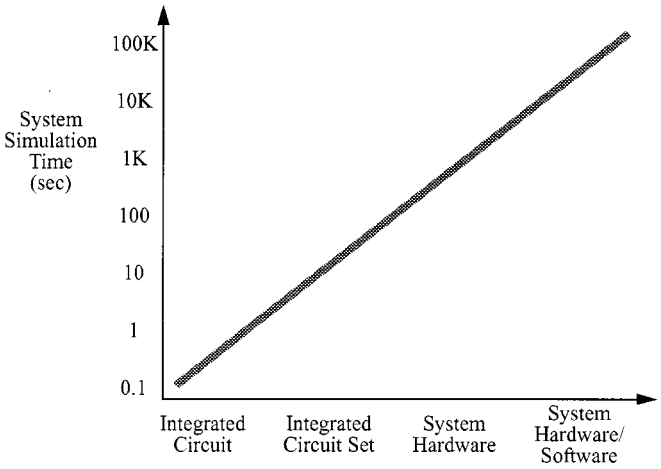


FIGURE 34.14 Simulation requirements.

event activity via messages. Messages are communicated using conservative or optimistic strategies. Optimistic message-passing strategies, such as *time warp* and *lazy cancellation,* make assumptions about future event activity to advance local device simulation. If the assumptions are correct, the processors operate more independently and better exploit parallel computation. However, if the assumptions are incorrect, then local device simulations may be forced to "roll back" to synchronize local device simulations [Bryant, 1979; Chandy and Misra, 1981].

Schedule-driven, cycle-driven, and message-driven simulations are software-based simulation acceleration techniques. Simulation can also be accelerated by relegating certain simulation activities to dedicated hardware. For example, *hardware modelers* can be attached to software simulators to accelerate the activity of device evaluation. As the name implies, hardware modeling uses actual hardware devices instead of software models to obtain stimulus/response information. Using actual hardware devices reduces the expense of generating and maintaining software models and provides an environment to support application software development. However, it is sometimes difficult for a slave hardware modeler to preserve accurate real-time device operating response characteristics within a master non-real-time software simulation environment. For example, some hardware devices may not be able to retain state information between invocations, so the hardware modeler must save the history of previous inputs and reapply them to bring the hardware device to the correct state to apply a new input.

Another technique for addressing the growing computational demands of simulation is via simulation engines. A simulation engine can be viewed as an extension of the simulation acceleration technique of hardware modeling. With a hardware modeler, the simulation algorithm executes in software and component evaluation executes in dedicated hardware. With a simulation engine, the simulation algorithm *and* component evaluation execute in dedicated hardware. Simulation engines are typically two to three orders of magnitude faster than software simulation [Takasaki et al., 1989].

Analog Simulation

Analog simulation involves time-domain analyses and frequency-domain analyses, which are generally conducted using some form of direct current (DC) simulation, diagrammed in Fig. 34.15. DC simulation determines the quiescent or steady-state operating point for a circuit, specifying **node voltages**, **branch** currents, input/output resistances, element sensitivities, and input/output gains [Chua and Lin, 1975; Nagel, 1975].

Several popular equation formulation schemes are summarized in Table 34.1. Equation formulation schemes generate a set of linear equations denoting relationships between circuit voltages and currents; these relationships are based on the physical principle of the conservation of energy expressed via **Kirchoff's current law** (KCL), **Kirchoff's voltage law** (KVL), and branch constitutive relationships (BCRs). A circuit having N nodes and B branches possesses $2B$ independent variables defining B branch voltages and B branch currents. These

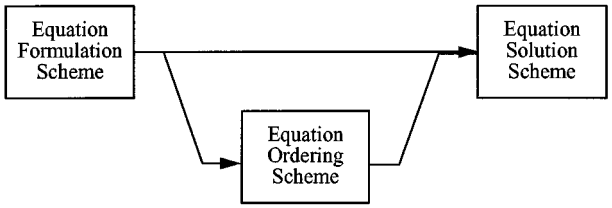


FIGURE 34.15 DC simulation.

TABLE 34.1 Common Circuit Equation Formulation Schemes

Equation Formulation Schemes	Desired Unknowns
Nodal analysis	Node voltages
Modified nodal analysis	Node voltages
	Dependent source currents
	Independent voltage source currents
Sparse tableau analysis	Node voltages
	Branch currents
	Branch voltages
Reduced tableau analysis	Node voltages
	Branch currents
Tree analysis	Tree branch voltages
Link analysis	Link branch currents

variables are governed by $2B$ linearly independent equations composed of $N-1$ KCL equations, $B-N+1$ KVL equations, and B BCR equations [Hachtel et al., 1971; Ho et al., 1975].

Equation-ordering schemes augment equation formulation schemes by reorganizing, modifying, and scaling the equations to improve the efficiency and/or accuracy of the subsequent equation solution scheme. More specifically, equation-ordering schemes seek to improve the "diagonal dominance" structure of the coefficient matrix by maximizing the number of "off-diagonal" zeros. Popular equation-ordering schemes include pivoting and row ordering (Markowitz) [Zlatev, 1980].

Finally, equation solution schemes determine the values for the independent variables that comply with the governing equations. There are basically two types of equation solution schemes: explicit and implicit. Explicit solution schemes, such as Gaussian elimination and/or LU factorization, determine independent variable values using closed-form, deterministic techniques. Implicit solution schemes, such as Gauss–Jacobi and Gauss–Seidel, determine independent variable values using iterative, nondeterministic techniques.

Emulation

Emulation, also called *computer-aided prototyping*, verifies a design by realizing the design in "preproduction" hardware and exercising the hardware. The term *preproduction* hardware means nonoptimized hardware providing the correct functional behavior, but not necessarily the correct performance. That is, emulation hardware may be slower, require more area, or dissipate more power than production hardware. At present, preproduction hardware commonly involves some form of **programmable logic devices** (PLDs), typically field-programmable **gate arrays** (FPGAs). PLDs provide generic combinational and sequential digital system logic that can be programmed to realize a wide variety of designs [Walters, 1991].

Emulation offers the advantage of providing prototype hardware early in the design cycle to check for errors or inconsistencies in initial functional specifications. Problems can be isolated and design modifications can be easily accommodated by reprogramming the logic devices. Emulation can support functional verification at computational rates much greater than conventional simulation. However, emulation does not generally support performance verification because, as explained above, prototype hardware typically does not operate at production clock rates.

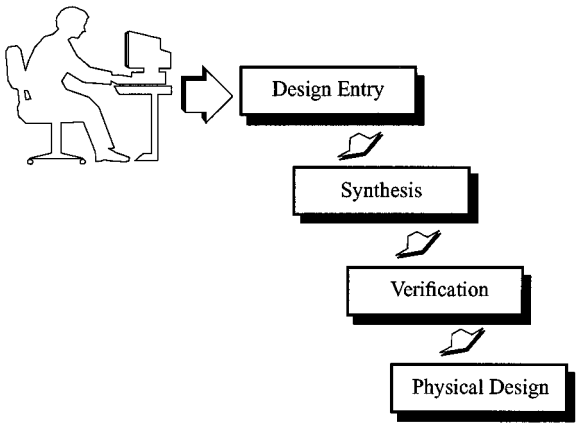


FIGURE 34.16 Design process — physical design.

34.5 Physical Design

Figure 34.16 shows that the physical design task generally follows the verification task. Having validated the function and performance of the detailed design during verification, physical design realizes the detailed design by translating logic into actual hardware. Physical design involves placement, routing, artwork generation, rules checking, and back annotation [Sait and Youseff, 1995].

Placement transforms a logical hierarchy into a physical hierarchy by defining how hardware elements are oriented and arranged relative to each other. Placement determines the overall size, i.e., area, a digital system will occupy. Two popular placement algorithms are *mincut* and *simulated annealing*. Mincut placement techniques group highly connected cells into clusters. Then, the clusters are sorted and arranged according to user-supplied priorities. Simulated annealing conducts a series of trial-and-error experiments by pseudorandomly moving cells and evaluating the resulting placements, again according to user-supplied priorities.

Routing defines the wires that establish the required port-to-port connections. Routing is often performed in two stages: global and local. Global routing assigns networks to major wiring regions, called tracks; local routing defines the actual wiring for each network within its assigned track. Two common classes of routing algorithms are *channel* and *maze*. Channel routing connects ports abutting the same track. Maze routing, also called switch-box routing, connects ports abutting different channels. Routing considers a variety of metrics, including timing skew, wire length, number of vias, and number of jogs (corners) [Spinks, 1985; Preas et al., 1988].

Rules checking verifies that the final layout of geometric shapes and their orientation complies with logical, electrical, and physical constraints. Logical rules verify that the implementation realizes the desired digital system. Electrical rules verify conformance to loading, noise margins, and fan-in/fan-out connectivity constraints. Finally, physical rules verify conformance to dimensional, spacing, and alignment constraints [Hollis, 1987].

34.6 Test

Figure 34.17 shows that **test** follows physical design. After physical design, the digital system is manufactured and test checks the resulting hardware for correct function and performance. Thus, the primary objective of test is to detect a faulty device by applying input test stimuli and observing expected results [Buckroyd, 1989; Weyerer and Goldemund, 1992].

The test task is difficult because designs are growing in complexity; more components provide more opportunity for manufacturing defects. Test is also challenged by new microelectronic fabrication processes. These new processes support higher levels of integration that provide fewer access points to probe internal electrical nodes and new failure modes that provide more opportunity for manufacturing defects.

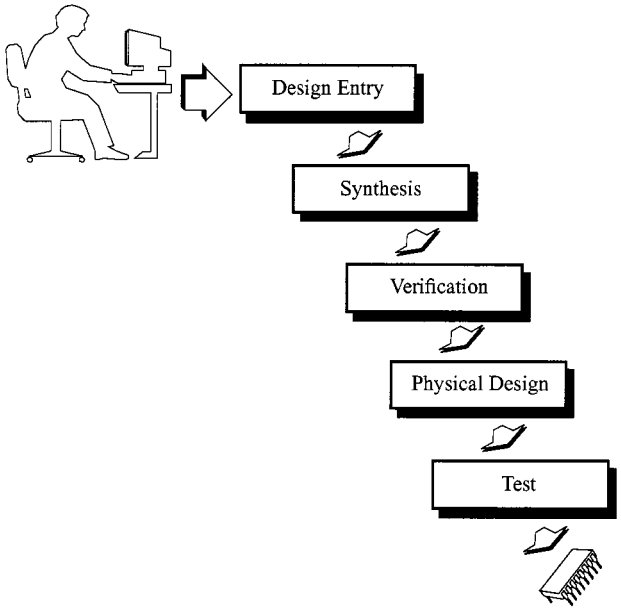


FIGURE 34.17 Design process — test.

Fault Modeling

What is a fault? A fault is a manufacturing or aging defect that causes a device to operate incorrectly or to fail. A sample listing of common integrated circuit physical faults are given below:

- Wiring faults,
- Dielectric faults,
- Threshold faults, and
- Soft faults.

Wiring faults are unwanted opens and shorts. Two wires or networks that should be electrically connected, but are not connected constitute an open. Two wires or networks that should not be electrically connected, but are connected constitute a short. Wiring faults can be caused by manufacturing defects, such as metallization and etching problems, or aging defects, such as corrosion and **electromigration**. Dielectric faults are electrical isolation defects that can be caused by **masking defects**, chemical impurities, material imperfections, or electrostatic discharge. Threshold faults occur when the turn-on and turn-off voltage potentials of electrical devices exceed allowed ranges. Soft faults occur when radiation exposure temporarily changes electrical charge distributions. Such changes can alter circuit voltage potentials, which can, in turn, change logical values, also called *dropping bits*. Radiation effects are called "soft" faults because the hardware is not permanently damaged [Zobrist, 1993].

To simplify the task of fault testing, the physical faults described above are translated into logical faults. Typically, a single logical fault covers several physical faults. A popular logical fault model is the *single stuck line* (SSL) fault model. The single stuck line fault model supports faults that denote wires permanently set to a logic 0, "stuck-at-0," or a logic 1, "stuck-at-1." Building on the single stuck line fault model, the *multiple stuck line* (MSL) fault model supports faults where multiple wires are stuck-at-0/stuck-at-1. Stuck fault models do not address all physical faults because not all physical faults result in signal lines permanently set to low or high voltages, i.e., stuck-at-0 or stuck-at-1 logic faults. Thus, other fault models have been developed to address specific failure mechanisms. For example, the *bridging* fault model addresses electrical shorts that cause unwanted coupling or spurious feedback loops.

Fault Testing

Once the physical faults that may cause device malfunction have been identified and categorized and how the physical faults relate to logical faults has been determined, the next task is to develop tests to detect these faults. When the tests are generated by a computer program, this activity is called *automatic test program generation* (ATPG). Examples of fault testing techniques are listed below:

- Stuck-at techniques,
- Scan techniques,
- Signature techniques,
- Coding techniques, and
- Electrical monitoring techniques.

Basic stuck-at fault testing techniques address combinational digital systems. Three of the most popular stuck-at fault testing techniques are the D algorithm, the Path-Oriented Decision Making (Podem) algorithm, and the Fan algorithm. These algorithms first identify a circuit fault, e.g., stuck-at-0 or stuck-at-1, and then try to generate an input stimulus that detects the fault and makes the fault visible at an output. Detecting a fault is called *fault sensitization* and making a fault visible is called *fault propagation*. To illustrate this process, consider the simple combinational design in Fig. 34.18 [Goel, 1981; Fujiwara and Shimono, 1983].

The combinational digital design is defective because a manufacturing defect has caused the output of the top **and** gate to be permanently tied to ground, i.e., stuck-at-0, using a positive logic convention. To sensitize the fault, the inputs A and B should both be set to 1, which should force the top **and** gate output to a 1 for a good circuit. To propagate the fault, the inputs C and D should both be set to 0, which should force the **xor** gate output to 1, again for a good circuit. Thus, if A = 1, B = 1, C = 0, and D = 0 in Fig. 34.18, then a good circuit would yield a 1, but the defective circuit yields a 0, which detects the stuck-at-0 fault at the top **and** gate output.

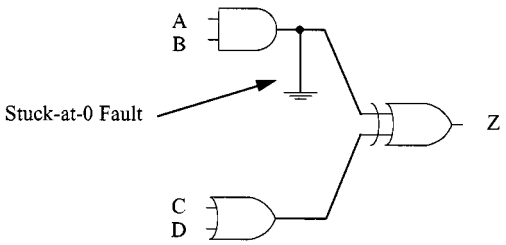


FIGURE 34.18 Combinational logic stuck-at fault testing.

Sequential ATP generation is more difficult than combinational ATPG because exercising or sensitizing a particular circuit path to detect the presence of a possible manufacturing fault may require a *sequence* of input test vectors. One technique for testing sequential digital systems is scan fault testing. Scan fault testing is an example of *design for testability* (DFT) because it modifies or constrains a design in a manner that facilitates fault testing. Scan techniques impose a logic design discipline that connects all state registers into one or more chains to form "scan rings," as shown in Fig. 34.19 [Eichelberger and Williams, 1977].

During normal device operation, the scan rings are disabled and the registers serve as conventional memory (state) storage elements. During test operation, the scan rings are enabled and stimulus test vectors are shifted into the memory elements to set the state of the digital system. The digital system is exercised for one clock cycle and then the results are shifted out of the scan ring to record the response.

A variation of scan DFT, called *boundary scan*, has been defined for testing integrated circuits on printed circuit boards (PCBs). Advancements in PCB manufacturing, such as fine-lead components, surface mount assembly, and **multichip modules**, have yielded high-density boards with fewer access points to probe individual pins. These PCBs are difficult to test. As the name implies, boundary scan imposes a design discipline for PCB components to enable the input/output pins of the components to be connected into scan chains. As an example, Fig. 34.20 shows a simple PCB containing two integrated circuits configured for boundary scan. Each integrated circuit contains scan registers between its input/output pins and its core logic to enable the PCB test bus to control and observe the behavior of individual integrated circuits [Parker, 1989].

Another DFT technique is signature analysis, also called *built-in self-test* (BIST). Signature testing techniques use additional logic, typically linear feedback shift registers, to generate automatically pseudorandom test vectors. The output responses are compressed into a single vector and compared with a known good vector. If the output response vector does not exactly match the known good vector, then the design is considered faulty.

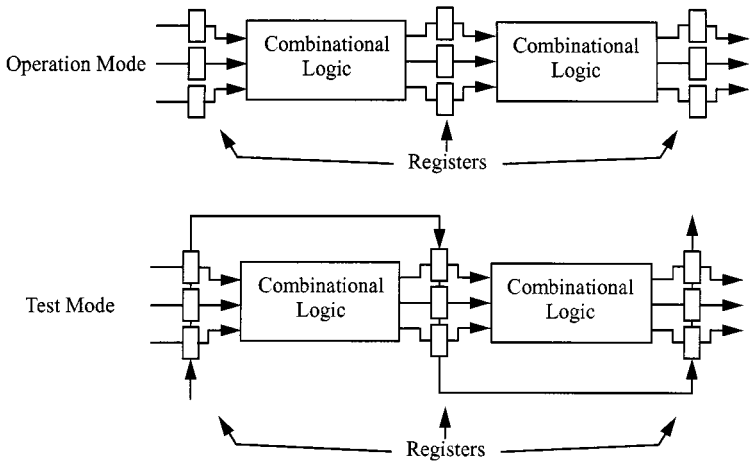


FIGURE 34.19 Scan-based DFT.

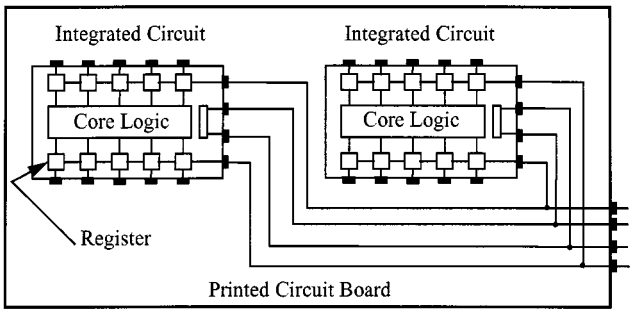


FIGURE 34.20 Boundary scan.

Matching the output response vector and a known good vector does not guarantee correct hardware; however, if enough pseudorandom test vectors are exercised, then the chances are acceptably small of obtaining a false positive result. Signature analysis is often used to test memories [Agrawal et al., 1993].

Coding test techniques encode signal information so that errors can be detected and possibly corrected. Although often implemented in software, coding techniques can also be implemented in hardware. For example, a simple coding technique called *parity checking* is often implemented in hardware. Parity checking adds an extra bit to multibit data. The parity bit is set such that the total number of logic 1s in the multibit data *and* parity bit is either an even number (even parity) or an odd number (odd parity). An error has occurred if an even-parity-encoded signal contains an odd number of logic 1s or if an odd-parity-encoded signal contains an even number of logic 1s. Coding techniques are used extensively to detect and correct transmission errors on system buses and networks, storage errors in system memory, and computational errors in processors [Peterson and Weldon, 1972].

Finally, electrical monitoring testing techniques, also called current/voltage testing, rely on the simple observation that an out-of-range current or voltage often indicates a defective or bad part. Possibly a short or open is present causing a particular input/output signal to have the wrong voltage or current. Current testing, or I_{ddq} *testing*, is particularly useful for digital systems using CMOS integrated circuit technology. Normally, CMOS circuits have very low static or quiescent currents. However, physical faults, such as **gate oxide** defects, can increase static current by several orders of magnitude. Such a substantial change in static current is straightforward to detect. The principal advantages of current testing are that the tests are simple and the fault models address detailed transistor-level defects. However, current testing requires that enough time be allotted between input stimuli to allow the circuit to reach a static state, which slows testing down and causes problems with circuits that cannot be tested at scaled clock rates.

34.7 Summary

DA technology offers the potential of serving as a powerful fulcrum in leveraging the skills of a designer against the growing demands of electronic system design and manufacturing. DA programs help to relieve the designer of the burden of tedious, repetitive tasks that can be labor-intensive and error prone.

DA technology can be broken down into several topical areas, such as design entry, synthesis, verification, physical design, and test. Each topical area has developed an extensive body of knowledge and experience.

Design entry defines a desired specification. Synthesis refines the initial design specification into a detailed design ready for implementation. Verification checks that the detailed design faithfully realizes the desired specification. Physical design defines the implementation, i.e., the actual hardware. Finally, test checks that the manufactured part is functionally and parametrically correct.

Defining Terms

Bipolar: Type of semiconductor transistor that involves both minority and majority carrier conduction mechanisms.

BiCMOS: Bipolar/complementary metal-oxide semiconductor. A logic family and form of microelectronic fabrication.

Branch: A circuit element between two nodes. Branch current is the current through the branch. Branch voltage is the potential difference between the nodes. The relationship between the branch current and voltage is defined by the branch constitutive relationship.

Capacitor: Two-terminal electronic device governed by the branch constitutive relationship, Charge = Capacitance $\times$ Voltage.

CMOS: Complementary metal-oxide semiconductor. A logic family and form of microelectronic fabrication.

Data Flow: Nonprocedural modeling style in which the textual order that statements are written has no bearing on the order in which they execute.

Design automation: Computer programs that assist engineers in performing digital system development.

Design entry: Area of DA addressing modeling analog and digital electronic systems. Design entry uses a hierarchy of models involving physical, electrical, logical, functional, and architectural abstractions.

Electromigration: Gradual erosion of metal due to excessive currents.

Fan-in/fan-out: Fan-in defines the maximum number of logic elements that may drive another logic element. Fan-out defines the maximum number of logic elements a logic element may drive.

Finite state machine: Sequential digital system. A finite state machine is classified as either Moore and Mealy.

Gate array: Application-specific integrated circuit implementation technique that realizes a digital system by programming the metal interconnect of a prefabricated array of gates.

Gate oxide: Dielectric insulating material between the gate and source/drain terminals of a MOS transistor.

Ground bounce: Transient condition when the potential of a ground network varies appreciably from its uniform static value.

Integrated circuit: Electronic circuit manufactured on a monolithic piece of semiconductor material, typically silicon.

Kirchoff's current law: The amount of current entering a circuit node equals the amount of current leaving a circuit node.

Kirchoff's voltage law: Any closed loop of circuit branch voltages sums to zero.

Masking defects: Defects in masking plate patterns used for integrated circuit lithography that result in errant material composition and/or placement.

Multichip modules: Multiple integrated circuits interconnected on a monolithic substrate.

Netlist: Collection of wires that are electrically connected to each other.

NMOS: N-type metal-oxide semiconductor. A logic family and form of microelectronic fabrication.

Node voltage: Potential of a circuit node relative to ground potential.

Programmable logic devices (PLDs): Generic logic devices that can be programmed to realize specific digital systems. PLDs include programmable logic arrays, programmable array logic, memories, and field-programmable gate arrays.

Resistor: Two-terminal electronic device governed by the branch constitutive relationship, Voltage = Resistance × Current.

Silicon compilation: Synthesis application that generates final physical design ready for silicon fabrication.

Simulation: Computer program that examines the dynamic semantics of a model of a digital system by applying a series of inputs and generating the corresponding outputs. Major types of simulation include schedule driven, cycle driven, and message driven.

Skew: Timing difference between two events that are supposed to occur simultaneously.

Standard cell: Application-specific integrated circuit implementation technique that realizes a digital system using a library of predefined (standard) logic cells.

Synthesis: Computer program that helps generate a digital/analog system design by transforming a high-level model of abstract behavior into a lower-level model of more-detailed behavior.

Test: Area of EDA that addresses detecting faulty hardware. Test involves stuck-at, scan, signature, coding, and monitoring techniques.

Timing analysis: Verifies timing behavior of electronic system including rise time, fall time, setup time, hold time, glitch detection, clock periods, race conditions, reflections, and cross talk.

Transistor: Electronic device that enables a small voltage and/or current to control a larger voltage and/or current. For analog systems, transistors serves as amplifiers. For digital systems, transistors serve as switches.

Verification: Area of EDA that addresses validating designs for correct function and expected performance. Verification involves timing analysis, simulation, emulation, and formal proofs.

VHDL: Hardware description language used as an international standard for communicating electronic systems information.

Via: Connection or contact between two materials that are otherwise electrically isolated.

Related Topics

23.2 Testing • 25.1 Integrated Circuit Technology • 25.3 Application-Specific Integrated Circuits

References

Design Automation

D. Barbe, Ed., *Very Large Scale Integration (VLSI) — Fundamentals and Applications*, New York: Springer-Verlag, 1980.

T. Dillinger, *VLSI Engineering*, Englewood Cliffs, N.J.: Prentice-Hall, 1988.

S. Sapiro, *Handbook of Design Automation*, Englewood Cliffs, Prentice-Hall, N.J.: 1986

S. Trimberger, *An Introduction to CAD for VLSI*, Calif.: Domancloud Publishers, 1990.

Design Entry

G. Birtwistle, and Subrahmanyan, P., *VLSI Specification, Verification, and Synthesis*, Boston: Kluwer Academic Publishers, 1988.

A. Dewey, "VHSIC hardware description language development program," *Proceedings Design Automation Conference*, June, 1983.

A. Dewey, "VHDL: towards a unified view of design," *IEEE Design and Test of Computers*, June, 1992.

A. Dewey, *Analysis and Design of Digital Systems with VHDL*, Boston: PWS Publishing, 1997.

J. Douglas-Young, *Complete Guide to Reading Schematic Diagrams*, Englewood Cliffs, N.J.: Prentice-Hall, 1988.

M. Pechet, Ed., *Handbook of Electrical Package Design*, New York: Marcel Dekker, 1981.

J. Peterson, *Petri Net Theory and Modeling of Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Synthesis

R. Ayres, *VLSI: Silicon Compilation and the Art of Automatic Microchip Design*, Englewood Cliffs, N.J.: Prentice-Hall, 1983.

Brayton et al., *Logic Minimization Algorithms for VLSI Synthesis*, Boston: Kluwer Academic Publishers, 1992.

R. Camposano, and Wolfe, W., *High-Level VLSI Synthesis*, Boston: Kluwer Academic Publishers, 1991.
D. Gajski, Ed., *Silicon Compilation*, Boston: Addison-Wesley, 1988.
D. Gajski, et al., *High-Level Synthesis — Introduction to Chip and System Design*, Boston: Kluwer Academic Publishers, 1992.
P. Paulin, and Knight, J., "Force-directed scheduling for the behavioral synthesis of ASIC's," *IEEE Design and Test of Computers*, October, 1989.
T. Sasao, Ed., *Logic Synthesis and Optimization*, Boston: Kluwer Academic Publishers, 1993.

Verification

R. Bryant, "Simulation on distributed systems," *Proceedings International Conference on Distributed Systems,* 1979.
K. Chandy, and Misra, J., "Asynchronous distributed simulation via a sequence of parallel computations," *Communications of the ACM*, April, 1981.
L. Chua, and Lin, P., *Computer-Aided Analysis of Electronic Circuits: Algorithms and Computational Techniques*, Englewood Cliffs, N.J.: Prentice-Hall, 1975.
G. Hachtel, Brayton, R., and Gustavson, F., "The sparse tableau approach to network analysis and design," *IEEE Transactions on Circuit Theory*, CT-18, 1971.
W. Hahn, and Fischer, K., "High performance computing for digital design simulation", *VLSI85*, New York: Elsevier Science Publishers, 1985.
C. Ho, Ruehli, A., and Brennan, P., "The modified nodal analysis approach to network analysis," *IEEE Transactions on Circuits and Systems*, 1975.
R. McHaney, *Computer Simulation: A Practical Perspective*, New York: Academic Press, 1991.
T. McWilliams, and Widdoes, L., "SCALD — structured computer aided logic design," *Proceedings Design Automation Conference*, June, 1978.
L. Nagel, SPICE2: A Computer Program to Simulate Semiconductor Circuits, Electronic Research Laboratory, ERL-M520, Berkeley: University of California, 1975.
U. Pooch, *Discrete Event Simulation: A Practical Approach*, Boca Raton, Fla.: CRC Press, 1993.
T. Sasiki, et al., "Hierarchical design and verification for large digital systems," in *Proceedings Design Automation Conference*, June, 1978.
S. Takasaki, Hirose, F., and Yamada, A., "Logic simulation engines in Japan," *IEEE Design and Test of Computers*, October, 1989.
S. Walters, "Computer-aided prototyping for ASIC-based synthesis," *IEEE Design and Test of Computers*, June, 1991.
Z. Zlatev, "On some pivotal strategies in Gaussian elimination by sparse technique," *SIAM Journal of Numerical Analysis*, vol. 17, no. 1, 1980.

Physical Design

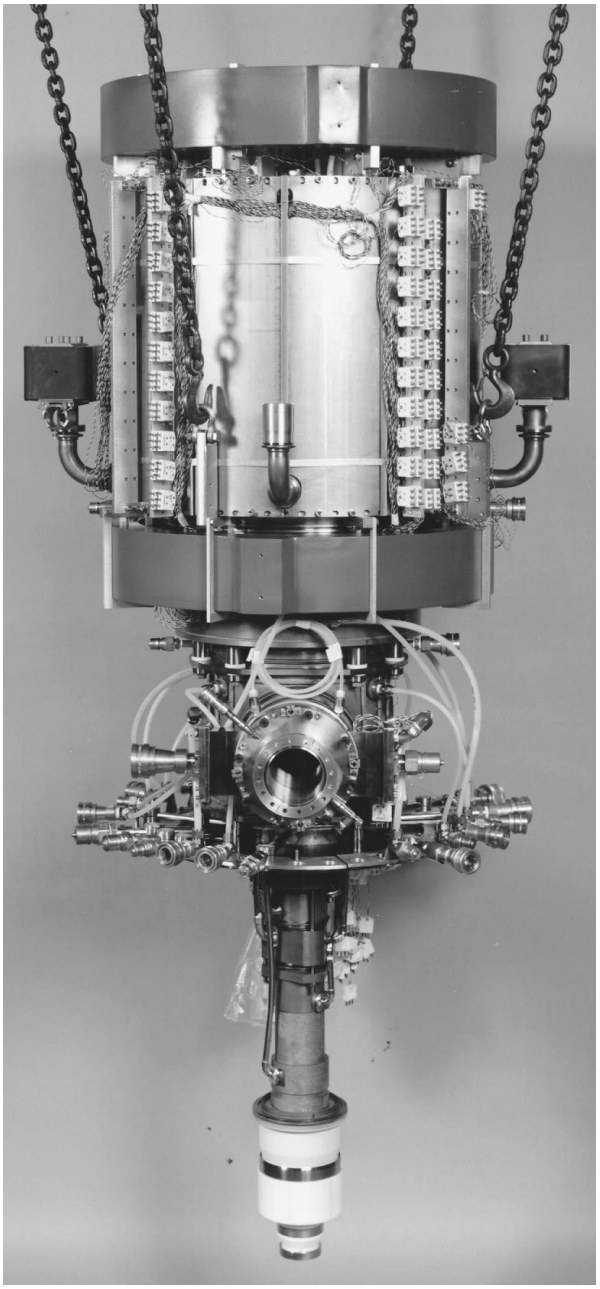
E. Hollis, *Design of VLSI Gate Array Integrated Circuits*, Englewood Cliffs, N.J.: Prentice-Hall, 1987.
B. Preas, B., Lorenzetti, M., and Ackland, B., Eds., *Physical Design Automation of VLSI Systems*, New York: Benjamin Cummings, 1988.
S. Sait and Youssef, H., *VLSI Physical Design Automation: Theory and Practice*, New York: McGraw-Hill, 1995.
B. Spinks, *Introduction to Integrated Circuit Layout*, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

Test

V. Agrawal, Kime, C., and Saluja, K., "A tutorial on built-in self-test," *IEEE Design and Test of Computers*, June, 1993.
A. Buckroyd, *Computer Integrated Testing*, New York: Wiley, 1989.
E. Eichelberger, and Williams, T., "A logic design structure for LSI testability," *Proceedings Design Automation Conference*, June, 1977.
H. Fujiwar, and Shimono, T., "On the acceleration of test generation algorithms," *IEEE Transactions on Computers*, December, 1983.
P. Goel, "An implicit enumeration algorithm to generate tests for combinational logic circuits," *IEEE Transactions on Computers*, March, 1981.

K. Parker, "The impact of boundary scan on board test," *IEEE Design and Test of Computers*, August, 1989.
W. Peterson and Weldon, E., *Error-Correcting Codes*, Boston: The MIT Press, 1972.
M. Weyerer and Goldemund, G., *Testability of Electronic Circuits*, Englewood Cliffs, N.J.: Prentice-Hall, 1992.
G. Zobrist, Ed., *VLSI Fault Modeling and Testing Technologies*, New York: Ablex Publishing Company, 1993.

Rawat, B. "Section IV – Electromagnetics"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



High power gyrotron from CPI. The 110-GHz gyrotron is the current world record holder for high frequency power generation. This gyrotron is used in electron cyclotron resonance heating (ECRH) by producing extremely high frequencies of microwaves which heat a hydrogen gas — "plasma" — to very high levels in experimental fusion reactors.

In the photo, the lower section is the electron gun region that generates extremely high-powered electron beam. Shown in the middle section of the gyrotron is an interactive window. The produced microwave power is transformed into a microwave beam that then passes through the high power interaction window. The large upper section is a fully instrumented collector of the "spent" electron beam. (Photo courtesy of Communications & Power Industries.)

IV

Electromagnetics

35 Electromagnetic Fields *J.A. Kong*
Maxwell Equations • Constitutive Relations • Wave Equations and Wave Solutions

36 Magnetism and Magnetic Fields *G. Bate, M.H. Kryder*
Magnetism • Magnetic Recording

37 Wave Propagation *M.N.O. Sadiku, K. Demarest*
Space Propagation • Waveguides

38 Antennas *N.J. Kolias, R.C. Compton, J.P. Fitch, D.M. Pozar*
Wire • Aperture • Microstrip Antennas

39 Microwave Devices *M.B. Steer, R.J. Trew*
Passive Microwave Devices • Active Microwave Devices

40 Compatibility *L.H. Hemming, V. Ungvichian, J.M. Roman, M.A. Uman, M. Rubinstein*
Grounding, Shielding, and Filtering • Spectrum, Specifications, and Measurement Techniques • Lightning

41 Radar *M.L. Belcher, J.T. Nessmith, J.C. Wiltse*
Pulse Radar • Continuous Wave Radar

42 Lightwave *S.O. Agbo, A.H. Cherin, B.K. Tariyal*
Lightwave Waveguides • Optical Fibers and Cables

43 Solid State Circuits *I.J. Bahl*
Amplifiers • Oscillators • Multipliers • Mixers • Control Circuits • Summary and Future Trends

44 Three-Dimensional Analysis *C.W. Trowbridge*
The Field Equations • Numerical Methods • Modern Design Environment

45 Computational Electromagnetics *E.K. Miller*
Background Discussion • Analytical Issues in Developing a Computer Model • Numerical Issues in Developing a Computer Model • Some Practical Considerations • Ways of Decreasing Computer Time • Validation, Error Checking, and Error Analysis

Banmali S. Rawat
University of Nevada, Reno

ELECTRIC AND MAGNETIC FORCES are among the five original forces in the universe. These forces are important as we are affected by them almost every instant. Electromagnetics is the combined effect of electric and magnetic fields. Today's scientific development to a great extent is based on the electromagnetic fields, their propagation, and varying effects under different boundary conditions. Very few subjects are understood as thoroughly as electromagnetics and have such wide applications. Electricity, telephones, radio, television, datalinks, medical electronics, radar, remote sensing, etc.—all have considerable impact on human life. Now that impact is being carried out further with optical fiber technology, which is also based on the concept of electromagnetic wave propagation. All of human society has been revolutionized by electromagnetics, but still our understanding is not complete. As H.G. Wells once wrote, and this is still true, "The past is but a beginning of a beginning, and all that is and has been is but the twilight of the dawn."

This section focuses on the basic electromagnetic field concepts, wave propagation, devices, circuits, and other applications. The electric fields which are produced by stationary or moving charges are described in Chapter 35. Maxwell's equations and their solutions under different boundary conditions help in determining the electric field components and resulting effects. The next chapter describes the magnetic fields and magnetic effects due to moving charges or current. These magnetic fields are also governed by Maxwell's equations and their solutions are obtained for different boundary conditions. Particular magnetic materials with an assemblage of ferromagnetic particles in a nonferromagnetic matrix are useful as audio or video tapes. This subject is investigated in Chapter 36 to provide insight into the recording mechanism of the music we hear all the time. The time-varying electromagnetic field propagation in space or in transmission lines provides the concept of radio communication as discussed in Chapter 37. Another article in the chapter analyzes the transmission of energy through waveguides and microstriplines. Microstriplines have become the basic building blocks for microwave integrated circuits (MICS). For the propagation of electromagnetic fields in space, properly matched antennas between generator and space are required, as described in Chapter 38. Wire and aperture antennas are also described.

The high-frequency or microwave-frequency electromagnetic field concepts are helpful in studying the microwave devices as discussed in Chapter 39. The electromagnetic compatibility (EMC) study in the following chapter is important for proper functioning of microwave devices and circuits. The important application of electromagnetic radiation in the form of radar, discussed in Chapter 41, is useful not only for defense but in remote sensing and weather forecasting also. The next chapter explains the propagation of light through waveguides and optical fibers/cables. Optical fiber technology is an emerging technology and is affecting every facet of human life. Microwave circuits are the practical realization of electromagnetic field concepts and are discussed in Chapter 42. With the arrival of sophisticated software packages and high-speed computers, now it is possible and worthwhile to do 3-D analysis and computer modeling of electromagnetic fields in the circuits or devices, as discussed in the last two chapters of this section. This is helpful in the accurate design of microwave components and circuits. All the topics mentioned in this introduction are discussed in detail in their respective chapters.

Nomenclature

Symbol	Quantity	Unit
A_e	actual effective aperture of antenna	m^2
A_{em}	maximum effective aperture of antenna	m^2
α	attenuation constant	neper/m
b	Doppler filter bandwidth	Hz
B	magnetic flux density	Wb/m^2
β	phase constant	rad/m
c	velocity of light in vacuum	2.998×10^8 m/s
D	electric displacement	C/m^2
D	divergent factor	
D	directivity of antenna	dB
δ	penetrating depth	m
E	electric field intensity	V/m
ϵ	permittivity	F/m
ϵ_0	8.854×10^{-12}	F/m
f_D	Doppler frequency	Hz
F	receiver noise figure	dB
g_m	transconductance	S
G	gain of antenna	dB
γ	propagation constant	m^{-1}
Γ	Fresnel reflection coefficient	
H	magnetic field intensity	A/m
η	intrinsic impedance	Ω
J	electric current density	A/m^2
J	electric charge density	C/m^2
k	wavenumber	
k	radiation efficiency factor	
L	antenna loss	dB
λ	wavelength	m
μ	permeability	H/m
μ_0	$4\pi \times 10^{-7}$	H/m
P	Poynting vector	W/m^2
P_T	average power	W
Ψ	grazing angle	degree
q	electronic charge	1.6×10^{-19} C
R	detection range of target	m
ρ_s	roughness coefficient	
$S(\theta)$	shadowing function	
U	unilateral power gain	dB

Kong, J.A. “Electromagnetic Fields”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

35

Electromagnetic Fields

35.1 Maxwell Equations
35.2 Constitutive Relations
Anisotropic and Bianisotropic Media • Biisotropic Media • Constitutive Matrices
35.3 Wave Equations and Wave Solutions
Wave Solution • Wave Vector $\bar{k}$ • Wavenumbers k

Jin Au Kong
Massachusetts Institute of Technology

35.1 Maxwell Equations[1]

The fundamental equations of electromagnetic theory were established by James Clerk Maxwell in 1873. In three-dimensional vector notation, the Maxwell equations are

$$\nabla \times \bar{E}(\bar{r},t) + \frac{\partial}{\partial t}\bar{B}(\bar{r},t) = 0 \tag{35.1}$$

$$\nabla \times \bar{H}(\bar{r},t) - \frac{\partial}{\partial t}\bar{D}(\bar{r},t) = \bar{J}(\bar{r},t) \tag{35.2}$$

$$\nabla \cdot \bar{B}(\bar{r},t) = 0 \tag{35.3}$$

$$\nabla \cdot \bar{D}(\bar{r},t) = \rho(\bar{r},t) \tag{35.4}$$

where $\bar{E}$, $\bar{B}$, $\bar{H}$, $\bar{D}$, $\bar{J}$, and ρ are real functions of position and time.

$$\begin{aligned}
\bar{E}(\bar{r},t) &= \text{electric field strength} && (\text{volts/m}) \\
\bar{B}(\bar{r},t) &= \text{magnetic flux density} && (\text{webers/m}^2) \\
\bar{H}(\bar{r},t) &= \text{magnetic field strength} && (\text{amperes/m}) \\
\bar{D}(\bar{r},t) &= \text{electric displacement} && (\text{coulombs/m}^2) \\
\bar{J}(\bar{r},t) &= \text{electric current density} && (\text{amperes/m}^2) \\
\rho(\bar{r},t) &= \text{electric charge density} && (\text{coulombs/m}^3)
\end{aligned}$$

[1]This chapter is an abridged version of Chapter 1 in *Electromagnetic Wave Theory* (J. A. Kong), New York: Wiley-Interscience, 1990.

Equation (35.1) is Faraday's induction law. Equation (35.2) is the generalized Ampere's circuit law. Equations (35.3) and (35.4) are Gauss' laws for **magnetic and electric fields.** Taking the divergence of (35.2) and introducing (35.4), we find that

$$\nabla \cdot \bar{J}(\bar{r},t) + \frac{\partial}{\partial t}\rho(\bar{r},t) = 0 \tag{35.5}$$

This is the conservation law for electric charge and current densities. Regarding (35.5) as a fundamental equation, we can use it to derive (35.4) by taking the divergence of (35.2). Equation (35.3) can also be derived by taking the divergence of (35.1) which gives $\partial(\nabla \cdot \bar{B}(\bar{r},t))/\partial t = 0$ or that $\nabla \cdot \bar{B}(\bar{r},t)$ is a constant independent of time. Such a constant, if not zero, then implies the existence of magnetic monopoles similar to free electric charges. Since magnetic monopoles have not been found to exist, this constant must be zero and we arrive at (35.3).

35.2 Constitutive Relations

The Maxwell equations are fundamental laws governing the behavior of electromagnetic fields in free space and in media. We have so far made no reference to the various material properties that provide connections to other disciplines of physics, such as plasma physics, continuum mechanics, solid-state physics, fluid dynamics, statistical physics, thermodynamics, biophysics, etc., all of which interact in one way or another with electromagnetic fields. We did not even mention the Lorentz force law, which constitutes a direct link to mechanics. It is time to state how we are going to account for this vast "outside" world. From the electromagnetic wave point of view, we shall be interested in how electromagnetic fields behave in the presence of media, whether the wave is diffracted, refracted, or scattered. Whatever happens to a medium, whether it is moved or deformed, is of secondary interest. Thus we shall characterize material media by the so-called constitutive relations that can be classified according to the various properties of the media.

The necessity of using constitutive relations to supplement the Maxwell equations is clear from the following mathematical observations. In most problems we shall assume that sources of electromagnetic fields are given. Thus $\bar{J}$ and ρ are known and they satisfy the conservation law (35.5). Let us examine the Maxwell equations and see if there are enough equations for the number of unknown quantities. There are a total of 12 scalar unknowns for the four field vectors $\bar{E}$, $\bar{H}$, $\bar{B}$, and $\bar{D}$. As we have learned, Eqs. (35.3) and (35.4) are not independent equations; they can be derived from Eqs. (35.1), (35.2), and (35.5). The independent equations are Eqs. (35.1) and (35.2), which constitute six scalar equations. Thus we need six more scalar equations. These are the constitutive relations.

The constitutive relations for an isotropic medium can be written simply as

$$\bar{D} = \epsilon\bar{E} \qquad \text{where } \epsilon = \text{permittivity} \tag{35.6a}$$

$$\bar{B} = \mu\bar{H} \qquad \text{where } \mu = \text{permittivity} \tag{35.6b}$$

By isotropy we mean that the field vector $\bar{E}$ is parallel to $\bar{D}$ and the field vector $\bar{H}$ is parallel to $\bar{B}$. In free space void of any matter, $\mu = \mu_o$ and $\epsilon = \epsilon_o$,

$$\mu_o = 4\pi \times 10^{-7} \qquad \text{henry/meter}$$

$$\epsilon_o \approx 8.85 \times 10^{-12} \qquad \text{farad/meter}$$

Inside a material medium, the permittivity ϵ is determined by the electrical properties of the medium and the permeability μ by the magnetic properties of the medium.

A dielectric material can be described by a free-space part and a part that is due to the material alone. The material part can be characterized by a polarization vector $\bar{P}$ such that $\bar{D} = \epsilon_o \bar{E} + \bar{P}$. The polarization $\bar{P}$ symbolizes the electric dipole moment per unit volume of the dielectric material. In the presence of an external electric field, the polarization vector may be caused by induced dipole moments, alignment of the permanent dipole moments of the medium, or migration of ionic charges.

A magnetic material can also be described by a free-space part and a part characterized by a magnetization vector $\bar{M}$ such that $\bar{B} = \mu_o \bar{H} + \mu_o \bar{M}$. A medium is diamagnetic if $\mu \leq \mu_o$ and paramagnetic if $\mu \geq \mu_o$. Diamagnetism is caused by induced magnetic moments that tend to oppose the externally applied magnetic field. Paramagnetism is due to alignment of magnetic moments. When placed in an inhomogeneous magnetic field, a diamagnetic material tends to move toward regions of weaker magnetic field and a paramagnetic material toward regions of stronger magnetic field. Ferromagnetism and antiferromagnetism are highly nonlinear effects. Ferromagnetic substances are characterized by spontaneous magnetization below the Curie temperature. The medium also depends on the history of applied fields, and in many instances the magnetization curve forms a hysteresis loop. In an antiferromagnetic material, the spins form sublattices that become spontaneously magnetized in an antiparallel arrangement below the Néel temperature.

Anisotropic and Bianisotropic Media

The constitutive relations for anisotropic media are usually written as

$$\bar{D} = \bar{\bar{\epsilon}} \cdot \bar{E} \quad \text{where } \bar{\bar{\epsilon}} = \text{permittivity tensor} \tag{35.7a}$$

$$\bar{B} = \bar{\bar{\mu}} \cdot \bar{H} \quad \text{where } \bar{\bar{\mu}} = \text{permeability tensor} \tag{35.7b}$$

The field vector $\bar{E}$ is no longer parallel to $\bar{D}$, and the field vector $\bar{H}$ is no longer parallel to $\bar{B}$. A medium is *electrically anisotropic* if it is described by the permittivity tensor $\bar{\bar{\epsilon}}$ and a scalar permeability μ, and *magnetically anisotropic* if it is described by the permeability tensor $\bar{\bar{\mu}}$ and a scalar permittivity ϵ. Note that a medium can be both electrically and magnetically anisotropic as described by both $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ in Eq. (35.7).

Crystals are described in general by symmetric permittivity tensors. There always exists a coordinate transformation that transforms a symmetric matrix into a diagonal matrix. In this coordinate system, called the *principal system,*

$$\bar{\bar{\epsilon}} = \begin{bmatrix} \epsilon_x & 0 & 0 \\ 0 & \epsilon_y & 0 \\ 0 & 0 & \epsilon_z \end{bmatrix} \tag{35.8}$$

The three coordinate axes are referred to as the principal axes of the crystal. For cubic crystals, ϵ_x $\bar{\bar{\epsilon}}_y$ $\bar{\bar{\epsilon}}_z$ and they are isotropic. In tetragonal, hexagonal, and rhombohedral crystals, two of the three parameters are equal. Such crystals are *uniaxial.* Here there is a two-dimensional degeneracy; the principal axis that exhibits this anisotropy is called the optic axis. For a uniaxial crystal with

$$\bar{\bar{\epsilon}} = \begin{bmatrix} \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon_z \end{bmatrix} \tag{35.9}$$

the z axis is the optic axis. The crystal is *positive uniaxial* if $\epsilon_z > \epsilon$; it is *negative uniaxial* if $\epsilon_z < \epsilon$. In orthorhombic, monoclinic, and triclinic crystals, all three crystallographic axes are unequal. We have $\epsilon_x \neq \epsilon_y \neq \epsilon_z$, and the medium is *biaxial.*

For isotropic or anisotropic media, the constitutive relations relate the two electric field vectors and the two magnetic field vectors by either a scalar or a tensor. Such media become polarized when placed in an electric field and become magnetized when placed in a magnetic field. A bianisotropic medium provides the cross coupling between the electric and magnetic fields. The constitutive relations for a bianisotropic medium can be written as

$$\overline{D} = \overline{\overline{\epsilon}} \cdot \overline{E} + \overline{\overline{\xi}} \cdot \overline{H} \tag{35.10a}$$

$$\overline{B} = \overline{\overline{\zeta}} \cdot \overline{E} + \overline{\overline{\epsilon}} \cdot \overline{H} \tag{35.10b}$$

When placed in an electric or a magnetic field, a bianisotropic medium becomes both polarized and magnetized.

Magnetoelectric materials, theoretically predicted by Dzyaloshinskii and by Landau and Lifshitz, were observed experimentally in 1960 by Astrov in antiferromagnetic chromium oxide. The constitutive relations that Dzyaloshinskii proposed for chromium oxide have the following form:

$$\overline{D} = \begin{bmatrix} \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon_z \end{bmatrix} \cdot \overline{E} + \begin{bmatrix} \xi & 0 & 0 \\ 0 & \xi & 0 \\ 0 & 0 & \xi_z \end{bmatrix} \cdot \overline{H} \tag{35.11a}$$

$$\overline{B} = \begin{bmatrix} \xi & 0 & 0 \\ 0 & \xi & 0 \\ 0 & 0 & \xi_z \end{bmatrix} \cdot \overline{E} + \begin{bmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu_z \end{bmatrix} \cdot \overline{H} \tag{35.11b}$$

It was then shown by Indenbom and by Birss that 58 magnetic crystal classes can exhibit the magnetoelectric effect. Rado proved that the effect is not restricted to antiferromagnetics; ferromagnetic gallium iron oxide is also magnetoelectric.

Biisotropic Media

In 1948, the gyrator was introduced by Tellegen as a new element, in addition to the resistor, the capacitor, the inductor, and the ideal transformer, for describing a network. To realize his new network element, Tellegen conceived of a medium possessing constitutive relations of the form

$$\overline{D} = \epsilon\overline{E} + \xi\overline{H} \tag{35.12a}$$

$$\overline{B} = \xi\overline{E} + \mu\overline{H} \tag{35.12b}$$

where $\xi^2/\mu\epsilon$ is nearly equal to 1. Tellegen considered that the model of the medium had elements possessing permanent electric and magnetic dipoles parallel or antiparallel to each other, so that an applied electric field that aligns the electric dipoles simultaneously aligns the magnetic dipoles, and a magnetic field that aligns the magnetic dipoles simultaneously aligns the electric dipoles. Tellegen also wrote general constitutive relations Eq. (35.10) and examined the symmetry properties by energy conservation.

Chiral media, which include many classes of sugar solutions, amino acids, DNA, and natural substances, have the following constitutive relations

$$\overline{D} = \epsilon \overline{E} - \chi \frac{\partial \overline{H}}{\partial t} \tag{35.13a}$$

$$\overline{B} = \mu \overline{H} + \chi \frac{\partial \overline{E}}{\partial t} \tag{35.13b}$$

where χ is the chiral parameter. Media characterized by the constitutive relations, Eqs. (35.12) and (35.13), are biisotropic media.

Media in motion were the first bianisotropic media to receive attention in electromagnetic theory. In 1888, Roentgen discovered that a moving dielectric becomes magnetized when it is placed in an electric field. In 1905, Wilson showed that a moving dielectric in a uniform magnetic field becomes electrically polarized. Almost any medium becomes bianisotropic when it is in motion.

The bianisotropic description of material has fundamental importance from the point of view of relativity. The principle of relativity postulates that all physical laws of nature must be characterized by mathematical equations that are form-invariant from one observer to the other. For electromagnetic theory, the Maxwell equations are form-invariant with respect to all observers, although the numerical values of the field quantities may vary from one observer to another. The constitutive relations are form-invariant when they are written in bianisotropic form.

Constitutive Matrices

Constitutive relations in the most general form can be written as

$$c\overline{D} = \overline{\overline{P}} \cdot \overline{E} + \overline{\overline{L}} \cdot c\overline{B} \tag{35.14a}$$

$$\overline{H} = \overline{\overline{M}} \cdot \overline{E} + \overline{\overline{Q}} \cdot c\overline{B} \tag{35.14b}$$

where $c = 3 \times 10^8$ m/s is the velocity of light in vacuum, and $\overline{\overline{P}}$, $\overline{\overline{Q}}$, $\overline{\overline{L}}$, and $\overline{\overline{M}}$ are all 3×3 matrices. Their elements are called *constitutive parameters*. In the definition of the constitutive relations, the constitutive matrices $\overline{\overline{L}}$ and $\overline{\overline{M}}$ relate electric and magnetic fields. When $\overline{\overline{L}}$ and $\overline{\overline{M}}$ are not identically zero, the medium is *bianisotropic*. When there is no coupling between electric and magnetic fields, $\overline{\overline{L}} = \overline{\overline{M}} = 0$ and the medium is *anisotropic*. For an anisotropic medium, if $\overline{\overline{P}} = c\epsilon \overline{\overline{I}}$ and $= Q = (1/c\mu)\ \overline{\overline{I}}$ with $\overline{\overline{I}}$ denoting the 3×3 unit matrix, the medium is *isotropic*. The reason that we write constitutive relations in the present form is based on relativistic considerations. First, the fields $\overline{E}$ and $c\,\overline{B}$ form a single tensor in four-dimensional space, and so do $c\overline{D}$ and $\overline{H}$. Second, constitutive relations written in the form Eq. (35.14) are Lorentz-covariant.

Equation (35.14) can be rewritten in the form

$$\begin{bmatrix} c\overline{D} \\ \overline{H} \end{bmatrix} = \overline{\overline{C}} \cdot \begin{bmatrix} \overline{E} \\ c\overline{B} \end{bmatrix} \tag{35.15a}$$

and $\overline{\overline{C}}$ is a 6×6 constitutive matrix:

$$\overline{\overline{C}} = \begin{bmatrix} \overline{\overline{P}} & \overline{\overline{L}} \\ \overline{\overline{M}} & \overline{\overline{Q}} \end{bmatrix} \tag{35.15b}$$

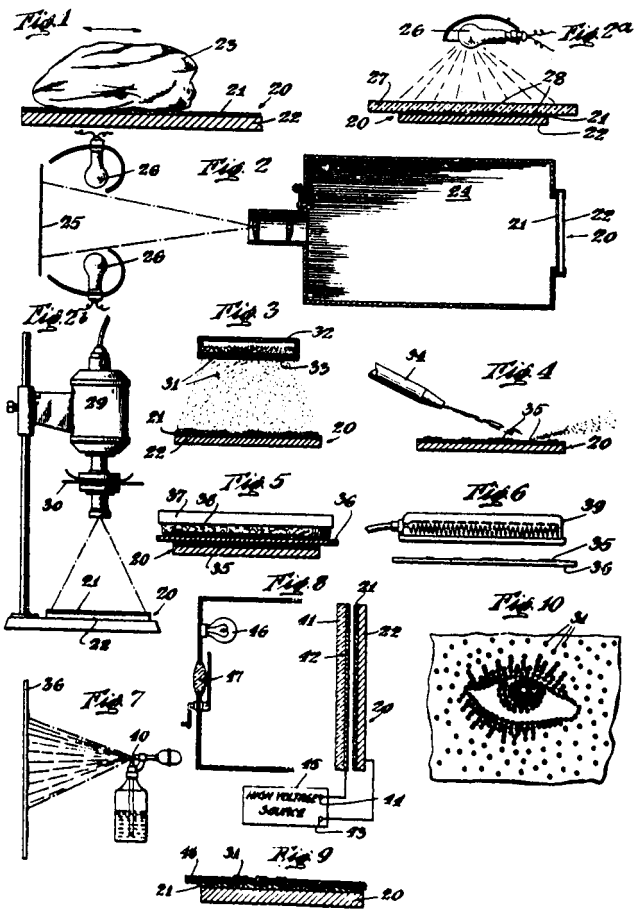
which has the dimension of admittance.

ELECTROPHOTOGRAPHY

Chester F. Carlson
Patented October 6, 1942
#2,297,691

An excerpt from Chester Carlson's patent application:

A feature of the present invention resides in the use of photoelectric or photoconductive materials for photographic purposes. In its preferred form the invention involves the use of materials which are insulators in the dark but become partial conductors when illuminated. These materials respond to light, being slightly conductive whenever they are illuminated and again becoming insulating when the light is cut off. They can be called photoconductive insulating materials.



Working in the patent department at the P.R. Mallory Company in the 1930s, Carlson became frustrated trying to obtain copies of patent drawings and specifications. Unlike others attempting chemical photographic methods, he used the principles of electrostatics to produce his first dry copy in 1938. A photoconductive plate connected to an electric charge was exposed to the desired image. The plate retained the electric charge on the dark areas of the image and lost it on the white areas. Dusting the plate with a powder reproduced the image. Xerox Corporation negotiated rights to the process in 1947 and introduced its first office copier in 1968.

The constitutive matrix $\overline{\overline{C}}$ may be functions of space–time coordinates, thermodynamical and continuum-mechanical variables, or electromagnetic field strengths. According to the functional dependence of $\overline{\overline{C}}$, we can classify the various media as (1) inhomogeneous if $\overline{\overline{C}}$ is a function of space coordinates, (2) nonstationary if $\overline{\overline{C}}$ is a function of time, (3) time-dispersive if $\overline{\overline{C}}$ is a function of time derivatives, (4) spatial-dispersive if $\overline{\overline{C}}$ is a function of spatial derivatives, (5) nonlinear if $\overline{\overline{C}}$ is a function of the electromagnetic field, and so forth. In the general case $\overline{\overline{C}}$ may be a function of integral-differential operators and coupled to fundamental equations of other physical disciplines.

35.3 Wave Equations and Wave Solutions

The Maxwell equations in differential form are valid at all times for every point in space. First we shall investigate solutions to the Maxwell equations in regions void of source, namely, in regions where $\overline{J} = \rho = 0$. This, of course, does not mean that there is no source anywhere in all space. Sources must exist outside the regions of interest in order to produce fields in these regions. From the source-free Maxwell equations, a wave equation for the electric field $\overline{E}$ can be easily derived for isotropic permittivity ϵ and permeability μ

$$\nabla^2 \overline{E} - \mu\epsilon \frac{\partial^2}{\partial t^2} \overline{E} = 0 \tag{35.16}$$

The Laplacian operator ∇^2 in a rectangular coordinate system is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

The wave Eq. (35.16) is a second-order partial differential equation of space and time coordinates *x*, *y*, *z*, and *t*.

Wave Solution

The simplest solution to Eq. (35.16) for the electric field $\overline{E}$ is

$$\overline{E} = \hat{x} E_0 \cos(kz - \omega t) = \hat{x} E_x(z, t) \tag{35.17}$$

Substituting Eqs. (35.17) in (35.16) we find that the following equation, called the dispersion relation, which relates ω and k, must be satisfied:

$$k^2 = \omega^2 \mu\epsilon \tag{35.18}$$

There are two points of view useful in the study of a space–time varying quantity such as $E_x(z,t)$. The first is to examine the time variation at fixed points in space. The second is to examine spatial variation at fixed times, a process that amounts to taking a series of pictures.

We first fix our attention to one particular point in space, say $z = 0$. We then have the electric vector $E_x(z,t) = E_0 \cos \omega t$. Plotted as a function of time, we find that the waveform repeats itself in time as $\omega t = 2m\pi$ for any integer m. The period is defined as the time T for which $\omega T = 2\pi$. The frequency f is defined as $f = 1/T$ which gives

$$f = \frac{\omega}{2\pi}$$

Since $\omega = 2\pi f$, ω is the angular frequency of the wave.

To examine wave behavior from the other point of view, we let $\omega t = 0$ and plot $E_x(z,t)$. The waveform repeats itself in space when $kz = 2m\pi$ for integer values of m. The wavelength λ is defined as the distance for which $k\lambda = 2\pi$. Thus $\lambda = 2\pi/k$, or

$$k = \frac{2\pi}{\lambda}$$

We call k the wavenumber which is equal to the number of wavelengths in a distance of 2π and has the dimension inverse length.

Wave Vector $\bar{k}$

The solution for the electric field $\bar{E}$ in Eq. (35.17) represents an electromagnetic wave propagating in the $\hat{z}$-direction. For a wave propagating in a general direction, we define a wave vector

$$\bar{k} = \hat{x}k_x + \hat{y}k_y + \hat{z}k_z \tag{35.19}$$

It is easily verified that the electric field

$$\bar{E}(\bar{r},t) = \bar{E}_0 \cos(k_x x + k_y y + k_z z - \omega t) \tag{35.20}$$

is a solution to Eq. (35.16), where $\bar{E}_0$ is a constant vector.

The dispersion relation corresponding to Eq. (35.18) is obtained by substituting Eq. (35.20) in Eq. (35.16) which yields

$$k_x^2 + k_y^2 + k_z^2 = \omega^2\mu\epsilon$$

We may write the solution, Eq. (35.20), in the following form:

$$\bar{E}(\bar{r}, t) = \bar{E}_0 \cos(\bar{k} \cdot \bar{r} - \omega t)$$

where

$$\bar{r} = \hat{x}x + \hat{y}y + \hat{x}x$$

is a position vector. The wave vector $\bar{k}$ is often referred to simply as the $\bar{k}$ vector.

Wavenumbers k

The wavenumber k is the magnitude of the wave vector $\bar{k}$ and is of more fundamental importance in electromagnetic wave theory than both of the more popular concepts of wavelength λ and frequency f. In Fig. 35.1, we illustrate the electromagnetic wave spectrum according to the free space wavenumber $k = k_o = \omega/c$. The corresponding values of frequency and wavelength are $f = ck_o/2\pi$ and $\lambda = 2\pi/k_o$. It is useful to define a fundamental unit K_o such that for free space $k_o = 1\mathrm{K}_o = 2\pi\ \mathrm{m}^{-1}$. Thus $k_o = A\ \mathrm{K}_o$ corresponds to $\lambda = 1/A$ m and $f = 3 \times 10^8\ A$ Hz. The photon energy in electronvolts is calculated from $\hbar\omega = \hbar ck$ where $\hbar = 1.05 \times 10^{-34}$ Joule-sec is Planck's constant divided by 2π and the electron charge is $q = 1.6 \times 10^{-19}$ C. Thus $\hbar\omega = (2\pi\hbar c/q)k_o \approx 1.26 \times 10^{-6}\ k_o$ and $k_o = A\ \mathrm{K}_o$ corresponds to $1.26 \times 10^{-6}\ A$ eV.

Defining Terms

Electric field: State of a region in which charged bodies are subject to forces by virtue of their charge, the force acting on a unit positive charge.

Magnetic field: State produced by electric charge in motion and evidenced by a force exerted on a moving charge in the field.

Magnetic flux: Summation obtained by integrating flux density over an area.

Magnetic flux density: Measure of the strength and direction of a magnetic field at a point.

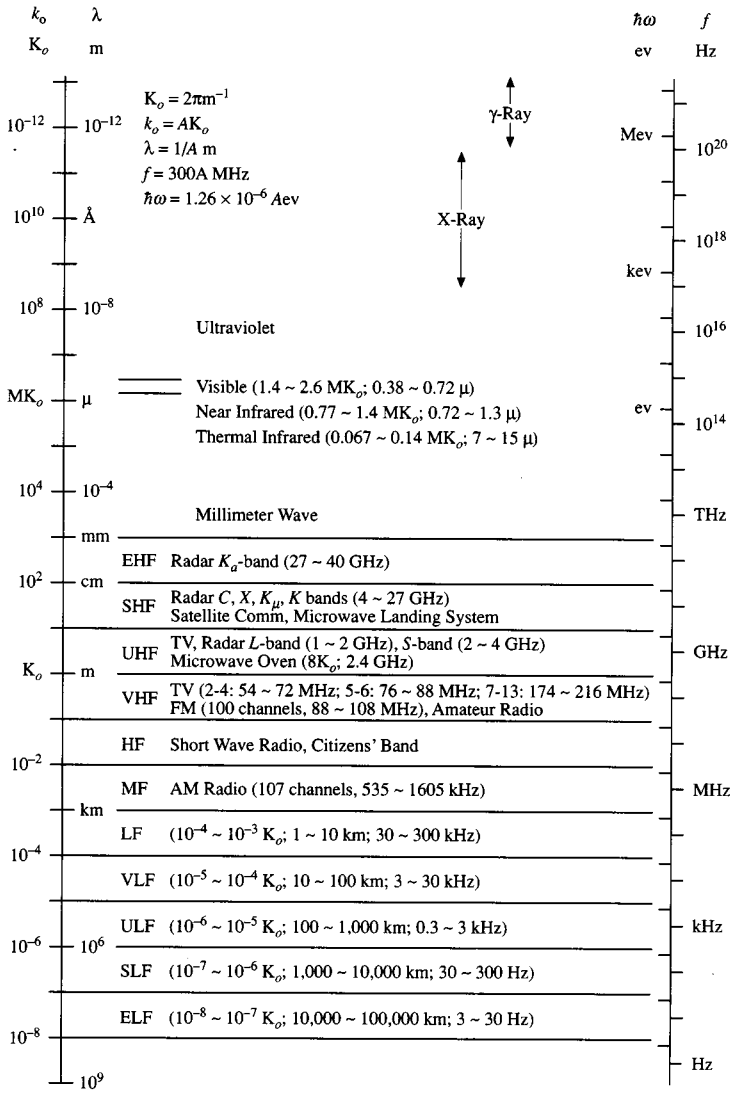


FIGURE 35.1 Electromagnetic wave spectrum.

Related Topics

39.1 Passive Microwave Devices • 44.2 The Field Equations

Reference

J.A. Kong, *Electromagnetic Wave Theory,* New York: Wiley-Interscience, 1990, chap. 1.

Further Information

IEEE Transactions on Microwave Theory and Techniques
IEEE Transactions on Antennas and Propagation

Bate, G., Kryder, M.H. “Magnetism and Magnetic Fields”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

36

Magnetism and Magnetic Fields

36.1 Magnetism
Static Magnetic Fields • Time-Dependent Electric and Magnetic Fields • Magnetic Flux Density • Relative Permeabilities • Forces on a Moving Charge • Time-Varying Magnetic Fields • Maxwell's Equations • Dia- and Paramagnetism • Ferromagnetism and Ferrimagnetism • Intrinsic Magnetic Properties • Extrinsic Magnetic Properties • Amorphous Magnetic Materials

36.2 Magnetic Recording
Fundamentals of Magnetic Recording • The Recording Process • The Readback Process • Magnetic Recording Media • Magnetic Recording Heads • Conclusions

Geoffrey Bate
Consultant in Information Storage Technology

Mark H. Kryder
Carnegie Mellon University

36.1 Magnetism

Geoffrey Bate

Static Magnetic Fields

To understand the phenomenon of magnetism we must also consider electricity and vice versa. A stationary electric charge produces, at a point a fixed distance from the charge, a static (i.e., time-invariant) electric field. A moving electric charge, i.e., a current, produces at the same point a time-dependent electric field and a magnetic field, d**H**, whose magnitude is constant if the electric current, I, represented by the moving electric charge, is constant.

Fields from Constant Currents

Figure 36.1 shows that the direction of the magnetic field is perpendicular both to the current I and to the line, **R**, from the element d**L** of the current to a point, P, where the magnetic field, d**H**, is being calculated or measured.

$$\mathrm{d}\mathbf{H} = I\,\mathrm{d}\mathbf{L} \times \mathbf{R}/4\pi R^3 \quad \text{A/m when I is in amps and d}\mathbf{L}\text{ and } R \text{ are in meters}$$

If the thumb of the right hand points in the direction of the current, then the fingers of the hand curl in the direction of the magnetic field. Thus, the stream lines of H, i.e., the lines representing at any point the direction of the H field, will be an infinite set of circles having the current as center. The magnitude of the field $H_\phi = I/2\pi R$ A/m. The line integral of H about any closed path around the current is $\oint \mathbf{H} \cdot \mathrm{d}\mathbf{L} = I$. This relationship (known as Ampère's circuital law) allows one to find formulas for the magnetic field strength for a variety of symmetrical coil geometries, e.g.,

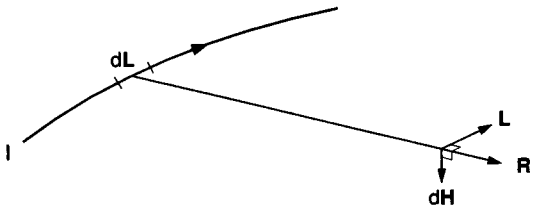


FIGURE 36.1 A current I flowing through a small segment d**L** of a wire produces at a distance **R** a magnetic field whose direction d**H** is perpendicular both to **R** and d**L**.

1. At a radius, ρ, between the conductors of a coaxial cable

$$H_\phi = I/2\pi\rho \text{ A/m}$$

2. Between two infinite current sheets in which the current, **K**, flows in opposite directions

$$\mathbf{H} = \mathbf{K} \times \mathbf{a}_n$$

   where $\mathbf{a}_n$ is the unit vector normal to the current sheets
3. Inside an infinitely long, straight solenoid of diameter d, having N turns closely wound

$$\mathbf{H} = NI/d \text{ A/m}$$

4. Well inside a toroid of radius ρ, having N closely wound turns

$$\mathbf{H} = NI/2\pi\rho \cdot \mathbf{a}_\phi \text{ A/m}$$

Applying Stokes' theorem to Ampère's circuital law we find the point form of the latter.

$$\nabla \times \mathbf{H} = \mathbf{J}$$

where **J** is the current density in amps per square meter.

Time-Dependent Electric and Magnetic Fields

A constant current I produces a constant magnetic field **H** which, in turn, polarizes the medium containing **H**. While we cannot obtain isolated magnetic poles, it is possible to separate the "poles" by a small distance to create a magnetic dipole (i.e., to *polarize* the medium), and the dipole moment (the product of the pole strength and the separation of the poles) per unit volume is defined as the *magnetization M*. The units are emu/cc in the cgs system and amps per meter in the SI system of units. Because it is usually easier to determine the mass of a sample than to determine its volume, we also have a magnetization per unit mass, σ, whose units are emu/g or Am^2/kg. The conversion factors between cgs and SI units in magnetism are shown in Table 36.1.

The effects of the static and time-varying currents may be summarized as follows:

Static	$[\mathbf{I}]_o \rightarrow [\mathbf{H}]_o \rightarrow [\mathbf{M}]_o$
	↙ motion
Time-varying	$[\mathbf{I}]_t \rightarrow [\mathbf{H}]_t \rightarrow [\mathbf{M}]_t$

where the suffixes "o" and "t" signify *static* and *time-dependent*, respectively.

TABLE 36.1 Units in Magnetism

Quality	Symbol	cgs Units	×	Factor	=	SI units
		$B = H + 4\pi M$				$B = \mu_o(H + M)$
Magnetic flux density	B	gauss (G)	×	10^{-4}	=	tesla (T), Wb/m^2
Magnetic flux	Φ	maxwell (Mx) G · cm^2	×	10^{-8}	=	webers (Wb)
Magnetic potential difference (magnetomotive force)	U	gilbert (Gb)	×	$10/4\pi$	=	ampere (A)
Magnetic field strength	H	oersted (Oe)	×	$10^3/4\pi$	=	A/m
Magnetization (per volume)	M	emu/cc	×	10^3	=	A · m
Magnetization (per mass)	σ	emu/g	×	1	=	A · m^2/kg
Magnetic moment	m	emu	×	10^{-3}	=	A · m^2
Susceptibility (volume)	χ	dimensionless	×	4π	=	dimensionless
Susceptibility (mass)	κ	dimensionless	×	4π	=	dimensionless
Permeability (vacuum)	μ_o	dimensionless	×	$4\pi.10^{-7}$	=	Wb/A · m
Permeability (material)	μ	dimensionless	×	$4\pi.10^{-7}$	=	Wb/A · m
Bohr magneton	μ_B	$= 0.927 \times 10^{-20}$ erg/Oe	×	10^{-3}	=	Am^2
Demagnetizing factor	N	dimensionless	×	$1/4\pi$	=	dimensionless

Magnetic Flux Density

In the case of electric fields there is in addition to **E** an electric flux density field **D,** the lines of which begin on positive charges and end on negative charges. D is measured in coulombs per square meter and is associated with the electric field **E** (V/m) by the relation $\mathbf{D} = \epsilon_r\epsilon_o\mathbf{E}$ where ϵ_o is the *permittivity* of free space ($\epsilon_o = 8.854 \times 10^{-12}$ F/m) and ϵ_r is the (dimensionless) dielectric constant.

For magnetic fields there is a magnetic flux density $\mathbf{B}(\text{Wb/m}^2) = \mu_r\mu_o\mathbf{H}$, where μ_o is the *permeability* of free space ($\mu_o = 4\pi \times 10^{-7}$ H/m) and μ_r is the (dimensionless) permeability. In contrast to the lines of the **D** field, lines of **B** are closed, having no beginning or ending. This is not surprising when we remember that while isolated positive and negative charges exist, no magnetic monopole has yet been discovered.

Relative Permeabilities

The range of the relative permeabilities covers about six orders of magnitude (Table 36.2) whereas the range of dielectric constants is only three orders of magnitude.

Forces on a Moving Charge

A charged particle, q, traveling with a velocity v and subjected to a magnetic field experiences a force

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}$$

This equation reveals how the Hall effect can be used to determine whether the majority current carriers in a sample of a semiconductor are (negatively charged) electrons flowing, say, in the negative direction or (positively charged) holes flowing in the positive direction. The (transverse) force (Fig. 36.2) will be in the same direction in either case, but the *sign* of the charge transported to the voltage probe will be positive for holes and negative for electrons.

In general, when both electric and magnetic fields are present, the force experienced by the carriers is given by

$$\mathbf{F} = q\,(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

The Hall effect is the basis of widely used and sensitive instruments for measuring the intensity of magnetic fields over a range of 10^{-5} to 2×10^6 A/m.

TABLE 36.2 Relative Permeability, μ_r, of Some Diamagnetic, Paramagnetic, and Ferromagnetic Materials

Material	μr	M_s, A/m^2
Diamagnetics		
Bismuth	0.999833	
Mercury	0.999968	
Silver	0.9999736	
Lead	0.9999831	
Copper	0.9999906	
Water	0.9999912	
Paraffin wax	0.99999942	
Paramagnetics		
Oxygen (s.t.p.)	1.000002	
Air	1.00000037	
Aluminum	1.000021	
Tungsten	1.00008	
Platinum	1.0003	
Manganese	1.001	
Ferromagnetics		
Purified iron: 99.96% Fe	280,000	2.158
Motor-grade iron: 99.6% Fe	5,000	2.12
Permalloy: 78.5% Ni, 21.5% Fe	70,000	2.00
Supermalloy: 79% Ni, 15% Fe, 5% Mo, 0.5% Mn	1,000,000	0.79
Permendur: 49% Fe, 49% Ca, 2% V	5,000	2.36
Ferrimagnetics		
Manganese–zinc ferrite	750	0.34
	1,200	0.36
Nickel–zinc ferrite	650	0.29

Source: F. Brailsford, *Physical Principles of Magnetism*, London: Van Nostrand, 1966. With permission.

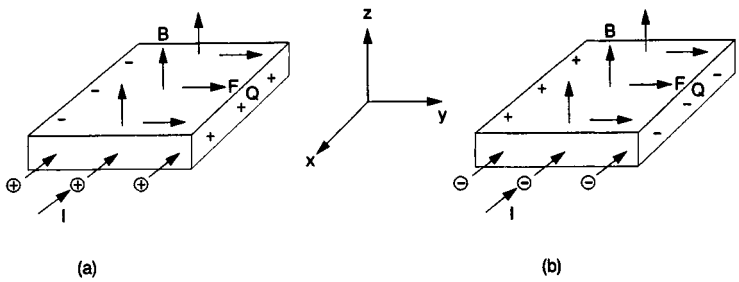


FIGURE 36.2 Hall effect. A magnetic field **B** applied to a block of semiconducting material through which a current I is flowing exerts a force $\mathbf{F} = \mathbf{V} \times \mathbf{B}$ on the current carriers (electrons or holes) and produces an electric charge on the right face of the block. The charge is positive if the carriers are holes and negative if the carriers are electrons.

Time-Varying Magnetic Fields

In 1831, 11 years after Oersted demonstrated that a current produced a magnetic field which could deflect a compass needle, Faraday succeeded in showing the converse effect—that a magnetic field could produce a current. The reason for the delay between the two discoveries was that it is only when a magnetic field is changing that an emf is produced.

$$\text{emf} = -\frac{d\Phi}{dt} \quad \text{V}$$

where $\Phi = BS$ = flux density (in gauss) × area S. The time-changing flux, $d\Phi/dt$, can happen as a result of

1. A changing magnetic field within a stationary circuit
2. A circuit moving through a steady magnetic field
3. A combination of 1 and 2

The electrical circuit may have N turns and then

$$\text{emf} = -N \frac{d\Phi}{dt}$$

We can write emf = $\mathbf{E} \cdot d\mathbf{L}$ and in the presence of changing magnetic fields or a moving electrical circuit $\mathbf{E} \cdot d\mathbf{L}$ is no longer required to be equal to 0 as it was for stationary fields and circuits.

Maxwell's Equations

Because the flux Φ can be written $\int \mathbf{B} \cdot d\mathbf{s}$ we have emf = $\mathbf{E} \cdot d\mathbf{L} = -d/dt\, \mathbf{B} \cdot d\mathbf{s}$, and by using Stokes' theorem

$$(\nabla \times \mathbf{E}) \cdot d\mathbf{s} = -d\mathbf{B}/dt\ d\mathbf{s}$$

or

$$\nabla \times \mathbf{E} = -d\mathbf{B}/dt$$

That is, a spatially changing *electric field* produces a time-changing *magnetic field*. This is one of Maxwell's equations linking electric and magnetic fields.

By a similar argument it can be shown that

$$\nabla \times \mathbf{H} = \mathbf{J} + d\mathbf{D}/dt$$

This is another of Maxwell's equations and shows a spatially changing *magnetic field* produces a time-changing *electric field*. The latter $d\mathbf{D}/dt$ can be treated as an electric current which flows through a dielectric, e.g., in a capacitor, when an alternating potential is applied across the plates. This current is called the *displacement current* to distinguish it from the conduction current which flows in conductors. The *conduction current* involves the movement of electrons from one electrode to the other through the conductor (usually a metal). The *displacement current* involves no translation of electrons or holes but rather an alternating polarization throughout the dielectric material which is between the plates of the capacitor.

From the last two equations we see a key conclusion of Maxwell: that in electromagnetic fields a time-varying magnetic field produces a spatially varying electric field and a time-varying electric field produces a spatially varying magnetic field.

Maxwell's equations in point form, then, are

$$\nabla \times \mathbf{E} = d\mathbf{B}/dt$$

$$\nabla \times \mathbf{H} = \mathbf{J} + d\mathbf{D}/dt$$

$$\nabla \cdot \mathbf{D} = \rho_v$$

$$\nabla \cdot \mathbf{B} = 0$$

These equations are supported by the following auxiliary equations:

$\mathbf{D} = \epsilon\mathbf{E}$ (displacement = permittivity × electric field intensity)

$\mathbf{B} = \mu\mathbf{H}$ (flux density = permeability × magnetic field intensity)

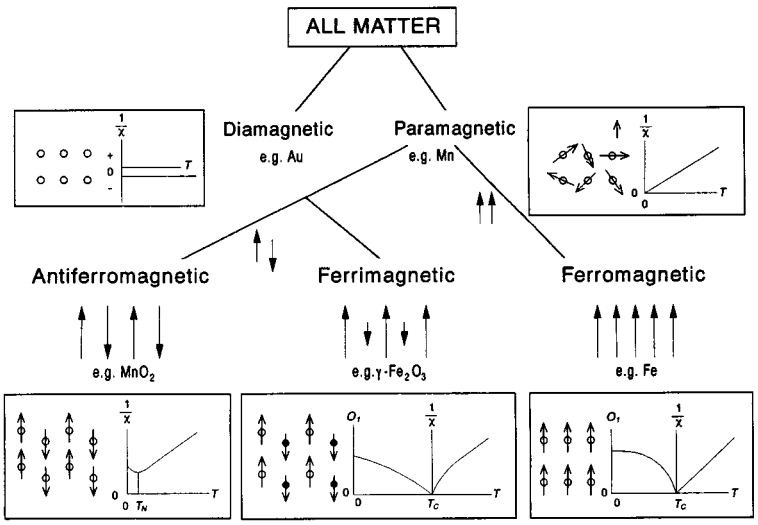


FIGURE 36.3 All matter consists of diamagnetic material (atoms having no permanent magnetic dipole moment) or paramagnetic material (atoms having magnetic dipole moment). Paramagnetic materials may be further divided into ferromagnetics, ferrimagnetics, and antiferromagnetics.

$$\mathbf{J} = \sigma\mathbf{E} \text{ (current density = conductivity } \times \text{ electric field strength)}$$

$$\mathbf{J} = \rho_v\mathbf{V} \text{ (current density = volume charge density } \times \text{ carrier velocity)}$$

$$\mathbf{D} = \boldsymbol{\epsilon}_o\mathrm{E} + \mathrm{P} \text{ (displacement as function of electric field and polarization)}$$

$$\mathbf{B} = \mu_o(\mathbf{H} + \mathbf{M}) \text{ (magnetic flux density as function of magnetic field strength and magnetization)}$$

$$\mathbf{P} = \chi_e\boldsymbol{\epsilon}_o\mathbf{E} \text{ (polarization = electric susceptibility } \times \text{ permittivity of free space } \times \text{ electrical field strength)}$$

$$\mathbf{M} = \chi_m\mu_o\mathbf{H} \text{ (magnetization = magnetic susceptibility } \times \text{ permeability of free space } \times \text{ magnetic field strength)}$$

The last two equations relate, respectively, the electric polarization **P** to the displacement $\mathbf{D} = \boldsymbol{\epsilon}_o\mathbf{E}$ and the magnetic moment **M** to the flux density $\mathbf{B} = \mu_o\mathbf{H}$. They apply only to "linear" materials, i.e., those for which **P** is linearly related to **E** and **M** to **H**. For magnetic materials we can say that nonlinear materials are usually of greater practical interest.

Dia- and Paramagnetism

The phenomenon of magnetism arises ultimately from moving electrical charges (electrons). The movement may be orbital around the nucleus or the other degree of freedom possessed by electrons which, by analogy with the motion of the planets, is referred to as *spin*. In technologically important materials, i.e., ferromagnetics and ferrimagnetics, spin is more important than orbital motion. Each arrow in Fig. 36.3 represents the *total spin* of an atom.

An atom may have a permanent magnetic moment, in which case it is referred to as belonging to a paramagnetic material, or the atom may be magnetized only when in the presence of a magnetic field, in which case it is called *diamagnetic*. Diamagnetics are magnetized in the *opposite direction* to that of the applied magnetic field, i.e., they display *negative* susceptibility (a measure of the induced magnetization per unit of applied magnetic field). Paramagnetics are magnetized in the *same* direction as the applied magnetic field, i.e., they

TABLE 36.3 The Occurrence of Ferromagnetism

	Cr	Mn	Fe	Co	Ni	Gd
Atomic number	24	25	26	27	28	64
Atomic spacing/diameter	1.30	1.47	1.63	1.82	1.97	1.57
Ferromagnetic moment/mass (Am^2/kg)						
At 293 K	—	—	217.75	161	54.39	0
At 0 K	—	—	221.89	162.5	57.50	250
Curie point, Θ_c K	—	—	1,043	1,400	631	289
Néel temp., Θ_n K	475	100	—	—	—	—

have *positive* susceptibility. *All* atoms are diamagnetic by virtue of their having electrons. *Some* atoms are also paramagnetic as well, but in this case they are called *paramagnetics* since paramagnetism is roughly a hundred times stronger than diamagnetism and overwhelms it. Faraday discovered that paramagnetics are attracted by a magnetic field and move toward the region of maximum field, whereas diamagnetics are repelled and move toward a field minimum.

The total magnetization of both paramagnetic and diamagnetic materials is zero in the absence of an applied field, i.e., they have zero **remanence**. Atomic paramagnetism is a necessary condition but not a sufficient condition for ferro- or ferrimagnetism, i.e., for materials having useful magnetic properties.

Ferromagnetism and Ferrimagnetism

To develop technologically useful materials, we need an additional force that ensures that the spins of the outermost (or almost outermost) electrons are mutually parallel. Slater showed that in iron, cobalt, and nickel this could happen if the distance apart of the atoms (D) was more than 1.5 times the diameter of the 3d electron shell (d). (These are the electrons, *near* the outside of atoms of iron, cobalt, and nickel, that are responsible for the strong paramagnetic moment of the atoms. Paramagnetism of the atoms is an essential prerequisite for ferro- or ferrimagnetism in a material.)

Slater's result suggested that, of these metals, iron, cobalt, nickel, and gadolinium should be ferromagnetic at room temperature, while chromium and manganese should not be ferromagnetic. This is in accordance with experiment. Gadolinium, one of the rare earth elements, is only weakly ferromagnetic in a cool room. Chromium and manganese in the elemental form narrowly miss being ferromagnetic. However, when manganese is alloyed with copper and aluminum ($Cu_{61}Mn_{24}Al_{15}$) to form what is known as a Heusler alloy [Crangle, 1962], it becomes ferromagnetic. The radius of the 3d electrons has not been changed by alloying, but the atomic spacing has been increased by a factor of 1.53/1.47. This small change is sufficient to make the difference between positive exchange, parallel spins, and ferromagnetism and negative exchange, antiparallel spins, and antiferromagnetism.

For all ferromagnetic materials there exists a temperature (the **Curie temperature**) above which the thermal disordering forces are stronger than the exchange forces that cause the atomic spins to be parallel. From Table 36.3 we see that in order of descending Curie temperature we have Co, Fe, Ni, Gd. From Fig. 36.4 we find that this is also the order of descending values of the exchange integral, suggesting that high positive values of the exchange integral are indicative of high Curie temperatures rather than high magnetic intensity in ferromagnetic materials.

Negative values of exchange result in an antiparallel arrangement of the spins of adjacent atoms and in antiferromagnetic materials (Fig. 36.3). Until 5 years ago, it was true to say that antiferromagnetism had no practical application. Thin films on antiferromagnetic materials are now used to provide the bias field which is used to linearize the response of some magnetoresistive reading heads in magnetic disk drives. Ferrimagnetism, also illustrated in Fig. 36.3, is much more widely used. It can be produced as soft, i.e., low **coercivity**, ferrites for use in magnetic recording and reading heads or in the core of transformers operating at frequencies up to tens of megahertz. High-coercivity, single-domain particles (which are discussed later) are used in very large quantities to make magnetic recording tapes and flexible disks γ-Fe_2O_3 and cobalt-impregnated iron oxides and to make barium ferrite, the most widely used material for permanent magnets.

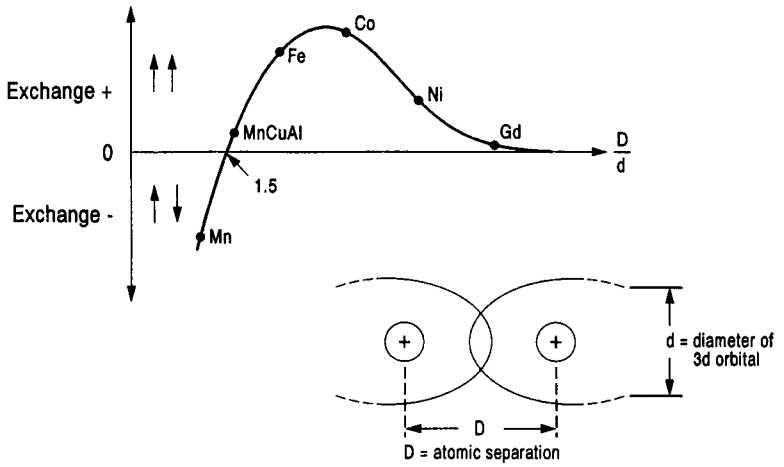


FIGURE 36.4 Quantum mechanical exchange forces cause a parallel arrangement of the spins of materials for which the ratio of atomic separation, D, is at least $1.5 \times d$, the diamter of the $3d$ orbital.

Intrinsic Magnetic Properties

Intrinsic magnetic properties are those properties that depend on the type of atoms and their composition and crystal structure, but not on the previous history of a particular sample. Examples of intrinsic magnetic properties are the *saturation* magnetization, Curie temperature, magnetocrystallic anisotropy, and magnetostriction.

Extrinsic magnetic properties depend on type, composition, and structure, but they also depend on the previous history of the sample, e.g., heat treatment. Examples of extrinsic magnetic properties include the technologically important properties of *remanent* magnetization, coercivity, and permeability. These properties can be substantially altered by heat treatment, quenching, cold-working the sample, or otherwise changing the size of the magnetic particle.

A ferromagnetic or ferrimagnetic material, on being heated, suffers a reduction of its magnetization (per unit mass, i.e., σ, and per unit volume, M). The slope of the curve of M_s vs. T increases with increasing temperature as shown in Fig. 36.5. This figure represents the conflict between the ordering tendency of the exchange interaction and the disordering effect of increasing temperature. At the Curie temperature, the order no longer exists and we have a paramagnetic material. The change from ferromagnetic or ferrimagnetic materials to paramagnetic is completely reversible on reducing the temperature to its initial value. Curie temperatures are always lower than melting points.

A single crystal of iron has the body-centered structure at room temperature. If the magnetization as a function of applied magnetic field is measured, the shape of the curve is found to depend on the direction of the field. This phenomenon is *magnetocrystalline anisotropy.* Iron has body-centered structure at room temperature, and the "easy" directions of magnetization are those directions parallel to the cube edges [100], [010], and [001] or, collectively, <100>. The hard direction of magnetization for iron is the body diagonal [111]. At higher temperatures, the anisotropy becomes smaller and disappears above 300°C.

Nickel crystals (face-centered cubic) have an easy direction of [111] and a hard direction of [100]. Cobalt has the hexagonal close-packed (HCP) structure and the hexagonal axis is the easy direction at room temperature.

Magnetocrystalline anisotropy plays a very important part in determining the coercivity of ferro- or ferrimagnetic materials, i.e., the field value at which the direction of magnetization is reversed.

Many magnetic materials change dimensions on becoming magnetized: the phenomenon is known as *magnetostriction* and can be positive, i.e., length increases, or negative. Magnetostriction plays an important role in determining the preferred direction of magnetization of *soft*, i.e., low H_c, films such as those of alloys of nickel and iron, known as *Permalloy.*

The origin of both magnetocrystalline anisotropy and magnetostriction is *spin-orbit* coupling. The magnitude of the magnetization of the film is controlled by the electron spin as usual, but the preferred direction of that

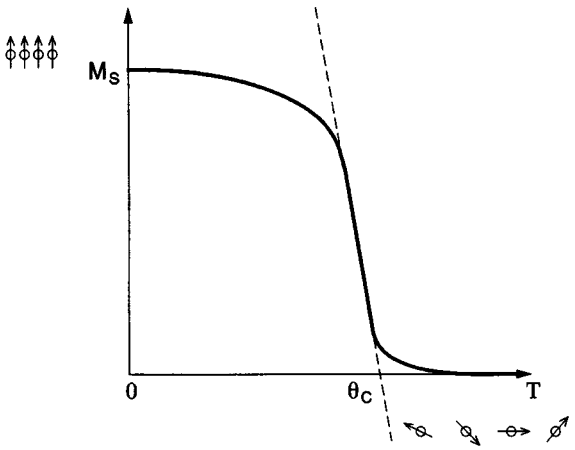


FIGURE 36.5 Ferro- and ferrimagnetic materials lose their spontaneous magnetic moment at temperatures above the Curie temperature, Θ_c.

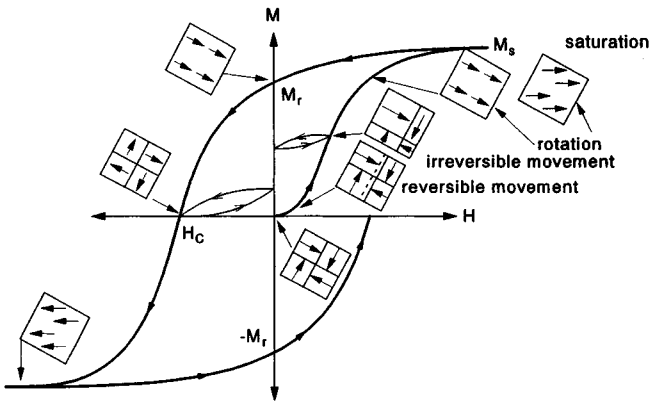


FIGURE 36.6 In soft magnetic materials domains form such that the total magnetization is zero. By applying small magnetic fields, domain walls move and the magnetization changes.

magnetization with respect to the crystal lattice is determined by the electron orbits which are large enough to interact with the atomic structure of the film.

Extrinsic Magnetic Properties

Extrinsic magnetic properties are those properties that depend not only on the shape and size of the sample, but also on the shape and size of the magnetic constituents of the sample. For example, if we measure the hysteresis loop like the one shown in Fig. 36.6 on a disk-shaped sample punched from a magnetic recording tape, the result will depend not only on the diameter and thickness of the disk coating but also on the distribution of shapes and sizes of the magnetic particles within the disk. They display hysteresis individually and collectively. For a soft magnetic material, i.e., one that might be used to make the laminations of a transformer, the dependence of magnetization, M, on the applied magnetic field, H, is also complex. Having once left a point described by the coordinates (H_1, M_1), it is not immediately clear how one might return to that point.

Alloys of nickel and iron, in which the nickel content is the greater, can be capable of a reversal of magnetization by the application of a magnetic field, H, which is weaker than the earth's magnetic field (0.5 Oe, 40 A/m) by a factor of five. (To avoid confusion caused by the geomagnetic field it would be necessary to screen the sample, for example, by surrounding it by a shield of equally soft material or by measuring the earth's field

TABLE 36.4 "Hard" and "Soft" Magnetic Materials

	High M_s	Low H_c	Low M_r	High μ
Soft				
Fe	1700 emu/cc	1 Oe	< 500	20,000
80 Ni 20 Fe	660	0.1	< 300	50,000
Mn Zn ferrite	400	0.02	< 200	5,000
$Co_{70}Fe_5Si_{15}B_{10}$	530	0.1	< 250	10,000
	High M_s	**High H_c**	**High M_r**	**T_c**
Hard				
Particles				
γ-Fe_2O_3	400	250–450	200–300	115–126
CrO_2	400	450–600	300	120
Fe	870–1100	1,100–1,500	435–550	768
$BaO.6Fe_2O_3$	238–370	800–3,000	143–260	320
Alloys				
$SmCo_5$	875	40,000	690	720
Sm_2Co_{17}	1,000	17,000	875	920
$Fe_{14}BNd_2$	1,020	12,000	980	310

and applying a field which is equal in magnitude but opposite in direction to the earth's field in order to cancel its effects.)

The magnetization of the sample in zero field may be macroscopically zero, but locally the material may be magnetized virtually to the saturation state. As shown in Fig. 36.6, which shows a greatly simplified domain structure, the net magnetization at the center of the loop is zero because the magnetization of the four "domains" cancels in pairs. A domain is a region (not necessarily square or even of a regular shape, although the shape often is regular in Ni–Fe thin films or sheets) over which the magnetization is constant in magnitude and direction. Thus, the sample in Fig. 36.6 consists of four domains, initially separated from each other by "domain walls." If a magnetic field is applied in the direction of $+H$, that domain will grow whose direction of magnetization is closest to the field direction and the domains will shrink if their magnetization is opposed to the field. For small applied fields, the movement of the walls is reversible, i.e., on reducing the applied field to zero, the original domain configuration will be obtained. Beyond a certain field the movement of the walls is irreversible, and eventually near the knee of the magnetization curve all the domain walls have been swept away by the applied field. The sample is not yet in the saturated state since the direction of M is not quite the same as the direction of the applied field. However, a small increase in the strength of the applied field finally achieves the saturated state by rotating the magnetization of the whole sample into the field direction.

On removing the applied field, the sample does not retrace the magnetization curve, and when the applied field is zero, we can see that a considerable amount of magnetization remains, M_r. Appropriately, this is referred to as the remanent state, and M_r is the *remanent magnetization*. By reversing the original direction of the applied field, domains reappear and the magnetization is eventually reduced to zero at the *coercive field*, H_c. It should be noticed that, at H_c, although the net magnetization is clearly zero, the individual domains may be magnetized in directions that are different from those at the starting point. Figure 36.6 shows an incomplete hysteresis loop. If the field H were increased beyond $+H_c$ the loop would be completed.

The differences between ideally magnetically soft materials (used in transformers and magnetic read/write heads) and magnetically hard materials (used in permanent magnets and in recording tapes and disks) are as follows:

Magnetically soft materials:	H_c 0;	M_r 0;	M_s high value
Magnetically hard materials:	H_c high value;	M_r/M_s ("squareness");	M_s high value

Examples are given in Table 36.4.

It is noticeable that the differences between hard and soft magnetic materials are confined to the extrinsic properties, M_r, H_c, and permeability, μ. The latter is related to M and H as follows:

$$\mathbf{B}\ (\mathrm{G}) = \mathbf{H}\ (\mathrm{Oe}) + 4\pi\mathbf{M}\ (\mathrm{emu/cc})$$

$$\mu = \mathbf{B}/\mathbf{H} = 1 + 4\pi\kappa\ (\mathrm{cgs\ units})$$

or

$$\mathbf{B}\ (\mathrm{Wb/m^2}) = \mu_o\ (\mathbf{H}\ \mathrm{A/m} + \mathbf{M}\ \mathrm{A/m})$$

$$= \mu_o\mu_r\mathbf{H} = \mu\mathbf{H}\ (\mathrm{SI\ units})$$

Domain walls form in order to minimize the magnetic energy of the sample. The magnetic energy is $\mu H^2/8\pi$ cgs units or $1/2\,\mu H^2$ J/m^3 (SI units) and clearly depends on **H**, the magnetic field emanating from the sample. In the *initial* domain configuration shown in Fig. 36.6, there is no net magnetization of the sample and thus no substantial **H** exists outside the sample and the magnetostatic energy is zero. Thus, the establishment of domains *reduces* the energy associated with H but it *increases* the energy needed to establish domain walls within the sample. A compromise is reached in which domain walls are formed until the establishment of one more wall would *increase*, rather than decrease, the total magnetic energy of the sample.

The wall energy depends on the area of the wall, i.e., L^2, while the energy associated with the external magnetic field depends on L^3, the volume of the sample. Clearly, as the size of single particles becomes small, terms in L^2 are more important than terms in L^3, and so for small magnetic particles, the formation of domain walls may not be energetically feasible and a *single-domain particle* results. These are found in the particles of iron oxide, cobalt-modified iron oxide, chromium dioxide, iron, or barium ferrite, which are used to make magnetic recording tapes, and in barium ferrite, samarium cobalt, and neodymium iron boron, which are used to make powerful permanent magnets. In the latter cases, the very high coercivities are caused by domain walls being pinned at grain boundaries between the main phase grains and finely precipitated secondary phases. This is an example of *nucleation-controlled coercivity.*

The amount of available energy that can be stored in a permanent magnet is the area of the largest rectangle that can be drawn in the second quadrant of the B vs. H hysteresis loop. The *energy product* has grown remarkably by a factor of about 50 since 1900 [Strnat, 1986]. We see from the graph in Fig. 36.7 of *intrinsic* coercive force, i.e., the coercive force obtained from the graph of M vs. H (in contrast to the smaller coercive force obtained by plotting B vs. H), that increases in H_c (rather than increases in M_r) have been responsible for almost all the improvement in the energy product.

The key attributes of technologically important magnetic materials are

1. Large, spontaneous atomic magnetic moments
2. Large, positive exchange integrals
3. Magnetic anisotropy and heterogeneity which are small for soft magnetic materials and large for hard magnetic materials.

In single-domain materials, the magnetic particles are so small that reversal of the magnetization can only occur by rotation of the magnetization vector. This rotation can be resisted by combinations of three anisotropies: crystalline anisotropy, shape anisotropy, and magnetoelastic anisotropy (which depends on the magnetostrictive properties of the material).

Crystalline Anisotropy

Crystalline anisotropy arises from the existence of easy and hard directions of magnetization within the crystal structure of the material. For example, in iron the {100} directions are easy directions while the {111} directions are hard. In nickel crystals the reverse is true. In cobalt the hexagonal {00.1} directions are easy and the {10.0} directions in the basal plane are hard.

Hard and easy directions in crystalline materials come about as a result of spin-orbit coupling. The spin, as usual, determines the occurrence of ferro- or ferrimagnetization, while the orbital motion of the electrons (3*d* in the case of Fe, Co, and Ni) responds to the structure of the crystal lattice.

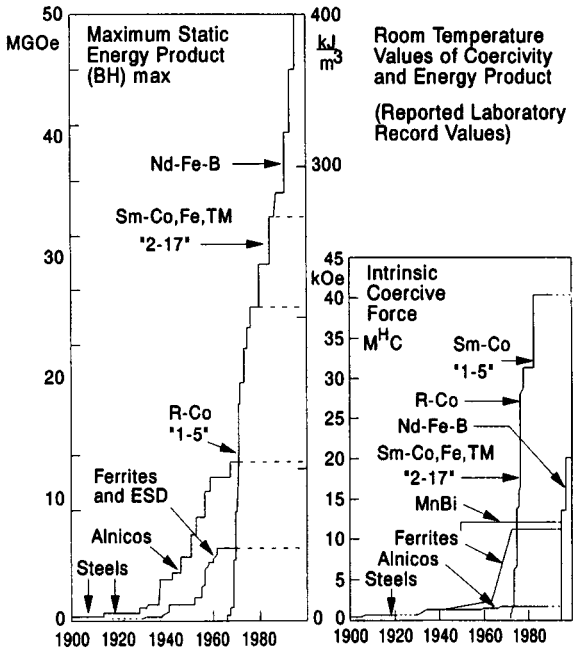


FIGURE 36.7 The development of magnetic materials for permanent magnets showing the increase in energy product and in intrinsic coercivity as a function of time.

The maximum value of coercivity resulting from crystalline anisotropy is given by $H_c = 2K_1/M_s$, where K_1 is the first magnetocrystalline anisotropy constant and M_s is the saturation magnetization.

Shape Anisotropy

A spherical particle has no shape anisotropy, i.e., all directions are equally easy (or hard). For particles (having low crystalline anisotropy) of any other shape, the longest dimension is the easy direction and the shortest dimension is the hardest direction of magnetization. Thus, a needle-shaped (acicular) particle will tend to be magnetized along the long dimension, whereas a particle in the form of a disk will have the axis of the disk as its hard direction, while any direction in the plane of the disk will be equally easy (assuming that shape is the dominant anisotropy).

For an acicular particle, the maximum value of the particle's switching field is [Stoner and Wohlfarth, 1948]

$$H_c = (N_b - N_a)\ M_s$$

where N_b is the demagnetizing factor in the shorter dimension and N_a is the factor for the long axis of the particle. When the ratio $b/a \rightarrow \infty$, then

$$\left.\begin{array}{l} N_a \rightarrow 0 \\ N_b \rightarrow 2\pi \end{array}\right\} N_a + N_b + N_c \equiv 4\pi$$

and H_c for iron > 10,000 Oe (7.95×10^5 A/m), higher than has been achieved in the laboratory for single-domain iron particles. Particles of iron having $H_c \leq 2000$ Oe are widely used in high-quality audio and video tapes.

The reason for the discrepancy is that the simplest single-domain model makes the assumption that the spins on all the atoms in a particle rotate in the same direction and at the same time, i.e., are coherent. This seems to be improbable since switching may begin at different places in the single-domain particle at the same time. Jacobs and Bean [1955] proposed an incoherent mode, *fanning*, in which different segments on a

TABLE 36.5 Maximum Coercivity (Oe) for Single-Domain Particles (Coherent Rotation)

	Iron	Cobalt	Nickel	γ-Fe2O3
Crystalline	250	3,000	70	230
Strain	300	300	2,000	<10
Shape (10:1)	5,300	4,400	1,550	2,450

longitudinal chain of atoms rotate in opposite directions. Shtrikman and Treves [1959] introduced another incoherent mode, *buckling*. These incoherent modes of magnetization reversal within single-domain particles not only predicted values of coercivity closer to the observed values, but also they could explain why the observed coercivity values for single-domain particles increased with decreasing particle size [Bate, 1980].

Shape anisotropy also plays an important role in determining the magnetization direction in thin magnetic films. It, of course, favors magnetization in the film plane.

Magnetoelastic Anisotropy

Spin-orbit coupling is also responsible for magnetostriction (the increase or decrease of the dimensions of a body on becoming magnetized or demagnetized). The magnetostriction coefficient λ_s = fractional change of a dimension of the body. It can be positive or negative, and it varies with changes in the direction and magnitude of the applied stress (or internal stress) and of the applied magnetic field. It is highly sensitive to composition, to structure, and to the previous history of the sample. The maximum coercivity is given by the formula $H_c = 3\lambda_s T/M_s$ where λ_s is the saturation magnetostriction coefficient, T is the tension, and M_s the saturation magnetization. Magnetostriction has been put to practical use in the generation of sonar waves for the detection of schools of fish or submarines.

For samples made of single-domain particles, the maximum coercivity for three ferromagnetic metals and one ferrimagnetic oxide (widely used in magnetic recording) is calculated using the preceding formula. Table 36.5 shows maximum coercivity (Oe) for single-domain particles (coherent rotation).

The assumption is made that all the spins rotate so that they remain parallel at all times. This is known as *coherent rotation*. In the case of γ-Fe_2O_3, an incoherent mode of reversal probably occurs since the maximum observed coercivity is only 350 Oe. Several incoherent modes have been proposed, e.g., chain-of-spheres fanning [Jacobs and Bean, 1955], curling [Shtrikman and Treves, 1959]. Their characteristics and differences are discussed by Bate [1980]. The coercivity of particles of γ-Fe_2O_3 is increased (in order to make recording tapes of extended frequency response) by precipitating cobalt hydroxide on the surface of the particles. After gentle warming, the cobalt is incorporated on the surface of the particles and increases the coercivity to 650 Oe (51 · 73 kA/m).

Figure 36.8 illustrates two additional extrinsic magnetic properties of importance. They are the remanence coercivity, H_r, and the switching field distribution (SFD). These are of particular importance in magnetic particles used in magnetic tapes (audio, video, or data) or magnetic disks. The coercivity, H_c, of a magnetic material is the value of the magnetic field (the major loop) at which $M = 0$. However, if the applied field is allowed to go to zero, a small magnetization remains. It is necessary to increase the applied field from $H_c > H_r$ (Fig. 36.8) to achieve $M_r = 0$. H_r is the *remanence coercivity* and is more relevant than is H_c in discussing the writing process in magnetic recording since it corresponds to the center of the remanent magnetization transition on the recording medium.

Particles, for example, in a magnetic recording do not all reverse their magnetization at the same field; there is a distribution of switching fields, which can be found by differentiating M with respect to H around the point H_c. The result is shown as the broken curve on Fig. 36.8, where the SFD = $\Delta H/H_c$, ΔH being the width at half the maximum of the curve. Typically, SFD = 0.2–0.3 for a high-quality particulate medium and smaller than this for thin-film recording media.

Figure 36.8 also shows the construction used to find the parameter S^*. The quantity $1 - S^*$ is found to be very close to $\Delta H/H_c$ and it is quicker to evaluate. Either of these parameters can be used to determine the distribution of switching fields, small values of which are required to achieve high recording densities.

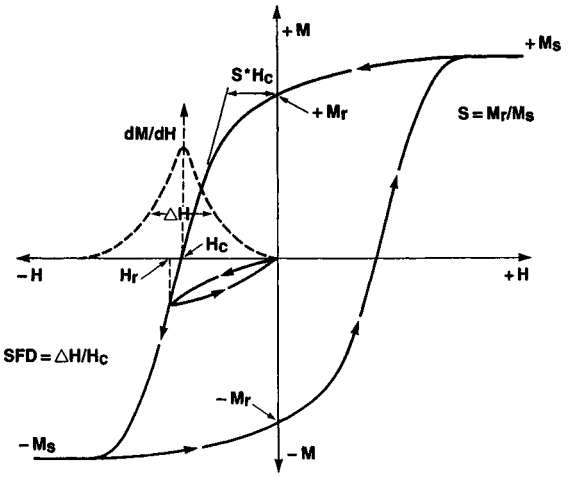


FIGURE 36.8 The switching field distribution (SFD) for a magnetic recording medium can be obtained in two ways: (1) SFD = $\Delta H/H_c$ or (2) SFD = $1 - S^*$.

Amorphous Magnetic Materials

Before 1960, all the known ferro- and ferrimagnetic materials were crystalline. Because the occurrence of these magnetic states is known to depend on short-range interaction between atoms, there is no reason why amorphous materials (which have *only* short-range order) should not have useful magnetic properties. This was found to be true in 1960, when thin, amorphous ribbons of $Au_{81}Si_{19}$ were made by rapidly cooling the molten alloy through the melting point and the lower glass transition temperature. Because there is always at least one crystalline phase more stable than the amorphous state, the problem is to invent a production method that yields the amorphous phase rather than the crystalline one.

Most methods involve cooling the molten mixture so rapidly that there is insufficient time for crystals to form. Cooling rates of 10^5–10^6 degrees per second are needed and can be achieved in several ways:

1. Pouring the molten mixture from a silica crucible onto the edge of a rapidly rotating copper wheel. This yields a ribbon of the amorphous alloy, typically 1 mm wide and 25 μm thick.
2. Depositing a thin film from a metal vapor or a solution of metal ions.
3. Irradiating a thin sample of the metal with high-energy particles.

Once an amorphous alloy is formed, it will remain indefinitely in the *glassy* state at room temperature. The problem is that only *thin* films are obtained, and large areas are required to make, for example, the core of a transformer.

There are three main groups of amorphous films:

1. Metal-metalloid alloys, e.g., $Au_{81}Si_{19}$
2. Late transition–early transition metal alloys, e.g., $Ni_{60}Nb_{40}$
3. Simple metal alloys, e.g., $Cu_{65}Al_{35}$

When normal metals freeze, crystallization begins at a fixed temperature, the *liquidus*, T_e. In amorphous alloys *configurational freezing* occurs at a lower temperature, the *glass temperature*, T_g, which is not as well defined as T_e. There is an abrupt increase in the time required for the rearrangement of the atoms, from 10^{-12} s for liquids to 10^5 s (a day) for glasses. Not surprisingly, this increase in atomic rearrangement time is associated with an abrupt increase in viscosity, from 10^{-2} poise for liquids, e.g., water or mercury, to 10^{15} poise for glasses.

The principal difference between magnetic glasses and ferromagnetic alloys is that the glasses are completely isotropic (all directions of magnetization are very easy directions), and consequently, considering only the magnetic properties, soft amorphous alloys are almost ideally suited for use in the core of power transformers

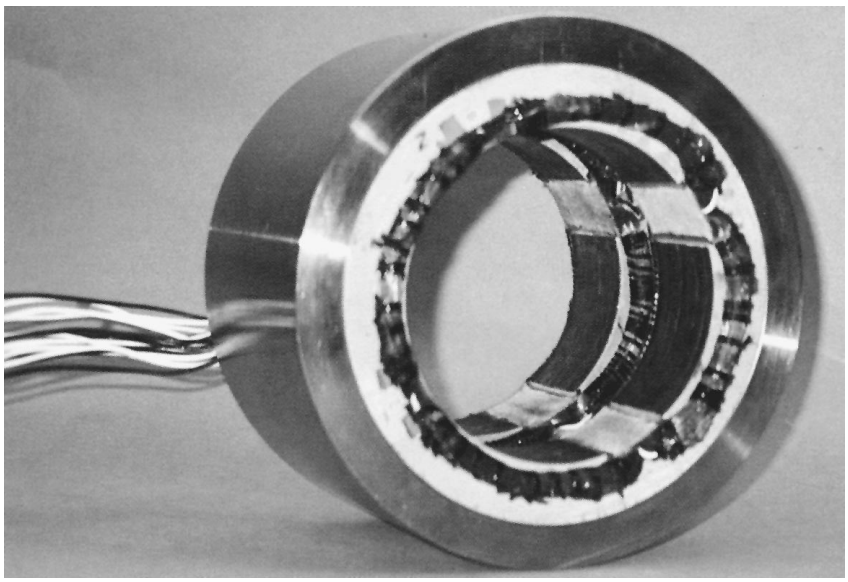
Magnetic Bearing

Magnetic bearings support moving machinery without physical contact. For example, they can levitate a rotating shaft and permit relative motion without friction or wear. Long considered a promising advancement, magnetic bearings are now in actual service in such industrial applications as electric power generation, petroleum refining, machine tool operation, and natural gas pipelines.

AVCON, Inc. worked initially with Lewis Research Center on the development of a magnetic bearing system for a cryogenic magnetic bearing test facility. The resulting AVCON development was extensively tested over a two-year span and these tests provided a wealth of data on the performance of magnetic bearings under severe conditions. In this program, AVCON developed the basic hybrid magnetic bearing approach in which both permanent magnets and electromagnets are employed to suspend a shaft; the permanent magnets provide suspension, the electromagnets provide control. Analyses of AVCON bearing tests showed that a hybrid magnetic bearing test was typically only one-third the weight, substantially smaller and dramatically less power-demanding than previous generations of magnetic bearings.

In 1993, Marshall Space Flight Center awarded AVCON a contract to fabricate a set of magnetic bearings, install them in a fixture representing a Space Shuttle main engine turbopump, and test them under simulated shuttle mission conditions.

AVCON has been able to develop a unique "homopolar" approach to permanent magnet type bearings that the company says are significantly smaller than prior designs, their control electronics are a fraction of the weight of previous systems, and power consumption is much lower than in all electromagnetic designs. Among other advantages cited are virtually zero friction and therefore no lubricant requirement, no wear, no vibration, longer service life, and very high reliability because single point failure modes are eliminated. (Courtesy of National Aeronautics and Space Administration.)



This AVCON magnetic bearing permits motion without friction or wear. (Photo courtesy of National Aeronautics and Space Administration.)

or magnetic recording heads, where almost zero remanence and coercivity are desired at frequencies up to megahertz. The limit on their performance seems to be the magnetic anisotropy which arises from strains generated during the manufacturing process.

When an amorphous material is required to store energy (as in a permanent magnet) or information (as in magnetic bubbles or thermomagneto-optic films), it must have magnetic anisotropy. This is generally produced by applying a magnetic field at high temperatures to the amorphous material. The field and temperature must be high enough to allow a local rearrangement of atoms to take place in order to create the desired degree of magnetocrystalline anisotropy.

Amorphous materials will apparently play increasingly important roles as magnetic materials. To accelerate their use, we need to have answers to the questions "What governs the formation of amorphous materials?" and "What is the origin of their anisotropy and magnetostriction?".

Defining Terms

Coercivity, H_c (Oe, A/m): The property of a magnetized body enabling it to resist reversal of its magnetization.

Compensation temperature, T_c (°C, K): The temperature at which the magnetization of a material comprising ferromagnetic atoms (e.g., Fe, Co, Ni) and rare earth atoms (e.g., Gd, Tb) becomes zero because the magnetization of the sublattice of ferromagnetic atoms is canceled by the opposing magnetization of the rare earth sublattice.

Curie temperature, Θ_c (°C, K): The temperature at which the spontaneous magnetization of a ferromagnetic or ferrimagnetic body becomes zero.

Remanence, M_r (emu/cc, A/m): The property of a magnetized body enabling it to retain its magnetization.

Related Topics

1.3 Transformers • 35.1 Maxwell Equations

References

G. Bate, in *Recording Materials in Ferromagnetic Materials,* vol. 2, Amsterdam: North-Holland, 1980, pp. 381–507.

G. Bate, *J. Magnetism and Magnetic Materials*, vol. 100, pp. 413–424, 1991.

F. Brailsford, *Physical Principles of Magnetism*, London: Van Nostrand, 1966.

J. Crangle, "Ferromagnetism and antiferromagnetism in non-ferrous metals and alloys," *Metallurgical Reviews,* pp. 133–174, 1962.

I.S. Jacobs and C.P. Bean, *Phys. Rev.*, vol. 100, p. 1060, 1955.

K. Moorjani and J.M.D. Coey, *Magnetic Glasses: Methods and Phenomena, Their Application in Science and Technology*, vol. 6, Amsterdam: Elsevier, 1984.

S. Shtrikman and D. Treves, *J. Phys. Radium*, vol. 20, p. 286, 1959.

J.C. Slater, *Phys. Rev.*, vol. 36, p. 57, 1930.

E.C. Stoner and E.P. Wohlfarth, *Phil. Trans. Roy. Soc.*, vol. A240, p. 599, 1948.

K.J. Strnat, *Proceedings of Symposium on Soft and Hard Magnetic Materials with Applications*, vol. 8617-005, Metals Park, Ohio: American Society of Metals, 1986.

Further Information

A substantial fraction of the papers published in English on the technologically important aspects of magnetism appear in the *IEEE Transactions on Magnetics* or in the *Journal of Magnetism and Magnetic Materials.*

The two major annual conferences are Intermag (proceedings published in the *IEEE Transactions on Magnetics*) and the Magnetism and Magnetic Materials Conference, MMM (proceedings published in the American Physical Society's *Journal of Applied Physics*).

36.2 Magnetic Recording

Mark H. Kryder

Magnetic recording is used in a wide variety of applications and formats, ranging from relatively low-density, low-cost floppy disk drives and audio recorders to high-density videocassette recorders, digital audio tape recorders, computer tape drives, rigid disk drives, and instrumentation recorders. The storage density of this technology has been advancing at a very rapid pace. With a storage density exceeding 1 Gbit/in.^{2}, magnetic recording media today can store the equivalent of about 50,000 pages of text on one square inch. This is more than 500,000 times the storage density on the RAMAC, which was introduced in 1957 by IBM as the first disk drive for storage of digital information. The original Seagate 5.25-inch magnetic disk drive, introduced in 1980, stored just 5 Mbytes. Today, instead of storing megabytes, 5.25-inch drives store tens of gigabytes, and drives as small as 2.5 inches store over a gigabyte.

This astounding rate of progress shows no sign of slowing. Fundamental limits to magnetic recording density are still several orders of magnitude away, and recent product announcements and laboratory demonstrations indicate the industry is accelerating the rate of progress rather than approaching practical limits. Recently IBM demonstrated the feasability of storing information at a density of 3 Gbit/in.^{2} [Tsang et al., 1996]. Similar advances can also be expected in audio and video recording.

Fundamentals of Magnetic Recording

Although magnetic recording is practiced in a wide variety of formats and serves a wide variety of applications, the fundamental principles by which it operates are similar in all cases. The fundamental magnetic recording configuration is illustrated in Fig. 36.9. The recording head consists of a toroidally shaped core of soft magnetic material with a few turns of conductor around it. The magnetic medium below the head could be either tape or disk, and the substrate could be either flexible (for tape and floppy disks) or rigid (for rigid disks). To record on the medium, current is applied to the coil around the core of the head, causing the high-permeability magnetic core to magnetize. Because of the gap in the recording head, magnetic flux emanates from the head and penetrates the medium. If the field produced by the head is sufficient to overcome the **coercive force** of the medium, the medium will be magnetized by the head field. Thus, a representation of the current waveform applied to the head is stored in the magnetization pattern in the medium.

Readout of previously recorded information is typically accomplished by using the head to sense the magnetic stray fields produced by the recorded patterns in the medium. The recorded patterns in the medium cause magnetic stray fields to emanate from the medium and to flow through the core of the head. Thus, if the medium is moved with respect to the head, the flux passing through the coil around the head will change in a manner which is representative of the recorded magnetization pattern in the medium. By Faraday's law of induction, a voltage representative of the recorded information is thus induced in the coil.

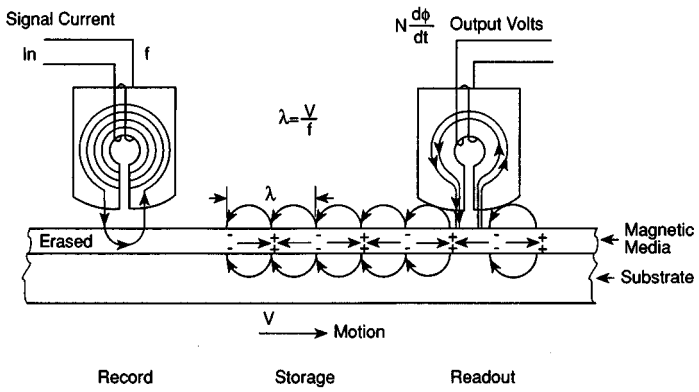


FIGURE 36.9 The fundamental magnetic recording configuration.

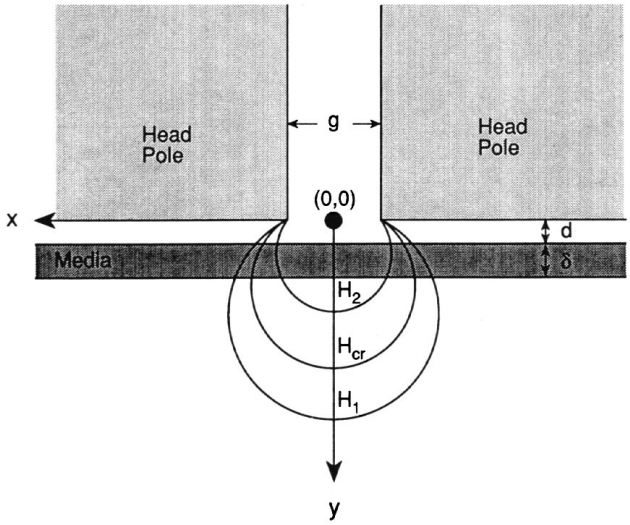


FIGURE 36.10 The constant longitudinal field contours in the gap region of a recording head.

The Recording Process

During recording the head is used to produce large magnetic fields which magnetize the medium. It was shown by Karlqvist [1954] that, in the case where the track width and length of the poles along the gap are both large compared to the gap length, the fields produced by a recording head could be described by

$$H_x = \frac{NI}{\pi g}\left[\tan^{-1}\left(\frac{x + g/2}{y}\right) - \tan^{-1}\left(\frac{x - g/2}{y}\right)\right] \tag{36.1a}$$

$$H_y = \frac{NI}{2\pi g}\ln\left[\frac{(x + g/2)^2 + y^2}{(x - g/2)^2 + y^2}\right] \tag{36.1b}$$

where H_x and H_y are the longitudinal and perpendicular components of field, as indicated by the coordinates in Fig. 36.10, N is the number of turns on the head, I is the current driving the head, and g is the gap width of the head. In this approximation, the contours of equal longitudinal field are described by circles which intersect the gap corners as shown in Fig. 36.10.

In digital or saturation recording, the recording head is driven with sufficiently large currents that a portion of the recording medium is driven into saturation. However, because of the gradient in the head fields, other portions of the medium see fields less than those required for saturation. This is illustrated in Fig. 36.10 where the contours for three different longitudinal fields are drawn. In this figure H_{cr} is the **remanence coercivity** or the field required to produce zero **remanent magnetization** in the medium after it was saturated in the opposite direction, and H_1 and H_2 are fields which would produce negative and positive remanent magnetization, respectively. Note that the head field gradient is the sharpest near the pole tips of the head. This means that smaller head-to-medium spacing and thinner medium both lead to narrower transitions being recorded.

Modeling the recording process involves convolving the head field contours with the very nonlinear and hysteretic magnetic properties of the recording medium. A typical M-H hysteresis loop for a longitudinal magnetic recording medium is shown in Fig. 36.11. Whether the medium has positive or negative magnetization depends upon not only the magnetic field applied but the past history of the magnetization. If the medium was previously saturated at $-M_s$, then when the magnetic field H is reduced to zero, the remanent magnetization will be $-M_r$; however, if it was previously saturated at $+M_s$, then the remanent magnetization would be $+M_r$.

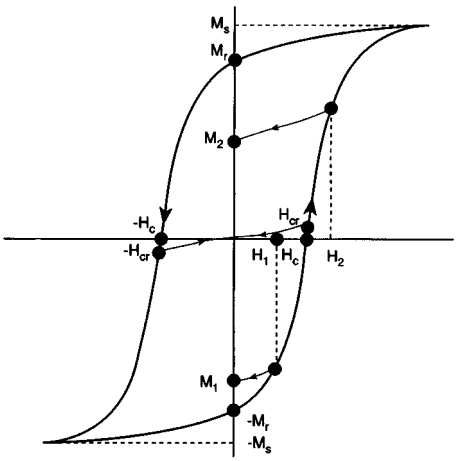


FIGURE 36.11 A remanent *M-H* hysteresis loop for a longitudinal recording medium.

Similarly, if the medium was initially saturated to $-M_s$, then magnetized by a field $+H_1$, and finally allowed to go to a remanent state, the magnetization would go to value M_1. This hysteretic behavior is the basis for the use of the medium for long-term storage of information but makes the recording process highly nonlinear.

An additional complicating factor in determining the actual recorded pattern is the **demagnetizing field** of the medium itself. As shown in Fig. 36.9, transitions in the recorded magnetization direction produce effective magnetostatic charge given by

$$\rho_M = -\nabla \cdot \vec{\mathbf{M}} \tag{36.2}$$

which in turn results in demagnetizing fields. The demagnetizing fields outside the medium are what is sensed by the head during readback, but demagnetizing fields also exist inside the medium and act to alter the total field seen by the medium during the recording process from that of the head field alone.

Taking into account the head field gradients, the nonlinear *M-H* loop characteristics of the medium and the demagnetizing fields, Williams and Comstock [1971] developed a model for the recording process. This model predicts the width of a recorded transition, in a material with a square hysteresis loop, to be

$$a = \sqrt{\frac{M_r \delta d}{\pi H_c}} \tag{36.3}$$

where δ is the medium thickness, d is the head-medium spacing, M_r is the remanent magnetization of the medium, and H_c is the coercivity of the medium. That the transition widens with the product $M_r \cdot \delta$ is a result of the fact that the demagnetizing fields increase linearly with this quantity. Similarly, the transition narrows as H_c is increased, because with high coercivity, the medium can resist the transition broadening due to the demagnetizing fields. The increase in transition width with d is due to the fact that a poorer head field gradient is obtained with larger head-medium spacing.

The nonlinearities of the recording process can be largely removed by a technique referred to as ac bias recording. This is frequently used in analog recording in audio and video recorders. In this technique, a high-frequency ac bias is added to the signal to be recorded. This ac bias signal is ramped from a value much larger than the coercivity of the medium to zero. This removes the hysteretic behavior of the medium and causes it to assume a magnetization state which represents the minimum energy state determined by the amount of field produced by the signal to be recorded.

The Readback Process

As opposed to the recording process, the readback process can usually be modeled as a linear process. This is because the changes in magnetization which occur in either the head or the medium during readback are typically small.

The most common way to model the readback process is to use the principle of reciprocity, which states that the flux produced by the head through a cross section of an element of the medium, normalized by the number of ampere turns of current driving the head, is equal to the flux produced in the head by the element of medium, normalized by the equivalent current required to produce the magnetization of that element. For a magnetic recording head, which produces the longitudinal field $H_x(x,y)$ when driven by NI ampere turns of current, this principle leads to the following expression for the voltage induced in the head by a recording medium with magnetization $M(x,y)$ and moving with velocity v relative to the head:

$$e = \frac{\mu_o W v}{I} \int_d^{d+\delta} \int_{-\infty}^{\infty} \frac{\partial M_x(x - \bar{x}, y)}{\partial \bar{x}} H_x(x, y) dx dy \tag{36.4}$$

where W is the track width of the head.

This expression shows that the readback voltage induced in the recording head is linearly dependent upon the magnitude of the magnetization in the recording medium being sensed and the relative head-to-medium velocity. The linearity of the readback process ensures that analog recordings such as those recorded on audio or video tapes are faithfully reproduced.

Magnetic Recording Media

A wide variety of magnetic recording media are available today. Different applications require different media, but furthermore, in many cases the same application will be able to utilize a variety of different competitive media.

Just a decade ago, essentially all recording media consisted of fine acicular magnetic particles embedded in a polymer and coated onto either flexible substrates such as mylar for floppy disks and tapes or onto rigid aluminum-alloy substrates for rigid disks. Today, although such particulate media are still widely used for tape and floppy disks, thin-film media have almost entirely taken over the rigid disk business, and metal-evaporated thin-film media has been introduced into the tape marketplace. Furthermore, many new particle types have been introduced.

The most common particulate recording media today are γ-Fe_2O_3, Co surface-modified γ-Fe_2O_3, CrO_2, and metal particle media. All of these particles are acicular in shape with aspect ratios on the order of 5 or 10 to 1. The particles are sufficiently small that it is energetically most favorable for them to remain in a single-domain saturated state. Because of demagnetizing effects caused by the acicular shape, the magnetization prefers to align along the long axis of the particle.

As was noted in the discussion of Eq. (36.3), to achieve higher recording densities requires media with higher coercivity. The coercivity of a particle is determined by the field required to cause the magnetization to switch by 180°. If the magnetization remained in a single-domain state during the switching process, then the coercivity should be given by

$$H_c = (N_a - N_b) M_s \tag{36.5}$$

where N_a and N_b are the demagnetizing factors in the directions transverse and parallel to the particle axis, respectively. In practice the coercivity is measured to be less than this. This has been explained as being a result of the fact the magnetization does not remain uniform during the switching process, but switches inhomogeneously [White, 1984]. In addition to the effects which the shape anisotropy of the particles has on the coercivity, crystalline anisotropy can also be used to control coercivity.

TABLE 36.6 Magnetic Material, Saturation Remanence $M_r(\infty)$, Coercivity H_c, Switching-Field Distribution Δh_r, and Number of Particles per Unit Volume, N, of Various Particulate Magnetic Recording Media

Application	Material	*Mr* (∞), kA/m (emu/cm³)	*Hc*, kA/m (4π Oe)	Δ*hr*	*N*, $10^3/\mu m^3$
Reel-to-reel audio tape	γ-Fe_2O_3	100–120	23–28	0.30–0.35	0.3
Audio tape IEC I	γ-Fe_2O_3	120–140	27–32	0.25–0.35	0.6
Audio tape IEC II	CrO_2	120–140	38–42	0.25–0.35	1.4
	γ-Fe_2O_3 + Co	120–140	45–52	0.25–0.35	0.6
Audio tape IEC IV	Fe	230–260	80–95	0.30–0.37	3
Professional video tape	γ-Fe_2O_3	75	24	0.4	0.1
	CrO_2	110	42	0.3	1.5
	γ-Fe_2O_3 + Co	90	52	0.35	1
Home video tape	CrO_2	110	45–50	0.35	2
	γ-Fe_2O_3 + Co	105	52–57	0.35	1
	Fe	220	110–120	0.38	4
Instrumentation tape	γ-Fe_2O_3	90	27	0.35	0.6
	γ-Fe_2O_3 + Co	105	56	0.50	0.8
Computer tape	γ-Fe_2O_3	87	23	0.30	0.16
	CrO_2	120	40	0.29	1.4
Flexible disk	γ-Fe_2O_3	56	27	0.34	0.3
	γ-Fe_2O_3 + Co	60	50	0.34	0.5
Computer disk	γ-Fe_2O_3	56	26–30	0.30	0.3
	γ-Fe_2O_3 + Co	60	44–55	0.30	0.5

The coercivity of the medium which is made from the particles is determined by the distribution of coercivities of the particles from which it is made, their orientation in the medium relative to the fields from the head, and their interactions among each other. The coercivities of a variety of particulate recording media are summarized in Table 36.6.

Although coercivity is indeed an important parameter for magnetic recording media, it is by no means the only one. Particle size affects the medium noise because, at any time, the head is sensing a fixed volume of the medium. Because the particles are quantized and there are statistical variations in their switching behavior, the medium power signal-to-noise ratio varies linearly with the number of particles contained in that volume. To reduce particulate medium noise, it is therefore generally desirable to use small particles.

There is a limit, however, to how small particles may be made and still remain stable. When the thermal energy kT is comparable in magnitude to the energy required to switch a particle, $M \cdot H_c$, the particle becomes unstable and may switch because of thermal excitation. This phenomenon is known as **superparamagnetism** and can lead to decay of recorded magnetization patterns over time.

The remanent magnetization of a medium is important because it directly affects the signal level during readback as shown by Eq. (36.4). The remanent magnetizations of several particulate media are listed in Table 36.6. Obtaining high remanent magnetization in particulate media requires the use of particles with high **saturation magnetization** and a high-volume packing fraction of particles in the polymer binder. Obtaining a high-volume packing fraction of particles in the binder, however, can lead to nonuniform distributions of particles and agglomerates of many particles, which switch together, also causing noise during readback.

Generally, then, to obtain good high-density particulate recording media it is desired to have adequate coercivity (to achieve the required recording density), small particles (for low noise), with a very narrow switching field distribution (to obtain a narrow transition), to have them oriented along the direction of recording (to obtain a large remanence), and to have them uniformly dispersed (to obtain low modulation noise), with high packing density (to obtain large signals).

Thin-film recording media generally have excellent magnetic properties for high-density recording. Because they are nearly 100% dense (voids at the grain boundaries reduce the density somewhat), they can be made to have the highest possible magnetization. Because of their high magnetization, they can be made extremely thin and still provide adequate signal during readback. This helps narrow the recorded transition since the head field gradient is sharper for thinner media, as was discussed in reference to Fig. 36.10.

Thin-film media can also be made extremely smooth. To achieve the smallest possible head-to-medium spacing and therefore the sharpest head field gradient and the least spacing loss, smooth media are required.

The coercivity of thin-film media can also be made very high. In volume production today are media with coercivities of 160 kA/m; however, media with coercivities to 250 kA/m have been made and appear promising [Velu and Lambeth, 1992]. Such high coercivities are adequate to achieve more than an order of magnitude higher recording density than today.

Numerical models indicate that noise in thin-film media increases when the grains in polycrystalline films are strongly exchange coupled [Zhu and Bertram, 1988]. Exchange coupled films tend to exhibit zigzag transitions, which produce considerable jitter in the transition position relative to the location where the record current in the head goes through zero. A variety of experimental studies have indicated that the introduction of nonmagnetic elements which segregate to the grain boundaries and careful control of the sputtering conditions to achieve a porous microstructure at the grain boundaries reduce such transition jitter [Chen and Yamashita, 1988].

Magnetic Recording Heads

Early recording heads consisted of toroids of magnetically soft ferrites, such as NiZn-ferrite and MnZn-ferrite, with a few turns of wire around them. For high-density recording applications, however, ferrite can no longer be used, because the saturation magnetization of ferrite is limited to about 400 kA/m. Saturation of the pole tips of a ferrite head begins to occur when the deep gap field in the head approaches one-half the saturation magnetization of the ferrite. Because the fields seen by a medium are one-half to one-quarter the deep gap field, media with coercivities above about 80 kA/m cannot be reliably written with a ferrite head. High-density thin-film disk media, metal particle media, and metal evaporated media, therefore, cannot be written with a ferrite head.

Magnetically soft alloys of metals such as Permalloy (NiFe) and Sendust (FeAlSi) have saturation magnetizations on the order of 800 kA/m, about twice that of ferrites, but because they are metallic may suffer from eddy current losses when operated at high frequencies. To overcome the limitations imposed by eddy currents, they are used in layers thinner than a skin depth at their operating frequency. To prevent saturation of the ferrite heads, the high magnetization metals are applied to the pole faces of the ferrite, making a so-called metal-in-gap or MiG recording head, as shown in Fig. 36.12. Since the corners of the pole faces are the first parts of a ferrite head to saturate, the high magnetization metals enable these MiG heads to be operated to nearly twice the field to which a ferrite head can be operated. Because the layer of metal is thin, it can furthermore be less than a skin depth, and eddy current losses do not limit performance at high frequencies.

Yet another solution to the saturation problem of ferrite heads is to use thin-film heads. Thin-film heads are made of Permalloy and are therefore metallic, but the films are made sufficiently thin that they are thinner than the skin depth and, consequently, the heads operate well at high frequencies. A diagram of a thin-film head is shown in Fig. 36.13. It consists of a bottom yoke of Permalloy, some insulating layers, a spiral conductor, and a top yoke of Permalloy, which is joined to the bottom yoke at the back gap but separated from it by a thin insulator at the recording gap. These thin-film heads are made using photolithography and microfabrication techniques similar to those used in the manufacture of semiconductor devices. The thin pole tips of these heads actually sharpen the head field function and, consequently, the pulse shape produced by an isolated transition, although at the expense of some undershoot, as illustrated in Fig. 36.14. Because thin-film heads are made by photolithographic techniques, they can be made extremely small and to have low inductance. This, too, helps extend the frequency of operation.

A relatively new head which is now being used for readback of information in high-density recording is the magnetoresistive (MR) head. MR heads are based on the phenomenon of **magnetoresistance**, in which the electrical resistance of a magnetic material is dependent upon the direction of magnetization in the material

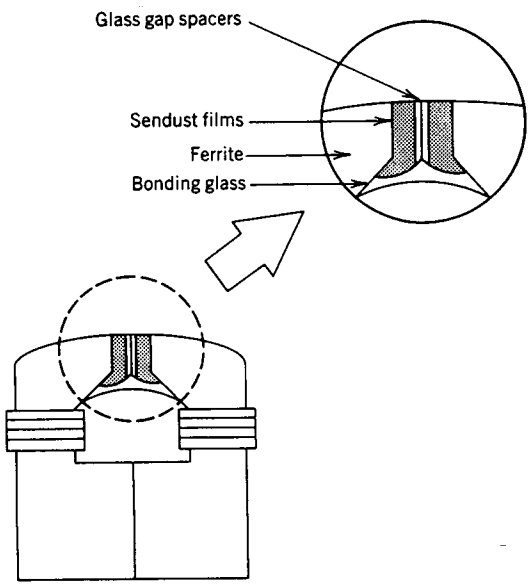


FIGURE 36.12 A diagram of a metal-in-gap or MiG recording head. (*Source:* A.S. Hoagland and J.E. Monson, *Digital Magnetic Recording*, 2nd ed., New York: Wiley-Interscience, 1991, p. 127. With permission.)

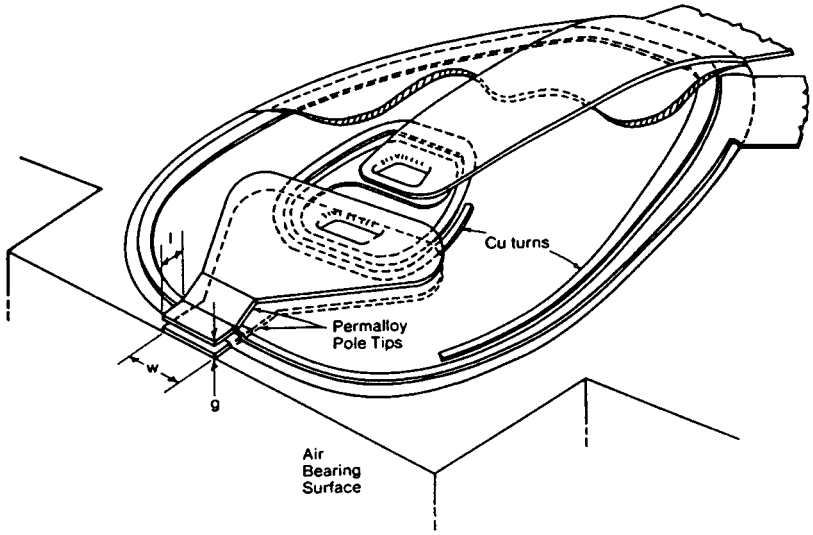


FIGURE 36.13 A thin-film head. (*Source:* R.M. White, Ed., *Introduction to Magnetic Recording*, New York: IEEE Press, p. 28. ©1985 IEEE.)

relative to the direction of current flow. An unshielded MR head is depicted in Fig. 36.15. Current flows in one end of the head and out the other. The resistivity of Permalloy from which the head is made varies as

$$\rho = \rho_o + \Delta\rho \cos^2\theta \tag{36.6}$$

where θ is the angle which the magnetization in the Permalloy makes relative to the direction of current flow, ρ_o is the isotropic resistivity, and $\Delta\rho$ is the magnetoresistivity. When the recording medium with a changing magnetization pattern moves under the MR head, the stray fields from the medium cause a change in the

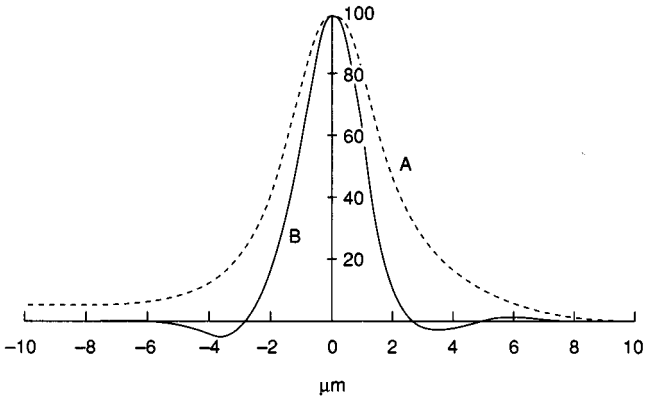


FIGURE 36.14 Pulse shapes for (curve A) long- and (curve B) short-pole heads normalized for equal amplitude. (*Source:* E.P. Valstyn and L.F. Shew, "Performance of single-turn film heads," *IEEETrans. Magnet.*, vol. MAG-9, no. 3, p. 317. ©1973 IEEE.)

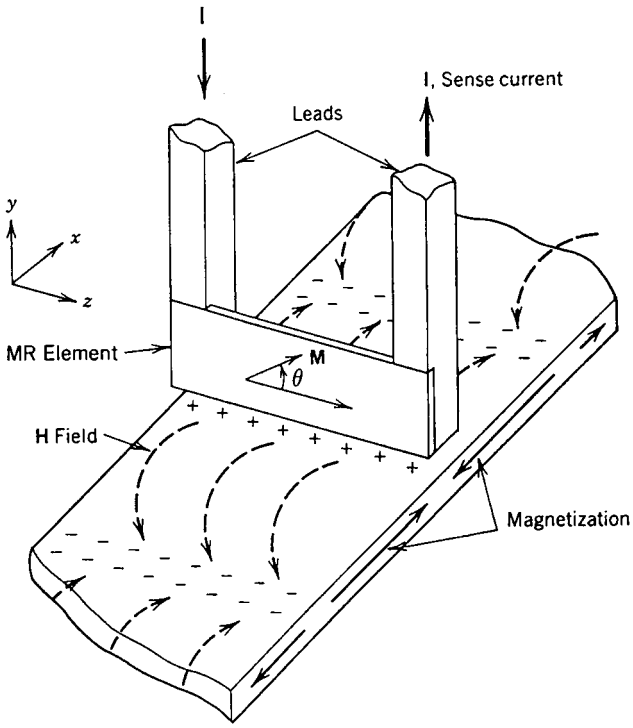


FIGURE 36.15 An unshielded magnetoresistive element. (*Source:* A.S. Hoagland and J.E. Monson, *Digital Magnetic Recording*, 2nd ed., New York: Wiley-Interscience, 1991, p. 131. With permission.)

direction of magnetization and, consequently, a change in resistance in the head. With a constant current source driving the head, the head will therefore exhibit a change in voltage across its terminals.

Magnetoresistive heads are typically more sensitive than inductive heads and therefore produce larger signal amplitudes during readback. The increased sensitivity and the fact that the read head is independent of the write head can be used to make a write/read head combination in which the write head writes a wider track than the read head senses. Thus, adjacent track interference is reduced during the readback process.

Another advantage of the MR head is that it senses magnetic flux ϕ, not the time rate of change of flux $d(\phi)/dt$ as an inductive head does. Consequently, whereas the inductive head output voltage is dependent upon the head-to-medium velocity as was shown by Eq. (36.4), the output voltage of an MR head is independent of velocity.

Conclusions

Magnetic recording today is used in a wide variety of formats for a large number of applications. Formats range from tape, which has the highest volumetric packing density and lowest cost per bit stored, to rigid disks, which provide fast access to a large volume of data. Applications include computer data storage, audio and video recording, and collecting data from scientific instruments.

The technology has increased storage density by more than a factor of 1,000,000 over the past 35 years since it was first used in a disk format for computer data storage; however, fundamental limits, set by superparamagnetism, are estimated yet to be a factor of more than 1,000 from where we are today. Furthermore, recent product announcements and developments in research labs suggest that the rate of progress is likely to accelerate. Storage densities of over 5 Gbit/in.^{2} are likely by the end of this decade and densities of 100 Gbit/in.^{2} appear likely in the early twenty-first century.

Defining Terms

Coercive force or coercivity: The magnetic field required to reduce the mean magnetization of a sample to zero after it was saturated in the opposite direction.

Demagnetizing field: The magnetic field produced by divergences in the magnetization of a magnetic sample.

Magnetoresistance: The resistance change produced in a magnetic sample when its magnetization is changed.

Remanence coercivity: The magnetic field required to produce zero remanent magnetization in a material after the material was saturated in the opposite direction.

Remanent magnetization: The magnetic moment per unit volume of a material in zero field.

Saturation magnetization: The magnetic moment per unit volume of a material when the magnetization in the sample is aligned (saturated) by a large magnetic field.

Superparamagnetism: A form of magnetism in which the spins in small particles are exchange coupled but may be collectively switched by thermal energy.

Related Topics

36.1 Magnetism • 80.2 Basic Disk System Architectures

References

T. Chen and T. Yamashita, "Physical origin of limits in the performance of thin-film longitudinal recording media," *IEEE Trans. Magnet.*, vol. MAG-24, p. 2700, 1988.

A.S. Hoagland and J.E. Monson, *Digital Magnetic Recording,* New York: John Wiley & Sons, 1991.

O. Karlqvist, "Calculation of the magnetic field in the ferromagnetic layer of a magnetic drum," *Trans. Roy. Inst. Technol., Stockholm*, No. 86, 1954. Reprinted in R. M. White, Ed., *Introduction to Magnetic Recording*, New York: IEEE Press, 1985.

E. Köster and T.C. Arnoldussen, "Recording media," in *Magnetic Recording,* C. D. Mee and E.D. Daniel, Eds., New York: McGraw-Hill, 1987.

M.-M. Tsang, H. Santini, T. Mccown, J. Lo, and R. Lee, "3 Gbit/in.^{2} recording demonstration with dual element heads of thin film drives," *IEEE Trans Magnet.,* MAG-32, p. 7, 1996.

E.P. Valstyn and L.F. Shew, "Performance of single-turn film heads," *IEEE Trans. Magnet.*, vol. MAG-9, p. 317, 1973.

E. Velu and D. Lambeth, "High Density Recording on SmCo/Cr Thin Film Media," Paper KA-01, Intermag Conference, St. Louis, April 1992; to be published in *IEEE Trans. Magnet.*, vol. MAG-28, 1992.

R.M. White, *Introduction to Magnetic Recording*, New York: IEEE Press, 1984, p. 14.

M.L. Williams and R.L. Comstock, "An analytical model of the write process in digital magnetic recording," *AIP Conf. Proc.*, part 1, no. 5, pp. 738–742, 1971.
J-G. Zhu and H.N. Bertram, "Recording and transition noise simulations in thin film media," *IEEE Trans. Magnet.*, vol. MAG-24, p. 2706, 1988.

Further Information

There are several books which provide additional information on magnetic and magneto-optic recording. They include the following:

R.M. White, *Introduction to Magnetic Recording,* New York: IEEE Press, 1984.
C. D. Mee and E. D. Daniel, *Magnetic Recording,* New York: McGraw-Hill, 1987.
A. S. Hoagland and J. E. Monson, *Digital Magnetic Recording,* New York: John Wiley & Sons, 1991.

Sadiku, M.N.O., Demarest, K. “Wave Propagation”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

37

Wave Propagation

Matthew N.O. Sadiku
Temple University

Kenneth Demarest
University of Kansas

37.1 Space Propagation
Propagation in Simple Media • Propagation in the Atmosphere
37.2 Waveguides
Waveguide Modes • Rectangular Waveguides • Circular Waveguides • Commercially Available Waveguides • Waveguide Losses • Mode Launching

37.1 Space Propagation

Matthew N. O. Sadiku

This section summarizes the basic principles of electromagnetic (EM) **wave propagation** in space. The principles essentially state how the characteristics of the earth and the atmosphere affect the propagation of EM waves. Understanding such principles is of practical interest to communication system engineers. Engineers cannot competently apply formulas or models for communication system design without an adequate knowledge of the propagation issue.

Propagation of an EM wave may be regarded as a means of transferring energy or information from one point (a transmitter) to another (a receiver). EM wave propagation is achieved through guided structures such as transmission lines and waveguides or through space. Wave propagation through waveguides and microstrip lines will be treated in Section 37.2. In this section, our major focus is on EM wave propagation in space and the power resident in the wave.

For a clear understanding of the phenomenon of EM wave propagation, it is expedient to break the discussion of propagation effects into categories represented by four broad frequency intervals [Collin, 1985]:

- Very low frequencies (VLF), 3–30 kHz
- Low-frequency (LF) band, 30–300 kHz
- High-frequency (HF) band, 3–30 MHz
- Above 50 MHz

In the first range, wave propagates as in a waveguide, using the earth's surface and the ionosphere as boundaries. Attenuation is comparatively low, and hence VLF propagation is useful for long-distance worldwide telegraphy and submarine communication. In the second frequency range, the availability of increased bandwidth makes standard AM broadcasting possible. Propagation in this band is by means of surface wave due to the presence of the ground. The third range is useful for long-range broadcasting services via sky wave reflection and refraction by the ionosphere. Basic problems in this band include fluctuations in the ionosphere and a limited usable frequency range. Frequencies above 50 MHz allow for line-of-sight space wave propagation, FM radio and TV channels, radar and navigation systems, and so on. In this band, due consideration must be given to reflection from the ground, refraction by the troposphere, scattering by atmospheric hydrometeors, and **multipath** effects of buildings, hills, trees, etc.

EM wave propagation can be described by two complementary models. The physicist attempts a theoretical model based on universal laws, which extends the field of application more widely than currently known. The engineer prefers an empirical model based on measurements, which can be used immediately. This section presents complementary standpoints by discussing theoretical factors affecting wave propagation and the semiempirical rules allowing handy engineering calculations. First, we consider wave propagation in idealistic simple media, with no obstacles. We later consider the more realistic case of wave propagation around the earth, as influenced by its curvature and by atmospheric conditions.

Propagation in Simple Media

The conventional propagation models, on which the basic calculation of radio links is based, result directly from Maxwell's equations:

$$\nabla \cdot \mathbf{D} = \rho_v \tag{37.1}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{37.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{37.3}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{37.4}$$

In these equations, **E** is electric field strength in volts per meter, **H** is magnetic field strength in amperes per meter, **D** is electric flux density in coulombs per square meter, **B** is magnetic flux density in webers per square meter, **J** is conduction current density in amperes per square meter, and ρ_v is electric charge density in coulombs per cubic meter. These equations go hand in hand with the constitutive equations for the medium:

$$\mathbf{D} = \epsilon \mathbf{E} \tag{37.5}$$

$$\mathbf{B} = \mu \mathbf{H} \tag{37.6}$$

$$\mathbf{J} = \sigma \mathbf{E} \tag{37.7}$$

where $\epsilon = \epsilon_o \epsilon_r$, $\mu = \mu_o \mu_r$, and σ are the permittivity, the permeability, and the conductivity of the medium, respectively.

Consider the general case of a lossy medium which is charge-free ($\rho_v = 0$). Assuming time-harmonic fields and suppressing the time factor $e^{j\omega t}$, Eqs. (37.1) to (37.7) can be manipulated to yield Helmholtz's wave equations

$$\nabla^2 \mathbf{E} - \gamma^2 \mathbf{E} = 0 \tag{37.8}$$

$$\nabla^2 \mathbf{H} - \gamma^2 \mathbf{H} = 0 \tag{37.9}$$

where $\gamma = \alpha + j\beta$ is the **propagation constant**, α is the *attenuation constant* in nepers per meter or decibels per meter, and β is the *phase constant* in radians per meter. Constants α and β are given by

$$\alpha = \omega \sqrt{\frac{\mu\epsilon}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\epsilon}\right)^2} - 1\right]} \tag{37.10}$$

$$\beta = \omega \sqrt{\frac{\mu\epsilon}{2}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2}+1\right]} \tag{37.11}$$

where $\omega = 2\pi f$ is the frequency of the wave. The wavelength λ and wave velocity u are given in terms of β as

$$\lambda = \frac{2\pi}{\beta} \tag{37.12}$$

$$u = \frac{\omega}{\beta} = f\lambda \tag{37.13}$$

Without loss of generality, if we assume that wave propagates in the z-direction and the wave is polarized in the x-direction, solving the wave equations (37.8) and (37.9) results in

$$\mathbf{E}(z,t) = E_o e^{-\alpha z}\cos(\omega t - \beta z)\mathbf{a}_x \tag{37.14}$$

$$\mathbf{H}(z,t) = \frac{E_o}{|\eta|} e^{-\alpha z}\cos(\omega t - \beta z - \theta_\eta)\mathbf{a}_y \tag{37.15}$$

where $\eta = |\eta|\angle\theta_\eta$ is the *intrinsic impedance* of the medium and is given by

$$|\eta| = \frac{\sqrt{\mu/\epsilon}}{\sqrt[4]{\left[1+\left(\frac{\sigma}{\omega\epsilon}\right)\right]^{1/4}}}, \qquad \tan 2\theta_\eta = \frac{\sigma}{\omega\epsilon}, \qquad 0 \le \theta_\eta \le 45^\circ \tag{37.16}$$

Equations (37.14) and (37.15) show that as the EM wave travels in the medium, its amplitude is attenuated according to $e^{-\alpha z}$, as illustrated in Fig. 37.1. The distance δ through which the wave amplitude is reduced by a factor of e^{-1} (about 37%) is called the *skin depth* or *penetration depth* of the medium, i.e.,

$$\delta = \frac{1}{\alpha} \tag{37.17}$$

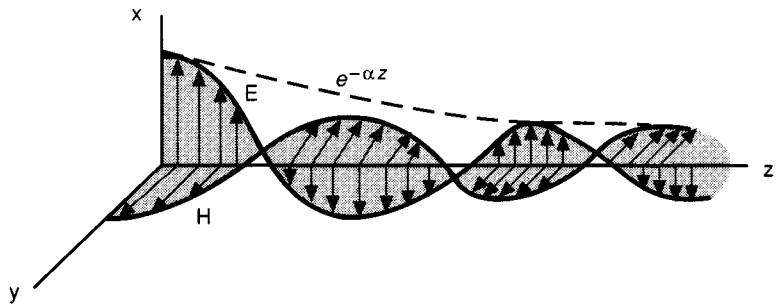


FIGURE 37.1 The magnetic and electric field components of a plane wave in a lossy medium.

The power density of the EM wave is obtained from the Poynting vector

$$\mathbf{P} = \mathbf{E} \times \mathbf{H} \tag{37.18}$$

with the time-average value of

$$\begin{aligned} P_{\text{ave}} &= \frac{1}{2}\,\text{Re}(\mathbf{E} \times \mathbf{H}^*) \\ &= \frac{E_o^2}{2\,|\eta|}\, e^{-2\alpha z} \cos\theta_\eta \mathbf{a}_z \end{aligned} \tag{37.19}$$

It should be noted from Eqs. (37.14) and (37.15) that **E** and **H** are everywhere perpendicular to each other and also to the direction of wave propagation. Thus, the wave described by Eqs. (37.14) and (37.15) is said to be *plane-polarized*, implying that the electric field is always parallel to the same plane (the *xz*-plane in this case) and is perpendicular to the direction of propagation. Also, as mentioned earlier, the wave decays as it travels in the *z*-direction because of loss. This loss is expressed in the *complex relative permittivity* of the medium

$$\epsilon_c = \epsilon_r' - j\epsilon_r'' = \epsilon_r\left(1 - j\,\frac{\sigma}{\omega\epsilon}\right) \tag{37.20}$$

and measured by the *loss tangent*, defined by

$$\tan\delta = \frac{\epsilon_r''}{\epsilon_r'} = \frac{\sigma}{\omega\epsilon} \tag{37.21}$$

The imaginary part $\epsilon_r'' = \sigma/\omega\epsilon_o$ corresponds to the losses in the medium. The refractive index of the medium *n* is given by

$$n = \sqrt{\epsilon_c} \tag{37.22}$$

Having considered the general case of wave propagation through a lossy medium, we now consider wave propagation in other types of media. A medium is said to be a good conductor if the loss tangent is large ($\sigma >> \omega\epsilon$) or a lossless or good dielectric if the loss tangent is very small ($\sigma << \omega\epsilon$). Thus, the characteristics of wave propagation through other types of media can be obtained as special cases of wave propagation in a lossy medium as follows:

1. Good conductors: $\sigma >> \omega\epsilon$, $\epsilon = \epsilon_o$, $\mu = \mu_o\mu_r$
2. Good dielectric: $\sigma << \omega\epsilon$, $\epsilon = \epsilon_o\epsilon_r$, $\mu = \mu_o\mu_r$
3. Free space: $\sigma = 0$, $\epsilon = \epsilon_o$, $\mu = \mu_o$

where $\epsilon_o = 8.854 \times 10^{-12}$ F/m is the free-space permittivity, and $\mu_o = 4\pi \times 10^{-7}$ H/m is the free-space permeability.

The conditions for each medium type are merely substituted in Eqs. (37.10) to (37.21) to obtain the wave properties for that medium. The formulas for calculating attenuation constant, phase constant, and intrinsic impedance for different media are summarized in Table 37.1.

TABLE 37.1 Attenuation Constant, Phase Constant, and Intrinsic Impedance for Different Media

	Lossy Medium	Good Conductor $\sigma/\omega\epsilon >> 1$	Good Dielectric $\sigma/\omega\epsilon << 1$	Free Space
Attenuation constant α	$\omega\sqrt{\frac{\mu\epsilon}{2}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2}-1\right]}$	$\sqrt{\frac{\omega\mu\sigma}{2}}$	$\simeq 0$	0
Phase constant β	$\omega\sqrt{\frac{\mu\epsilon}{2}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2}+1\right]}$	$\sqrt{\frac{\omega\mu\sigma}{2}}$	$\omega\sqrt{\mu\epsilon}$	$\omega\sqrt{\mu_o\epsilon_o}$
Intrinsic impedance η	$\sqrt{\frac{j\omega\mu}{\sigma+j\omega\epsilon}}$	$\sqrt{\frac{\omega\mu}{2\sigma}}(1+j)$	$\sqrt{\frac{\mu}{\epsilon}}$	377

The classical model of a wave propagation presented in this subsection helps us understand some basic concepts of EM wave propagation and the various parameters that play a part in determining the motion of a wave from the transmitter to the receiver. We now apply the ideas to the particular case of wave propagation in the atmosphere.

Propagation in the Atmosphere

Wave propagation hardly occurs under the idealized conditions assumed in the previous subsection. For most communication links, the analysis must be modified to account for the presence of the earth, the ionosphere, and atmospheric precipitates such as fog, raindrops, snow, and hail. This will be done in this subsection.

The major regions of the earth's atmosphere that are of importance in radio wave propagation are the troposphere and the ionosphere. At radar frequencies (approximately 100 MHz to 300 GHz), the troposphere is by far the most important. It is the lower atmosphere consisting of a nonionized region extending from the earth's surface up to about 15 km. The ionosphere is the earth's upper atmosphere in the altitude region from 50 km to one earth radius (6370 km). Sufficient ionization exists in this region to influence wave propagation.

Wave propagation over the surface of the earth may assume one of the following three principal modes:

- Surface wave propagation along the surface of the earth
- Space wave propagation through the lower atmosphere
- Sky wave propagation by reflection from the upper atmosphere

These modes are portrayed in Fig. 37.2. The sky wave is directed toward the ionosphere, which bends the propagation path back toward the earth under certain conditions in a limited frequency range (0–50 MHz approximately). The surface wave is directed along the surface over which the wave is propagated. The space wave consists of the direct wave and the reflected wave. The direct wave travels from the transmitter to the receiver in nearly a straight path, while the reflected wave is due to ground reflection. The space wave obeys the optical laws in that direct and reflected wave components contribute to the total wave. Although the sky and surface waves are important in many applications, we will only consider space waves in this section.

Figure 37.3 depicts the electromagnetic energy transmission between two antennas in space. As the wave radiates from the transmitting antenna and propagates in space, its power density decreases, as expressed ideally in Eq. (37.19). Assuming that the antennas are in free space, the power received by the receiving antenna is given by the *Friis transmission equation* [Liu and Fang, 1988]:

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t \tag{37.23}$$

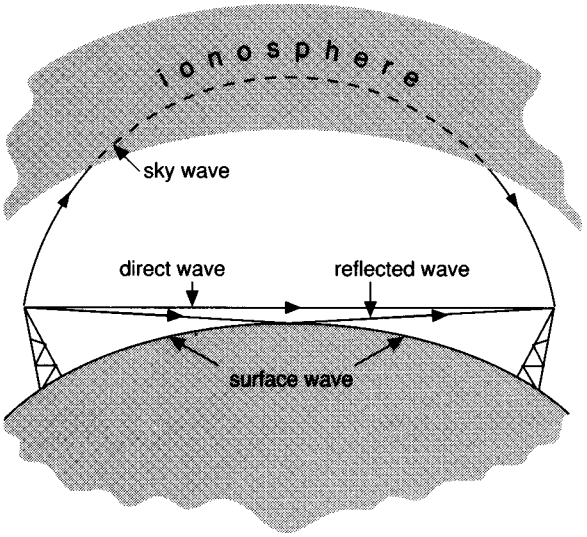


FIGURE 37.2 Modes of wave propagation.

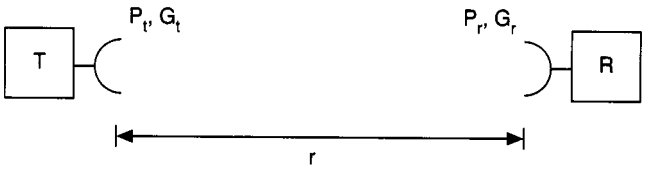


FIGURE 37.3 Transmitting and receiving antennas in free space.

where the subscripts t and r, respectively, refer to transmitting and receiving antennas. In Eq. (37.23), P is the power in watts, G is the antenna gain (dimensionless), r is the distance between the antennas in meters, and λ is the wavelength in meters. The Friis equation relates the power received by one antenna to the power transmitted by the other provided that the two antennas are separated by $r > 2d^2/\lambda$, where d is the largest dimension of either antenna. Thus, the Friis equation applies only when the two antennas are in the far-field of each other. In case the propagation path is not in free space, a correction factor F is included to account for the effect of the medium. This factor, known as the **propagation factor**, is simply the ratio of the electric field intensity E_m in the medium to the electric field intensity E_o in free space, i.e.,

$$F = \frac{E_m}{E_o} \tag{37.24}$$

The magnitude of F is always less than unity since E_m is always less than E_o. Thus, for a lossy medium, Eq. (37.23) becomes

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t \left|F\right|^2 \tag{37.25}$$

For practical reasons, Eqs. (37.23) and (37.25) are commonly expressed in the logarithmic form. If all terms are expressed in decibels (dB), Eq. (37.25) can be written in the logarithmic form as

$$P_r = P_t + G_r + G_t - L_o - L_m \tag{37.26}$$

where P is power in decibels referred to 1 W (or simply dBW), G is gain in decibels, L_o is free-space loss in decibels, and L_m is loss in decibels due to the medium.

The free-space loss is obtained from standard monograph or directly from

$$L_o = 20 \log \left(\frac{4\pi r}{\lambda} \right) \tag{37.27}$$

while the loss due to the medium is given by

$$L_m = -20 \log |F| \tag{37.28}$$

Our major concern in the rest of the section is to determine L_o and L_m for two important cases of space propagation that differ considerably from the free-space conditions.

Effect of the Earth

The phenomenon of multipath propagation causes significant departures from free-space conditions. The term *multipath* denotes the possibility of EM wave propagation along various paths from the transmitter to the receiver. In multipath propagation of an EM wave over the earth's surface, two such paths exist: a direct path and a path via reflection and diffractions from the interface between the atmosphere and the earth. A simplified geometry of the multipath situation is shown in Fig. 37.4. The reflected and diffracted component is commonly separated into two parts, one specular (or coherent) and the other diffuse (or incoherent), that can be separately analyzed. The specular component is well defined in terms of its amplitude, phase, and incident direction. Its main characteristic is its conformance to Snell's law for reflection, which requires that the angles of incidence and reflection be equal and coplanar. It is a plane wave and, as such, is uniquely specified by its direction. The diffuse component, however, arises out of the random nature of the scattering surface and, as such, is nondeterministic. It is not a plane wave and does not obey Snell's law for reflection. It does not come from a given direction but from a continuum.

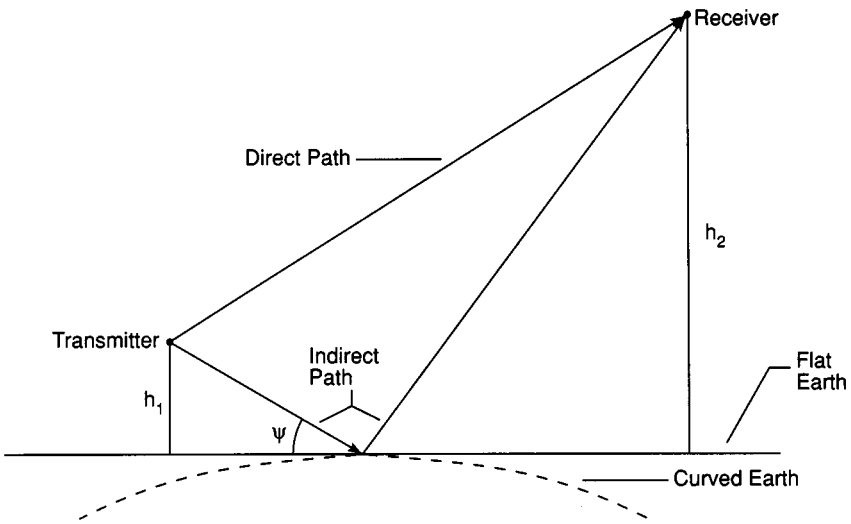


FIGURE 37.4 Multipath geometry.

The loss factor F that accounts for the departures from free-space conditions is given by

$$F = 1 + \Gamma \rho_s D\, S(\theta) e^{-j\Delta} \tag{37.29}$$

where Γ is the Fresnel reflection coefficient, ρ_s is the roughness coefficient, D is the divergence factor, $S(\theta)$ is the shadowing function, and Δ is the phase angle corresponding to the path difference. We now account for each of these terms.

The Fresnel reflection coefficient Γ accounts for the electrical properties of the earth's surface. Because the earth is a lossy medium, the value of the reflection coefficient depends on the complex relative permittivity ϵ_c of the surface, the grazing angle ψ, and the wave polarization. It is given by

$$\Gamma = \frac{\sin\psi - z}{\sin\psi + z} \tag{37.30}$$

where

$$z = \sqrt{\epsilon_c - \cos^2\psi} \qquad \text{for horizontal polarization} \tag{37.31}$$

$$z = \frac{\sqrt{\epsilon_c - \cos^2\psi}}{\epsilon_c} \qquad \text{for vertical polarization} \tag{37.32}$$

$$\epsilon_c = \epsilon_r - j\frac{\sigma}{\omega\epsilon_o} = \epsilon_r - j60\sigma\lambda \tag{37.33}$$

ϵ_r and σ are the dielectric constant and conductivity of the surface; ω and λ are the frequency and wavelength of the incident wave; and ψ is the grazing angle. It is apparent that $0 < |\Gamma| < 1$.

To account for the spreading (or divergence) of the reflected rays because of the earth's curvature, we introduce the divergence factor D. The curvature has a tendency to spread out the reflected energy more than a corresponding flat surface. The divergence factor is defined as the ratio of the reflected field from curved surface to the reflected field from flat surface [Kerr, 1951]. Using the geometry of Fig. 37.5, D is given by

$$D \simeq \left(1 + \frac{2G_1G_2}{a_eG\sin\psi}\right)^{-1/2} \tag{37.34}$$

where $G = G_1 + G_2$ is the total ground range and $a_e = 6370$ km is the effective earth radius. Given the transmitter height h_1, the receiver height h_2, and the total ground range G, we can determine G_1, G_2, and ψ. If we define

$$p = \frac{2}{\sqrt{3}}\left[a_e(h_1 + h_2) + \frac{G^2}{4}\right]^{1/2} \tag{37.35}$$

$$\alpha = \cos^{-1}\left[\frac{2a_e(h_1 - h_2)G}{p^3}\right] \tag{37.36}$$

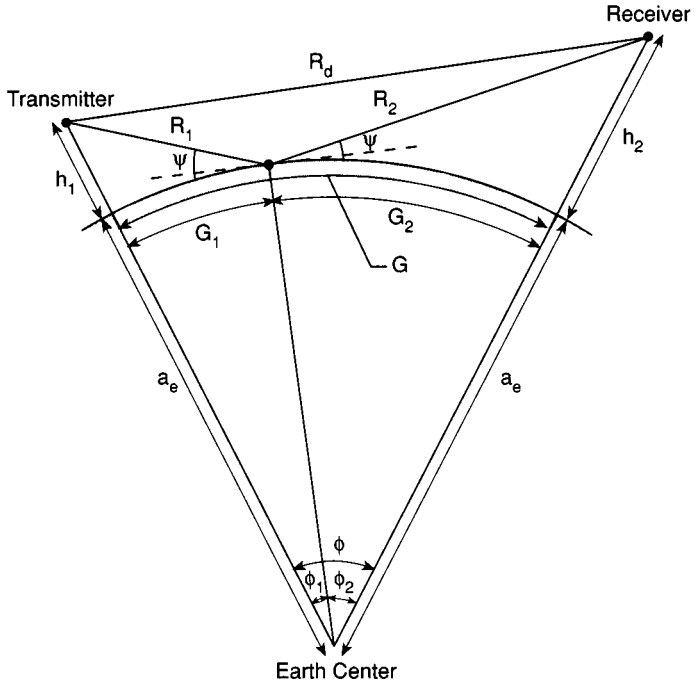


FIGURE 37.5 Geometry of spherical earth reflection.

and assume $h_1 \leq h_2$, $G_1 \leq G_2$, using small angle approximation yields [Blake, 1986]

$$G_1 = \frac{G}{2} + p \cos\left(\frac{\pi + \alpha}{3}\right) \tag{37.37}$$

$$G_2 = G - G_1 \tag{37.38}$$

$$\phi_i = \frac{G_i}{a_e}, \qquad i = 1, 2 \tag{37.39}$$

$$R_i = [h_i^2 + 4a_e(a_e + h_i)\sin^2(\phi_i/2)]^{1/2}, \qquad i = 1, 2 \tag{37.40}$$

The grazing angle is given by

$$\psi = \sin^{-1}\left[\frac{2a_e h_1 + h_1^2 - R_1^2}{2a_e R_1}\right] \tag{37.41}$$

or

$$\psi = \sin^{-1}\left[\frac{2a_e h_1 + h_1^2 + R_1^2}{2(a_e + h_1)R_1}\right] - \phi_1 \tag{37.42}$$

Although D varies from 0 to 1, in practice D is a significant factor at low grazing angle ψ.

The phase angle corresponding to the path difference between direct and reflected waves is given by

$$\Delta = \frac{2\pi}{\lambda}(R_1 + R_2 - R_d) \tag{37.43}$$

The roughness coefficient ρ_s takes care of the fact that the earth's surface is not sufficiently smooth to produce specular (mirrorlike) reflection except at a very low grazing angle. The earth's surface has a height distribution that is random in nature. The randomness arises out of the hills, structures, vegetation, and ocean waves. It is found that the distribution of the heights of the earth's surface is usually the Gaussian or normal distribution of probability theory. If σ_h is the standard deviation of the normal distribution of heights, we define the roughness parameters

$$g = \frac{\sigma_h \sin \psi}{\lambda} \tag{37.44}$$

If $g < 1/8$, specular reflection is dominant; if $g > 1/8$, diffuse scattering results. This criterion, known as *Rayleigh criterion*, should only be used as a guideline since the dividing line between a specular and diffuse reflection or between a smooth and a rough surface is not well defined [Beckman and Spizzichino, 1963]. The roughness is taken into account by the roughness coefficient ($0 < \rho_s < 1$), which is the ratio of the field strength after reflection with roughness taken into account to that which would be received if the surface were smooth. The roughness coefficient is given by

$$\rho_s = \exp[-2(2\pi g)^2] \tag{37.45}$$

The shadowing function $S(\theta)$ is important at a low grazing angle. It considers the effect of geometric shadowing—the fact that the incident wave cannot illuminate parts of the earth's surface shadowed by higher parts. In a geometric approach, where diffraction and multiple scattering effects are neglected, the reflecting surface will consist of well-defined zones of illumination and shadow. As there will be no field on a shadowed portion of the surface, the analysis should include only the illuminated portions of the surface. The phenomenon of shadowing of a stationary surface was first investigated by Beckman in 1965 and subsequently refined by Smith [1967] and others. A pictorial representation of rough surfaces illuminated at angle of incidence θ ($= 90° - \psi$) is shown in Fig. 37.6. It is evident from the figure that the shadowing function $S(\theta)$ is equal to unity when $\theta = 0$ and zero when $\theta = \pi/2$. According to Smith [1967],

$$S(\theta) \simeq \frac{\left[1 - \frac{1}{2}\operatorname{erfc}(a)\right]}{1 + 2B} \tag{37.46}$$

where $\operatorname{erfc}(x)$ is the complementary error function,

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \tag{37.47}$$

and

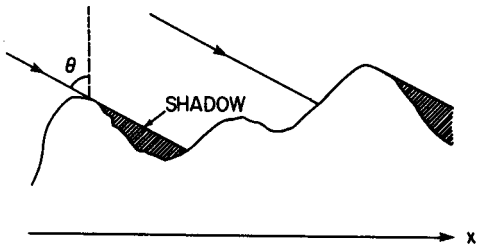


FIGURE 37.6 Rough surface illuminated at an angle of incidence θ.

$$B = \frac{1}{4a}\left[\frac{1}{\sqrt{\pi}}e^{a^2} - a\,\mathrm{erfc}(a)\right] \tag{37.48}$$

$$a = \frac{\cot\theta}{2s} \tag{37.49}$$

$$s = \frac{\sigma_h}{\sigma_l} = \text{rms surface slope} \tag{37.50}$$

In Eq. (37.50) σ_h is the rms roughness height and σ_l is the correlation length. Alternative models for $S(\theta)$ are available in the literature. Using Eqs. (37.30) to (37.50), the loss factor in Eq. (37.29) can be calculated. Thus

$$L_o = 20\log\left(\frac{4\pi R_d}{\lambda}\right) \tag{37.51}$$

$$L_m = -20\log\left[1 + \Gamma\,\rho_s D\,S(\theta)e^{-j\Delta}\right] \tag{37.52}$$

Effect of Atmospheric Hydrometeors

The effect of atmospheric hydrometeors on satellite–earth propagation is of major concern at microwave frequencies. The problem of scattering of electromagnetic waves by atmospheric hydrometeors has attracted much interest since the late 1940s. The main hydrometeors that exist for long duration and have the greatest interaction with microwaves are rain and snow. At frequencies above 10 GHz, rain has been recognized as the most fundamental obstacle on the earth–space path. Rain has been known to cause attenuation, phase difference, and depolarization of radio waves. For analog signals, the effect of rain is more significant above 10 GHz, while for digital signals, rain effects can be significant down to 3 GHz. Attenuation of microwaves because of precipitation becomes severe owing to increased scattering and beam energy absorption by raindrops, thus impairing terrestrial as well as earth–satellite communication links. Cross-polarization distortion due to rain has also engaged the attention of researchers. This is of particular interest when frequency reuse employing signals with orthogonal polarizations is used for doubling the capacity of a communication system. A thorough review on the interaction of microwaves with hydrometeors has been given by Oguchi [1983].

The loss due to a rain-filled medium is given by

$$L_m = \gamma(R)\,\ell_e(\mathrm{R})\,\mathrm{p}(\mathrm{R}) \tag{37.53}$$

where γ is attenuation per unit length at rain rate R, ℓ is the equivalent path length at rain rate R, and p(R) is the probability in percentage of rainfall rate R.

Attenuation is a function of the cumulative rain-rate distribution, drop-size distribution, refractive index of water, temperature, and other variables. A rigorous calculation of $\gamma(R)$ incorporating raindrop-size distribution, velocity of raindrops, and refractive index of water can be found in Sadiku [1992]. For practical engineering purposes, what is needed is a simple formula relating attenuation to rain parameters. Such is found in the aR^b empirical relationship, which has been used to calculate rain attenuation directly [Collin, 1985], i.e.,

$$\gamma(R) = aR^b \text{ dB/km} \tag{37.54}$$

where R is the rain rate and a and b are constants. At 0°C, the values of a and b are related to frequency f in gigahertz as follows:

$$a = G_a f^{Ea} \tag{37.55}$$

where $G_a = 6.39 \times 10^{-5}$, $E_a = 2.03$, for $f < 2.9$ GHz; $G_a = 4.21 \times 10^{-5}$, $E_a = 2.42$, for $2.9 \text{ GHz} \le f \le 54$ GHz; $G_a = 4.09 \times 10^{-2}$, $E_a = 0.699$, for $54 \text{ GHz} \le f < 100$ GHz; $G_a = 3.38$, $E_a = -0.151$, for $180 \text{ GHz} < f$; and

$$b = G_b f^{Eb} \tag{37.56}$$

where $G_b = 0.851$, $E_b = 0.158$, for $f < 8.5$ GHz; $G_b = 1.41$, $E_b = -0.0779$, for $8.5 \text{ GHz} \le f < 25$ GHz; $G_b = 2.63$, $E_b = -0.272$, for $25 \text{ GHz} \le f < 164$ GHz; $G_b = 0.616$, $E_b = 0.0126$, for $164 \text{ GHz} \le f$.

The effective length $\ell_e(R)$ through the medium is needed since rain intensity is not uniform over the path. Its actual value depends on the particular area of interest and therefore has a number of representations [Liu and Fang, 1988]. Based on data collected in western Europe and eastern North America, the effective path length has been approximated as [Hyde, 1984]

$$\ell_e(R) = [0.00741R^{0.766} + (0.232 - 0.00018R)\sin\theta]^{-1} \tag{37.57}$$

where θ is the elevation angle.

The cumulative probability in percentage of rainfall rate R is given by [Hyde, 1984]

$$p(R) = \frac{M}{87.66}[0.03\beta e^{-0.03R} + 0.2(1-\beta)(e^{-0.258R} + 1.86e^{-1.63R})] \tag{37.58}$$

where M is the mean annual rainfall accumulation in millimeters and β is the Rice–Holmberg thunderstorm ratio.

The effect of other hydrometeors such as water vapor, fog, hail, snow, and ice is governed by similar fundamental principles as the effect of rain [Collin, 1985]. In most cases, however, their effects are at least an order of magnitude less than the effect of rain.

Other Effects

Besides hydrometeors, the atmosphere has the composition given in Table 37.2. While attenuation of EM waves by hydrometeors may result from both absorption and scattering, gases act only as absorbers. Although some of these gases do not absorb microwaves, some possess permanent electric and/or magnetic dipole moment and play some part in

TABLE 37.2 Composition of Dry Atmosphere from Sea Level to about 90 km

Constituent	Percent by Volume	Percent by Weight
Nitrogen	78.088	75.527
Oxygen	20.949	23.143
Argon	0.93	1.282
Carbon dioxide	0.03	0.0456
Neon	1.8×10^{-3}	1.25×10^{-3}
Helium	5.24×10^{-4}	7.24×10^{-5}
Methane	1.4×10^{-4}	7.75×10^{-5}
Krypton	1.14×10^{-4}	3.30×10^{-4}
Nitrous oxide	5×10^{-5}	7.60×10^{-5}
Xenon	8.6×10^{-6}	3.90×10^{-5}
Hydrogen	5×10^{-5}	3.48×10^{-6}

Source: D.C. Livingston, *The Physics of Microwave Propagation*, Englewood Cliffs, N.J.: Prentice-Hall, 1970, p. 11. With permission.

microwave absorption. For example, nitrogen molecules do not possess permanent electric or magnetic dipole moment and therefore play no part in microwave absorption. Oxygen has a small magnetic moment, which enables it to display weak absorption lines in the centimeter and millimeter wave regions. Water vapor is a molecular gas with a permanent electric dipole moment. It is more responsive to excitation by an EM field than is oxygen.

Defining Terms

Multipath: Propagation of electromagnetic waves along various paths from the transmitter to the receiver.

Propagation constant: The negative of the partial logarithmic derivative, with respect to the distance in the direction of the wave normal, of the phasor quantity describing a traveling wave in a homogeneous medium.

Propagation factor: The ratio of the electric field intensity in a medium to its value if the propagation took place in free space.

Wave propagation: The transfer of energy by electromagnetic radiation.

Related Topic

35.1 Maxwell Equations

References

P. Beckman and A. Spizzichino, *The Scattering of Electromagnetic Waves from Random Surfaces*, New York: Macmillan, 1963.

L.V. Blake, *Radar Range-Performance Analysis*, Norwood, Mass.: Artech House, 1986, pp. 253–271.

R.E. Collin, *Antennas and Radiowave Propagation*, New York: McGraw-Hill, 1985, pp. 339–456.

G. Hyde, "Microwave propagation," in *Antenna Engineering Handbook*, 2nd ed., R.C. Johnson and H. Jasik, Eds., New York: McGraw-Hill, 1984, pp. 45.1–45.17.

D.E. Kerr, *Propagation of Short Radio Waves*, New York: McGraw-Hill (republished by Peter Peregrinus, London, 1987), 1951, pp. 396–444.

C.H. Liu and D.J. Fang, "Propagation," in *Antenna Handbook: Theory, Applications, and Design*, Y.T. Lo and S.W. Lee, Eds., New York: Van Nostrand Reinhold, 1988, pp. 29.1–29.56.

T. Oguchi, "Electromagnetic wave propagation and scattering in rain and other hydrometeors," *Proc. IEEE*, vol. 71, pp. 1029–1078, 1983.

M.N.O. Sadiku, *Numerical Techniques in Electromagnetics*, Boca Raton, Fla.: CRC Press, 1992, pp. 96–116.

B.G. Smith, "Geometrical shadowing of a random rough surface," *IEEE Trans. Ant. Prog.*, vol. 15, pp. 668–671, 1967.

Further Information

There are several sources of information dealing with the theory and practice of wave propagation in space. Some of these are in the reference section. Journals such as *Radio Science, IEE Proceedings Part H,* and *IEEE Transactions on Antennas and Propagation* are devoted to EM wave propagation. *Radio Science* is available from the American Geophysical Union, 2000 Florida Avenue NW, Washington DC 20009; *IEE Proceedings Part H* from IEE Publishing Department, Michael Faraday House, 6 Hills Way, Stevenage, Herts, SG1 2AY, U.K.; and *IEEE Transactions on Antennas and Propagation* from IEEE, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

Other mechanisms that can affect EM wave propagation in space, not discussed in this section, include clouds, dust, and the ionosphere. The effect of the ionosphere is discussed in detail in standard texts.

37.2 Waveguides

Kenneth Demarest

Waveguide Modes

Any structure that guides electromagnetic waves can be considered a **waveguide**. Most often, however, this term refers to closed metal cylinders that maintain the same cross-sectional dimensions over long distances. Such a structure is shown in Fig. 37.7, which consists of a metal cylinder filled with a dielectric. When filled with low-loss dielectrics (such as air), waveguides typically exhibit lower losses than transmission lines, which makes them useful for transporting RF energy over relatively long distances. They are most often used for frequencies ranging from 1 to 150 GHz.

Every type of waveguide has an infinite number of distinct electromagnetic field configurations that can exist inside it. Each of these configurations is called a **waveguide mode**. The characteristics of these modes depend upon the cross-sectional dimensions of the conducting cylinder, the type of dielectric material inside the waveguide, and the frequency of operation.

Waveguide modes are typically classed according to the nature of the electric and magnetic field components E_z and H_z. These components are called the longitudinal components of the fields. Several types of modes are possible in waveguides:

TE modes: Transverse-electric modes, sometimes called H modes. These modes have $E_z = 0$ at all points within the waveguide, which means that the electric field vector is always perpendicular (i.e., transverse) to the waveguide axis. These modes are always possible in waveguides with uniform dielectrics.

TM modes: Transverse-magnetic modes, sometimes called E modes. These modes have $H_z = 0$ at all points within the waveguide, which means that the magnetic field vector is perpendicular to the waveguide axis. Like TE modes, they are always possible in waveguides with uniform dielectrics.

EH modes: EH modes are hybrid modes in which neither E_z nor H_z are zero, but the characteristics of the transverse fields are controlled more by E_z than H_z. These modes are often possible in waveguides with inhomogeneous dielectrics.

HE modes: HE modes are hybrid modes in which neither E_z nor H_z are zero, but the characteristics of the transverse fields are controlled more by H_z than E_z. Like EH modes, these modes are often possible in waveguides with inhomogeneous dielectrics.

TEM modes: Transverse-electromagnetic modes, often called transmission line modes. These modes can exist only when a second conductor exists within the waveguide, such as a center conductor on a coaxial cable. Because these modes cannot exist in single, closed conductor structures, they are not waveguide modes.

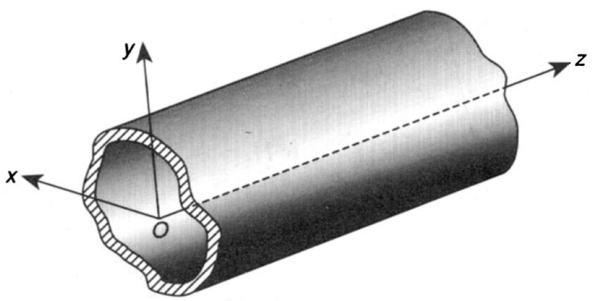


FIGURE 37.7 A uniform waveguide with arbitrary cross section.

Waveguide modes are most easily determined by first computing the longitudinal field components, E_z and H_z, that can be supported by the waveguide. From these, the transverse components (such as E_x and E_y) can easily be found simply by taking spatial derivatives of the longitudinal fields [Collin, 1992].

When the waveguide properties are constant along the z axis, E_z and H_z vary in the longitudinal direction as E_z, $H_z \propto \exp(\omega t - \gamma z)$, where $\omega = 2\pi f$ is the radian frequency of operation and γ is a complex number of the form

$$\gamma = \alpha + j\beta \tag{37.59}$$

The parameters γ, α, and β are called the propagation, attenuation, and phase constants, respectively, and $j = \sqrt{-1}$. When there are no metal or dielectric losses, γ is always either purely real or imaginary. When γ is real, E_z and H_z have constant phase and decay exponentially with increasing z. When γ is imaginary, E_z and H_z vary in phase with increasing z but do not decay in amplitude. When this occurs, the fields are said to be propagating.

When the dielectric is uniform (i.e., homogeneous), E_z and H_z satisfy the scalar wave equation at all points within the waveguide:

$$\nabla_t^2 E_z + h^2 E_z = 0 \tag{37.60}$$

and

$$\nabla_t^2 H_z + h^2 H_z = 0 \tag{37.61}$$

where

$$h^2 = (2\pi f)^2 \mu\epsilon + \gamma^2 = k^2 + \gamma^2 \tag{37.62}$$

Here, μ and ϵ are the permeability and permittivity of the dielectric media, respectively, and $k = 2\pi f\sqrt{\mu\epsilon}$ is the wavenumber of the dielectric. The operator ∇_t^2 is called the transverse Laplacian operator. In Cartesian coordinates,

$$\nabla_t^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

Most of the properties of the allowed modes in real waveguides can usually be found by assuming that the walls are perfectly conducting. Under this condition, $E_z = 0$ and $\partial H_z/\partial p = 0$ at the waveguide walls, where p is the direction perpendicular to the waveguide wall. When these conditions are imposed upon the general solutions of Eqs. (37.60) and (37.61), it is found that only certain values of h are allowed. These values are called the *modal eigenvalues* and are determined by the cross-sectional shape of the waveguide. Using Eq. (37.62), the propagation constant γ for each mode varies with frequency according to

$$\gamma = \alpha + j\beta = h\sqrt{1 - \left(\frac{f}{f_c}\right)^2} \tag{37.63}$$

where

$$f_c = \frac{h}{2\pi\sqrt{\mu\epsilon}} \tag{37.64}$$

The modal parameter f_c has units hertz and is called the **cut-off frequency** of the mode it is associated with. According to Eq. (37.63), when $f > f_c$, the propagation constant γ is imaginary and thus the mode is propagating. On the other hand, when $f < f_c$, γ is real, which means that the fields decay exponentially with increasing values of z. Modes operated at frequencies below their cut-off frequency are not able to propagate energy over long distances and are called evanescent modes.

The dominant mode of a waveguide is the one with the lowest cut-off frequency. Although higher-order modes are often useful for a variety of specialized uses of waveguides, signal distortion is usually minimized when a waveguide is operated in the frequency range where only the dominant mode exists. This range of frequencies is called the *dominant rang*e of the waveguide.

The distance over which the fields of propagating modes repeat themselves is called the **guide wavelength** λ_g. From Eq. (37.63), it can be shown that λ_g always varies with frequency according to

$$\lambda_g = \frac{\lambda_o}{\sqrt{1 - \left(\frac{f_c}{f}\right)^2}} \tag{37.65}$$

where $\lambda_o = 1/(f\sqrt{\mu\epsilon})$ is the wavelength of a plane wave of the same frequency in an infinite sample of the waveguide dielectric. For $f >> f_c$, $\lambda_g \approx \lambda_o$. Also, $\lambda_g \to \infty$ as $f \to f_c$, which is one reason why it is usually undesirable to operate a waveguide mode near modal cut-off frequencies.

Although waveguide modes are not plane waves, the ratio of their transverse electric and magnetic field magnitudes is constant throughout the cross section of the waveguide, just as for plane waves. This ratio is called the modal **wave impedance** and has the following values for TE and TM modes:

$$Z_{TE} = \frac{E_T}{H_T} = \frac{j\omega\mu}{\gamma} \tag{37.66}$$

and

$$Z_{TM} = \frac{E_T}{H_T} = \frac{\gamma}{j\omega\epsilon} \tag{37.67}$$

where E_T and H_T are the magnitudes of the transverse electric and magnetic fields, respectively. In the limit as $f \to \infty$, both Z_{TE} and Z_{TM} approach $\sqrt{\mu\epsilon}$, which is the intrinsic impedance of the dielectric medium. On the other hand, as $f \to f_c$, $Z_{TE} \to \infty$ and $Z_{TM} \to 0$, which means that the transverse electric fields are dominant in TE modes near cut-off and vice versa for TM modes.

Rectangular Waveguides

A rectangular waveguide is shown in Fig. 37.8. The conducting walls are formed such that the inner surfaces form a rectangular cross section, with dimensions a and b along the x and y coordinate axes, respectively.

If the walls are perfectly conducting and the dielectric material is lossless, the field components for the TE_{mn} modes are given by

$$E_x = H_0 \frac{j\omega\mu}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.68}$$

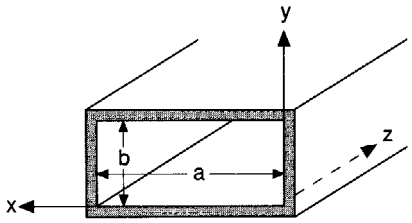


FIGURE 37.8 A rectangular waveguide.

$$E_y = -H_0 \frac{j\omega\mu}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.69}$$

$$E_z = 0$$

$$H_x = H_0 \frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.70}$$

$$H_y = H_0 \frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.71}$$

$$H_z = H_0 \cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.72}$$

where

$$h_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} = 2\pi f_{c_{mn}}\sqrt{\mu\epsilon} \tag{37.73}$$

For the TM_{mn} modes, m and n can be any positive integer value, including zero, as long as both are not zero.

The field components for the TM_{mn} modes are

$$E_x = -E_0 \frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.74}$$

$$E_y = -E_0 \frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.75}$$

$$E_z = E_0 \sin\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.76}$$

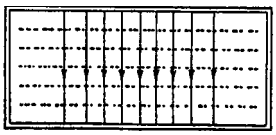


FIGURE 37.9 Field configurations for the TE_{10} (dominant) mode of a rectangular waveguide. Solid lines, E; dashed lines, H. (*Source:* Adapted from N. Marcuvitz, *Waveguide Handbook*, 2nd ed., London: Peter Peregrinus Ltd., and New York: McGraw-Hill, 1986, p. 63. With permission.)

TABLE 37.3 Cut-off Frequencies of the Lowest-Order Rectangular Waveguide Modes (Referenced to the Cut-off Frequency of the Dominant Mode) for a Rectangular Waveguide with $a/b = 2.1$

$f_c/f_{c_{10}}$	Modes
1.0	TE_{10}
2.0	TE_{20}
2.1	TE_{01}
2.326	TE_{11}, TM_{11}
2.9	TE_{21}, TM_{21}
3.0	TE_{30}
3.662	TE_{31}, TM_{31}
4.0	TE_{40}

$$H_x = E_0 \frac{j\omega\epsilon}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.77}$$

$$H_y = -E_0 \frac{j\omega\epsilon}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - \gamma_{mn}z) \tag{37.78}$$

$$H_z = 0 \tag{37.79}$$

where the values of h_{mn} and f_{cmn} are given by Eq. (37.73). For the TM_{mn} modes, m and n can be any positive integer value except zero.

The dominant mode in a rectangular waveguide is the TE_{10} mode, which has a cut-off frequency

$$f_{c_{10}} = \frac{1}{2a\sqrt{\mu\epsilon}} = \frac{c}{2a} \tag{37.80}$$

where c is the speed of light in the dielectric media. The modal field patterns for this mode are shown in Fig. 37.9.

Table 37.3 shows the cut-off frequencies of the lowest-order rectangular waveguide modes (as referenced to the cut-off frequency of the dominant mode) when $a/b = 2.1$. The modal field patterns for several lower-order modes are shown in Fig. 37.10.

Circular Waveguides

A circular waveguide with inner radius a is shown in Fig. 37.11. Here the axis of the waveguide is aligned with the z axis of a circular-cylindrical coordinate system, where ρ and ϕ are the radial and azimuthal coordinates,

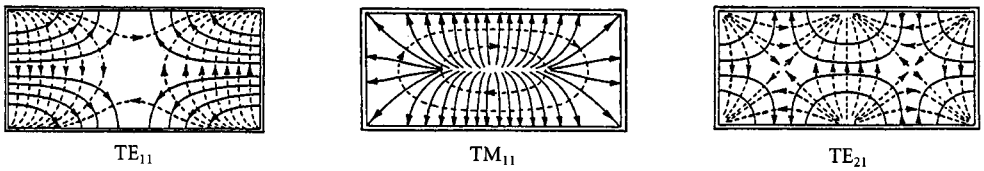


FIGURE 37.10 Field configurations for the TE_{11}, TM_{11}, and the TE_{21} modes. Solid lines, E; dashed lines, H. (*Source:* Adapted from N. Marcuvitz, *Waveguide Handbook*, 2nd. ed., London: Peter Peregrinus Ltd., and New York: McGraw-Hill, 1986, p. 59. With permission.)

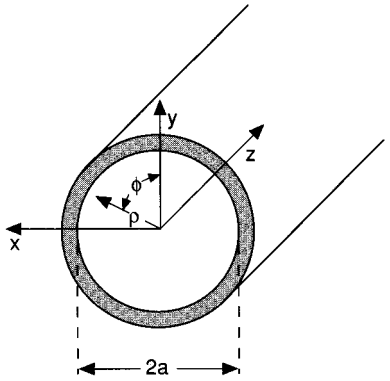


FIGURE 37.11 A circular waveguide.

respectively. If the walls are perfectly conducting and the dielectric material is lossless, the equations for the TE_{nm} modes are

$$E_\rho = H_0 \frac{j\omega\mu n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.81}$$

$$E_\phi = H_0 \frac{j\omega\mu}{h_{nm}} J_n'(h_{nm}\rho) \cos(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.82}$$

$$E_z = 0 \tag{37.83}$$

$$H_\rho = -H_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho) \cos(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.84}$$

$$H_\phi = H_0 \frac{\gamma_{nm}}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.85}$$

$$H_z = H_0 J_n(h_{nm}\rho) \cos(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.86}$$

where n is any positive valued integer, including zero, and $J_n(x)$ and $J_n'(x)$ are the regular Bessel function of order n and its first derivative, respectively. The allowed values of the modal eigenvalues h_{nm} satisfy

$$J_n'(h_{nm}\, a) = 0 \tag{37.87}$$

where m signifies the root number of Eq. (37.87). By convention, $1 < m < \infty$, where $m = 1$ indicates the smallest root.

The equations that define the TM_{nm} modes in circular waveguides are

$$E_\rho = -E_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho) \cos(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.88}$$

$$E_\phi = E_0 \frac{\gamma_{nm}}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.89}$$

$$E_z = E_0 J_n(h_{nm}\rho) \cos(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.90}$$

$$H_\rho = -E_0 \frac{j\omega\epsilon n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.91}$$

$$H_\phi = -E_0 \frac{j\omega\epsilon}{h_{nm}} J_n'(h_{nm}\rho) \cos(n\phi) \exp(j\omega t - \gamma_{nm} z) \tag{37.92}$$

$$H_z = 0 \tag{37.93}$$

where n is any positive valued integer, including zero. For the TM_{nm} modes, the values of the modal eigenvalues are solutions of

$$J_n(h_{nm} a) = 0 \tag{37.94}$$

where m signifies the root number of Eq. (37.94). As in the case of the TE modes, $1 < m < \infty$.

The dominant mode in a circular waveguide is the TE_{11} mode, which has a cut-off frequency given by

$$f_{c_{11}} = \frac{0.293}{a\sqrt{\mu\epsilon}} \tag{37.95}$$

The configuration of the electric and magnetic fields of this mode is shown in Fig. 37.12.

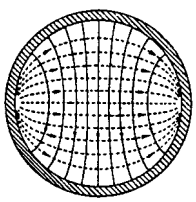


FIGURE 37.12 Field configuration for the TE_{11} (dominant) mode of a circular waveguide. Solid lines, E; dashed lines, H. (*Source:* Adapted from N. Marcuvitz, *Waveguide Handbook*, 2nd ed., London: Peter Peregrinus Ltd., and New York: McGraw-Hill, 1986, p. 68. With permission.)

TABLE 37.4 Cut-off Frequencies of the Lowest-Order Circular Waveguide Modes, Referenced to the Cut-off Frequency of the Dominant Mode

$f_c/f_{c_{11}}$	Modes
1.0	TE_{11}
1.307	TM_{01}
1.66	TE_{21}
2.083	TE_{01}, TM_{11}
2.283	TE_{31}
2.791	TM_{21}
2.89	TE_{41}
3.0	TE_{12}

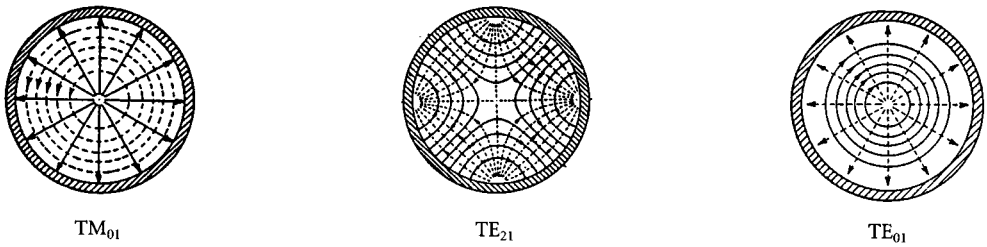


FIGURE 37.13 Field configurations for the TM_{01}, TE_{21}, and TE_{01} circular waveguide modes. Solid lines, E; dashed lines, H. (*Source:* Adapted from N. Marcuvitz, *Waveguide Handbook*, 2nd ed., London: Peter Peregrinus Ltd., and New York: McGraw-Hill, 1986, p. 71. With permission.)

Table 37.4 shows the cut-off frequencies of the lowest-order modes for circular waveguides, referenced to the cut-off frequency of the dominant mode. The modal field patterns for several lower-order modes are shown in Fig. 37.13.

Commercially Available Waveguides

The dimensions of standard rectangular waveguides are given in Table 37.5.

In addition to rectangular and circular waveguides, several other waveguide types are commonly used in microwave applications. Among these are ridge waveguides and elliptical waveguides. The modes of elliptical waveguides can be expressed in terms of Mathieu functions [Kretzschmar, 1970] and are similar to those of circular waveguides but are less perturbed by minor twists and bends of the waveguide. This property makes them attractive for coupling to antennas.

Single-ridge and double-ridge waveguides are shown in Fig. 37.14. The modes of these waveguides bear similarities to those of rectangular guides, but can only be derived numerically [Montgomery, 1971]. Ridge waveguides are useful because their dominant ranges exceed those of rectangular waveguides. However, this range increase is obtained at the expense of higher losses.

Waveguides are also available in a number of constructions, including rigid, semirigid, and flexible. In applications where it is not necessary for the waveguide to bend, rigid construction is always the best since it exhibits the lowest loss. In general, the more flexible the waveguide construction, the higher the loss.

Waveguide Losses

There are two mechanisms that cause losses in waveguides: dielectric losses and metal losses. In both cases, these losses cause the amplitudes of the propagating modes to decay as $\exp(-\alpha z)$, where α is the attenuation constant, measured in units of nepers/meter. Typically, the attenuation constant is considered as the sum of

TABLE 37.5 Standard Rectangular Waveguides

EIA[a] Designation WR[b]( )	Physical Dimensions				Cut-off Frequency for Air-filled Waveguide, GHz	Recommended Frequency Range for TE_{10} Mode, GHZ
	Inside, cm (in.)		Outside, cm (in.)			
	Width	Height	Width	Height		
2300	58.420 (23.000)	29.210 (11.500)	59.055 (23.250)	29.845 (11.750)	0.257	0.32–0.49
2100	53.340 (21.000)	26.670 (10.500)	53.973 (21.250)	27.305 (10.750)	0.281	0.35–0.53
1800	45.720 (18.000)	22.860 (9.000)	46.350 (18.250)	23.495 (9.250)	0.328	0.41–0.62
1500	38.100 (15.000)	19.050 (7.500)	38.735 (15.250)	19.685 (7.750)	0.394	0.49–0.75
1150	29.210 (11.500)	14.605 (5.750)	29.845 (11.750)	15.240 (6.000)	0.514	0.64–0.98
975	24.765 (9.750)	12.383 (4.875)	25.400 (10.000)	13.018 (5.125)	0.606	0.76–1.15
770	19.550 (7.700)	9.779 (3.850)	20.244 (7.970)	10.414 (4.100)	0.767	0.96–1.46
650	16.510 (6.500)	8.255 (3.250)	16.916 (6.660)	8.661 (3.410)	0.909	1.14–1.73
510	12.954 (5.100)	6.477 (2.500)	13.360 (5.260)	6.883 (2.710)	1.158	1.45–2.20
430	10.922 (4.300)	5.461 (2.150)	11.328 (4.460)	5.867 (2.310)	1.373	1.72–2.61
340	8.636 (3.400)	4.318 (1.700)	9.042 (3.560)	4.724 (1.860)	1.737	2.17–3.30
284	7.214 (2.840)	3.404 (1.340)	7.620 (3.000)	3.810 (1.500)	2.079	2.60–3.95
229	5.817 (2.290)	2.908 (1.145)	6.142 (2.418)	3.233 (1.273)	2.579	3.22–4.90
187	4.755 (1.872)	2.215 (0.872)	5.080 (2.000)	2.540 (1.000)	3.155	3.94–5.99
159	4.039 (1.590)	2.019 (0.795)	4.364 (1.718)	2.344 (0.923)	3.714	4.64–7.05
137	3.485 (1.372)	1.580 (0.622)	3.810 (1.500)	1.905 (0.750)	4.304	5.38–8.17
112	2.850 (1.122)	1.262 (0.497)	3.175 (1.250)	1.588 (0.625)	5.263	6.57–9.99
90	2.286 (0.900)	1.016 (0.400)	2.540 (1.000)	1.270 (0.500)	6.562	8.20–12.50
75	1.905 (0.750)	0.953 (0.375)	2.159 (0.850)	1.207 (0.475)	7.874	9.84–15.00
62	1.580 (0.622)	0.790 (0.311)	1.783 (0.702)	0.993 (0.391)	9.494	11.90–18.00
51	1.295 (0.510)	0.648 (0.255)	1.499 (0.590)	0.851 (0.335)	11.583	14.50–22.00
42	1.067 (0.420)	0.432 (0.170)	1.270 (0.500)	0.635 (0.250)	14.058	17.60–26.70
34	0.864 (0.340)	0.432 (0.170)	1.067 (0.420)	0.635 (0.250)	17.361	21.70–33.00
28	0.711 (0.280)	0.356 (0.140)	0.914 (0.360)	0.559 (0.220)	21.097	26.40–40.00
22	0.569 (0.224)	0.284 (0.112)	0.772 (0.304)	0.488 (0.192)	26.362	32.90–50.10
19	0.478 (0.188)	0.239 (0.094)	0.681 (0.268)	0.442 (0.174)	31.381	39.20–59.60
15	0.376 (0.148)	0.188 (0.074)	0.579 (0.228)	0.391 (0.154)	39.894	49.80–75.80

TABLE 37.5 (continued) Standard Rectangular Waveguides

EIA[a] Designation WR[b]( )	Physical Dimensions				Cut-off Frequency for Air-filled Waveguide, GHz	Recommended Frequency Range for TE_{10} Mode, GHZ
	Inside, cm (in.)		Outside, cm (in.)			
	Width	Height	Width	Height		
12	0.310 (0.122)	0.155 (0.061)	0.513 (0.202)	0.358 (0.141)	48.387	60.50–91.90
10	0.254 (0.100)	0.127 (0.050)	0.457 (0.180)	0.330 (0.130)	59.055	73.80–112.00
8	0.203 (0.080)	0.102 (0.040)	0.406 (0.160)	0.305 (0.120)	73.892	92.20–140.00
7	0.165 (0.065)	0.084 (0.033)	0.343 (0.135)	0.262 (0.103)	90.909	114.00–173.00
5	0.130 (0.051)	0.066 (0.026)	0.257 (0.101)	0.193 (0.076)	115.385	145.00–220.00
4	0.109 (0.043)	0.056 (0.022)	0.211 (0.083)	0.157 (0.062)	137.615	172.00–261.00
3	0.086 (0.034)	0.043 (0.017)	0.163 (0.064)	0.119 (0.047)	174.419	217.00–333.00

[a]Electronic Industry Association.
[b]Rectangular waveguide.

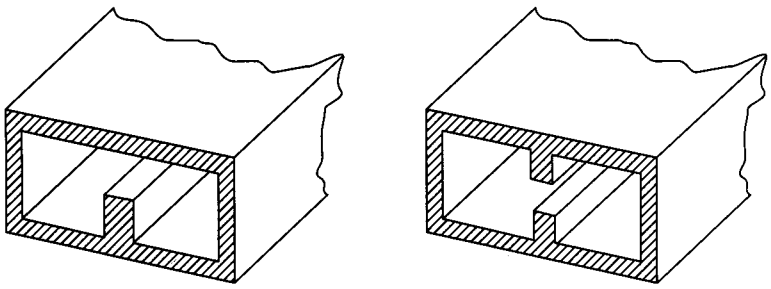


FIGURE 37.14 Single- and double-ridged waveguides.

two components: $\alpha = \alpha_{die} + \alpha_{met}$, where α_{die} and α_{met} are the dielectric and metal attenuation constants, respectively.

The attenuation constant α_{die} can be found directly from Eq. (37.63) simply by generalizing the dielectric wavenumber k to include the effect of the dielectric conductivity σ. For a lossy dielectric, the wavenumber is given by $k^2 = \omega^2\mu\epsilon[1 + (\sigma/j\omega\epsilon)]$. Thus, from Eqs. (37.62) and (37.63) the attenuation constant α_{die} due to dielectric losses is given by

$$\alpha_{die} = \text{real}\left[\sqrt{h^2 - \omega^2\mu\epsilon\left(1 + \frac{\sigma}{j\omega\epsilon}\right)}\right] \tag{37.96}$$

where the allowed values of h are given by Eq. (37.73) for rectangular modes and Eqs. (37.87) and (37.94) for circular modes.

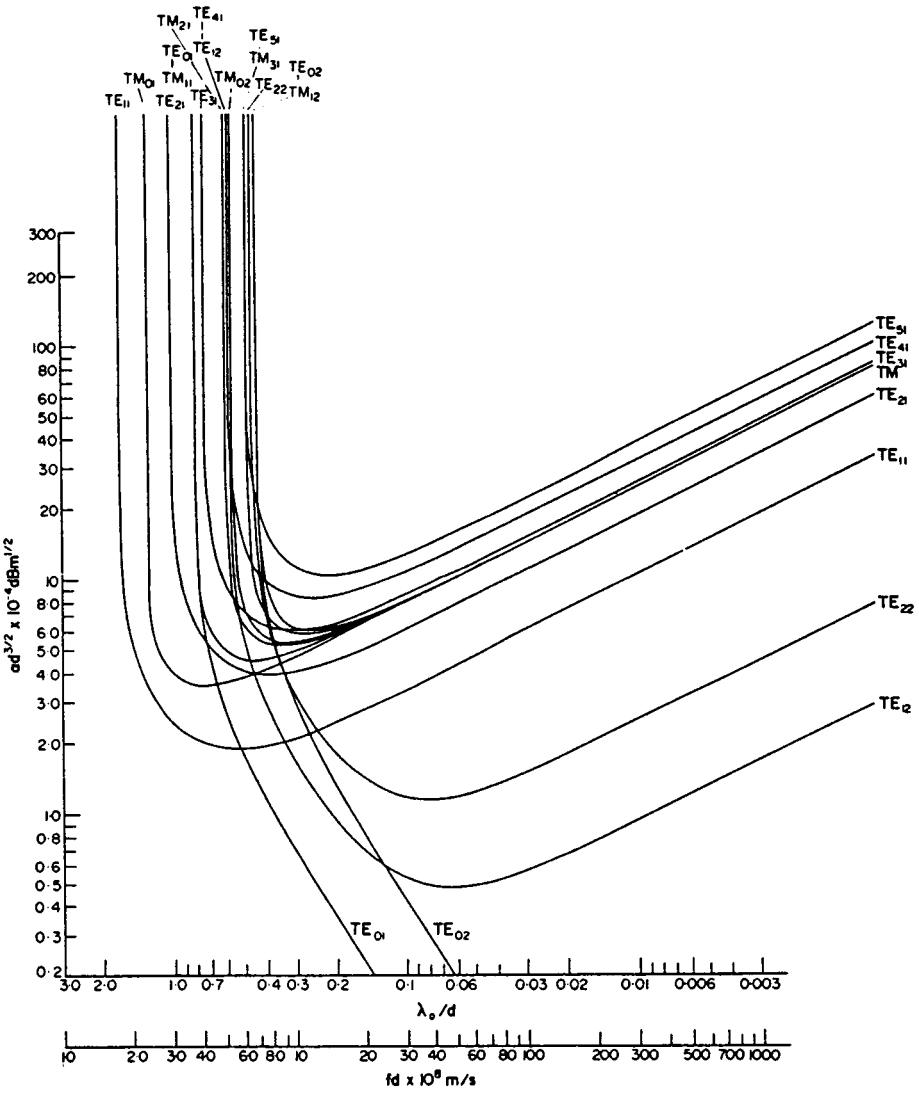


FIGURE 37.15 Values of metallic attenuation constant α for the first few waveguide modes in a circular waveguide of diameter d, plotted against normalized wavelength. (*Source:* A.J. Baden Fuller, *Microwaves,* 2nd ed., New York: Pergamon Press, 1979, p. 138. With permission.)

The metal loss constant α_{met} is usually obtained by assuming that the wall conductivity is high enough to have only a negligible effect on the transverse properties of the modal field patterns. Using this assumption, the power loss in the walls per unit distance along the waveguide can then be calculated to obtain α_{met} [Marcuvitz, 1986]. Figure 37.15 shows the metal attenuation constants for several circular waveguide modes, each normalized to the resistivity R_s of the walls, where $R_s = \sqrt{(\pi f \mu / \sigma)}$ and where μ and σ are the permeability and conductivity of the metal walls, respectively. As can be seen from this figure, the TE_{0m} modes exhibit particularly low loss at frequencies significantly above their cut-off frequencies, making them useful for transporting microwave energy over large distances.

Mode Launching

When coupling electromagnetic energy into a waveguide, it is important to ensure that the desired modes are excited and that reflections back to the source are minimized. Similar concerns must be considered when

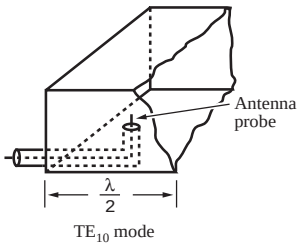


FIGURE 37.16 Coaxial to rectangular waveguide transition that couples the transmission line mode to the dominant waveguide mode

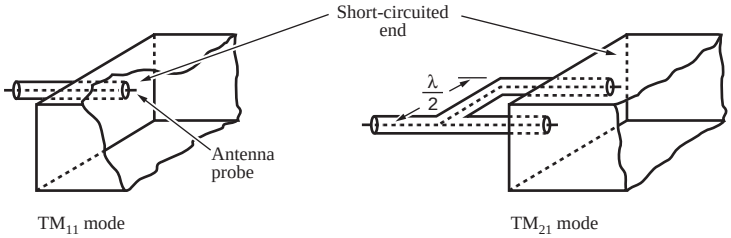


FIGURE 37.17 Coaxial to rectangular waveguide transitions that couple the transmission line mode to the TM_{11} and TM_{21} waveguide modes.

coupling energy from a waveguide to a transmission line or circuit element. This is achieved by using launching (or coupling) structures that allow strong coupling between the desired modes on both structures.

Figure 37.16 shows a mode launching structure for coaxial cable to rectangular waveguide transitions.

This structure provides good coupling between the TEM (transmission line) mode on a coaxial cable and the TE_{10} mode in the waveguide because the antenna probe excites a strong transverse electric field in the center of the waveguide, directed between the broad walls. The distance between the probe and the short circuit back wall is chosen to be approximately $\lambda/4$, which allows the TE_{10} mode launched in this direction to reflect off the short circuit and arrive in phase with the mode launched towards the right.

Launching structures can also be devised to launch higher-order modes. Mode launchers that couple the transmission line mode on a coaxial cable to the TM_{11} and TM_{21} waveguide mode are shown in Fig. 37.17.

Defining Terms

Cut-off frequency: The minimum frequency at which a waveguide mode will propagate energy with little or no attenuation.

Guide wavelength: The distance over which the fields of propagating modes repeat themselves in a waveguide.

Waveguide: A closed metal cylinder, filled with a dielectric, used to transport electromagnetic energy over short or long distances.

Waveguide modes: Unique electromagnetic field configurations supported by a waveguide that have distinct electrical characteristics.

Wave impedance: The ratio of the transverse electric and magnetic fields inside a waveguide.

Related Topics

35.1 Maxwell Equations • 39.1 Passive Microwave Devices • 42.1 Lightwave Waveguides

References

A. J. Baden Fuller, *Microwaves*, 2nd ed., New York: Pergamon Press, 1979.
R. E. Collin, *Foundations for Microwave Engineering*, 2nd ed., New York: McGraw-Hill, 1992.
J. Kretzschmar, "Wave propagation in hollow conducting elliptical waveguides," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-18, no. 9, pp. 547–554, Sept. 1970.
S. Y. Liao, *Microwave Devices and Circuits*, 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1990.
N. Marcuvitz, *Waveguide Handbook*, 2nd ed., London: Peter Peregrinus Ltd., 1986.
J. Montgomery, "On the complete eigenvalue solution of ridged waveguide," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-19, no. 6, pp. 457–555, June 1971.

Further Information

There are many textbooks and handbooks that cover the subject of waveguides in great detail. In addition to the references cited above, others include

L. Lewin, *Theory of Waveguides*, New York: John Wiley, 1975.
Reference Data for Radio Engineers, Howard W. Sams Co., 1975.
R. E. Collin, *Field Theory of Guided Waves*, 2nd ed., Piscataway, N.J.: IEEE Press, 1991.
F. Gardiol, *Introduction to Microwaves*, Dedham, Mass.: Artech House, 1984.
S. Ramo, J. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, New York: John Wiley, 1965.

Kolias, N.J., Compton, R.C., Fitch, J.P., Pozar, D.M. “Antennas”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

38

Antennas

N.J. Kolias
Raytheon Company

R.C. Compton
Cornell University

J. Patrick Fitch
Lawrence Livermore Laboratory

David M. Pozar
University of Massachusetts at Amherst

38.1 Wire
Short Dipole • Directivity • Magnetic Dipole • Input Impedance • Arbitrary Wire Antennas • Resonant Half-Wavelength Antenna • End Loading • Arrays of Wire Antennas • Analysis of General Arrays • Arrays of Identical Elements • Equally Spaced Linear Arrays • Planar (2-D) Arrays • Yagi–Uda Arrays • Log-Periodic Dipole Arrays

38.2 Aperture
The Oscillator or Discrete Radiator • Synthetic Apertures • Geometric Designs • Continuous Current Distributions (Fourier Transform) • Antenna Parameters

38.3 Microstrip Antennas
Introduction • Basic Microstrip Antenna Element • Feeding Techniques for Microstrip Antennas • Microstrip Antenna Arrays • Computer-Aided Design for Microstrip Antennas

38.1 Wire

N.J. Kolias and R.C. Compton

Antennas have been widely used in communication systems since the early 1900s. Over this span of time scientists and engineers have developed a vast number of different antennas. The radiative properties of each of these antennas are described by an antenna pattern. This is a plot, as a function of direction, of the power P_r per unit solid angle Ω radiated by the antenna. The antenna pattern, also called the **radiation pattern,** is usually plotted in spherical coordinates θ and φ. Often two orthogonal cross sections are plotted, one where the *E*-field lies in the plane of the slice (called the *E*-plane) and one where the *H*-field lies in the plane of the slice (called the *H*-plane).

Short Dipole

Antenna patterns for a short dipole are plotted in Fig. 38.1. In these plots the radial distance from the origin to the curve is proportional to the radiated power. Antenna plots are usually either on linear scales or decibel scales (10 log power).

The antenna pattern for a short dipole may be determined by first calculating the vector potential **A** [Collin, 1985; Balanis, 1982; Harrington, 1961; Lorrain and Corson, 1970]. Using Collin's notation, the vector potential in spherical coordinates is given by

$$\mathbf{A} = \mu_0 I\, dl\, \frac{e^{-jk_0 r}}{4\pi r} (\mathbf{a}_r \cos\theta - \mathbf{a}_\theta \sin\theta) \tag{38.1}$$

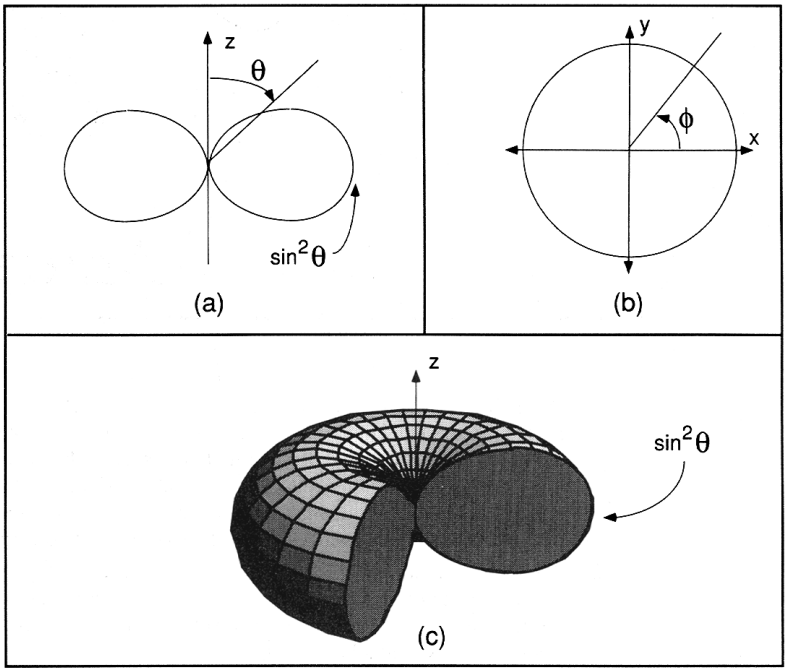


FIGURE 38.1 Radiation pattern for a short dipole of length dl ($dl << \lambda_0$). These are plots of power density on linear scales. (a) *E*-plane; (b) *H*-plane; (c) three-dimensional view with cutout.

where $k_0 = 2\pi/\lambda_0$, and I is the current, assumed uniform, in the short dipole of length dl ($dl << \lambda_0$). Here the assumed time dependence $e^{j\omega t}$ has not been explicitly shown. The electric and magnetic fields may then be determined using

$$\mathbf{E} = -j\omega\mathbf{A} + \frac{\nabla\nabla\cdot\mathbf{A}}{j\omega\mu_0\varepsilon_0} \qquad \mathbf{H} = \frac{1}{\mu_0}\nabla\times\mathbf{A} \tag{38.2}$$

The radiated fields are obtained by calculating these fields in the so-called *far-field region* where $r >> \lambda$. Doing this for the short dipole yields

$$\begin{aligned}\mathbf{E} &= jZ_0 Idlk_0 \sin\theta\,\frac{e^{-jk_0 r}}{4\pi r}\,\mathbf{a}_\theta \\ \mathbf{H} &= jIdlk_0 \sin\theta\,\frac{e^{-jk_0 r}}{4\pi r}\,\mathbf{a}_\varphi\end{aligned} \tag{38.3}$$

where $Z_0 = \sqrt{\mu_0/\varepsilon_0}$. The average radiated power per unit solid angle Ω can then be found to be

$$\frac{\Delta P_r(\theta,\varphi)}{\Delta\Omega} = \frac{1}{2}r^2\Re\mathrm{e}\{\mathbf{E}\times\mathbf{H}^*\cdot\mathbf{a}_r\} = |I|^2 Z_0(dl)^2 k_0^2\,\frac{\sin^2\theta}{32\pi^2} \tag{38.4}$$

Directivity

The **directivity** $D(\theta,\varphi)$ and **gain** $G(\theta,\varphi)$ of an antenna are defined as

$$\begin{aligned} D(\theta,\varphi) &= \frac{\text{Radiated power per solid angle}}{\text{Total radiated power}/4\pi} = \frac{\Delta P_r(\theta,\varphi)/\Delta\Omega}{P_r/4\pi} \\ G(\theta,\varphi) &= \frac{\text{Radiated power per solid angle}}{\text{Total input power}/4\pi} = \frac{\Delta P_r(\theta,\varphi)/\Delta\Omega}{P_{\text{in}}/4\pi} \end{aligned} \tag{38.5}$$

Antenna efficiency, η, is given by

$$\eta \equiv \frac{P_r}{P_{\text{in}}} = \frac{G(\theta,\varphi)}{D(\theta,\varphi)} \tag{38.6}$$

For many antennas $\eta \approx 1$ and so the words *gain* and *directivity* can be used interchangeably. For the short dipole

$$D(\theta,\varphi) = \frac{3}{2}\sin^2\theta \tag{38.7}$$

The maximum directivity of the short dipole is 3/2. This single number is often abbreviated as the antenna directivity. By comparison, for an imaginary isotropic antenna which radiates equally in all directions, $D(\theta,\varphi) = 1$. The product of the maximum directivity with the total radiated power is called the *effective isotropic radiated power* (EIRP). It is the total radiated power that would be required for an isotropic radiator to produce the same signal as the original antenna in the direction of maximum directivity.

Magnetic Dipole

A small loop of current produces a *magnetic dipole*. The far fields for the magnetic dipole are dual to those of the electric dipole. They have the same angular dependence as the fields of the electric dipole, but the polarization orientations of **E** and **H** are interchanged.

$$\begin{aligned} \mathbf{H} &= -Mk_0^2 \sin\theta \frac{e^{-jk_0 r}}{4\pi r}\mathbf{a}_\theta \\ \mathbf{E} &= MZ_0 k_0^2 \sin\theta \frac{e^{-jk_0 r}}{4\pi r}\mathbf{a}_\varphi \end{aligned} \tag{38.8}$$

where $M = \pi\ r_0^2\ I$ for a loop with radius r_0 and uniform current I.

Input Impedance

At a given frequency the impedance at the feedpoint of an antenna can be represented as $Z_a = R_a + jX_a$. The real part of Z_a (known as the input resistance) corresponds to radiated fields plus losses, while the imaginary part (known as the input reactance) arises from stored evanescent fields. The radiation resistance is obtained from $R_a = 2P_r/|I|^2$ where P_r is the total radiated power and I is the input current at the antenna terminals. For electrically small electric and magnetic dipoles with uniform currents

$$R_a = 80\pi^2\left(\frac{dl}{\lambda_0}\right)^2 \quad \text{electric dipole}$$
$$R_a = 320\pi^6\left(\frac{r_0}{\lambda_0}\right)^4 \quad \text{magnetic dipole} \tag{38.9}$$

The reactive component of Z_a can be determined from $X_a = 4\omega(W_m\text{-}W_e)/|I|^2$ where W_m is the average magnetic energy and W_e is the average electric energy stored in the near-zone evanescent fields. The reflection coefficient, Γ, of the antenna is just

$$\Gamma = \frac{Z_a - Z_0}{Z_a + Z_0} \tag{38.10}$$

where Z_0 is the characteristic impedance of the system used to measure the reflection coefficient.

Arbitrary Wire Antennas

An arbitrary wire antenna can be considered as a sum of small current dipole elements. The vector potential for each of these elements can be determined in the same way as for the short dipole. The total vector potential is then the sum over all these infinitesimal contributions and the resulting E in the far field can be found to be

$$\mathbf{E}(r) = jk_0Z_0\frac{e^{-jk_0r}}{4\pi r}\int_c[(\mathbf{a}_r\cdot\mathbf{a})\mathbf{a}_r - \mathbf{a}]I(l')e^{jk_0\mathbf{a}_r\cdot\mathbf{r}'}dl' \tag{38.11}$$

where the integral is over the contour C of the wire, $\mathbf{a}$ is a unit vector tangential to the wire, and $\mathbf{r}'$ is the radial vector to the infinitesimal current element.

Resonant Half-Wavelength Antenna

The resonant half-wavelength antenna (commonly called the half-wave dipole) is used widely in antenna systems. Factors contributing to its popularity are its well-understood radiation pattern, its simple construction, its high efficiency, and its capability for easy impedance matching.

The electric and magnetic fields for the half-wave dipole can be calculated by substituting its current distribution, $I = I_0\cos(k_0z)$, into Eq. (38.11) to obtain

$$\mathbf{E} = jZ_0I_0\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\frac{e^{-jk_0r}}{2\pi r}\mathbf{a}_\theta$$
$$\mathbf{H} = jI_0\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\frac{e^{-jk_0r}}{2\pi r}\mathbf{a}_\varphi \tag{38.12}$$

The total radiated power, P_r, can be determined from the electric and magnetic fields by integrating the expression $1/2\ \Re\mathrm{e}\{\mathbf{E}\times\mathbf{H}^*\cdot\mathbf{a}_r\}$ over a surface of radius r. Carrying out this integration yields $P_r = 36.565\ |I_0|^2$. The radiation resistance of the half-wave dipole can then be determined from

$$R_a = \frac{2P_r}{|I_0|^2} \approx 73\,\Omega \tag{38.13}$$

This radiation resistance is considerably higher than the radiation resistance of a short dipole. For example, if we have a dipole of length 0.01λ, its radiation resistance will be approximately $0.08\,\Omega$ (from Eq. 38.9). This resistance is probably comparable to the ohmic resistance of the dipole, thereby resulting in a low efficiency. The half-wave dipole, having a much higher radiation resistance, will have much higher efficiency. The higher resistance of the half-wave dipole also makes impedance matching easier.

End Loading

At many frequencies of interest, for example, the broadcast band, a half-wavelength becomes unreasonably long. Figure 38.2 shows a way of increasing the effective length of the dipole without making it longer. Here, additional wires have been added to the ends of the dipoles. These wires increase the end capacitance of the dipole, thereby increasing the effective electrical length.

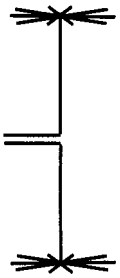


FIGURE 38.2 Using end loading to increase the effective electrical length of an electric dipole.

Arrays of Wire Antennas

Often it is advantageous to have several antennas operating together in an **array.** Arrays of antennas can be made to produce highly directional radiation patterns. Also, small antennas can be used in an array to obtain the level of performance of a large antenna at a fraction of the area.

The radiation pattern of an array depends on the number and type of antennas used, the spacing in the array, and the relative phase and magnitude of the excitation currents. The ability to control the phase of the exciting currents in each element of the array allows one to electronically scan the main radiated beam. An array that varies the phases of the exciting currents to scan the radiation pattern through space is called an electronically scanned **phased array.** Phased arrays are used extensively in radar applications.

Analysis of General Arrays

To obtain analytical expressions for the radiation fields due to an array one must first look at the fields produced by a single array element. For an isolated radiating element positioned as in Fig. 38.3, the electric field at a far-field point P is given by

$$\mathbf{E}_i = a_i \mathbf{K}_i(\theta, \varphi) e^{j[k_0(\mathbf{R}_i \cdot \mathbf{i}_p) - \alpha_i]} \tag{38.14}$$

where $\mathbf{K}_i(\theta,\varphi)$ is the electric field pattern of the individual element, $a_i\, e^{-j\alpha i}$ is the excitation of the individual element, $\mathbf{R}_i$ is the position vector from the phase reference point to the element, $\mathbf{i}_p$ is a unit vector pointing toward the far-field point P, and k_0 is the free space wave vector.

Now, for an array of N of these arbitrary radiating elements the total E-field at position P is given by the vector sum

$$\mathbf{E}_{\text{tot}} = \sum_{i=0}^{N-1} \mathbf{E}_i = \sum_{i=0}^{N-1} a_i \mathbf{K}_i(\theta, \varphi) e^{j[k_0(\mathbf{R}_i \cdot \mathbf{i}_p) - \alpha_i]} \tag{38.15}$$

This equation may be used to calculate the total field for an array of antennas where the mutual coupling between the array elements can be neglected. For most practical antennas, however, there is mutual coupling,

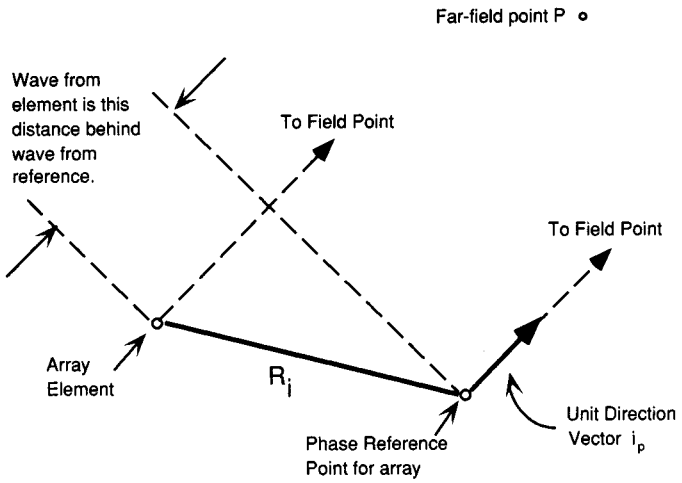


FIGURE 38.3 Diagram for determining the far field due to radiation from a single array element. (*Source: Reference Data for Radio Engineers,* Indianapolis: Howard W. Sams & Co., 1975, chap. 27–22. With permission.)

and the individual patterns will change when the element is placed in the array. Thus, Eq. (38.15) should be used with care.

Arrays of Identical Elements

If all the radiating elements of an array are identical, then $\mathbf{K}_i(\theta,\varphi)$ will be the same for each element and Eq. (38.15) can be rewritten as

$$\mathbf{E}_{\text{tot}} = \mathbf{K}(\theta,\varphi)\sum_{i=0}^{N-1} a_i e^{j[k_0(\mathbf{R}_i\cdot\mathbf{i}_p)-\alpha_i]} \tag{38.16}$$

This can also be written as

$$\mathbf{E}_{\text{tot}} = \mathbf{K}(\theta,\varphi)\, f(\theta,\varphi) \qquad \text{where} \quad f(\theta,\varphi) = \sum_{i=0}^{N-1} a_i e^{j[k_0(\mathbf{R}_i\cdot\mathbf{i}_p)-\alpha_i]} \tag{38.17}$$

The function $f(\theta,\varphi)$ is normally called the array factor or the array polynomial. Thus, one can find $\mathbf{E}_{\text{tot}}$ by just multiplying the individual element's electric field pattern, $\mathbf{K}(\theta,\varphi)$, by the array factor, $f(\theta,\varphi)$. This process is often referred to as pattern multiplication.

The average radiated power per unit solid angle is proportional to the square of $\mathbf{E}_{\text{tot}}$. Thus, for an array of identical elements

$$\frac{\Delta P_r(\theta,\varphi)}{\Delta\Omega} \sim \left|\mathbf{K}(\theta,\varphi)\right|^2 \left|f(\theta,\varphi)\right|^2 \tag{38.18}$$

Equally Spaced Linear Arrays

An important special case occurs when the array elements are identical and are arranged on a straight line with equal element spacing, *d,* as shown in Fig. 38.4. If a linear phase progression, α, is assumed for the excitation currents of the elements, then the total field at position *P* in Fig. 38.4 will be

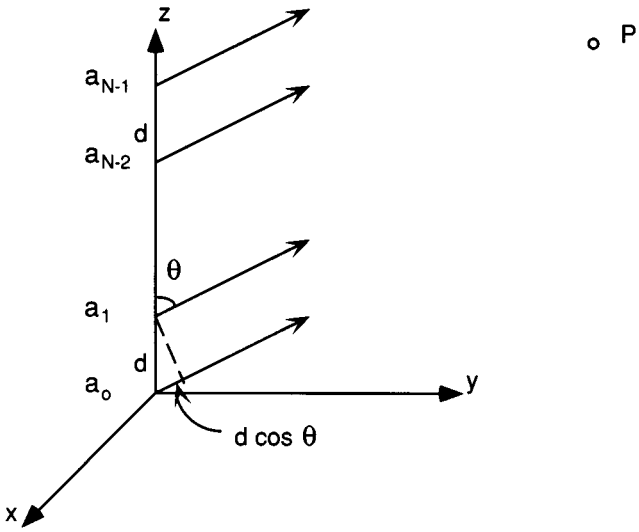


FIGURE 38.4 A linear array of equally spaced elements.

$$\begin{aligned} \mathbf{E}_{\text{tot}} &= \mathbf{K}(\theta,\varphi)\sum_{n=0}^{N-1} a_n e^{jn(k_0 d\cos\theta-\alpha)} \\ &= \mathbf{K}(\theta,\varphi)\sum_{n=0}^{N-1} a_n e^{jn\psi} = \mathbf{K}(\theta,\varphi) f(\psi) \end{aligned} \tag{38.19}$$

where $\psi = k_0 d\cos\theta - \alpha$.

Broadside Arrays

Suppose that, in the linear array of Fig. 38.4, all the excitation currents are equal in magnitude and phase ($a_0 = a_1 = \ldots = a_{N-1}$ and $\alpha = 0$). The array factor, $f(\psi)$, then becomes

$$f(\psi) = a_0\sum_{n=0}^{N-1} e^{jn\psi} = a_0\,\frac{1-e^{jN\psi}}{1-e^{j\psi}} \tag{38.20}$$

This can be simplified to obtain the normalized form

$$f'(\psi) = \left|\frac{f(\psi)}{a_0 N}\right| = \left|\frac{\sin\dfrac{N\psi}{2}}{N\sin\dfrac{\psi}{2}}\right| \tag{38.21}$$

Note that $f'(\psi)$ is maximum when $\psi = 0$. For our case, with $\alpha = 0$, we have $\psi = k_0 d\cos\theta$. Thus $f'(\psi)$ will be maximized when $\theta = \pi/2$. This direction is perpendicular to the axis of the array (see Fig. 38.4), and so the resulting array is called a broadside array.

Phased Arrays

By adjusting the phase of the elements of the array it is possible to vary the direction of the maximum of the array's radiation pattern. For arrays where all the excitation currents are equal in magnitude but not necessarily phase, the array factor is a maximum when $\psi = 0$. From the definition of ψ, one can see that at the pattern maximum

$$k_0 d \cos\theta = \alpha$$

Thus, the direction of the array factor maximum is given by

$$\theta = \cos^{-1}\left(\frac{\alpha}{k_0 d}\right) \tag{38.21b}$$

Note that if one is able to control the phase delay, α, the direction of the maximum can be scanned without physically moving the antenna.

Planar (2-D) Arrays

Suppose there are M linear arrays, all identical to the one pictured in Fig. 38.4, lying in the yz-plane with element spacing d in both the y and the z direction. Using the origin as the phase reference point, the array factor can be determined to be

$$f(\theta, \varphi) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} a_{mn} e^{[jn(k_0\, d \cos\theta - \alpha_z) + jm(k_0\, d \sin\theta \sin\varphi - \alpha_y)]} \tag{38.22}$$

where α_y and α_z are the phase differences between the adjacent elements in the y and z directions, respectively. The formula can be derived by considering the 2-D array to be a 1-D array of subarrays, where each subarray has an antenna pattern given by Eq. (38.19).

If all the elements of the 2-D array have excitation currents equal in magnitude and phase (all the a_{mn} are equal and $\alpha_z = \alpha_y = 0$), then the array will be a broadside array and will have a normalized array factor given by

$$f'(\theta, \varphi) = \frac{\sin\left(\frac{Nk_0 d}{2}\cos\theta\right)}{N \sin\left(\frac{k_0 d}{2}\cos\theta\right)} \frac{\sin\left(\frac{Mk_0 d}{2}\sin\theta\sin\varphi\right)}{M \sin\left(\frac{k_0 d}{2}\sin\theta\sin\varphi\right)} \tag{38.23}$$

Yagi–Uda Arrays

The Yagi–Uda array can be found on rooftops all over the world—the standard TV antenna is a Yagi–Uda array. The Yagi–Uda array avoids the problem of needing to control the feeding currents to all of the array elements by driving only one element. The other elements in the Yagi–Uda array are excited by near-field coupling from the driven element.

The basic three-element Yagi–Uda array is shown in Fig. 38.5. The array consists of a driven antenna of length l_1, a reflector element of length l_2, and a director element of length l_3. Typically, the director element is shorter than the driven element by 5% or more, while the reflector element is longer than the driven element by 5% or more [Stutzman and Thiele, 1981]. The radiation pattern for the array in Fig. 38.5 will have a maximum in the $+z$ direction.

One can increase the gain of the Yagi–Uda array by adding additional director elements. Adding additional reflector elements, however, has little effect because the field behind the first reflector element is small.

Yagi–Uda arrays typically have directivities between 10 and 100, depending on the number of directors [Ramo et al., 1984]. TV antennas usually have several directors.

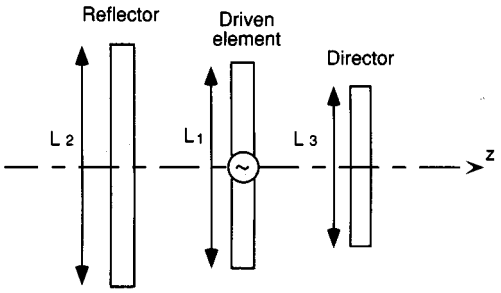


FIGURE 38.5 Three-element Yagi–Uda antenna. (*Source:* Shintaro Uda and Yasuto Mushiake, *Yagi–Uda Antenna,* Sendai, Japan: Sasaki Printing and Publishing Company, 1954, p. 100. With permission.)

Log-Periodic Dipole Arrays

Another variation of wire antenna arrays is the log-periodic dipole array. The log-periodic is popular in applications that require a broadband, frequency-independent antenna. An antenna will be independent of frequency if its dimensions, when measured in wavelengths, remain constant for all frequencies. If, however, an antenna is designed so that its characteristic dimensions are periodic with the logarithm of the frequency, and if the characteristic dimensions do not vary too much over a period of time, then the antenna will be essentially frequency independent. This is the basis for the log-periodic dipole array, shown in Fig. 38.6.

In Fig 38.6, the ratio of successive element positions equals the ratio of successive dipole lengths. This ratio is often called the scaling factor of the log-periodic array and is denoted by

$$\tau = \frac{z_{n+1}}{z_n} = \frac{L_{n+1}}{L_n} \tag{38.24}$$

Also note that there is a mechanical phase reversal between successive elements in the array caused by the crossing over of the interconnecting feed lines. This phase reversal is necessary to obtain the proper phasing between adjacent array elements.

To get an idea of the operating range of the log-periodic antenna, note that for a given frequency within the operating range of the antenna, there will be one dipole in the array that is half-wave resonant or is nearly so. This half-wave resonant dipole and its immediate neighbors are called the active region of the log-periodic array. As the operating frequency changes, the active region shifts to a different part of the log-periodic. Hence, the frequency range for the log-periodic array is roughly given by the frequencies at which the longest and shortest dipoles in the array are half-wave resonant (wavelengths such that $2L_N < \lambda < 2L_1$) [Stutzman and Thiele, 1981].

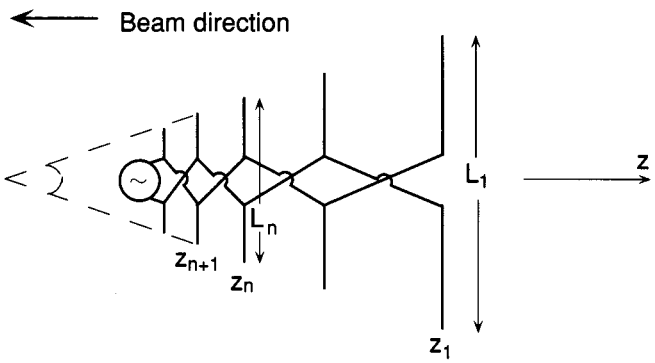


FIGURE 38.6 The log-periodic dipole array. (*Source:* D.G. Isbell, "Log periodic dipole arrays," *IRE Transactions on Antennas and Propagation,* vol. AP-8, p. 262, 1960. With permission.)

Defining Terms

Antenna gain: The ratio of the actual radiated power per solid angle to the radiated power per solid angle that would result if the total input power were radiated isotropically.

Array: Several antennas arranged together in space and interconnected to produce a desired radiation pattern.

Directivity: The ratio of the actual radiated power per solid angle to the radiated power per solid angle that would result if the radiated power was radiated isotropically. Oftentimes the word *directivity* is used to refer to the maximum directivity.

Phased array: An array in which the phases of the exciting currents are varied to scan the radiation pattern through space.

Radiation pattern: A plot as a function of direction of the power per unit solid angle radiated in a given polarization by an antenna. The terms *radiation pattern* and *antenna pattern* can be used interchangeably.

Related Topics

37.1 Space Propagation • 69.2 Radio

References

C.A. Balanis, *Antenna Theory Analysis and Design*, New York: Harper and Row, 1982.

R. Carrel, "The design of log-periodic dipole antennas," *IRE International Convention Record* (part 1), 1961, pp. 61–75.

R.E. Collin, *Antennas and Radiowave Propagation*, New York: McGraw-Hill, 1985.

R.F. Harrington, *Time Harmonic Electromagnetic Fields*, New York: McGraw-Hill, 1961.

D.E. Isbell, "Log periodic dipole arrays," *IRE Transactions on Antennas and Propagation,* vol. AP-8, pp. 260–267, 1960.

P. Lorrain and D.R. Corson, *Electromagnetic Fields and Waves*, San Francisco: W.H. Freeman, 1970.

S. Ramo, J.R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, New York: John Wiley & Sons, 1984.

W.L. Stutzman and G.A. Thiele, *Antenna Theory and Design*, New York: John Wiley & Sons, 1981.

S. Uda and Y. Mushiake, *Yagi–Uda Antenna*, Sendai, Japan: Sasaki Printing and Publishing Company, 1954.

Further Information

For general-interest articles on antennas the reader is directed to the *IEEE Antennas and Propagation Magazine.* In addition to providing up-to-date articles on current issues in the antenna field, this magazine also provides easy-to-read tutorials. For the latest research advances in the antenna field the reader is referred to the *IEEE Transactions on Antennas and Propagation.* In addition, a number of very good textbooks are devoted to antennas. The books by Collin and by Stutzman and Thiele were especially useful in the preparation of this section.

38.2 Aperture

J. Patrick Fitch

The main purpose of an **antenna** is to control a wave front at the boundary between two media: a source (or receiver) and the medium of propagation. The source can be a fiber, cable, waveguide, or other transmission line. The medium of propagation may be air, vacuum, water, concrete, metal, or tissue, depending on the application. Antenna aperture design is used in acoustic, optic, and electromagnetic systems for imaging, communications, radar, and spectroscopy applications.

There are many classes of antennas: wire, horn, slot, notch, reflector, lens, and **array**, to name a few (see Fig. 38.7). Within each class is a variety of subclasses. For instance, the horn antenna can be pyramidal or conical. The horn can also have flaring in only one direction (sectoral horn), asymmetric components, shaped

High-Speed Space Data Communications

TSI/TelSys Inc., Columbia, Maryland, is a company formed to commercialize NASA high-data-rate telemetry technology originally developed at Goddard Space Flight Center's Microelectronic Systems Branch. Today, TSI/TelSys Inc. designs, manufactures, markets, and supports a broad range of commercial satellite telecommunications gateway products. These technologies and products support two-way, high-speed space data communications for telemetry, satellite remote sensing, and high-data-rate communications applications. The satellite antenna shown above is part of a system used for high-speed data transmissions. (Courtesy of National Aeronautics and Space Administration.)

edges, or a compound design of sectoral and pyramidal combined. For all antennas, the relevant design and analysis will depend on antenna aperture size and shape, the center wavelength λ, and the distance from the aperture to a point of interest (the range, R). This section covers discrete **oscillators**, arrays of oscillators, synthetic apertures, geometric design, Fourier analysis, and parameters of some typical antennas. The emphasis is on microwave-type designs.

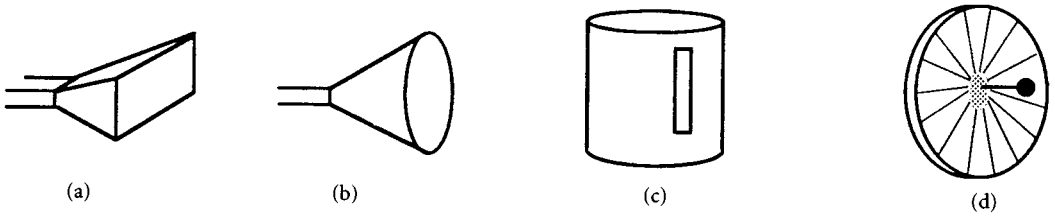


FIGURE 38.7 Examples of several types of antennas: (a) pyramidal horn, (b) conical horn, (c) axial slot on a cylinder, and (d) parabolic reflector.

The Oscillator or Discrete Radiator

The basic building block for antenna analysis is a linear conductor. Movement of electrons (current) in the conductor induces an electromagnetic field. When the electron motion is oscillatory—e.g., a dipole with periodic electron motion, the induced electric field, E, is proportional to $\cos(\omega t - kx + \phi)$, where ω is radian frequency of oscillation, t is time, k is wave number, x is distance from the oscillator, and ϕ is the phase associated with this oscillator (relative to the time and spatial coordinate origins). When the analysis is restricted to a fixed position x, the electric field can be expressed as

$$E(t) = A\cos(\omega t + \phi) \tag{38.25}$$

where the phase term ϕ now includes the kx term, and all of the constants of proportionality are included in the amplitude A. Basically, the assumption is that oscillating currents produce oscillating fields. The description of a receiving antenna is analogous: an oscillating field induces a periodic current in the conductor.

The field from a pair of oscillators separated in phase by δ radians is

$$E_\delta(t) = A_1\cos(\omega t + \phi) + A_2\cos(\omega t + \phi + \delta) \tag{38.26}$$

Using phasor notation, $\tilde{E}_\delta$, the cosines are converted to complex exponentials and the radial frequency term, ωt, is suppressed,

$$\tilde{E}_\delta(t) = A_1 e^{i\phi} + A_2 e^{i(\phi+\delta)} \tag{38.27}$$

The amplitude of the sinusoidal modulation $E_\delta(t)$ can be calculated as $|\tilde{E}_\delta|$. The intensity is

$$I = |\tilde{E}_\delta|^2 = |A_1|^2 + |A_2|^2 + 2A_1A_2\cos(\delta) \tag{38.28}$$

When the oscillators are of the same amplitude, $A = A_1 = A_2$, then

$$\begin{aligned} E_\delta(t) &= A\cos(\omega t + \phi) + A\cos(\omega t + \phi + \delta) \\ &= 2A\cos\left(\frac{\delta}{2}\right)\cos\left(\omega t + \phi + \frac{\delta}{2}\right) \end{aligned} \tag{38.29}$$

For a series of n equal amplitude oscillators with equal phase spacing

$$E_{n\delta}(t) = \sum_{j=0}^{n-1} A\cos(\omega t + \phi + j\delta) \tag{38.30}$$

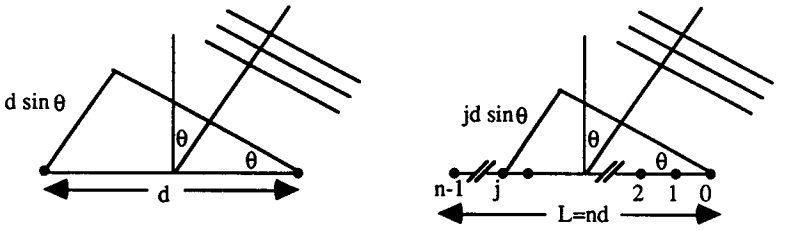


FIGURE 38.8 A two-element and an n-element array with equal spacing between elements. The propagation length difference between elements is $d\sin\theta$, which corresponds to a phase difference of $kd\sin\theta$, where k is the wave number $2\pi/\lambda$. The length L corresponds to a continuous aperture of length nd with the sample positions beginning $d/2$ from the ends.

By using phasor arithmetic the intensity is given as

$$I_{n\delta}(t) = \left|\tilde{E}_{n\delta}\right|^2 = \left|Ae^{i\phi}\sum_{j=0}^{n-1} e^{ij\delta}\right|^2 = A^2\left|\frac{1-e^{in\delta}}{1-e^{i\delta}}\right|^2 = I_0\,\frac{1-\cos(n\delta)}{1-\cos(\delta)} = I_0\,\frac{\sin^2(n\delta/2)}{\sin^2(\delta/2)} \tag{38.31}$$

where $I_0 = n^{-2}$ to normalize the intensity pattern at $\delta = 0$.

For an incoming plane wave which is tilted at an angle θ from the normal, the relative phase difference between two oscillators is $kd\sin\theta$, where d is the distance between oscillators and k is the wave number $2\pi/\lambda$ (see Fig. 38.8). For three evenly spaced oscillators, the phase difference between the end oscillators is $2kd\sin\theta$. In general, the end-to-end phase difference for n evenly spaced oscillators is $(n-1)kd\sin\theta$. This formulation is identical to the phase representation in Eq. (38.30) with $\delta = kd\sin\theta$. Therefore, the intensity as a function of incidence angle θ for an evenly spaced array of n elements is

$$I_{nL}(\theta) = I_0\,\frac{\sin^2\left(\frac{1}{2}knd\sin\theta\right)}{\sin^2\left(\frac{1}{2}kd\sin\theta\right)} = I_0\,\frac{\sin^2\left(\frac{1}{2}kL\sin\theta\right)}{\sin^2\left(\frac{1}{2n}kL\sin\theta\right)} = I_0\,\frac{\sin^2\left(\frac{\pi L}{\lambda}\sin\theta\right)}{\sin^2\left(\frac{\pi L}{n\lambda}\sin\theta\right)} \tag{38.32}$$

where $L = nd$ corresponds to the physical dimension (length) of the aperture of oscillators. The zeros of this function occur at $kL\sin\theta = 2m\pi$, for any nonzero integer m. Equivalently, the zeros occur when $\sin\theta = m\lambda/L$. When the element spacing d is less than a wavelength, the number of zeros for $0 < \theta < \pi/2$ is given by the largest integer M such that $M \le L/\lambda$. Therefore, the ratio of wavelength to largest dimension, λ/L, determines both the location (in θ space) and the number of zeros in the intensity pattern when $d \le \lambda$. The number of oscillators controls the amplitude of the side lobes.

For $n = 1$, the intensity is constant—i.e., independent of angle. For $\lambda > L$, both the numerator and denominator of Eq. (38.32) have no zeros and as the length of an array shortens (relative to a wavelength), the intensity pattern converges to a constant ($n = 1$ case). As shown in Fig. 38.9, a separation of $\lambda/4$ has an intensity rolloff less than 1 dB over $\pi/2$ radians (a $\lambda/2$ separation rolls off 3 dB). This implies that placing antenna elements closer than $\lambda/4$ does not significantly change the intensity pattern. Many microwave antennas exploit this and use a mesh or parallel wire (for polarization sensitivity) design rather than covering the entire aperture with conductor. This reduces both weight and sensitivity to wind loading. Note that the analysis has not accounted for phase variations from position errors in the element placement where the required accuracy is typically better than $\lambda/10$.

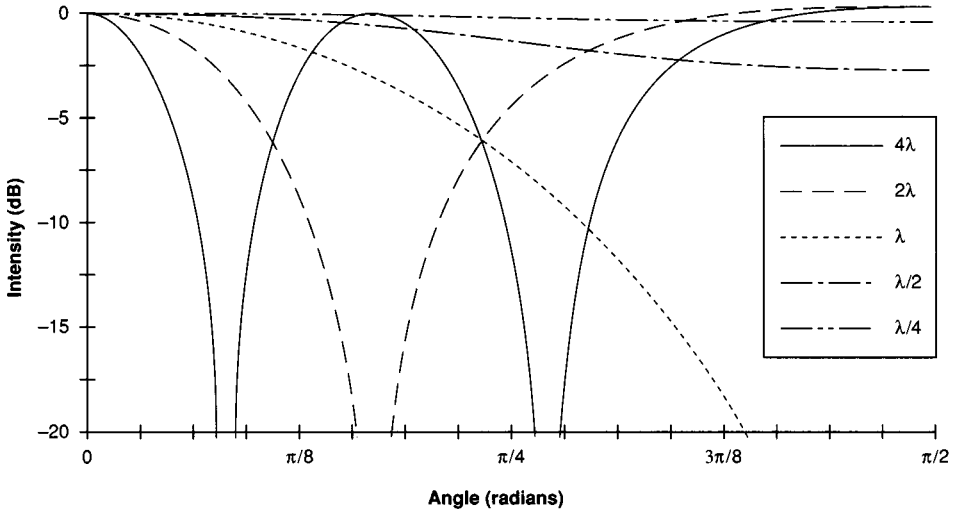


FIGURE 38.9 Normalized intensity pattern in decibels (10 log(I)) for a two-element antenna with spacing 4λ, 2λ, λ, λ/2, and λ/4 between the elements.

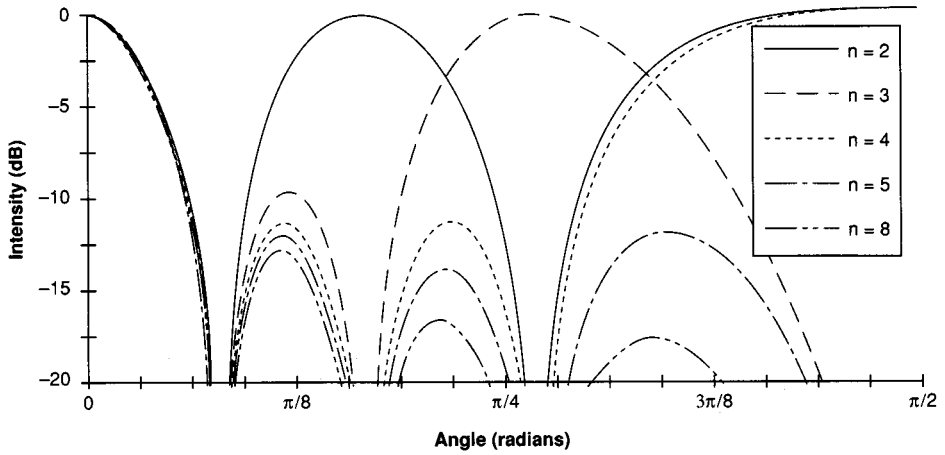
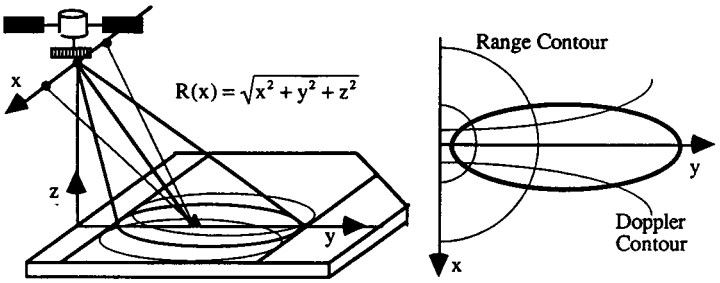


FIGURE 38.10 Normalized intensity pattern in decibels (10 log(I)) for a length 4λ array with 2, 3, 4, 5, and 8 elements.

For $L >> \lambda$, $\sin\theta \approx \theta$, which implies that the first zero is at $\theta = \lambda/L$. The location of the first zero is known as the Rayleigh resolution criteria. That is, two plane waves separated by at least λ/L radians can be discriminated. For imaging applications, this corresponds roughly to the smallest detectable feature size. As shown in Fig. 38.10, the first zero occurs at approximately $\lambda/L = 0.25$ radians (the Rayleigh resolution). Note that there is no side lobe suppression until $d \leq \lambda$, when the location of the zeros becomes fixed. Having more than eight array elements (separation of less than a quarter wavelength) only moderately reduces the height of the maximum side lobe.

Synthetic Apertures

In applications such as air- and space-based radar, size and weight constraints prohibit the use of very large antennas. For instance, if the L-band (23.5-cm wavelength) radar imaging system on the Seasat satellite (800-km altitude, launched in 1978) had a minimum resolution specification of 23.5 m, then, using the Rayleigh resolution criteria, the aperture would need to be 8 km long. In order to attain the desired resolution, an aperture is "synthesized" from data collected with a physically small (10 m) antenna traversing an 8-km flight



(a) Data collection geometry (b) Coordinates in radar footprint

FIGURE 38.11 Synthetic aperture radar geometry and nearly orthogonal partitioning of the footprint by range (circular) and Doppler frequency (hyperbolic) contours.

path. Basically, by using a stable oscillator on the spacecraft, both amplitude and phase are recorded, which allows postprocessing algorithms to combine the individual echoes in a manner analogous to an antenna array. From an antenna perspective, an individual scattering element produces a different round trip propagation path based on the position of the physical antenna—a synthetic antenna array. Using the geometry described in Fig. 38.11, the phase is

$$\phi(x) = \frac{2\pi}{\lambda} 2R(x) = \frac{2\pi}{\lambda} 2\sqrt{x^2 + y^2 + z^2} \tag{38.33}$$

It is convenient to assume a straight-line flight path along the x-axis, a planar earth (x, y plane), and a constant velocity, v, with range and cross-range components $v_r(x)$ and $v_c(x)$, respectively. In many radar applications the broad side distance to the center of the footprint, R, is much larger than the size of the footprint. This allows the distance $R(x)$ to be expanded about R resulting in

$$\phi(t) = \frac{2\pi}{\lambda} 2R(vt) = 2\pi\left\{\frac{2R}{\lambda} + \frac{2v_r}{\lambda} t + \frac{v_c^2}{\lambda R} t^2\right\} \tag{38.34}$$

The first term in Eq. (38.34) is a constant phase offset corresponding to the center of beam range bin and can be ignored from a resolution viewpoint. The second term, $2v_r/\lambda$, is the Doppler frequency shift due to the relative (radial) velocity between antenna and scattering element. The third term represents a quadratic correction of the linear flight path to approximate the constant range sphere from a scattering element. It is worth noting that synthetic aperture systems do not require the assumptions used here, but accurate position and motion compensation is required.

For an antenna with cross range dimension D and a scattering element at range R, the largest synthetic aperture that can be formed is of dimension $\lambda R/D$ (the width of the footprint). Because this data collection scenario is for round trip propagation, the phase shift at each collecting location is twice the shift at the edges of a single physical antenna. Therefore at a range R, the synthetic aperture resolution is

$$\frac{\lambda R}{D_{SA}} = \frac{\lambda R}{2\lambda R/D} = \frac{D}{2} \tag{38.35}$$

The standard radar interpretation for synthetic apertures is that information coded in the Doppler frequency shift can be decoded to produce high-resolution images. It is worth noting that the synthetic aperture can be formed even with no motion (zero Doppler shift). For the no-motion case the antenna array interpretation is appropriate. This approach has been used for acoustic signal processing in nondestructive evaluation systems as

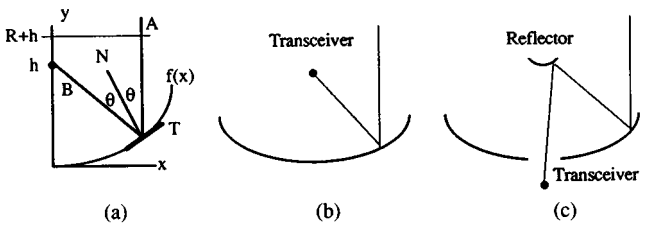


FIGURE 38.12 Parabolic reflector systems: (a) geometry for determining the function with a constant path length and specular reflection, (b) single-bounce parabolic reflector, (c) two-bounce reflector with a parabolic primary and hyperbolic secondary (Cassegrain).

well as wave migration codes for seismic signal processing. When there is motion, the Doppler term in the expansion of the range dominates the phase shift and therefore becomes the useful metric for predicting resolution.

Geometric Designs

The phase difference in a linear array was caused by the spatial separation and allowed the discrimination of plane waves arriving at different angles. Desired phase patterns can be determined by using analytic geometry to position the elements. For example, if coherent superposition across a wave front is desired, the wave front can be directed (reflected, refracted, or diffracted) to the receiver in phase. For a planar wave front, this corresponds to a constant path length from any point on the reference plane to the receiver. Using the geometry in Fig. 38.12, the sum of the two lengths $(x, R + h)$ to (x, y) and (x, y) to $(0, h)$ must be a constant independent of x—which is $R + 2h$ for this geometry. This constraint on the length is

$$R + h - y + \sqrt{x^2 + (h - y)^2} = R + 2h \quad \text{or} \quad x^2 = 4hy \tag{38.36}$$

This is the equation for a parabola. Losses would be minimized if the wave front were specularly reflected to the transceiver. Specular reflection occurs when the angles between the normal vector **N** [or equivalently the tangent vector $\mathbf{T} = (x, f'(x))] = (1, x/2h)$ and the vectors $\mathbf{A} = (0, -1)$ and $\mathbf{B} = (-x, h - y)$ are equal. This is the same as equality of the inner products of the normalized vectors, which is shown by

$$\hat{\mathbf{T}} \cdot \hat{\mathbf{A}} = \frac{(2h, x)}{\sqrt{x^2 + 4h^2}} \cdot (0, -1) = \frac{-x}{\sqrt{x^2 + 4h^2}} \tag{38.37}$$

$$\hat{\mathbf{T}} \cdot \hat{\mathbf{B}} = \frac{(2h, x)}{\sqrt{x^2 + 4h^2}} \cdot \frac{(-x, h - y)}{\sqrt{x^2 + (h - y)^2}} = \frac{-x(x^2 + 4h^2)}{(x^2 + 4h^2)^{3/2}} = \frac{-x}{\sqrt{x^2 + 4h^2}} \tag{38.38}$$

The constant path length and high gain make the parabolic antenna popular at many wavelengths including microwave and visible. More than one reflecting surface is allowed in the design. The surfaces are typically conical sections and may be designed to reduce a particular distortion or to provide better functionality. Compound designs often allow the active elements to be more accessible and eliminate long transmission lines. A two-bounce reflector with a parabolic primary and a hyperbolic secondary is known as a Cassegrain system. In all reflector systems it is important to account for the blockage ("shadow" of the feed, secondary reflector, and support structures) as well as the spillover (radiation propagating past the intended reflecting surface).

Continuous Current Distributions (Fourier Transform)

Ideally, antennas would be designed using solutions to Maxwell's equations. Unfortunately, in most cases exact analytic and numerical solutions to Maxwell's equations are difficult to obtain. Under certain conditions,

approximations can be introduced that allow solution to the wave equations. Approximating spherical wave fronts as quadratics has been shown for the synthetic aperture application and is valid when the propagation distance is greater than $(\pi L^2/4\lambda)^{1/3}$, where L is the aperture size. In general, this is known as the **Fresnel** or **near-field** approximation. When the propagation distance is at least $2L^2/\lambda$, the angular radiation pattern can be approximated as independent of distance from the aperture. This pattern is known as the normalized **far-field** or **Fraunhofer** distribution, $E(\theta)$, and is related to the normalized current distributed across an antenna aperture, $i(x)$, by a Fourier transform:

$$E(u) = \int i(x')e^{i2\pi ux'}dx' \tag{38.39}$$

where $u = \sin\theta$ and $x' = x/\lambda$.

Applying the Fraunhofer approximation to a line source of length L

$$E_L(u = \sin\theta) = \int_{-L/2\lambda}^{L/2\lambda} e^{i2\pi ux'}dx' = \frac{\sin\left(\frac{\pi L}{\lambda}u\right)}{\frac{\pi L}{\lambda}u} = \frac{\sin\left(\frac{\pi L}{\lambda}\sin\theta\right)}{\frac{\pi L}{\lambda}\sin\theta} \tag{38.40}$$

which is Eq. (38.32) when $n >> L/\lambda$. As with discrete arrays, the ratio L/λ is the important design parameter: $\sin\theta = \lambda/L$ is the first zero (no zeros for $\lambda > L$) and the number of zeros is the largest integer M such that $M \le L/\lambda$.

In two dimensions, a rectangular aperture with uniform current distribution produces

$$E_R(u_1, u_2) = \frac{\sin\left(\frac{\pi}{\lambda}u_1L_1\right)}{\frac{\pi}{\lambda}u_1L_1}\,\frac{\sin\left(\frac{\pi}{\lambda}u_2L_2\right)}{\frac{\pi}{\lambda}u_2L_2} \quad \text{and} \quad I_R(u_1, u_2) = \left|E_L(u_1)\right|^2\left|E_L(u_2)\right|^2 \tag{38.41}$$

The field and intensity given in Eq. (38.41) are normalized. In practice, the field is proportional to the aperture area and inversely proportional to the wavelength and propagation distance.

The normalized far-field intensity distribution for a uniform current on a circular aperture is a circularly symmetric function given by

$$I_C(u) = \left[\frac{2J_1\left(\frac{\pi}{\lambda}uL\right)}{\frac{\pi}{\lambda}uL}\right]^2 \tag{38.42}$$

where J_1 is the Bessel function of the first kind, order one. This far-field intensity is called the Airy pattern. As with the rectangular aperture, the far-field intensity is proportional to the square of the area and inversely proportional to the square of the wavelength and the propagation distance. The first zero (Rayleigh resolution criteria) of the Airy pattern occurs for $uL/\lambda = 1.22$ or $\sin\theta = 1.22\lambda/L$. As with linear and rectangular apertures, the resolution scales with λ/L.

Figure 38.13 shows a slice through the normalized far-field intensity of both a rectangular aperture and a circular aperture. The linearity of the Fourier transform allows apertures to be represented as the superposition of subapertures. The primary reflector, the obscurations from the support structures, and the secondary reflector

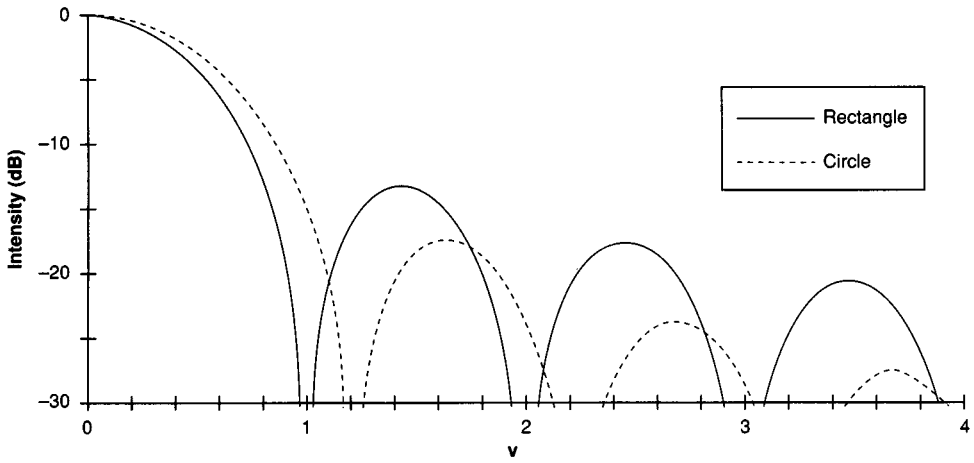


FIGURE 38.13 Normalized intensity pattern in decibels ($10 \log[I(v = uL/\lambda)]$) for a rectangular and a circular antenna aperture with uniform current distributions.

TABLE 38.1 Directivity and Gain of Some Higher Frequency Antennas

Antenna Type	Directivity[a]	Gain[a]
Uniform rectangular aperture	$\frac{4\pi}{\lambda^2} L_x L_y$	$\frac{4\pi}{\lambda^2} L_x L_y$
Large square aperture	$12.6\left(\frac{L}{\lambda}\right)^2$	$7.7\left(\frac{L}{\lambda}\right)^2$
Large circular aperture (parabolic reflector)	$9.87\left(\frac{D}{\lambda}\right)^2$	$7\left(\frac{D}{\lambda}\right)^2$
Pyramidal horn	$\left(\frac{4\pi}{\lambda^2}\right)L_x L_y$	$0.5\left(\frac{4\pi}{\lambda^2}\right)L_x L_y$

[a] Directivity and gain are relative to a half-wave dipole.

of a Cassegrain-type antenna can be modeled. Numerical evaluation of the Fourier transform permits straightforward calculation of the intensity patterns, even for nonuniform current distributions.

Antenna Parameters

Direct solutions to Maxwell's equations or solutions dependent on approximations provide the analytic tools for designing antennas. Ultimately, the analysis must be confirmed with experiment. Increasingly sensitive radar and other antenna applications have resulted in much more attention to edge effects (from the primary aperture, secondary, and/or support structures). The geometric theory of diffraction as well as direct Maxwell solvers are making important contributions.

With the diversity of possible antenna designs, a collection of design rules of thumb are useful. The **directivity** and **gain** for a few popular antenna designs are given in Table 38.1. Directivity is the ratio of the maximum to average radiation intensity. The gain is defined as the ratio of the maximum radiation intensity from the subject antenna to the maximum radiation intensity from a reference antenna with the same power input. The directivity, D, and gain, G, of an antenna can be expressed as

$$D = \left(\frac{4\pi}{\lambda^2}\right)A_{em} \qquad \text{and} \qquad G = \left(\frac{4\pi}{\lambda^2}\right)A_e \tag{38.43}$$

where A_{em} is the maximum effective aperture and A_e is the actual effective aperture of the antenna. Because of losses in the system, $A_e = kA_{em}$, where k is the radiation efficiency factor. The gain equals the directivity when there are no losses ($k = 1$), but is less than the directivity if there are any losses in the antenna ($k < 1$), that is, $G = kD$.

As an example, consider the parabolic reflector antenna where efficiency degradation includes

- Ohmic losses are small ($k = 1$)
- Aperture taper efficiency ($k = 0.975$)
- Spillover (feed) efficiency ($k = 0.8$)
- Phase errors in aperture field ($k = 0.996$ to 1)
- Antenna blockage efficiency ($k = 0.99$)
- Spar blockage efficiency ($k = 0.994$)

Each antenna system requires a customized analysis of the system losses in order to accurately model performance.

Defining Terms

Antenna: A physical device for transmitting or receiving propagating waves.

Aperture antenna: An antenna with a physical opening, hole, or slit. Contrast with a wire antenna.

Array antenna: An antenna system performing as a single aperture but composed of antenna subsystems.

Directivity: The ratio of the maximum to average radiation intensity.

Fraunhofer or far field: The propagation region where the normalized angular radiation pattern is independent of distance from the source. This typically occurs when the distance from the source is at least $2L^2/\lambda$, where L is the largest dimension of the antenna.

Fresnel or near field: The propagation region where the normalized radiation pattern can be calculated using quadratic approximations to the spherical Huygens' wavelet surfaces. The pattern can depend on distance from the source and is usually valid for distances greater than $(\pi/4\lambda)^{1/3}L^{2/3}$, where L is the largest dimension of the antenna.

Gain: The ratio of the maximum radiation intensity from the subject antenna to the maximum radiation intensity from a reference antenna with the same power input. Typical references are a lossless isotropic source and a lossless half-wave dipole.

Oscillator: A physical device that uses the periodic motion within the material to create propagating waves. In electromagnetics, an oscillator can be a conductor with a periodic current distribution.

Reactive near field: The region close to an antenna where the reactive components of the electromagnetic fields from charges on the antenna structure are very large compared to the radiating fields. Considered negligible at distances greater than a wavelength from the source (decay as the square or cube of distance). Reactive field is important at antenna edges and for electrically small antennas.

Related Topic

37.1 Space Propagation

References

R. Feynman, R.B. Leighton, and M.L. Sands, *The Feynman Lectures on Physics,* Reading, Mass.: Addison-Wesley, 1989.

J.P. Fitch, *Synthetic Aperture Radar,* New York: Springer-Verlag, 1988.

J.W. Goodman, *Introduction to Fourier Optics,* New York: McGraw-Hill, 1968.

H. Jasik, *Antenna Engineering Handbook,* New York: McGraw-Hill, 1961.

R.W.P. King and G.S. Smith, *Antennas in Matter,* Cambridge: MIT Press, 1981.

J.D. Krause, *Antennas,* New York: McGraw-Hill, 1950.

Y.T. Lo and S.W. Lee, *Antenna Handbook,* New York: Van Nostrand Reinhold, 1988.

A.W. Rudge, K. Milne, A.D. Olver, and P. Knight, *The Handbook of Antenna Design,* London: Peter Peregrinus, 1982.
M. Skolnik, *Radar Handbook,* New York: McGraw-Hill, 1990.
B.D. Steinberg, *Principles of Aperture & Array System Design,* New York: John Wiley & Sons, 1976.

Further Information

The monthly *IEEE Transactions on Antennas and Propagation* as well as the proceedings of the annual *IEEE Antennas and Propagation International Symposium* provide information about recent developments in this field. Other publications of interest include the *IEEE Transactions on Microwave Theory and Techniques* and the *IEEE Transactions on Aerospace and Electronic Systems.*

Readers may also be interested in the "IEEE Standard Test Procedures for Antennas," The Institute for Electrical and Electronics Engineers, Inc., ANSI IEEE Std. 149-1979, 1979.

38.3 Microstrip Antennas

David M. Pozar

Introduction

Microstrip antenna technology has been the most rapidly developing topic in the antenna field in the last 15 years, receiving the creative attentions of academic, industrial, and government engineers and researchers throughout the world. As a result, microstrip antennas have quickly evolved from a research novelty to commercial reality, with applications in a wide variety of microwave systems. Rapidly developing markets in personal communications systems (PCS), mobile satellite communications, direct broadcast television (DBS), wireless local-area networks (WLANs), and intelligent vehicle highway systems (IVHS) suggest that the demand for microstrip antennas and arrays will increase even further.

Although microstrip antennas have proved to be a significant advance in the established field of antenna technology, it is interesting to note that it is usually their nonelectrical characteristics that make microstrip antennas preferred over other types of radiators. Microstrip antennas have a low profile and are light in weight, they can be made conformal, and they are well suited to integration with **microwave integrated circuits (MICs)**. If the expense of materials and fabrication is not prohibitive, they can also be low in cost. When compared with traditional antenna elements such as wire or aperture antennas, however, the electrical performance of the basic microstrip antenna or array suffers from a number of serious drawbacks, including very narrow **bandwidth**, high feed network losses, poor cross polarization, and low power-handling capacity. Intensive research and development has demonstrated that most of these drawbacks can be avoided, or at least alleviated to some extent, with innovative variations and extensions to the basic microstrip element [James and Hall, 1989; Pozar and Schaubert, 1995]. Some of the basic features of microstrip antennas are listed below:

- Low profile form factor
- Potential light weight
- Potential low cost
- Potential conformability with mounting structure
- Easy integration with planar circuitry
- Capability for linear, dual, and circular polarizations
- Versatile feed geometries

Basic Microstrip Antenna Element

The basic microstrip antenna element is derived from a $\lambda_g/2$ **microstrip** transmission **line** resonator [Pozar, 1990]. It consists of a thin metallic conducting patch etched on a grounded dielectric substrate, as shown in Fig. 38.14. This example is shown with a coaxial probe feed, but other feeds are possible, as discussed below.

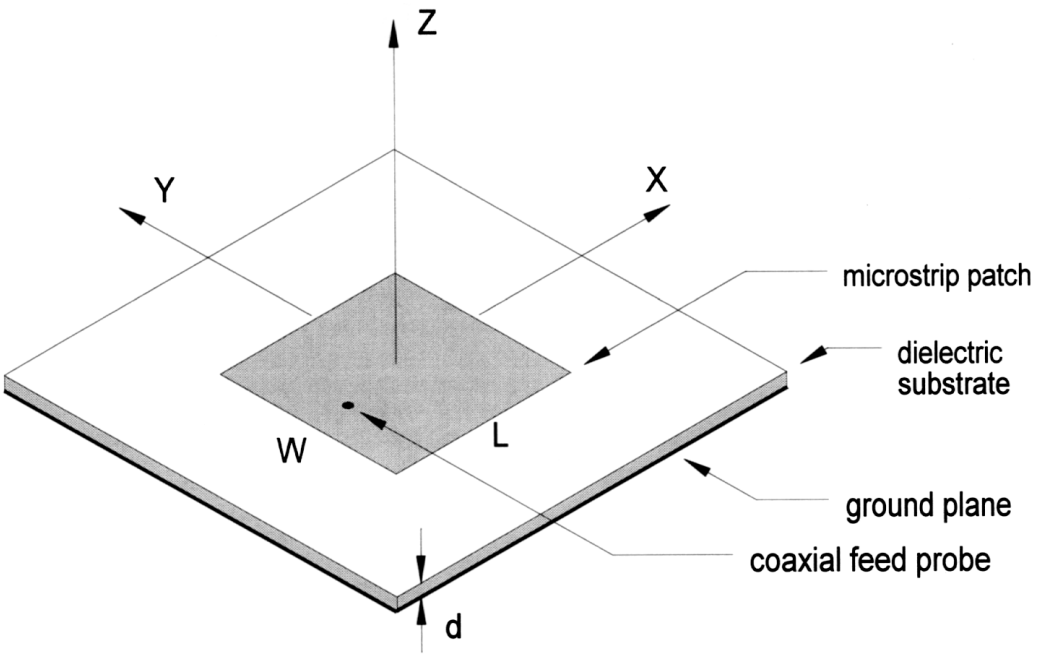


FIGURE 38.14 Geometry of rectangular coaxial probe-fed microstrip antenna.

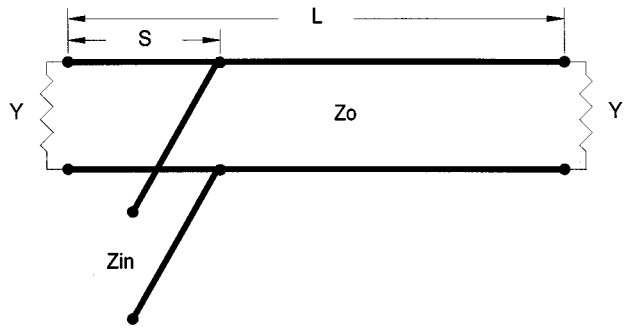


FIGURE 38.15 Transmission line circuit model for a rectangular microstrip antenna. The feed point is positioned a distance *s* from the radiating edge of the patch.

The patch has a length *L* along the *x*-axis, and width *W* along the *y*-axis. The dielectric substrate has a thickness *d* and a dielectric constant ε_r, and is backed with a conducting ground plane. With a coaxial probe feed, the outer conductor of the coaxial line is connected to the ground plane, and the inner conductor is attached to the patch element. The position of the feed point relative to the edge of the patch controls the input impedance level of the antenna. In operation, the length of the patch element is approximately $\lambda_g/2$, forming an open-circuit resonator. Because the patch is relatively wide, the patch edges at $x = -L/2$ and $L/2$ effectively form slot apertures which radiate in phase to form a broadside radiation pattern.

Many analytical models have been developed for the impedance and radiation properties of microstrip antennas [James and Hall, 1989], but most of the qualitative behavior of the element can be demonstrated using the relatively simple transmission line model. As shown in Fig. 38.15, the patch element is modeled as a length, *L*, of microstrip transmission line of characteristic impedance Z_0. The characteristic impedance of the line can be found using simple approximations [Pozar, 1990] and is a function of the width, *W*, of the line as well as the substrate thickness and dielectric constant. The ends of the transmission line are terminated in

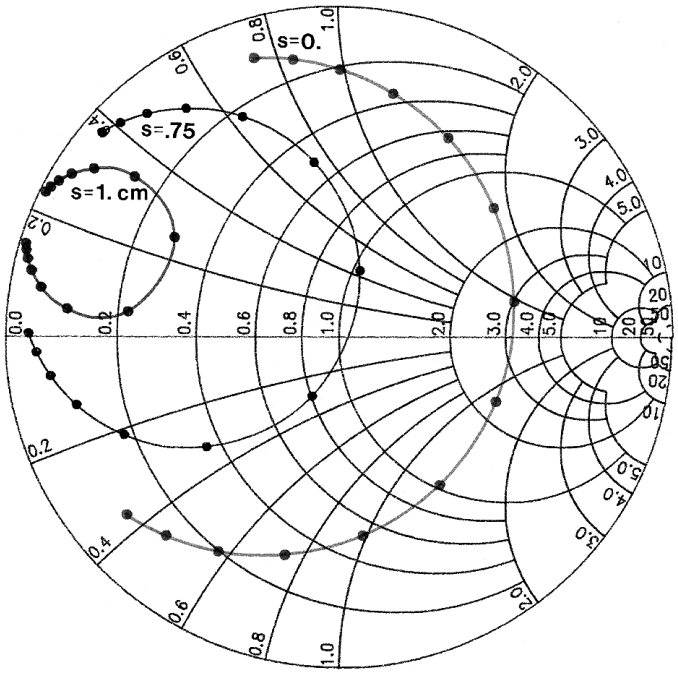


FIGURE 38.16 Smith chart plot of the input impedance of a probe-fed rectangular microstrip antenna vs. frequency, for three different feed positions. Patch parameters are $L = 2.5$ cm, $W = 3.0$ cm, $\varepsilon_r = 2.2$, $d = 0.79$ cm. Frequency sweep runs from 3.6 to 4.25 GHz, in steps of 50 MHz.

admittances, $Y = G + jB$, where the conductance serves to model the radiation from the ends of the line, and the susceptance serves to model the effective length extension of the line (due to fringing fields). Several approximations are available for calculating the end admittances [James and Hall, 1989], with a typical result for $d << \lambda_0$ given as

$$Y = G + jB = \frac{k_0 W}{2\eta_0}\left[1 + j\left(1 - 0.64 \ln k_0 d\right)\right] \tag{38.44}$$

where $k_0 = 2\pi/\lambda_0$ and $\eta = \sqrt{\frac{\mu_0}{\varepsilon_0}}$. The susceptance B is typically positive, implying a capacitive end effect. This means that the resonant length of the patch will be slightly less than $\lambda_g/2$. If the feed probe is located a distance s from the edge of the patch, the input impedance seen by the probe can be calculated using basic transmission line theory from the circuit of Fig. 38.15. Resonance is defined as the frequency at which the imaginary part of the input impedance is zero.

As a result of the symmetry of the transmission line resonator, the voltage along the transmission line will have maxima at the ends and a null at the center of the line. This implies that the input impedance will be maximum when the feed point is at the edge of the patch, and will decrease to zero as the feed is moved to the center of the patch. Fig. 38.16 shows a Smith chart plot of the input impedance of a coaxial probe-fed microstrip antenna vs. frequency, for three different probe positions. Observe that the input impedance locus decreases as the feed point moves toward the center of the patch. Also, observe that the impedance locus becomes more inductive as the feed point moves toward the center of the patch.

The far-field radiation patterns can also be derived from the transmission line model by treating the radiating edges at $x = -L/2$ and $L/2$ as equivalent slots. In the coordinate system of Fig. 38.14, the normalized far-zone fields of the rectangular patch can be expressed as

$$E_\theta = E_0 \frac{\sin \alpha}{\alpha} \cos \beta \cos \phi \tag{38.45a}$$

$$E_\phi = E_0 \frac{\sin \alpha}{\beta} \cos \beta \cos \theta \sin \phi \tag{38.45b}$$

where

$$\alpha = \frac{k_0 W}{2} \sin \theta \sin \phi$$

$$\beta = \frac{k_0 L}{2} \sin \theta \cos \phi$$

and θ and ϕ are spherical coordinates. These patterns have maxima broadside ($\theta = 0$) to the patch, with 3-dB **beamwidths** typically in the range of 90° to 120°. Typical *E*-plane ($\phi = 0$) and *H*-plane ($\phi = 90°$) microstrip antenna radiation patterns are shown in Fig. 38.17.

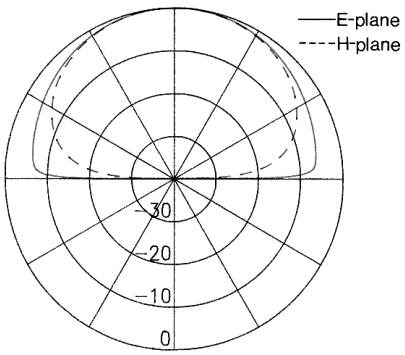


FIGURE 38.17 *E*- and H-plane far-field radiation patterns of the rectangular microstrip antenna of Fig. 38.16.

Microstrip antenna elements have a number of useful and interesting features, but probably the most serious limitation of this technology is the narrow bandwidth of the basic element. While antenna elements such as dipoles, slots, and waveguide horns have operating bandwidths ranging from 15 to 50%, the traditional microstrip patch element typically has an impedance bandwidth of only a few percent. Fig. 38.18 shows the impedance bandwidth vs. substrate thickness for a rectangular microstrip antenna with substrate permittivities of 2.2 and 10.2. Observe from the figure that bandwidth decreases as the substrate becomes thinner and as the dielectric constant increases. Both of these trends are explained as a result of the increased *Q* of the resonator, basically due to the fact that the patch current is in close proximity to its negative image in the substrate ground plane. In terms of bandwidth, it is preferable to use a thick antenna substrate, with a low dielectric constant. But because of inductive loading and possible spurious radiation from coplanar microstrip circuitry, the thickness of a microstrip antenna substrate is typically limited to 0.02λ or less. This illustrates the essential compromise associated with the microstrip antenna concept, as it is not possible to obtain optimum performance from both a microstrip antenna and microstrip circuitry on a single dielectric substrate. These two functions are distinct electromagnetically, since the bound fields associated with nonradiating circuitry obviate efficient radiation.

While the bandwidth of the basic element is limited, considerable research and development during the last 15 years has led to a number of creative and novel techniques for the enhancement of microstrip antenna bandwidth, so that impedance bandwidths ranging from 10 to 40% can now be achieved [James and Hall, 1989; Pozar and Schaubert, 1995]. While there have been dozens of proposed techniques for the enhancement of microstrip antenna bandwidth, they can all be categorized according to three canonical approaches:

- Impedance matching using matching network
- Introducing dual resonance with stacked or parasitic elements
- Reducing efficiency by adding lossy elements

The reader is referred to the literature for more details on specific techniques for bandwidth improvement.

Figure 38.18 also shows the efficiency of the antenna, defined as

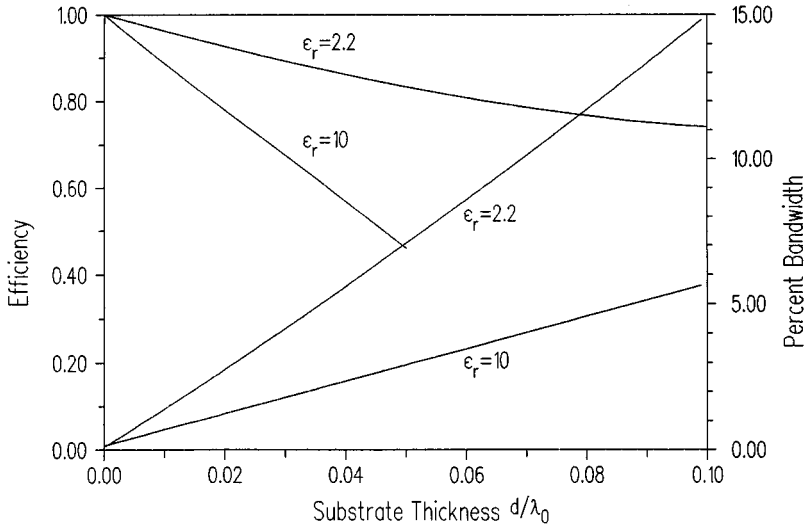


FIGURE 38.18 Impedance bandwidth and radiation efficiency of a microstrip antenna vs. substrate thickness, for two values of substrate permittivity.

$$e = \frac{P_{\text{rad}}}{P_{\text{rad}} + P_{\text{loss}}} \tag{38.46}$$

where P_{rad} is the radiated power and P_{loss} is the power lost in the antenna. Losses in a microstrip antenna occur in three ways: conductor loss, dielectric loss, and surface wave loss. Unless the substrate is extremely thin, conductor loss is generally negligible. For quality microwave substrates (loss tangent ≤ 0.002), dielectric loss is also relatively small. Surface waves, which are fields bound to the dielectric substrate that propagate along its surface, often account for the dominant loss mechanism for microstrip antennas. As can be seen in Fig. 38.18, efficiency decreases with increasing substrate thickness and dielectric constant, again suggesting the use of low-dielectric-constant substrates. The overall radiation efficiency of a microstrip antenna on a low-dielectric substrate is typically 95%, or better.

Besides rectangular patch elements, it is possible to use a variety of other patch shapes as resonant radiating elements. For purposes of polarization purity and analytical simplicity, however, it is usually preferable to use rectangular, square, or circular elements. Linear polarization is best obtained with rectangular elements, while dual linear or circular polarization can be obtained with square or circular patch elements [James and Hall, 1989; Pozar and Schaubert, 1995].

Feeding Techniques for Microstrip Antennas

While Fig. 38.14 shows a coaxial probe-fed microstrip antenna element, it is also possible to feed the patch element by several other methods. Fig. 38.19a shows a rectangular patch element fed with a microstrip transmission line coplanar with the patch element. The amount of inset of the feed line controls the input impedance level at resonance, in a manner analogous to the positioning of the coax probe feed for impedance control. The equivalent circuit of the antenna near resonance is also shown in the figure. The patch appears as a parallel RLC resonant circuit, with a series inductance that represents the near-field effect of the microstrip feed line. (The same equivalent circuit applies to the probe-fed microstrip antenna.) Both the probe feed and the line feed excite the patch element through coupling between the equivalent J_z electric current of the feed and the E_z directed field of the patch resonator, which has a maximum below the center of the patch.

The direct-contacting coax probe and inset microstrip line feeds have the advantage of simplicity, but suffer from some disadvantages. First, bandwidth is limited because of the requirement of a thin substrate, as discussed above. In addition, the inherent *E*-plane asymmetry of these feeds generates higher-order modes which lead

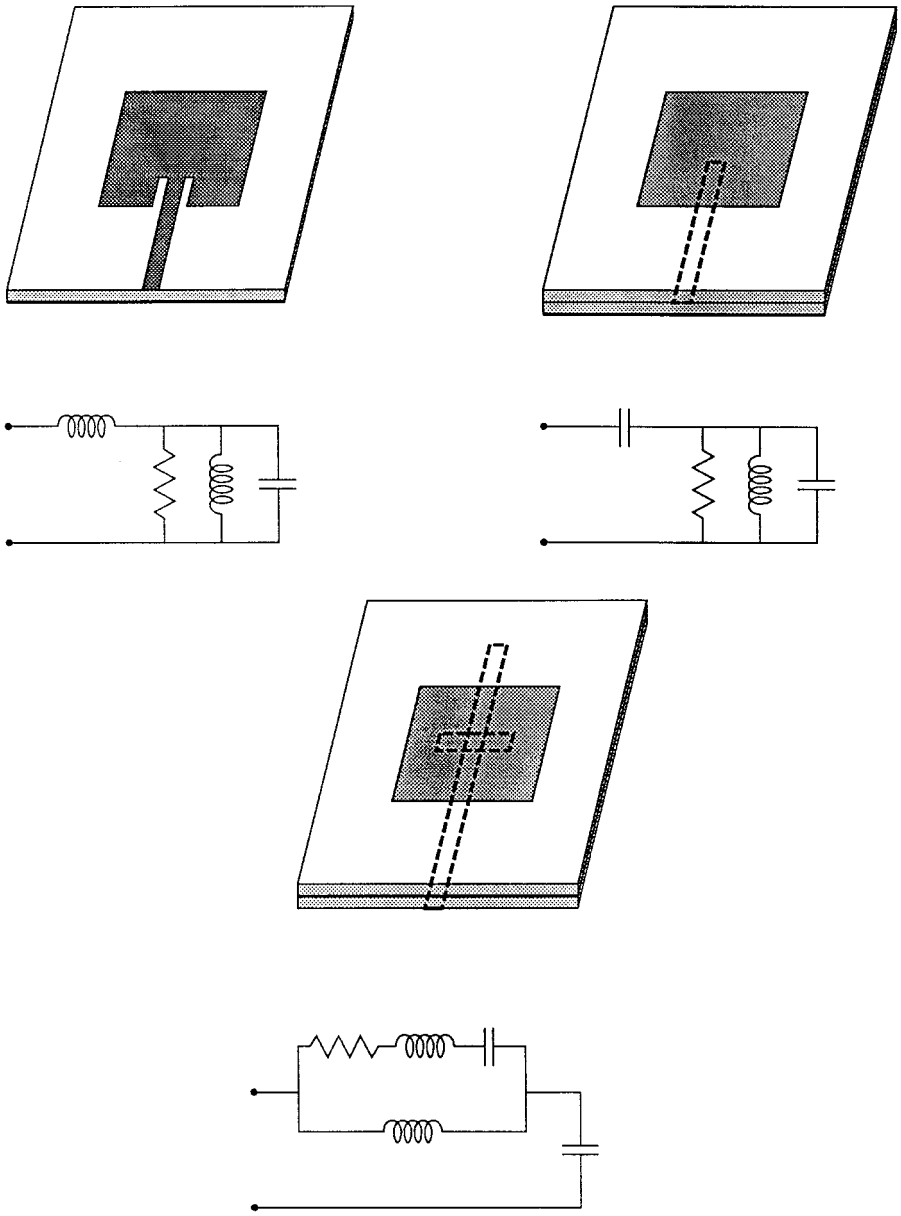


FIGURE 38.19 Three types of feeding methods for rectangular microstrip antennas, and their associated equivalent circuits: (a) patch fed with an inset microstrip transmission line, (b) patch fed by proximity coupling to a microstrip transmission line, (c) patch fed by aperture coupling to a microstrip transmission line.

to cross polarization. And, in the case of the coax feed, the need for soldering can decrease reliability and increase cost if a large number of elements are involved.

It is also possible to feed a microstrip antenna element using noncontacting feeds of various forms. Fig. 38.19b shows a proximity feed, where a two-layer substrate houses an embedded microstrip transmission feed line, with the radiating patch located on the top of a substrate layer placed over the microstrip feed line. The feed line is terminated in an open-circuited stub below the patch. Proximity coupling (often referred to in the literature by the less-descriptive term *electromagnetic coupling*) has the advantage of allowing the patch to reside on a relatively thick substrate, for enhanced bandwidth, while the feed line sees an effectively thinner substrate,

which is preferred to minimize spurious radiation and coupling. Fabrication is a bit more difficult than the single-layer coax or line feed, because of the requirement of bonding and aligning two substrates. The equivalent circuit of the proximity-coupled element is shown in Fig. 38.19b, where the series capacitor is indicative of the capacitive nature of the coupling between the open-ended microstrip line and the patch element.

Another type of noncontacting feed is the aperture-coupled microstrip antenna shown in Fig. 38.19c [James and Hall, 1989; Pozar and Schaubert, 1995]. This configuration consists of two parallel substrates separated by a ground plane. A microstrip feed line on the bottom of the bottom substrate is coupled through a small aperture (typically a thin slot) in the ground plane to a microstrip patch element on the top of the top substrate. This arrangement allows a thin, high-dielectric-constant substrate to be used for the feed line, and a thicker, low-dielectric-constant substrate to be used for the radiating element. In this way, the two-layer design of the aperture-coupled element allows the substrates to be optimized for the distinct functions of circuit components and radiating elements. In addition, the ground plane provides isolation between the radiating aperture and possible spurious radiation or coupling from the feed network. An important aspect of the aperture-coupled approach is that the coupling aperture is below resonant size, so that the backlobe radiated by the slot is typically 15 to 20 dB below the main forward beam.

The aperture-coupled geometry affords several degrees of freedom for control of the electrical properties of the antenna. The slot size (length) primarily determines the coupling level and, hence, the input impedance. Tightest coupling occurs when the slot is centered below the patch, with the input impedance decreasing as the slot size is decreased. As with any microstrip antenna, the resonant frequency is controlled primarily by the length of the patch. The feed line stub length can be used to adjust the reactance loading of the element, and the antenna substrate thickness and dielectric constant has a direct effect on the bandwidth of the element. It is also possible to use the slot to provide a double tuning effect to increase significantly impedance bandwidth. Such aperture-coupled antennas have been demonstrated with bandwidths up to 40% [Pozar and Schaubert, 1995]. Another unique feature of the aperture coupled patch is that the principal plane patterns have theoretically zero cross polarization because of the symmetry of the element.

Microstrip Antenna Arrays

One of the most useful features of microstrip antenna technology is the ease with which **array antennas** can be constructed, since the feed network can be fabricated with microstrip transmission lines and microstrip circuit components at the same time as the microstrip radiating elements. This eliminates the cumbersome and expensive coaxial or waveguide feed networks that are necessary for other types of arrays. In fact, microstrip technology offers such versatility in design that a wide variety of series-fed, corporate-fed, fixed-beam, scanning, multilayer, and polarization agile microstrip arrays have been demonstrated, many examples of which can be found in the literature [James and Hall, 1989; Pozar and Schaubert, 1995].

One of the most convenient architectures for microstrip arrays is the single-layer design where the microstrip feed lines are printed on the same substrate layer as the radiating patch elements. This results in a simple, low-profile, inexpensive, and easily fabricated antenna assembly. An example of a 2×4 microstrip array using this type of configuration is shown in Fig. 38.20. The microstrip feed network consists of a main feed line driving three levels of coplanar two-way power dividers, which in turn drive eight edge-fed patches. This is an example of a corporate feed network, in contrast to a series type of feed where array elements are tapped off of a single microstrip line. The corporate feed provides good bandwidth and allows precise control of element excitation, but requires considerable substrate area and can be lossy. A series feed can be very compact and efficient, but its bandwidth is typically limited to a few percent. Both types of feeds can be used for single and dual polarizations. More flexibility for feed network layout can be obtained by using a two-sided aperture-coupled patch geometry. This allows the feed network to be isolated from the radiating aperture by the ground plane, and the extra substrate area can be very useful for arrays that require dual polarization or dual-frequency operation. Similar features can be obtained by using feedthrough pins or vias, but the added fabricational complexity of solder connections can be formidable in a large array.

A serious limitation of microstrip array technology is that array gain is limited by the relatively high losses of microstrip transmission lines. While array directivity increases with the area of the radiating aperture, the losses of the feed network increase exponentially with array size. This is especially serious at higher frequencies,

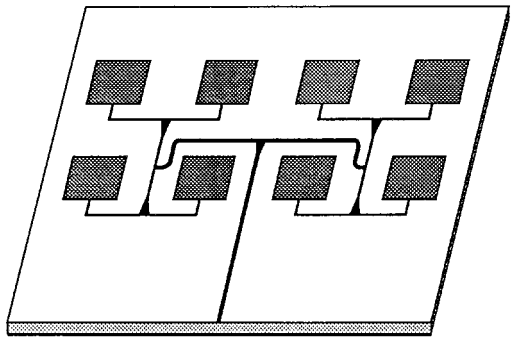


FIGURE 38.20 A corporate-fed eight-element microstrip array antenna.

where it is usually impractical to design a microstrip array with a coplanar feed network for gains in excess of 28 to 32 dB. Off-board feed networks or multilayer designs can be used to partially circumvent this problem, but at the expense of simplicity and cost.

Computer-Aided Design for Microstrip Antennas

Ideally, antenna CAD software would combine a user-friendly interface with a computationally efficient set of accurate and versatile theoretical models. While software with such features has reached a fairly high level of refinement for the analysis of low-frequency and microwave circuit analysis and optimization, the development of microstrip antenna CAD software lags far behind. One reason is the economic reality that the market for antenna software is relatively small, which perhaps explains why there is very little commercially available antenna CAD software of any type. Microstrip antenna CAD software development has also been slow because of the fact that such antennas are relatively new, receiving serious attention only during the last 15 years. Furthermore, microstrip antenna geometries are relatively difficult to model because of the presence of dielectric inhomogeneities and a wide variety of feeding techniques and other geometric features. This last consideration makes the development of a general-purpose microstrip antenna analysis package extremely difficult.

It may come as a surprise to the newcomer to practical antenna development, but it must be realized that many microstrip antenna designs have been successfully completed with little or no CAD support. There are, however, many situations where antennas and arrays can be designed more effectively, with better performance and less experimental iteration, when proper CAD software tools are available. And there are situations involving large arrays of microstrip elements which critically rely on the use of CAD software for successful design. Thus, CAD software is not absolutely necessary for all facets of microstrip antenna design work, but good software tools can be very useful for dealing with the more-complicated microstrip geometries. Another point that seems to be especially true for antenna design in general is that CAD software, no matter how versatile or accurate, cannot substitute for experience and understanding of the fundamentals of antenna operation. Further discussion of microstrip antenna CAD issues can be found in [Pozar and Schaubert, 1995].

Defining Terms

Array antenna: A repetitive grouping of basic antenna elements functioning as a single antenna with improved gain or pattern characteristics.

Bandwidth: The fractional frequency range over which the impedance match or pattern qualities of an antenna meet a required specification.

Beamwidth: The angular width of the main beam of an antenna, typically measured at either –3 or –10 dB below beam maximum.

Microstrip line: A planar transmission line consisting of a conducting strip printed (or etched) on a grounded dielectric substrate.

Microwave integrated circuit (MIC): A microwave or RF subsystem formed by the monolithic or hybrid integration of active devices, transmission lines, and related components.

Related Topics

37.1 Space Propagation • 37.2 Waveguides

References

J. R. James and P. S. Hall, Eds., *Handbook of Microstrip Antennas,* London: Peter Peregrinus (IEE), 1989.

D. M. Pozar, *Microwave Engineering,* Reading, Mass: Addison-Wesley, 1990.

D. M. Pozar and D. H. Schaubert, *Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays,* New York: IEEE Press, 1995.

Further Information

The most up-to-date information for developments in the field of microstrip antennas can be found in the technical journals. These include the *IEEE Transactions on Antennas and Propagation*, the *IEE Proceedings, Part H*, and *Electronics Letters.* There are also a large number of symposiums and conferences on antennas that usually emphasize practical microstrip antenna technology, such as the *IEEE International Symposium on Antennas and Propagation*, the *International Symposium on Antennas and Propagation* (ISAP–Japan), and the *International Conference on Antennas and Propagation* (ICAP–Great Britain). Coverage of basic antenna theory and design can be found in the previous sections in this chapter, as well as the references listed there.

Steer, M.B., Trew, R.J. “Microwave Devices”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

39

Microwave Devices

39.1 Passive Microwave Devices
Characterization of Passive Elements • Transmission Line Sections • Discontinuities • Impedance Transformers • Terminations • Attenuators • Microwave Resonators • Tuning Elements • Hybrid Circuits and Directional Couplers • Filters • Ferrite Components • Passive Semiconductor Devices

39.2 Active Microwave Devices
Semiconductor Material Properties • Two-Terminal Active Microwave Devices • Three-Terminal Active Microwave Devices

Michael B. Steer
North Carolina State University

Robert J. Trew
Case Western Reserve University

39.1 Passive Microwave Devices

Michael B. Steer

Wavelengths in air at microwave and millimeter-wave frequencies range from 1 m at 300 MHz to 1 mm at 300 GHz and are comparable to the physical dimensions of fabricated electrical components. For this reason circuit components commonly used at lower frequencies, such as resistors, capacitors, and inductors, are not readily available above 10 GHz. The available microwave frequency lumped elements have dimensions of around 1 mm. The relationship between the wavelength and physical dimensions enables new classes of distributed components to be constructed that have no analogy at lower frequencies. Components are realized by disturbing the field structure on a transmission line, resulting in energy storage and thus reactive effects. Electric (E) field disturbances have a capacitive effect and the magnetic (H) field disturbances appear inductive. Microwave components are fabricated in waveguide, coaxial lines, and strip lines. The majority of circuits are constructed using strip lines as the cost is relatively low and they are highly reproducible due to the photolithographic techniques used. Fabrication of waveguide components requires precision machining but they can tolerate higher power levels and are more easily realized at millimeter-wave frequencies (30–300 GHz) than either coaxial or microstrip components.

Characterization of Passive Elements

Passive microwave elements are defined in terms of their reflection and transmission properties for an incident wave of electric field or voltage. Scattering (S) parameters are based on traveling waves and so naturally describe these properties. As well they are the only ones that can be measured directly at microwave frequencies. S parameters are defined in terms of root power waves which in turn are defined using forward and backward traveling voltage waves. Consider the N port network of Fig. 39.1 where the nth port has a reference transmission line of characteristic impedance Z_{0n} and of infinitesimal length. The transmission line at the nth port serves to separate the forward and backward traveling voltage (V_n^+ and V_n^-) and current (I_n^+ and I_n^-) waves. The reference characteristic impedance matrix, $\mathbf{Z}_0$ is a diagonal matrix, $\mathbf{Z}_0 = \text{diag}(Z_{01} \ldots Z_{0n} \ldots Z_{0N})$, and the root power waves at the nth port, a_n and b_n, are defined by

$$a_n = V_n^+ \Big/ \sqrt{Z_{0n}} \qquad \text{and} \qquad b_n = V_n^- \sqrt{Z_{0n}} \tag{39.1}$$

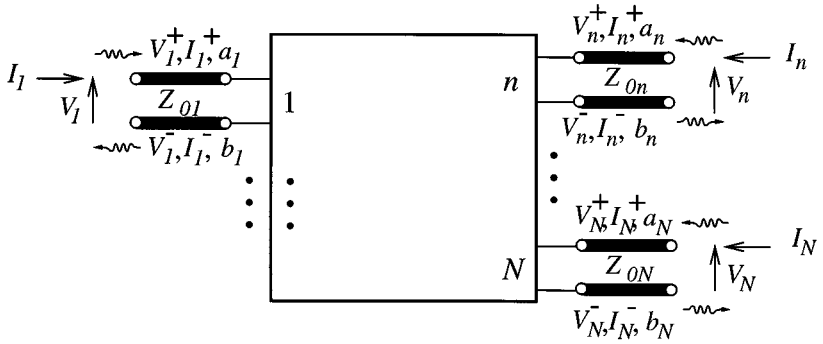


FIGURE 39.1 N port network with reference transmission lines used in defining S parameters.

In matrix form

$$\mathbf{a} = \mathbf{Z}_0^{-1/2}\mathbf{V}^+ = \mathbf{V}_0^{1/2}\mathbf{V}^+, \qquad \mathbf{b} = \mathbf{Z}_0^{-1/2}\mathbf{V}^- = \mathbf{Y}_0^{1/2}\mathbf{V}^-, \tag{39.2}$$

$$\mathbf{V}^+ = \mathbf{Z}_0^{1/2}\mathbf{a} = \mathbf{Y}_0^{-1/2}\mathbf{a} \qquad \text{and} \qquad \mathbf{V}^- = \mathbf{Z}_0^{1/2}\mathbf{b} = \mathbf{Y}_0^{-1/2}\mathbf{b} \tag{39.3}$$

where

$$\mathbf{a} = \left[a_1 \ldots a_n \ldots a_N\right]^{\mathrm{T}}, \qquad \mathbf{b} = \left[b_1 \ldots b_n \ldots b_N\right]^{\mathrm{T}}, \tag{39.4}$$

$$\mathbf{V}^+ = \left[V_1^+ \ldots V_n^+ \ldots V_N^+\right]^{\mathrm{T}}, \qquad \mathbf{V}^- = \left[V_1^- \ldots V_n^- \ldots V_N^-\right]^{\mathrm{T}}. \tag{39.5}$$

and the characteristic admittance matrix $\mathbf{Y}_0$ and $\mathbf{Z}_0^{-1}$. Now S parameters can be formally defined:

$$\mathbf{b} = \mathbf{Sa} \tag{39.6}$$

Thus, $\mathbf{Y}_0^{1/2}\mathbf{V}^- = \mathbf{S}\mathbf{Y}_0^{1/2}\mathbf{V}^+$ and so $\mathbf{V}^- = \mathbf{Y}_0^{-1/2}\mathbf{S}\mathbf{Y}_0^{1/2}\mathbf{V}^+$. This reduces to $\mathbf{V}^- = \mathbf{S}\mathbf{V}^+$ when all of the reference transmission lines have the same characteristic impedance.

S parameters can be related to other network parameters after first considering the relationship of total port voltage $\mathbf{V} = [V_1 \ldots V_n \ldots V_N]^{\mathrm{T}}$ and current $\mathbf{I} = [I_1 \ldots I_n \ldots I_N]^{\mathrm{T}}$ to forward and backward voltage and current waves:

$$\mathbf{V} = \mathbf{V}^+ + \mathbf{V}^- \qquad \text{and} \qquad \mathbf{I} = \mathbf{I}^+ + \mathbf{I}^- \tag{39.7}$$

where $\mathbf{I}^+ = \mathbf{Y}_0\mathbf{V}^+ = \mathbf{Y}_0^{1/2}\mathbf{a}$ and $\mathbf{I}^- = -\mathbf{Y}_0\mathbf{V}^- = -\mathbf{Y}_0^{1/2}\mathbf{b}$. The development of the relationship between S parameters and other network parameters is illustrated by considering Y parameters defined by

$$\mathbf{I} = \mathbf{YV} \tag{39.8}$$

Using traveling waves this becomes

$$\mathbf{I}^+ + \mathbf{I}^- = \mathbf{Y}\left(\mathbf{V}^+ + \mathbf{V}^-\right) \tag{39.9}$$

$$\mathbf{Y}_0\left(\mathbf{V}^+ - \mathbf{V}^-\right) = \mathbf{Y}\left(\mathbf{V}^+ + \mathbf{V}^-\right) \tag{39.10}$$

$$\mathbf{Y}_0\left(\mathbf{1} - \mathbf{Y}_0^{-1/2}\mathbf{S}\mathbf{Y}_0^{1/2}\right)\mathbf{V}^+ = \mathbf{Y}\left(\mathbf{1} + \mathbf{Y}_0^{-1/2}\mathbf{S}\mathbf{Y}_0^{1/2}\right)\mathbf{V}^+ \tag{39.11}$$

$$\mathbf{Y} = \mathbf{Y}_0\left(\mathbf{1} - \mathbf{Y}_0^{-1/2}\mathbf{S}\mathbf{Y}_0^{1/2}\right)\left(\mathbf{1} + \mathbf{Y}_0^{-1/2}\mathbf{S}\mathbf{Y}_0^{1/2}\right)^{-1} \tag{39.12}$$

Alternatively (39.10) can be rearranged as

$$\left(\mathbf{Y}_0 + \mathbf{Y}\right)\mathbf{V}^- = \left(\mathbf{Y}_0 - \mathbf{Y}\right)\mathbf{V}^+ \tag{39.13}$$

$$\mathbf{V}^- = \left(\mathbf{Y}_0 + \mathbf{Y}\right)^{-1}\left(\mathbf{Y}_0 - \mathbf{Y}\right)\mathbf{V}^+ \tag{39.14}$$

$$\mathbf{Y}_0^{-1/2}\mathbf{b} = \left(\mathbf{Y}_0 + \mathbf{Y}\right)^{-1}\left(\mathbf{Y}_0 - \mathbf{Y}\right)\mathbf{Y}_0^{-1/2}\mathbf{a} \tag{39.15}$$

Comparing this to the definition of S parameters, (39.6), leads to

$$\mathbf{S} = \mathbf{Y}_0^{1/2}\left(\mathbf{Y}_0 + \mathbf{Y}\right)^{-1}\left(\mathbf{Y}_0 - \mathbf{Y}\right)\mathbf{Y}_0^{-1/2} \tag{39.16}$$

For the usual case where all of the reference transmission lines have the same characteristic impedance $Z_0 = 1/Y_0$, $\mathbf{Y} = Y_0(\mathbf{1} - \mathbf{S})(\mathbf{1} + \mathbf{S})^{-1}$ and $\mathbf{S} = (\mathbf{Y}_0 + \mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y})$.

The most common situation involving conversion to and from S parameters is for a two port with both ports having a common reference characteristic impedance Z_0. Table 39.1 lists the most common conversions. S parameters require that the reference impedances be specified. If they are not it is assumed that it is 50 **Ω**. They are commonly plotted on Smith Charts — polar plots with lines of constant resistance and reactance [Vendelin *et al.*].

In Fig. 39.2(a) a travelling voltage wave with phasor $\mathbf{V}_1^+$ is incident at port 1 of a two-port passive element. A voltage $\mathbf{V}_1^-$ is reflected and $\mathbf{V}_2^-$ is transmitted. $\mathbf{V}_2^-$ is then reflected by Z_L to produce $\mathbf{V}_2^+$. $\mathbf{V}_2^+$ is zero if $Z_L = Z_0$. The input voltage reflection coefficient

$$\Gamma_1 = \mathbf{V}_1^-/\mathbf{V}_1^+ = s_{11} + s_{12}s_{21}/\left(1 - s_{22}\Gamma_L\right),$$

transmission coefficient

$$\mathrm{T} = \mathbf{V}_2^-/\mathbf{V}_1^+$$

and the load reflection coefficient

$$\Gamma_L = \left(Z_L - Z_0\right)/\left(Z_L + Z_0\right)$$

More convenient measures of reflection and transmission performance are the **return loss** and **insertion loss** as they are relative measures of power in transmitted and reflected signals. In decibels

$$\text{RETURN LOSS} = -20\log\Gamma_1\ (dB) \qquad \text{INSERTION LOSS} = -20\log\mathrm{T}\ (dB)$$

The input impedance at port 1, Z_{in}, is related to Γ by

$$Z_{\text{in}} = Z_0\left(1 + \Gamma_1/1 - \Gamma_1\right)$$

TABLE 39.1 Two-Port S Parameter Conversion Chart for Impedance, Z, Admittance, Y, and Hybrid, H, Parameters

	S	In Terms of **S**
Z	$z'_{11} = z_{11}/Z_0 \quad z'_{12} = z_{12}/Z_0$	$z'_{21} = z_{21}/Z_0 \quad z'_{22} = z_{22}/Z_0$
	$\delta = (Z'_{11}+1)(Z'_{22}+1) - Z'_{12}Z'_{21}$	$\delta = (1-S_{11})(1-S_{22}) - S_{12}S_{21}$
	$S_{11} = \left[(Z'_{11}-1)(Z'_{22}+1) - Z'_{12}Z'_{21}\right]/\delta$	$z'_{11} = \left[(1+S_{11})(1-S_{22}) + S_{12}S_{21}\right]/\delta$
	$S_{12} = 2Z'_{12}/\delta$	$Z'_{12} = 2S_{12}/\delta$
	$S_{21} = 2Z'_{21}/\delta$	$Z'_{21} = 2S_{21}/\delta$
	$S_{22} = \left[(Z'_{11}+1)(Z'_{22}-1) - Z'_{12}Z'_{21}\right]/\delta$	$Z'_{22} = \left[(1-S_{11})(1+S_{22}) + S_{12}S_{21}\right]/\delta$
Y	$Y'_{11} = Y_{11}Z_0 \quad Y'_{12} = Y_{12}Z_0$	$Y'_{21} = Y'_{21}Z_0 \quad Y'_{22} = Y_{22}Z_0$
	$\delta = (1+Y'_{11})(1+Y'_{22}) = Y'_{12}Y'_{21}$	$\delta = (1+S_{11})(1+S_{22}) - S_{12}S_{21}$
	$S_{11} = \left[(1-Y'_{11})(1+Y'_{22}) + Y'_{12}Y'_{21}\right]/\delta$	$Y'_{11} = \left[(1-S_{11})(1+S_{22}) + S_{12}S_{21}\right]/\delta$
	$S_{12} = -2Y'_{12}/\delta$	$Y'_{12} = -2S_{12}/\delta$
	$S_{21} = -2Y'_{21}/\delta$	$Y'_{21} = -2S_{21}/\delta$
	$S_{22} = \left[(1+Y'_{11})(1-Y'_{22}) + Y'_{12}Y'_{21}\right]/\delta$	$Y'_{22} = \left[(1+S_{11})(1-S_{22}) + S_{12}S_{21}\right]/\delta$
H	$H'_{11} = H_{11}/Z_0 \quad H'_{12} = H_{12}$	$H'_{21} = H_{21} \quad H'_{22} = H_{22}Z_0$
	$\delta = (1+H'_{11})(1+H'_{22}) - H'_{12}H'_{21}$	$\delta = (1-S_{11})(1+S_{22}) + S_{12}S_{21}$
	$S_{11} = \left[(H'_{11}-1)(H'_{22}+1) - H'_{12}H'_{21}\right]/\delta$	$H'_{11} = \left[(1+S_{11})(1+S_{22}) - S_{12}S_{21}\right]/\delta$
	$S_{12} = 2H'_{12}/\delta$	$H'_{12} = 2S_{12}/\delta$
	$S_{21} = -2H'_{21}/\delta$	$H'_{21} = -2S_{21}/\delta$
	$S_{22} = \left[(1+H'_{11})(1-H'_{22}) + H'_{12}H'_{21}\right]/\delta$	$H'_{22} = \left[(1-S_{11})(1-S_{22}) - S_{12}S_{21}\right]/\delta$

Note: The Z′, Y′ and H′ parameters are normalized to Z_0.

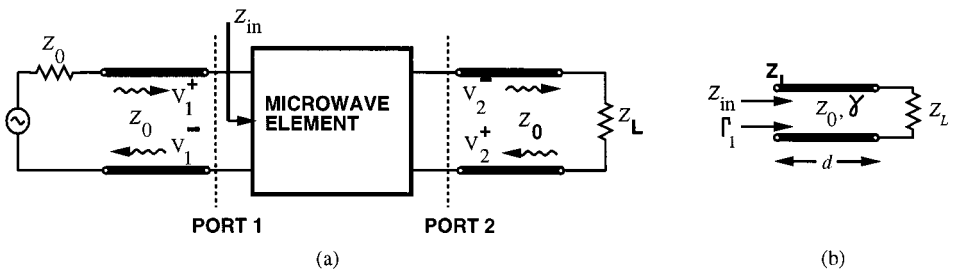


FIGURE 39.2 Incident, reflected and transmitted traveling voltage waves at (a) a passive microwave element and (b) a transmission line.

The reflection characteristics are also described by the voltage standing wave ratio (VSWR), a quantity that can be measured using relatively simple equipment. The VSWR is the ratio of the maximum voltage amplitude on the imput transmission line $\left(\left|V_1^+\right| + \left|V_1^-\right|\right)$ to the minimum voltage amplitude $\left(\left|V_1^+\right| - \left|V_1^-\right|\right)$. Thus,

$$\text{VSWR} = \left(1 + \left|\Gamma_1\right|\right)\Big/\left(1 - \left|\Gamma_1\right|\right)$$

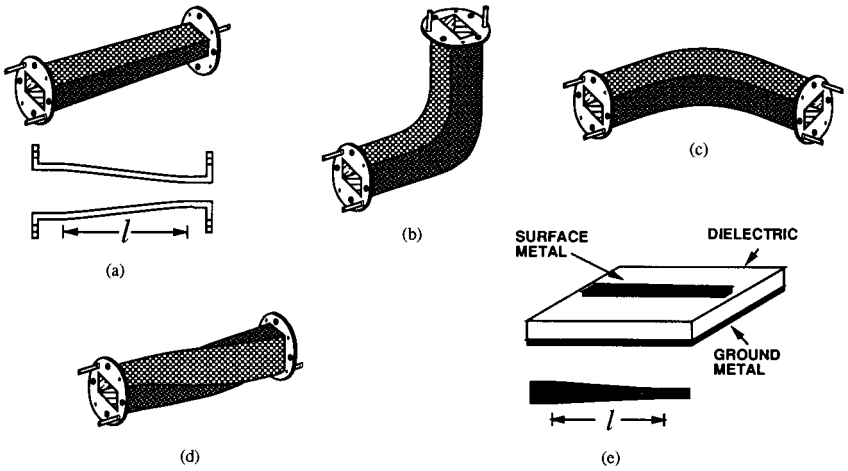


FIGURE 39.3 Sections of transmission lines used for interconnecting components: (a) waveguide tapered section, (b) waveguide E-plane bend, (c) waveguide H-plane bend, (d) waveguide twist, and (e) microstrip taper.

Most passive devices, with the notable exception of ferrite devices, are reciprocal and so $S_{pq} = S_{qp}$. A loss-less passive device also satisfies the unitary condition: $\Sigma_p |S_{pq}|^2 = 1$, which is a statement of power conservation indicating that all power is either reflected or transmitted.

Most microwave circuits are designed to minimize the reflected energy and maximize transmission at least over the frequency range of operation. Thus, the return loss is high and the VSWR ≈ 1 for well-designed circuits.

A terminated transmission line such as that in Fig. 39.2(b) has an input impedance

$$Z_{\text{in}} = Z_0 \frac{Z_L + jZ_0 \tanh \gamma d}{Z_0 + jZ_L \tanh \gamma d}$$

Thus, a short section ($\gamma_d << 1$) of a short circuited ($Z_L = 0$) transmission line looks like an inductor and a capacitor if it is open circuited ($Z_L = \infty$). When the line is a half wavelength long, an open circuit is presented at the input to the line if the other end is short circuited.

Transmission Line Sections

The simplest microwave circuit element is a uniform section of transmission line which can be used to introduce a time delay or a frequency-dependent phase shift. Other line segments for interconnections include bends, corners, twists, and transitions between lines of different dimensions (see Fig. 39.3). The dimensions and shapes are designed to minimize reflections and so maximize return loss and minimize insertion loss.

Discontinuities

The waveguide discontinuities shown in Fig. 39.4(a)–(f) illustrate most clearly the use of E and H field disturbances to realize capacitive and inductive components. An E-plane discontinuity [Fig. 39.4(a)] can be modeled approximately by a frequency-dependent capacitor. H-plane discontinuities [Figs. 39.4(b) and (c)] resemble inductors as does the circular iris of Fig. 39.4(d). The resonant waveguide iris of Fig. 39.4(e) disturbs both the E and H fields and can be modeled by a parallel LC resonant circuit near the frequency of resonance. Posts in waveguide are used both as reactive elements [Fig. 39.4(f)] and to mount active devices [Fig. 39.4(g)]. The equivalent circuits of microstrip discontinuities [Figs. 39.4(k)–(o)] are again modeled by capacitive elements if the E field is interrupted and by inductive elements if the H field (or current) is interrupted. The stub shown in Fig. 39.4(j) presents a short circuit to the through transmission line when the length of the stub is $\lambda_g/4$. When the stubs are electrically short ($<< \lambda_g/4$) they introduce shunt capacitances in the through transmission line.

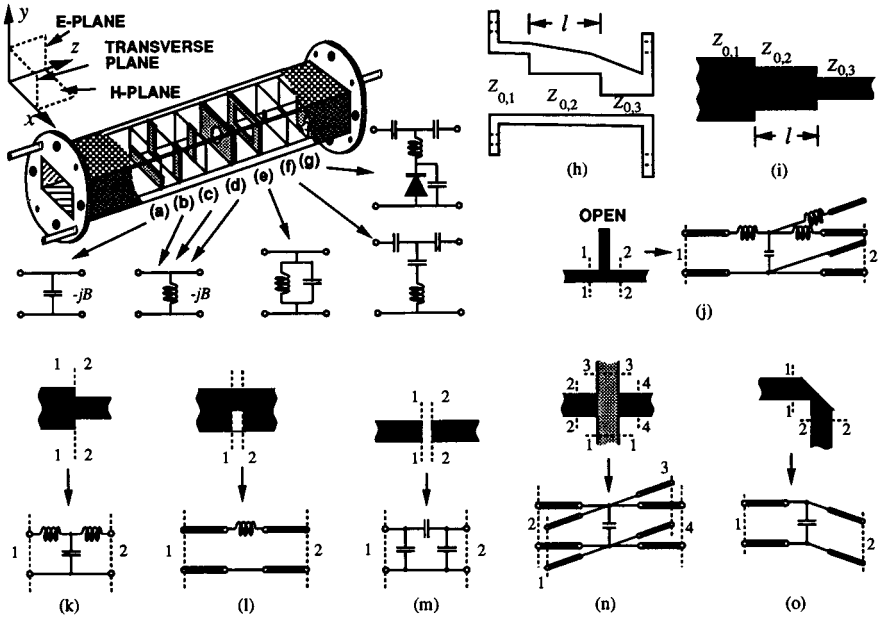


FIGURE 39.4 Discontinuities. Waveguide discontinuities: (a) capacitive E-plane discontinuity, (b) inductive H-plane discontinuity, (c) symmetrical inductive H-plane discontinuity, (d) inductive post discontinuity, (e) resonant window discontinuity, (f) capacitive post discontinuity, (g) diode post mount, and (h) quarter-wave impedance transformer. Microstrip discontinuities: (i) quarter-wave impedance transformer, (j) open microstrip stub, (k) step, (l) notch, (m) gap, (n) crossover, and (o) bend.

Impedance Transformers

Impedance transformers are used to interface two sections of line with different **characteristic impedances**. The smoothest transition and the one with the broadest bandwidth is a tapered line as shown in Fig. 39.3(a) and (e). This element tends to be very long and so step terminations called quarter-wave impedance transformers [see Fig. 39.4(h) and (i)] are sometimes used although their bandwidth is relatively small centered on the frequency at which $l = \lambda_g/4$. Ideally, $Z_{0,2} = \sqrt{Z_{0,1}Z_{0,3}}$.

Terminations

In a termination, power is absorbed by a length of lossy material at the end of a shorted piece of transmission line [Fig. 39.5 (a) and (c)]. This type of termination is called a matched load as power is absorbed and reflections are very small irrespective of the characteristic impedance of the transmission line. This is generally preferred as the characteristic impedance of transmission lines varies with frequency, particularly so for waveguides. When the characteristic impedance of a line does not vary much with frequency, as is the case with a coaxial line, a simpler smaller termination can be realized by placing a resistor to ground [Fig. 39.5(b)].

Attenuators

Attenuators reduce the level of a signal traveling along a transmission line. The basic construction is to make the line lossy but with a characteristic impedance approximating that of the connecting lines so as to reduce reflections. The line is made lossy by introducing a resistive vane in the case of a waveguide [Fig. 39.5(d)], replacing part of the outer conductor of a coaxial line by resistive material [Fig. 39.5(e)], or covering the line by resistive material in the case of a microstrip line [Fig. 39.5(f)]. If the amount of lossy material introduced into the transmission line is controlled, a variable attenuator is achieved, e.g., Fig. 39.5(d).

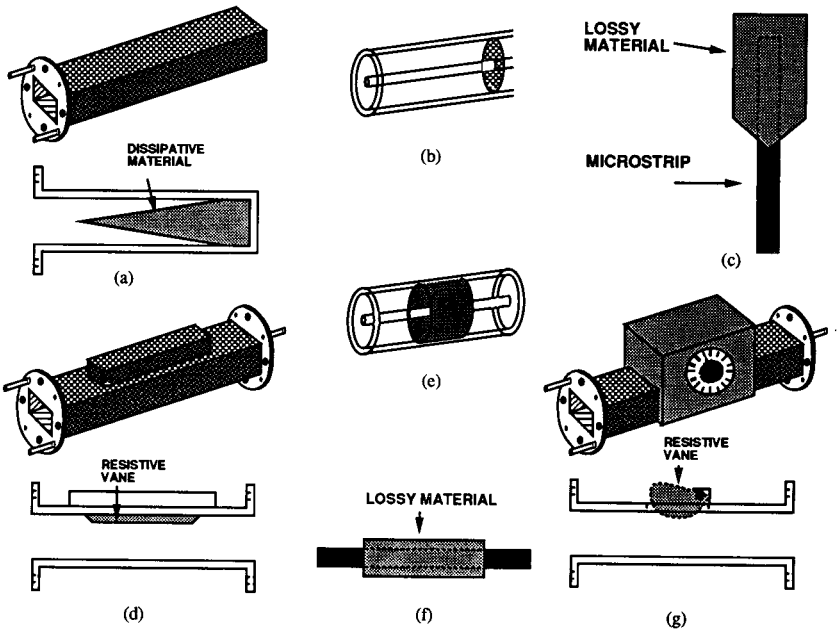


FIGURE 39.5 Terminations and attenuators: (a) waveguide matched load, (b) coaxial line resistive termination, (c) microstrip matched load, (d) waveguide fixed attenuator, (e) coaxial fixed attenuator, (f) microstrip attenuator, and (g) waveguide variable attenuator.

Microwave Resonators

In a lumped element resonant circuit, stored energy is transferred between an inductor which stores magnetic energy and a capacitor which stores electric energy, and back again every period. Microwave resonators function the same way, exchanging energy stored in electric and magnetic forms but with the energy stored spatially. Resonators are described in terms of their quality factor

$$Q = 2\pi f_0 \left(\frac{\text{Maximum energy stored in the resonator at } f_0}{\text{Power lost in the cavity}} \right) \tag{39.17}$$

where f_0 is the resonant frequency. The Q is reduced and thus the resonator bandwidth is increased by the power lost due to coupling to the external circuit so that the loaded Q

$$\begin{aligned} Q_L &= 2\pi f_0 \left(\frac{\text{Maximum energy stored in the resonator at } f_0}{\text{Power lost in the cavity and to the external circuit}} \right) \\ &= \frac{1}{1/Q + 1/Q_{\text{ext}}} \end{aligned} \tag{39.18}$$

where Q_{ext} is called the external Q. Q_L accounts for the power extracted from the resonant circuit and is typically large. For the simple response shown in Fig. 39.6(a) the half power (3 dB) bandwidth is f_0/Q_L.

Near resonance the response of a microwave resonator is very similar to the resonance response of a parallel or series R, L, C resonant circuit [Fig. 39.6(f) and (g)]. These equivalent circuits can be used over a narrow frequency range.

Several types of resonators are shown in Fig. 39.6. Figure 39.6(b) is a rectangular cavity resonator coupled to an external coaxial line by a small coupling loop. Figure 39.6(c) is a microstrip patch reflection resonator.

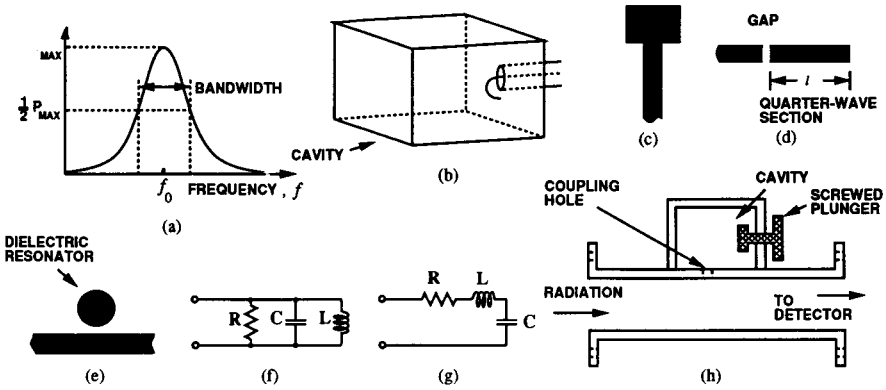


FIGURE 39.6 Microwave resonators: (a) resonator response, (b) rectangular cavity resonator, (c) microstrip patch resonator, (d) microstrip gap-coupled reflection resonator, (e) transmission dielectric transmission resonator in microstrip, (f) parallel equivalent circuits, (g) series equivalent circuits, and (h) waveguide wavemeter.

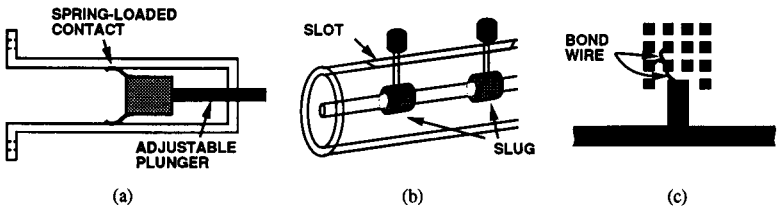


FIGURE 39.7 Tuning elements: (a) waveguide sliding short circuit, (b) coaxial line slug tuner, (c) microstrip stub with tuning pads.

This resonator has large coupling to the external circuit. The coupling can be reduced and photolithographically controlled by introducing a gap as shown in Fig. 39.6(d) for a microstrip gap-coupled transmission line reflection resonator. The Q of a resonator can be dramatically increased by using a high dielectric constant material as shown in Fig. 39.6(e) for a dielectric transmission resonator in microstrip. One simple application of a cavity resonator is the waveguide wavemeter [Fig. 39.6(h)]. Here the resonant frequency of a rectangular cavity is varied by changing the physical dimensions of the cavity with a null of the detector indicating that the frequency corresponds to the resonant cavity frequency.

Tuning Elements

In rectangular waveguide the basic adjustable tuning element is the sliding short shown in Fig. 39.7(a). Varying the position of the short will change resonance frequencies of cavities. It can be combined with hybrid tees to achieve a variety of tuning functions. The post in Fig. 39.4(f) can be replaced by a screw to obtain a screw tuner which is commonly used in waveguide filters. Sliding short circuits can be used in coaxial lines and in conjunction with branching elements to obtain stub tuners. Coaxial slug tuners are also used to provide adjustable matching at the input and output of active circuits. The slug is movable and changes the characteristic impedance of the transmission line. It is more difficult to achieve variable tuning in passive microstrip circuits. One solution is to provide a number of pads as shown in Fig. 39.7(c) which, in this case, can be bonded to the stub to obtain an adjustable stub length. Variable amounts of phase shift can be inserted by using a variable length of line called a line stretcher, or by a line with a variable propagation constant. One type of waveguide variable phase shifter is similar to the variable attenuator of Fig. 39.5(d) with the resistive material replaced by a low-loss dielectric.

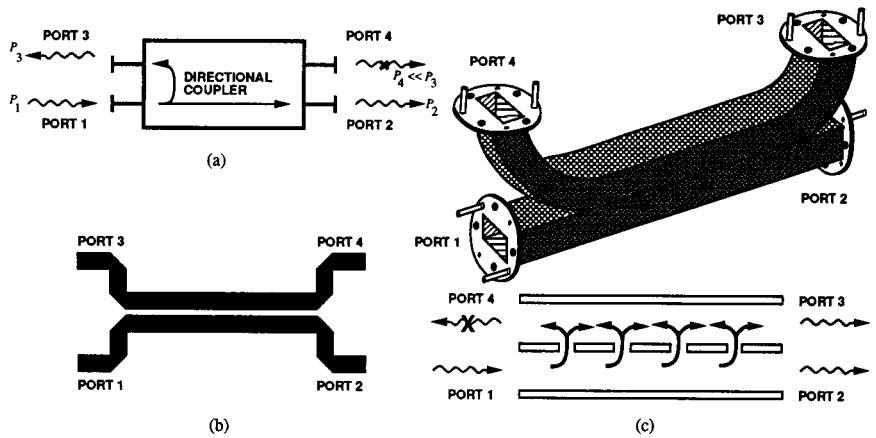


FIGURE 39.8 Directional couplers: (a) schematic, (b) backward-coupling microstrip directional coupler, (c) forward-coupling waveguide directional coupler.

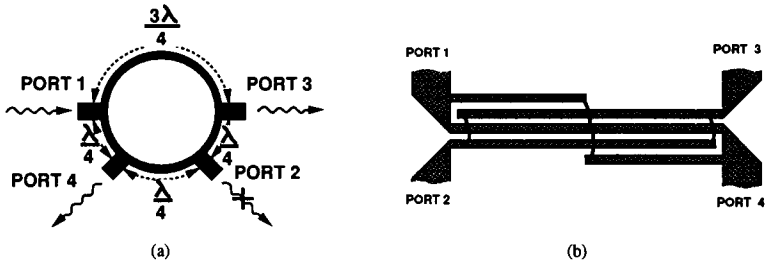


FIGURE 39.9 Microstrip hybrids: (a) rat race hybrid and (b) Lange coupler.

Hybrid Circuits and Directional Couplers

Hybrid circuits are multiport components which preferentially route a signal incident at one port to the other ports. This property is called directivity. One type of hybrid is called a directional coupler, the schematic of which is shown in Fig. 39.8(a). Here the signal incident at port 1 is coupled to ports 2 and 3 while very little is coupled to port 4. Similarly, a signal incident at port 2 is coupled to ports 1 and 4 but very little power appears at port 3. The feature that distinguishes a directional coupler from other types of hybrids is that the power at the output ports (here ports 2 and 3) is different. The performance of a directional coupler is specified by three parameters:

$$\begin{aligned} \text{Coupling factor} &= P_1/P_3 \\ \text{Directivity} &= P_3/P_4 \\ \text{Isolation} &= P_1/P_4 \end{aligned} \tag{39.19}$$

Microstrip and waveguide realizations of directional couplers are shown in Figs. 39.8(b) and (c) where the microstrip coupler couples in the backward direction and the waveguide coupler couples in the forward direction. The powers at the output ports of the hybrids shown in Fig. 39.9 are equal and so the hybrids serve to split a signal into half as well as having directional sensitivity.

Filters

Filters are combinations of microwave passive elements designed to have a specified frequency response. Typically, a topology of a filter is chosen based on established lumped element filter design theory. Then computer-aided design techniques are used to optimize the response of the circuit to the desired response.

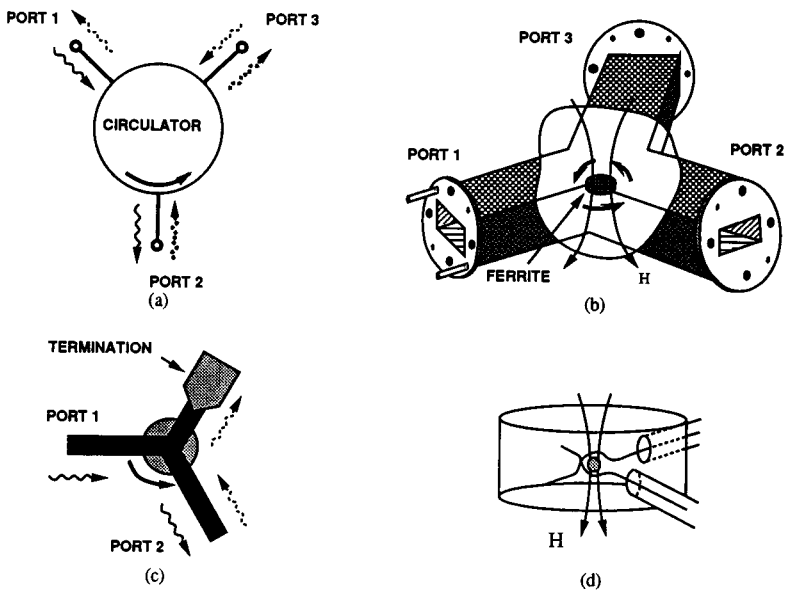


FIGURE 39.10 Ferrite components: (a) schematic of a circulator, (b) a waveguide circulator, (c) a microstrip isolator, and (d) a YIG tuned bandpass filter.

Ferrite Components

Ferrite components are nonreciprocal in that the insertion loss for a wave traveling from port A to port B is not the same as that from port B to port A.

Circulators and Isolators

The most important type of ferrite component is a circulator [Fig. 39.10(a) and (b)]. The essential element of a circulator is a piece of ferrite which when magnetized becomes nonreciprocal, preferring progression of electromagnetic fields in one circular direction. An ideal circulator has the scattering matrix

$$[S] = \begin{bmatrix} 0 & 0 & S_{13} \\ S_{21} & 0 & 0 \\ 0 & S_{32} & 0 \end{bmatrix} \tag{39.20}$$

In addition to the insertion and return losses, the performance of a circulator is described by its isolation which is its insertion loss in the undesired direction. An isolator is just a three-port circulator with one of the ports terminated in a matched load as shown in the microstrip realization of Fig. 39.10(c). It is used in a transmission line to pass power in one direction but not in the reverse direction. It is commonly used to protect the output of equipment from high reflected signals. The heart of isolators and circulators is the nonreciprocal element. Electronic versions have been developed for MMICs. A four-port version is called a duplexer and is used in radar systems and to separate the received and transmitted signals in a transceiver.

YIG Tuned Resonator

A magnetized YIG (yttrium iron garnet) sphere, shown in Fig. 39.10(d), provides coupling between two lines over a very narrow bandwidth. The center frequency of this bandpass filter can be adjusted by varying the magnetizing field.

Passive Semiconductor Devices

A semiconductor diode can be modeled by a voltage-dependent resistor and capacitor in shunt. Thus an applied dc voltage can be used to change the value of a passive circuit element. Diodes optimized to produce a voltage variable capacitor are called varactors. In detector circuits a diode's voltage variable resistance is used to achieve rectification and, through design, produce a dc voltage proportional to the power of an incident microwave signal. A controllable variable resistance is used in a PIN diode to realize an electronic switch.

Defining Terms

Characteristic impedance: Ratio of the voltage and current on a transmission line when there are no reflections.
Insertion loss: Power lost when a signal passes through a device.
Reference impedance: Impedance to which scattering parameters are referenced.
Return loss: Power lost upon reflection from a device.
Voltage standing wave ratio (VSWR): Ratio of the maximum voltage amplitude on a line to the minimum voltage ampitude.

Related Topics

35.3 Wave Equations and Wave Solutions • 37.2 Waveguides • 57.3 Applications of Magnetooptic Effects

Reference

G.D. Vendelin, A.M. Pavio, and U.L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, New York: Wiley, 1990.

Further Information

The following books provide good overviews of passive microwave components: *Microwave Engineering Passive Circuits* by P.A. Rizzi, Prentice-Hall, Englewood Cliffs, N.J., 1988; *Microwave Devices and Circuits* by S.Y. Liao, 3rd ed., Prentice-Hall, Englewood Cliffs, N.J., 1990; *Microwave Theory, Components and Devices* by J.A. Seeger, Prentice-Hall, Englewood Cliffs, N.J., 1986; *Microwave Technology* by E. Pehl, Artech House, Dedham, Mass., 1985; *Microwave Engineering and Systems Applications* by E.A. Wolff and R. Kaul, Wiley, New York, 1988; and *Microwave Engineering* by T.K. Ishii, 2nd ed., Harcourt Brace Jovanovich, Orlando, Fla., 1989. *Microwave Circuit Design Using Linear and Nonlinear Techniques* by G.D. Vendelin, A.M. Pavio, and U.L. Rohde, Wiley, New York, 1990, provides a comprehensive treatment of computer-aided design techniques for both passive and active microwave circuits. *Microwave Transistor Amplifiers: Analysis and Design,* 2nd ed., by G. Gonzalez, Prentice-Hall, Englewood Cliffs, N.J., 1996.

The monthly journals *IEEE Transactions on Microwave Theory Techniques, IEEE Microwave and Guided Wave Letters,* and *IEEE Transactions on Antennas and Propagation* publish articles on modeling and design of microwave passive circuit components. Articles in the first two journals are more circuit and component oriented while the third focuses on field theoretic analysis. These are published by The Institute of Electrical and Electronics Engineers, Inc. For subscription or ordering contact: IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, New Jersey 08855-1331.

Articles can also be found in the biweekly magazine *Electronics Letters* and the bimonthly magazine *IEE Proceedings Part H—Microwave, Optics and Antennas.* Both are published by the Institute of Electrical Engineers and subscription inquiries should be sent to IEE Publication Sales, P.O. Box 96, Stenage, Herts. SG1 2SD, United Kingdom. Telephone number (0438) 313311.

The *International Journal of Microwave and Millimeter-Wave Computer-Aided Engineering* is a quarterly journal devoted to the computer-aided design aspects of microwave circuits and has articles on component modeling and computer-aided design techniques. It has a large number of review-type articles. For subscription information contact John Wiley & Sons, Inc., Periodicals Division, P.O. Box 7247-8491, Philadelphia, Pennsylvania 19170-8491.

39.2 Active Microwave Devices

Robert J. Trew

Active devices that can supply gain at microwave frequencies can be fabricated from a variety of semiconductor materials. The availability of such devices permits a wide variety of system components to be designed and fabricated. Systems are generally constructed from components such as filters, amplifiers, oscillators, mixers, phase shifters, switches, etc. Active devices are primarily required for the oscillator and amplifier components. For these functions, devices that can supply current, voltage, or power gain at the frequency of interest are embedded in circuits that are designed to provide the device with the proper environment to create the desired response. The operation of the component is dictated, therefore, by both the capabilities of the active device and its embedding circuit.

It is common to fabricate microwave integrated circuits using both hybrid and monolithic techniques. In the hybrid approach, discrete active devices are mounted in RF circuits that can be fabricated from waveguides or transmission lines fabricated using coaxial, microstrip, stripline, coplanar waveguide, or other such media. Monolithic circuits are fabricated with both the active device and the RF circuit fabricated in the same semiconductor chip. Interconnection lines and the embedding RF circuit are generally fabricated using microstrip or coplanar waveguide transmission lines.

Active microwave devices can be fabricated as **two-terminal devices** (diodes) or **three-terminal devices** (transistors). Generally, three-terminal devices are preferred for most applications since the third terminal provides a convenient means to control the RF performance of the device. The third terminal allows for inherent isolation between the input and output RF circuit. Amplifiers and oscillators can easily be designed by providing circuits with proper stabilization or feedback characteristics. Amplifiers and oscillators can also be designed using two-terminal devices (diodes), but input/output isolation is more difficult to achieve since only one RF port is available. In this case it is generally necessary to use RF isolators or circulators.

The most commonly used two-terminal active devices consist of Gunn, tunnel, and IMPATT diodes. These devices can be designed to provide useful gain from low gigahertz frequencies to high millimeter-wave frequencies. Three-terminal devices consist of bipolar (BJT), heterojunction bipolar (HBT), and field-effect transistors (MESFETs and HEMTs). These devices can also be operated from UHF to millimeter-wave frequencies.

Semiconductor Material Properties

Active device operation is strongly dependent upon the charge transport characteristics of the semiconductor materials from which the device is fabricated. Semiconductor materials can be grown in single crystals with very high purity. The electrical conductivity of the crystal can be precisely controlled by introduction of minute quantities of dopant impurities. When these impurities are electrically activated, they permit precise values of current flow through the crystals to be controlled by potentials applied to contacts, placed upon the crystals. By clever positioning of the metal contacts, various types of semiconductor devices are fabricated. In this section we will briefly discuss the important material characteristics.

Semiconductor material parameters of interest for device fabrication consist of those involved in charge transport through the crystal, as well as thermal and mechanical properties of the semiconductor. The charge transport properties describe the ease with which free charge can flow through the material. For example, the velocity-electric field characteristics for several commonly used semiconductors are shown in Fig. 39.11. At low values of electric field, the charge transport is ohmic and the charge velocity is directly proportional to the magnitude of the electric field. The proportionality constant is called the mobility and has units of cm^2/V-s. Above a critical value for the electric field, the charge velocity saturates and either becomes constant (e.g., Si) or decreases with increasing field (e.g., GaAs). Both of these behaviors have implications for device fabrication, especially for devices intended for high-frequency operation. Generally, a high velocity is desired since current is directly proportional to velocity. Also, a low value for the saturation electric field is desirable since this implies a high-charge mobility. High mobility implies low resistivity and, therefore, low values for parasitic and access resistances for semiconductor devices.

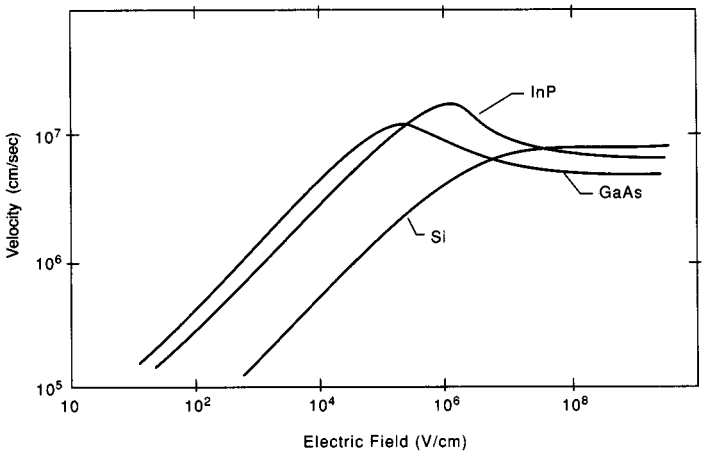


FIGURE 39.11 Electron velocity versus electric field for several semiconductors. This figure shows the electron velocity in several common semiconductors as a funcation of electric field strength. At low electric field the electron velocity is ohmic, as indicated by the linear characteristic. At higher electric field strength the electron velocity saturates and becomes nonlinear. Compound semiconductors such as GaAs and InP have highly nonlinear behavior at large electric fields.

TABLE 39.2 Material Parameters for Several Semiconductors

Semiconductor	E_g (eV)	ε_r	κ (W/cm-K) @300 K	E_c (V/cm)	$\tau_{minority}$ (s)
Si	1.12	11.9	1.5	3×10^5	2.5×10^{-3}
GaAs	1.42	12.5	0.54	4×10^5	$\sim 10^{-8}$
InP	1.34	12.4	0.67	4.5×10^5	$\sim 10^{-8}$
α-SiC	2.86	10.0	4	$(1–5) \times 10^6$	$\sim(1–10) \times 10^{-9}$
β-SiC	2.2	9.7	4	$(1–5) \times 10^6$	$\sim(1–10) \times 10^{-9}$

The decreasing electron velocity with electric field characteristic for compound semiconductors such as GaAs and InP makes possible active two-terminal devices called transferred electron devices (TEDs) or Gunn diodes. The negative slope of the velocity versus electric field characteristic implies a decreasing current with increasing voltage. That is, the device has a negative resistance. When a properly sized piece of these materials is biased and placed in a resonant cavity, the device will be unstable up to very high frequencies. By proper selection of embedding impedances oscillators or amplifiers can be constructed.

Other semiconductor materials parameters of interest include thermal, dielectric constant, energy bandgap, electric breakdown characteristics, and minority carrier lifetime. The thermal conductivity of the material is important because it describes how easily heat can be extracted from the device. The thermal conductivity has units of W/cm-K. Generally, high thermal conductivity is desirable. Compound semiconductors, such as GaAs and InP, have relatively poor thermal conductivity compared to elemental semiconductors such as Si. Materials such as SiC have excellent thermal conductivity and have uses in high-power electronic devices. The dielectric constant is important since it affects the size of the semiconductor device. The larger the dielectric constant, the smaller the device. Electric breakdown characteristics are important since breakdown limits the magnitudes of the dc and RF voltages that can be applied to the device. This is turn limits the RF power that can be handled by the device. The electric breakdown for the material is generally described by the critical value of electric field that produces avalanche ionization. Minority carrier lifetime is important for bipolar devices, such as pn junction diodes, rectifiers, and bipolar junction transistors (BJTs). A low value for minority carrier lifetime is desirable for devices such as diode temperature sensors and switches where low reverse bias leakage current is desirable. A long minority carrier lifetime is desirable for devices such as bipolar transistors. For materials such as Si and SiC, the minority carrier lifetime can be varied by controlled impurity doping. A comparison of some of the important material parameters for several common semiconductors is presented in Table 39.2.

Two-Terminal Active Microwave Devices

The IMPATT diode, transferred electron device, and tunnel diode are the most commonly used two-terminal devices. These devices can operate from the low microwave through high millimeter-wave frequencies. They were the first semiconductor devices that could provide useful RF power levels at microwave and millimeter-wave frequencies. The three devices are similar in that they are fabricated from blocks of semiconductors and require two electrodes (anode and cathode) for supplying dc bias. The same electrodes are used for the RF port, and since only two electrodes are available, the devices must be operated as a **one-port network.** This is generally accomplished by mounting the semiconductor in a pin-type package. The package can then be positioned in an RF circuit or resonant cavity and the top and bottom pins on the package used as the dc and RF electrical contacts. This arrangement works quite well and packaged devices can be operated up to about 90–100 GHz. For higher-frequency operation, the devices are generally mounted directly into circuits using microstrip or some other similar technology.

All three devices operate as negative immittance components. That is, their active characteristics can be described as either a negative resistance or a negative conductance. Which description to use is determined by the physical operating principles of the particular device.

Tunnel Diodes

Tunnel diodes [Sze, 1981] generate active characteristics by a mechanism involving the physical tunneling of electrons between energy bands in highly doped semiconductors. For example, if a pn junction diode is heavily doped, the conduction and valence bands will be located in close proximity and **charge carriers** can tunnel through the electrostatic barrier separating the p-type and n-type regions, rather than be thermionically emitted over the barrier as generally occurs in this type of diode. When the diode is biased (either forward or reverse bias) current immediately flows and junction conduction is basically ohmic. In the forward bias direction, conduction occurs until the applied bias forces the conduction and valence bands to separate. The tunnel current then decreases and normal junction conduction occurs. In the forward bias region where the tunnel current is decreasing with increasing bias voltage, a negative immittance characteristic is generated. The immittance is called "N-type" because the I-V characteristic "looks like" the letter N. This type of active element is short-circuit stable and is described by a negative conductance in shunt with a capacitance. Tunnel diodes are limited in operation frequency by the time it takes for charge carriers to tunnel through the junction. Since this time is very short (on the order of 10^{-12} s) operation frequency can be very high, approaching 1000 GHz. Tunnel diodes have been operated at hundreds of gigahertz, limited by practical packaging and parasitic impedance considerations. The RF power available from a tunnel diode is limited (hundreds of milliwatts level) since the maximum RF voltage swing that can be applied across the junction is limited by the forward turn-on characteristics of the device (typically 0.6–0.9 V). Increased RF power can only be obtained by increasing device area to increase RF current, but device area is limited by operation frequency according to an inverse scaling law. Tunnel diodes have moderate dc-to-RF conversion efficiency (<10%), very low **noise figures,** and are useful in low-noise systems applications, such as microwave and millimeter-wave receivers.

Transferred Electron Devices

Transferred electron devices (i.e., Gunn diodes) [Bosch and Engelmann, 1975] also have N-type active characteristics and can be modeled as a negative conductance in parallel with a capacitance. Device operation, however, is based upon a fundamentally different principle. The negative conductance derives from the complex conduction band structure of certain compound semiconductors, such as GaAs and InP. In these direct bandgap materials the central (or Γ) conduction band is in close energy-momentum proximity to secondary, higher-order conduction bands (i.e., the X and L valleys). The electron effective mass is determined by the shape of the conduction bands, and the effective mass is "light" in the Γ valley but "heavy" in the higher-order X and L valleys. When the crystal is biased, current flow is initially due to electrons in the light effective mass Γ valley and conduction is ohmic. However, as the bias field is increased, an increasing proportion of the free electrons are transferred into the X and L valleys where the electrons have heavier effective mass. The increased effective mass slows down the electrons, with a corresponding decrease in conduction current through the crystal. The net result is that the crystal displays a region of applied bias voltages where current decreases with increasing voltage. That is, a negative conductance is generated. The device is unstable and, when placed in an RF circuit

or resonant cavity, oscillators or amplifiers can be fabricated. The device is not actually a diode since no pn or Schottky junction is used. The phenomenon is a characteristic of the bulk material and the special structure of the conduction bands in certain compound semiconductors. Most semiconductors do not have the conduction band structure necessary for the transferred electron effect. The term *Gunn diode* is actually a misnomer since the device is not a diode. TEDs are widely used in oscillators from the microwave through high millimeter-wave frequency bands. They have good RF output power capability (milliwatts to watts level), moderate efficiency (<20%), and excellent noise and bandwidth capability. Octave band tunable oscillators are easily fabricated using devices such as YIG (yttrium iron garnet) resonators or varactors as the tuning element. Most commercially available solid-state sources for 60- to 100-GHz operation generally use InP TEDs.

IMPATT Diodes

IMPATT (impact avalanche transit time) diodes [Bhartia and Bahl, 1984] are fabricated from pn or Schottky junctions. A typical pn junction device is shown in Fig. 39.12. For optimum RF performance the diode is separated, by use of specially designed layers of controlled impurity doping, into avalanche and drift regions. In operation the diode is reverse biased into avalanche breakdown. Due to the very sensitive I-V characteristic, it is best to bias the diode using a constant current source in which the magnitude of the current is limited. When the diode is placed in a microwave resonant circuit, RF voltage fluctuations in the bias circuit grow and are forced into a narrow frequency range by the impedance characteristics of the resonant circuit. Due to the avalanche process the RF current across the avalanche region lags the RF voltage by 90 degrees. This inductive delay is not sufficient, by itself, to produce active characteristics. However, when the 90 degrees phase shift is added to that arising from an additional inductive delay caused by the transit time of the carriers drifting through the remainder of the diode external to the avalanche region, a phase shift between the RF voltage and current greater than 90 degrees is obtained. A Fourier analysis of the resulting waveforms reveals a device impedance with a negative real part. That is, the device is active and can be used to generate or amplify RF signals. The device impedance has an "S-type" active characteristic and the device equivalent circuit consists of a negative resistance in series with an inductor. The device has significant pn junction capacitance that must be considered, and a complete equivalent circuit would include the device capacitance in parallel with the series negative resistance-inductance elements. For optimum performance the drift region is designed so that the electric field throughout the RF cycle is sufficiently high to produce velocity saturation for the charge carriers. In order to achieve this, it is common to design complex structures consisting of alternating layers of highly doped and lightly doped semiconductor regions. These structures are called "high-low," "low-high-low," or "Read" diodes, after the man who first proposed their use. They can also be fabricated in a back-to-back arrangement to form double-drift structures. These devices are particularly attractive for millimeter-wave applications. IMPATT diodes can be fabricated from most semiconductors, but are generally fabricated from Si or GaAs. The devices are capable of good RF output power (mW to W) and good dc-to-RF conversion efficiency (~10–20%). They operate well into the millimeter-wave region and have been operated as high as 340 GHz. They have moderate bandwidth capability, but have relatively poor noise performance due to the impact ionization process.

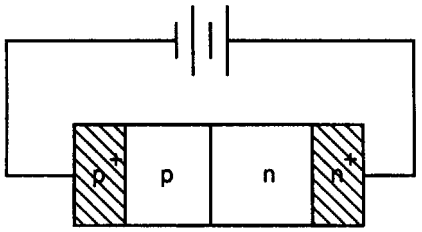


FIGURE 39.12 Diagram showing the structure for a typical pn junction IMPATT diode. This particular diode is called a double-drift device because avalanche breakdown occurs at the pn junction, which is located in the middle of the device. When operated in breakdown, electrons would travel through the n-type region towards the positive terminal of the bias source and holes would travel through the p-type region towards the negative terminal of the source. The diode, therefore, operates as two diodes connected in a back-to-back configuration. The frequency capability of the device is directly proportional to the width of the n and p regions.

Although the two-terminal active devices are used in many electronic systems, the one-port characteristic can introduce significant complexity into circuit design. Isolators and circulators are generally required, and these components are often large and bulky. They are often fabricated from magnetic materials, which can introduce thermal sensitivities. For these reasons three-terminal devices have replaced two-terminal devices in many practical applications. Generally, if two-terminal and three-terminal devices with comparable capability

are available, the three-terminal device offers a more attractive design solution and will be selected. Two-terminal devices are generally only used when a comparable three-terminal device is not available. For this reason IMPATT and TED devices are used in millimeter-wave applications, where they retain an advantage in providing good RF power. Tunnel diodes are not often used, except in a few special applications where their low-noise and wide-bandwidth performance can be used to advantage.

Three-Terminal Active Microwave Devices

The high-frequency performance of three-terminal semiconductor devices has improved dramatically during the past two decades. Twenty years ago transistors that could provide useful **gain** at frequencies above 10 GHz were a laboratory curiosity. Today, such devices are readily available, and state-of-the-art transistors operate well above 100 GHz. This dramatic improvement has been achieved by advances in semiconductor growth technology, coupled with improved device design and fabrication techniques. Semiconductor materials technology continues to improve and new device structures that offer improved high-frequency performance are continually being reported.

In this section we will discuss the two most commonly employed transistors for microwave applications, the metal-semiconductor field-effect transistor (MESFET) [Liechti, 1976] and the bipolar transistor (BJT) [Cooke, 1971]. These two transistors are commonly employed in practical microwave systems as amplifiers, oscillators, and gain blocks. The transistors have replaced many two-terminal devices due to their improved performance and ease of use. Transistors are readily integrated into both hybrid and monolithic integrated circuit environments (MICs). This, in turn, has resulted in significantly reduced size, weight, and dc power consumption, as well as increased reliability and mean time to failure for systems that use these components. Transistors are easily biased and the **two-port network** configuration leads naturally to inherent separation between input and output networks.

Field-Effect Transistors

A cross-sectional view of a microwave MESFET is shown in Fig. 39.13. The device is conceptionally very simple. The MESFET has two ohmic contacts (the source and drain) separated by some distance, usually in the range of 3 to 10 μm. A rectifying Schottky contact (the gate) is located between the two ohmic contacts. Typically, the gate length is on the order of 0.1 to 2 μm for modern microwave devices. The width of the device scales with frequency and typically ranges from about 1 to 10 μm for power microwave devices to 50 μm for millimeter-wave devices. All three contacts are located on the surface of a thin conducting layer (the channel) which is located on top of a high-resistivity, nonconductive substrate to form the device. The channel region is typically very thin (on the order of 0.1–0.3 μm) and is fabricated by epitaxial growth or ion implantation. In operation, the drain contact is biased at a specified potential (positive drain potential for an n-channel device) and the source is grounded. The flow of current through the conducting channel is controlled by negative dc and superimposed RF potentials applied to the gate, which modulate the channel current and provide RF gain. The current flow is composed of only one type of charge carrier (generally electrons) and the device is termed *unipolar*. The MESFET can be fabricated from a variety of semiconductors, but is generally fabricated from GaAs. MESFETs fabricated from Si do not work at high frequencies as well as those fabricated from GaAs due to lower electron mobility in Si (e.g., μ_n ~6000 cm²/V-s for GaAs and 1450 cm²/V-s for Si). The lower electron mobility in Si produces high source resistance, which seriously degrades the high-frequency gain possible from the device.

MESFETs can be optimized for small-signal, low-noise operation or for large-signal, RF power applications. Generally, low-noise operation requires short gate lengths, relatively narrow gate widths, and highly doped channels. Power devices generally have longer gate lengths, much wider gate widths, and lower doped channels. Low-noise devices can be fabricated that operate with good gain (~10 dB) and low noise figure (<3 dB) to above 100 GHz. Power devices can provide RF power levels on the order of watts (W) up to over 20 GHz.

The current gain of the MESFET is indicated by the f_T of the device, sometimes called the gain-bandwidth product. This parameter is defined as the frequency at which the short-circuited current gain is reduced to unity and can be expressed as

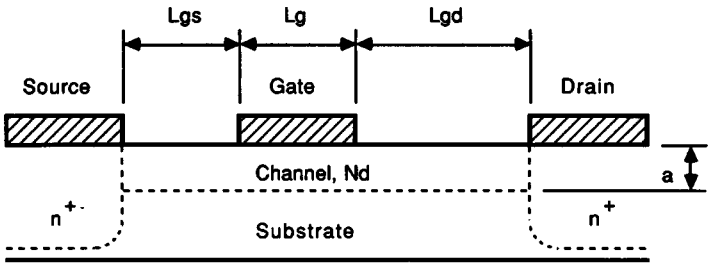


FIGURE 39.13 Cross-sectional view of a microwave MESFET. The cross-hatched areas indicate metal electrodes placed upon the semiconductor to provide for electrical connections. The areas indicated as n^+ are highly doped, highly conducting regions to reduce ohmic access resistances. The channel contains the region of current flow and the substrate is highly resistive and nonconducting so that the current flow is confined to the channel.

$$f_T = \frac{g_m}{2\pi C_{gs}} \tag{39.21}$$

where g_m is the device transconductance (a measure of gain capability) and C_{gs} is the gate source capacitance. High f_T is desirable and this is achieved with highly doped channels and low capacitance gates. The RF power gain is also of interest and this performance can be indicated by the unilateral power gain defined as

$$U = \frac{1}{4}\left(\frac{f_T}{f}\right)^2 \frac{R_{ds}}{R_g} \tag{39.22}$$

where U is the unilateral power gain, f is the operating frequency, R_{ds} is the drain-source resistance, and R_g is the gate resistance. As this expression indicates, large power gain requires a high f_T, and a large R_{ds}/R_g ratio. The highest frequency at which the device could be expected to produce power gain can be defined from the frequency, f, at which U goes to zero. This frequency is called the maximum frequency of oscillation, or $f_{\max}$, and is defined as

$$f_{\max} = \frac{f_T}{2}\sqrt{\frac{R_{ds}}{R_g}} \tag{39.23}$$

A different form of field-effect transistor can be fabricated by inserting a highly doped, wider-bandgap semiconductor between the conducting channel and the gate electrode [Drummond et al., 1986]. The conducting channel is then fabricated from undoped semiconductor. The discontinuity in energy bandgaps between the two semiconductors, if properly designed, results in free charge transfer from the highly doped, wide-bandgap semiconductor into the undoped, lower-bandgap channel semiconductor. The charge accumulates at the interface and creates a two-dimensional electron gas (2DEG). The sheet charge is essentially two-dimensional and allows current to flow between the source and drain electrodes. The amount of charge in the 2DEG can be controlled by the potential applied to the gate electrode. In this manner the current flow through the device can be modulated by the gate and gain results. Since the charge flows at the interface between the two materials, but is confined in the undoped channel semiconductor, very little impurity scattering occurs and extremely high charge carrier mobility results. The device, therefore, has very high transconductance and is capable of very high frequency operation and very low noise figure operation. This type of device is called a high electron mobility transistor (HEMT). HEMTs can be fabricated from material systems such as AlGaAs/GaAs or AlInAs/GaInAs/InP. The latter material system produces devices that have f_T's above 300 GHz and have produced noise figures of about 1 dB at 100 GHz.

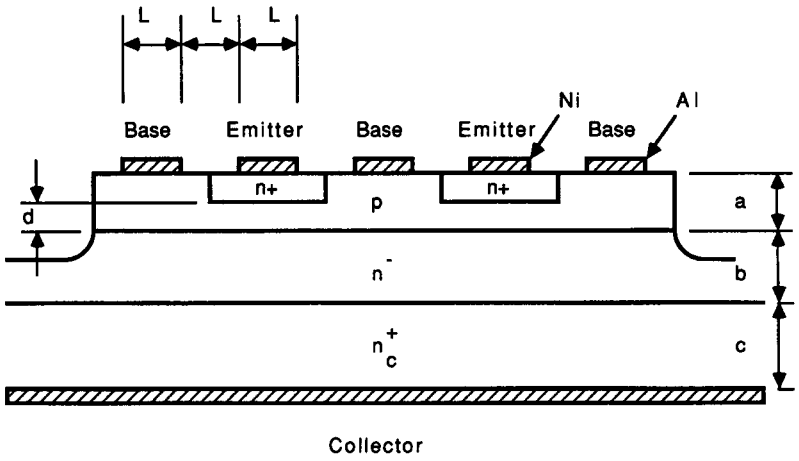


FIGURE 39.14 Cross-sectional view of a microwave bipolar transistor (BJT). The cross-hatched areas indicate metal electrodes. The electrode pattern on the surface is interdigitated with the base electrodes connected together at one end and the emitter electrodes connected together at the other end. Due to the interdigitated structure, there will always be one more base electrode than the number of emitter electrodes. The n^+, p, n^-, and n_c^+ designations indicate the impurity doping type and relative concentration level. This device has the collector electrode on the bottom of the device.

Bipolar Transistors

A cross-sectional view of a bipolar transistor is shown in Fig. 39.14. The bipolar transistor consists of back-to-back pn junctions arranged in a sandwich structure. The three regions are designated the emitter, base, and collector. This type of device differs from the field-effect transistors in that both electrons and holes are involved in the current transport process (thus the designation *bipolar*). Two structures are possible: pnp or npn, depending upon the conductivity type common to both pn junctions. Generally, for microwave applications the npn structure is used since device operation is controlled by electron flow. In general, electron transport is faster than that for holes, and npn transistors are capable of superior high-frequency performance compared to comparable pnp transistors. In operation, the base-emitter pn junction is forward biased and the collector-base pn junction is reverse biased. When an RF signal is applied to the base-emitter junction the junction allows a current to be injected into the base region. The current in the base region consists of minority charge carriers (i.e., carriers with the opposite polarity compared to the base material—electrons for an npn transistor). These charge carriers then diffuse across the base region to the base-collector junction, where they are swept across the junction by the large reverse bias electric field. The reverse bias electric field in the base-collector region is generally made sufficiently large that the carriers travel at their saturation velocity. The transit time of the charge carriers across this region is small, except for millimeter-wave transistors where the base-collector region transit time can be a significant fraction of the total time required for a charge carrier to travel from the emitter through the collector. The operation of the transistor is primarily controlled by the ability of the minority charge carriers to diffuse across the base region. For this reason microwave transistors are designed with narrow base regions in order to minimize the time required for the carriers to travel through this region. The base region transit time is generally the limiting factor in determining the high-frequency capability of the transistor. The gain of the transistor is also significantly affected by minority carrier behavior in the base region. The density of minority carriers is significantly smaller than the density of majority carriers (majority carrier density is approximately equal to the impurity doping density) for typical operating conditions and the probability that the minority charge will recombine with a majority carrier is high. If recombination occurs, the minority charge cannot reach the base-collector junction but appears as base current. This, in turn, reduces the current gain capability of the transistor. Narrow base regions reduce the semiconductor volume where recombination can occur and, therefore, result in increased gain. Modern microwave transistors typically have base regions on the order of 0.1–0.25 μm.

The frequency response of a bipolar transistor can be determined by an analysis of the total time it takes for a charge carrier to travel from the emitter through the collector. The total time can be expressed as

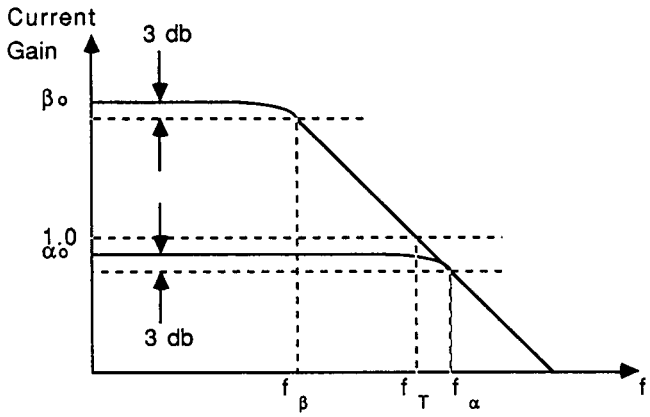


FIGURE 39.15 Current gains versus frequency for bipolar transistors. The common-emitter and common-base current gains are designated as β and α, respectively. The subscript "o" indicates the dc value. The gains decrease with frequency above a certain value. The frequencies where the gains are reduced by 3 dB from their dc values are indicated as the CE and CB cutoff frequencies, f_β and f_α, respectively. The frequency at which the CE current gain is reduced to unity is defined as the gain-bandwidth product f_T. Note that the CB current gain is restricted to values less than unity and that the CE current gain has values that significantly exceed unity.

$$\tau_{ec} = \tau_e + \tau_b + \tau_c + \tau'_c \tag{39.24}$$

where τ_{ec} is the total emitter-collector transit time, τ_e is the base-emitter junction capacitance charging time, τ_b is the base region transit time, τ_c is the base-collector junction capacitance charging time, and τ_c' is the base-collector region transit time. The total emitter-base time is related to the gain-bandwidth capability of the transistor according to the relation

$$f_T = \frac{1}{2\pi\tau_{ec}} \tag{39.25}$$

Since the bipolar transistor has three terminals, it can be operated in various configurations, depending upon the electrode selected as the common terminal. The two most commonly employed are the common emitter (CE) and the common base (CB) configurations, although the common collector (CC) configuration can also be used. Small-signal amplifiers generally use the CE configuration and power amplifiers often use the CB configuration.

The current gain for a bipolar transistor is shown in Fig. 39.15. The current gains of the transistor operated in the CE and CB configurations are called β and α, respectively. As indicated in the figure, the CE current gain β is much larger than the CB current gain α, which is limited to values less than unity. For modern microwave transistors $\alpha_o \sim 0.98$–0.99 and $\beta_o \sim 50$–60.

A measure of the RF power gain for the transistor is indicated by the unilateral power gain, which can be expressed as

$$U \cong \frac{\alpha_o}{16\pi^2 r_b C_c f^2 \left(\tau_{ec} + \dfrac{r_e C_c}{a_o}\right)} \tag{39.26}$$

where U is the power gain, α_o is the dc CB current gain, r_b is the base resistance, C_c is the collector capacitance, τ_{ec} is the total emitter-to-collector transit-time, and r_e is the emitter resistance. The frequency at which U is

Tracking a Microwave-Cavity Resonance by Use of Vibrations

A microwave heating apparatus that comprises a resonant cavity excited by a magnetron has been equipped with an automated tuning system (see figure) to maintain resonance. Resonance is a desirable condition because it maximizes the transfer of power to the material sample that one seeks to heat. Typically, the cavity becomes detuned from resonance during heating of the sample because the permittivity of the sample changes with temperature, altering the electromagnetic field in the cavity. A system like this that automatically compensates for the detuning effect can enhance efficiency and productivity in microwave processing of materials.

This tuning system, like others, is based on the old-fashioned radio tuning principle, in which one brackets a resonance by manual back-and-forth actuation of a tuning device while seeking an optimum quantitative or qualitative measure of the signal. In both traditional manual tuning and in other automated resonance-tracking systems, the tuning adjustments usually vary the frequencies, but in this case, one does not have the option of frequency tuning because the frequency of oscillation of the magnetron is not adjustable and is nominally constant.

The back-and-forth tuning action needed to locate the resonance is provided by an auxiliary tuning device, which includes a hollow metal rod that is mounted on a loudspeaker outside the cavity and that protrudes into the cavity, preferably at a position of maximum electric field. An audio oscillator drives the loudspeaker at a convenient frequency between 30 and 100 Hz, causing the rod to vibrate and thereby impose a slight modulation on the microwave field in the cavity. A diode across the cavity from the vibrating rod detects the amplitude modulation on the electric field.

Both the signal from the audio oscillator and the amplified output of the diode are fed to the mixer at saturating amplitudes, so the low-pass filtered output of the mixer depends only on the difference between the phases of these two signals. This phase difference is a measure of the deviation from resonance. Thus, the output of the mixer constitutes an error signal that indicates the adjustment needed to restore the cavity to resonance. This signal is fed to the motor that drives the plunger to obtain the required tuning adjustment.

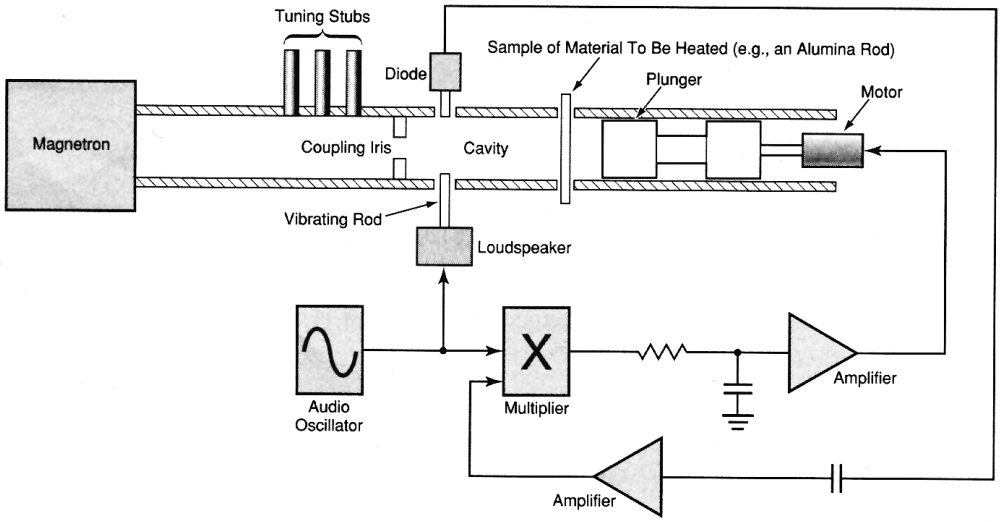
reduced to unity (f_{max}) is the maximum frequency at which the device will have active characteristics. This frequency is

$$f_{max} = \left[\frac{f_T}{8\pi r_b C_c}\right]^{1/2} \tag{39.27}$$

In order to maximize the high-frequency performance of a transistor, it is necessary to design the device so that it has high current gain (f_T), low base resistance (r_b), and low collector capacitance (C_c).

Bipolar transistors operating to about 20 GHz are generally fabricated from Si. These devices are easily fabricated and low cost. They are useful in moderate gain and low to high RF power applications. The have relatively high noise figure that varies from about 1 dB at 1 GHz to about 4–5 dB at 10 GHz.

An improved high-frequency bipolar transistor can be fabricated using heterostructures of compound semiconductors, such as AlGaAs/GaAs [Kroemer, 1982]. These devices have their emitters fabricated from a wide-bandgap semiconductor (such as AlGaAs) and the remainder of the device fabricated from the lower-bandgap semiconductor (GaAs). The wide-bandgap emitter results in improved charge injection efficiency across the base-emitter junction into the base region and much improved RF performance. While the operation of standard



The vibrating rod modulates the electric field in the cavity. The phase difference between the vibrations and the modulation provides an indication of deviation from resonance and is used to control repositioning of the plunger to maintain resonance. (Reprinted with permission from *NASA Tech Briefs,* 20(10), 54, 1996.)

The effectiveness of the automated tuning system was demonstrated in experiments in which the microwave apparatus was used to heat an alumina rod. The apparatus was operated with a forward microwave power of 200 W in two cases; one with and one without automatic tuning. Without automatic tuning, the rod attained an asymptotic temperature of about 600°C in 5 min. With automatic tuning, the temperature of the rod exceeded 900°C (and was still increasing) in less than 2 min.

This work was done by Martin Barmatz and Ofer Iny of Caltech for NASA's Jet Propulsion Laboratory. (Reprinted with permission from *NASA Tech Briefs,* 20(10), 54, 1996.)

Si bipolar transistors is limited to frequencies less than about 40 GHz, the heterojunction bipolar transistors (HBTs) can operate in excess of 100 GHz. They are useful in both low-noise and high RF power applications. The heterostructure concept has recently been applied in Si-based devices using heterostructures using SiGe/Si compounds. These devices show consider promise for high-frequency applications and the transistors have demonstrated RF performance comparable to that obtained from the AlGaAs/GaAs HBTs.

Comparison of Bipolar Transistor and MESFET Noise Figures

In low-noise applications, GaAs MESFETs are generally preferred to Si bipolar transistors. The MESFET demonstrates a lower noise figure than the bipolar transistor throughout the microwave frequency range, and the advantage increases with frequency. This advantage is demonstrated by a comparison of the expressions for the minimum noise figure for the two devices. The bipolar transistor has a minimum noise figure that can be expressed as

$$F_{\min} \cong 1 + bf^2\left[1 + \sqrt{1 + \frac{2}{bf^2}}\right] \tag{39.28}$$

where $F_{\min}$ is the noise figure and

$$b = \frac{40 I_c r_b}{f_T^2} \tag{39.29}$$

where I_c is the collector current and the other terms are as previously defined. The minimum noise figure for the MESFET is

$$F_{\min} \cong 1 + mf \tag{39.30}$$

where

$$m = \frac{2.5}{f_T} \sqrt{g_m (R_g + R_s)} \tag{39.31}$$

where g_m is the MESFET **transconductance**, R_g is the gate resistance, and R_s is the source resistance.

Comparing these expressions shows that the minimum noise figure increases with frequency quadratically for bipolar transistors and linearly for MESFETs. Therefore, as operating frequency increases, the MESFET demonstrates increasingly superior noise figure performance as compared to Si bipolar transistors.

Conclusions

Various active solid-state devices that are useful at microwave and millimeter-wave frequencies have been discussed. Both two-terminal and three-terminal devices were included. The most commonly used two-terminal devices are tunnel diodes, transferred-electron devices, and IMPATT diodes. Three-terminal devices consist of various forms of field-effect transistors and bipolar transistors. Recent advances employ heterostructures using combinations of different semiconductors to produce devices with improved RF performance, especially for high-frequency applications. Both two-terminal and three-terminal devices can provide useful gain at frequencies in excess of 100 GHz. Further improvements are likely as fabrication technology continues to improve.

Defining Terms

Active device: A device that can convert energy from a dc bias source to a signal at an RF frequency. Active devices are required in oscillators and amplifiers.

Charge carriers: Units of electrical charge that when moving produce current flow. In a semiconductor two types of charge carriers exist: electrons and holes. Electrons carry unit negative charge and have an effective mass that is determined by the shape of the conduction band in energy-momentum space. The effective mass of an electron in a semiconductor is generally significantly less than an electron in free space. Holes have unit positive charge. Holes have an effective mass that is determined by the shape of the valence band in energy-momentum space. The effective mass of a hole is generally significantly larger than that for an electron. For this reason electrons generally move much faster than holes when an electric field is applied to the semiconductor.

Gain: A measure of the ability of a network to increase the energy level of a signal. Gain is generally measured in decibels. For voltage or current gain: G (dB) = 20 log($S_{\text{out}}/S_{\text{in}}$), where S is the RF voltage or current out of and into the network. For power gain G (dB) = 10 log($P_{\text{out}}/P_{\text{in}}$). If the network has net loss, the gain will be negative.

Noise figure: A measure of the noise added by a network to an RF signal passing through it. Noise figure can be defined in terms of signal-to-noise ratios at the input and output ports of the network. Noise figure is generally measured in decibels and can be defined as F (dB) = 10 log[$(S/N)_{\text{in}}/(S/N)_{\text{out}}$].

One-port network: An electrical network that has only one RF port. This port must be used as both the input and output to the network. Two-terminal devices result in one-port networks.

Three-terminal device: An electronic device that has three contacts, such as a transistor.

Transconductance: A measure of the gain capability of a transistor. It is defined as the change in output current as a function of a change in input voltage.

Two-port network: An electrical network that has separate RF ports for the input and output. Three-terminal devices can be configured into two-port networks.

Two-terminal device: An electronic device, such as a diode, that has two contacts. The contacts are usually termed the cathode and anode.

Related Topic

37.2 Waveguides

References

P.B. Bhartia and I.J. Bahl, *Millimeter Wave Engineering and Applications*, New York: Wiley-Interscience, 1984.

B.G. Bosch and R.W. Engelmann, *Gunn-Effect Electronics*, New York: Halsted Press, 1975.

H.F. Cooke, "Microwave transistors: Theory and design," *Proc. IEEE*, vol. 59, pp. 1163–1181, Aug. 1971.

T.J. Drummond, W.T. Masselink, and H. Morkoc, "Modulation-doped GaAs/AlGaAs heterojunction field-effect transistors: MODFET's," *Proc. IEEE*, vol. 74, pp. 773–822, June 1986.

H. Kroemer, "Heterostructure bipolar transistors and integrated circuits," *Proc. IEEE*, vol. 70, pp. 13–25, Jan. 1982.

C.A. Liechti, "Microwave field-effect transistors—1976," *IEEE Trans. Microwave Theory and Tech.*, vol. MTT-24, pp. 128–149, June 1976.

S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley-Interscience, 1981.

Further Information

Additional details on the various devices discussed in this chapter can be found in the following books:

I. Bahl and P. Bhartia, *Microwave Solid State Circuit Design*, New York: Wiley-Interscience, 1988.

M. Shur, *Physics of Semiconductor Devices*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

S.M. Sze, *High-Speed Semiconductor Devices*, New York: Wiley-Interscience, 1990.

S. Tiwari, *Compound Semiconductor Device Physics*, San Diego: Academic Press, 1992.

S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Englewood Cliffs, N.J.: Prentice-Hall, 1989.

Hemming, L.H., Ungvichian, V., Roman, J.M., Uman, M.A., Rubinstein, M.
"Compatibility"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

40

Compatibility

Leland H. Hemming
McDonnell Douglas Helicopter Systems

Vichate Ungvichian
Florida Atlantic University

John M. Roman
Telematics

Martin A. Uman
University of Florida, Gainesville

Marcos Rubinstein
Swiss PTT

40.1 Grounding, Shielding, and Filtering
Grounding • Shielding • Filtering

40.2 Spectrum, Specifications, and Measurement Techniques
Electromagnetic Spectrum • Specifications • Measurement Procedures

40.3 Lightning
Terminology and Physics • Lightning Occurrence Statistics • Electric and Magnetic Fields • Modeling of the Return Stroke • Lightning-Overhead Wire Interactions

40.1 Grounding, Shielding, and Filtering

Leland H. Hemming

Electromagnetic interference (EMI) is defined to exist when undesirable voltages or currents are present to influence adversely the performance of an electronic circuit or system. Interference can be within the system (intrasystem), or it can be between systems (intersystem). The system is the equipment or circuit over which one exercises design or management control.

The cause of an EMI problem is an unplanned coupling between a source and a receptor by means of a transmission path. Transmission paths may be conducted or radiated. Conducted interference occurs by means of metallic paths. Radiated interference occurs by means of near- and far- field coupling. These different paths are illustrated in Fig. 40.1.

The control of EMI is best achieved by applying good interference control principles during the design process. These involve the selection of signal levels, impedance levels, frequencies, and circuit configurations that minimize conducted and radiated interference. In addition, signal levels should be selected to be as low as possible, while being consistent with the required signal-to-noise ratio. Impedance levels should be chosen to minimize undesirable capacitive and inductive coupling.

The frequency spectral content should be designed for the specific needs of the circuit, minimizing interference by constraining signals to desired paths, eliminating undesired paths, and separating signals from interference. Interference control is also achieved by physically separating leads carrying currents from different sources.

For optimum control, the three major methods of EMI suppression—grounding, shielding, and filtering—should be incorporated early in the design process. The control of EMI is first achieved by proper grounding, then by good shielding design, and finally by filtering.

Grounding is the process of electrically establishing a low impedance path between two or more points in a system. An ideal ground plane is a zero potential, zero impedance body that can be used as reference for all signals in the system. Associated with grounding is *bonding*, which is the establishment of a low impedance path between two metal surfaces.

Shielding is the process of confining radiated energy to the bounds of a specific volume or preventing radiated energy from reaching a specific volume. *Filtering* is the process of eliminating conducted interference by controlling the spectral content of the conducted path. Filtering is the last step in the EMI design process.

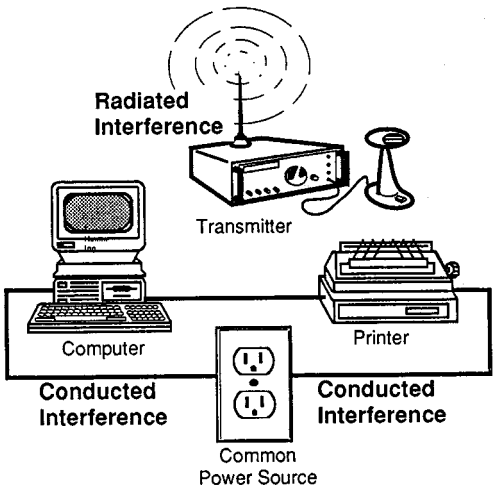


FIGURE 40.1 Electromagnetic interference is caused by uncontrolled conductive paths and radiated near/far fields.

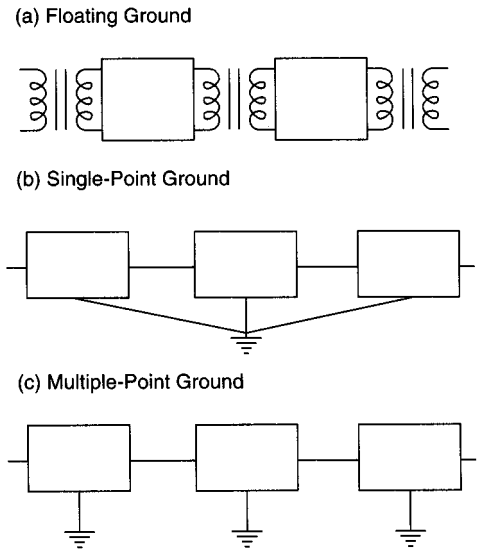


FIGURE 40.2 The type of ground system used must be selected carefully.

Grounding

Grounding Principles

The three fundamental grounding techniques—floating, single-point, and multiple-point—are illustrated in Fig. 40.2.

Floating grounds are used to isolate circuits or equipment from a common ground plane. Static charges are a hazard with this type of ground. Dangerous voltages may develop or a noise-producing discharge might occur. Generally, bleeder resistors are used to control the static problem. Floating grounds are useful only at low frequencies where capacitive coupling paths are negligible.

The single-point ground is a single physical point in a circuit. By connecting all grounds to a common point, no interference will be produced in the equipment because the configuration does not result in potential differences across the equipment. At high frequencies care must be taken to prevent capacitive coupling, which will result in interference.

A multipoint ground system exists when each ground connection is made directly to the ground plane at the closest available point on it, thus minimizing ground lead lengths. A large conductive body is chosen for the ground. Care must be taken to avoid ground loops.

Circuit grounding design is dependent on the function of each type of circuit. In unbalanced systems, care must be taken to reduce the potential of common mode noise. Differential devices are commonly used to suppress this form of noise. The use of high circuit impedances should be minimized. Where it cannot be avoided, all interconnecting leads should be shielded, with the shield well grounded. Power supply grounding must be done properly to minimize load inducted noise on a power supply bus. When electromechanical relays are used in a system, it is best that they be provided with their own power supplies.

Cable shield grounding must be designed based upon the frequency range, impedance levels, (whether balanced or unbalanced) and operating voltage and/or current. Cross talk between cables is a major problem and must be carefully considered during the design process.

Building facility grounds must be provided for electrical faults, signal, and lightning. The fault protection (green wire) subsystem is for the protection of personnel and equipment from the hazards of electrical power faults and static charge buildup. The lightning protection system consists of air terminals (lightning rods), heavy duty down-conductors, and ground rods. The **signal reference subsystem** provides a ground for signal circuits to control static charges and noise and to establish a common reference between signals and loads.

Earth grounds may consist of vertical rods, horizontal grids or radials, plates, or incidental electrodes such as utility pipes or buried tanks. The latter must be constructed and tested to meet the design requirements of the facility.

Grounding Design Guidelines

The following design guidelines represent good practice but should be applied subject to the detailed design objectives of the system.

Fundamental Concepts

- Use single-point grounding for circuit dimensions less than 0.03 λ (wavelength) and multipoint grounding for dimensions greater than 0.15 λ.
- The type of grounding for circuit dimensions between 0.03 and 0.15 λ depends on the physical arrangement of the ground leads as well as the conducted emission and conducted susceptibility limits of the circuits to be grounded. Hybrid grounds may be needed for circuits that must handle a broad portion of the frequency spectrum.
- Apply floating ground isolation techniques (i.e., transformers) if ground loop problems occur.
- Keep all ground leads as short as possible.
- Design ground reference planes so that they have high electrical conductivity and can be maintained easily to retain good conductivity.

Safety Considerations

- Connect test equipment grounds directly to the grounds of the equipment being tested.
- Make certain the ground connections can handle fault currents that might flow unexpectedly.

Circuit Grounding

- Maintain separate circuit ground systems for signal returns, signal shield returns, power system returns, and chassis or case grounds. These returns then can be tied together at a single ground reference point.
- For circuits that produce large, abrupt current variations, provide a separate grounding system, or provide a separate return lead to the ground to reduce transient coupling into other circuits.
- Isolate the grounds of low-level circuits from all other grounds.
- Where signal and power leads must cross, make the crossing so that the wires are perpendicular to each other.
- Use balanced differential circuitry to minimize the effects of ground circuit interference.
- For circuits whose maximum dimension is significantly less than λ/4, use tightly twisted wires (either shielded or unshielded, depending on the application) that are single-point grounded to minimize equipment susceptibility.

Cable Grounding

- Avoid pigtails when terminating cable shields.
- When coaxial cable is needed for signal transmission, use the shield as the signal return and ground at the generator end for low-frequency circuits. Use multipoint grounding of the shield for high-frequency circuits.
- Provide multiple shields for low-level transmission lines. Single-point grounding of each shield is recommended.

Shielding

The control of near- and far-field coupling (radiation) is accomplished using shielding techniques. The first step in the design of a shield is to determine what undesired field level may exist at a point with no shielding and what the tolerable field level is. The difference between the two then is the needed **shielding effectiveness.**

This section discusses the shielding effectiveness of various solid and nonsolid materials and their application to various shielding situations. **Penetrations** and their design are discussed so that the required shielding effectiveness is maintained. Finally, common shielding effectiveness testing methods are reviewed.

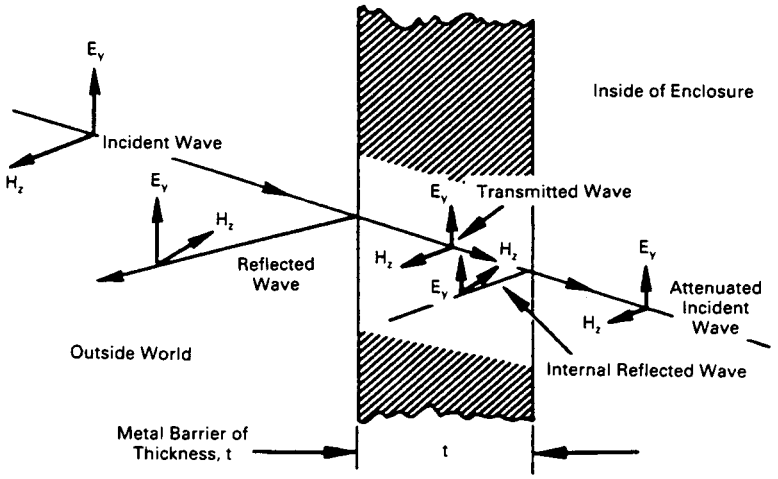


FIGURE 40.3 Shielding effectiveness is the result of three loss mechanisms.

Enclosure Theory

The attenuation provided by a shield results from three loss mechanisms as illustrated in Fig. 40.3.

1. Incident energy is reflected (R) by the surface of the shield because of the impedance discontinuity of the air–metal boundary. This mechanism does not require a particular material thickness but simply an impedance discontinuity.
2. Energy that does cross the boundary (not reflected) is attenuated (A) in passing through the shield.
3. The energy that reaches the opposite face of the shield encounters another air–metal boundary and thus some of it is reflected (B) back into the shield. This term is only significant when $A < 15$ dB and is generally neglected because the barrier thickness is generally great enough to exceed the 15-dB loss rule of thumb.

Thus:

$$S = R + A + B \text{ dB} \tag{40.1}$$

Absorption loss is independent of the type of wave (electric/magnetic) and is given by

$$A = 1.314(f\mu_r\sigma_r)^{1/2}d \text{ dB} \tag{40.2}$$

where d is shield thickness in centimeters, μ_r is relative permeability, f is frequency in Hz, and σ_r is conductivity of metal relative to that of copper. Typical absorption loss is provided in Table 40.1.

Reflection loss is a function of the intrinsic impedance of the metal boundary with respect to the wave impedance, and therefore, three conditions exist: near-field magnetic, near-field electric, and plane wave.

The relationship for low-impedance (magnetic field) source is

$$R = 20 \log_{10}\{[1.173(\mu_r/f\sigma_r)^{1/2}/D] + 0.0535\ D(f\sigma_r/\mu_r)^{1/2} + 0.354 \text{ dB}\} \tag{40.3}$$

where D is distance to source in meters. For a plane wave source the reflection loss is

$$R = 168 - 10 \log_{10}(f\mu_r/\sigma_r) \text{ dB} \tag{40.4}$$

For a high-impedance (electric field) source the reflection loss R is

$$R = 362 - 20 \log_{10}[(\mu_r f^3/\sigma_r)^{1/2} D] \text{ dB} \tag{40.5}$$

TABLE 40.1 Absorption Loss Is a Function of Type of Material and Frequency (Loss Shown is at 150 kHz)

Metal	Relative Conductivity	Relative Permeability	Absorption Loss *A*, dB/mm
Silver	1.05	1	52
Copper—annealed	1.00	1	51
Copper—hard drawn	0.97	1	50
Gold	0.70	1	42
Aluminum	0.61	1	40
Magnesium	0.38	1	31
Zinc	0.29	1	28
Brass	0.26	1	26
Cadmium	0.23	1	24
Nickel	0.20	1	23
Phosphor–bronze	0.18	1	22
Iron	0.17	1000	650
Tin	0.15	1	20
Steel, SAE1045	0.10	1000	500
Beryllium	0.10	1	16
Lead	0.08	1	14
Hypernik	0.06	80000	3500[a]
Monel	0.04	1	10
Mu-metal	0.03	80000	2500[a]
Permalloy	0.03	80000	2500[a]
Steel, stainless	0.02	1000	220[a]

[a]Assuming that material is not saturated.
Source: MIL-HB-419A.

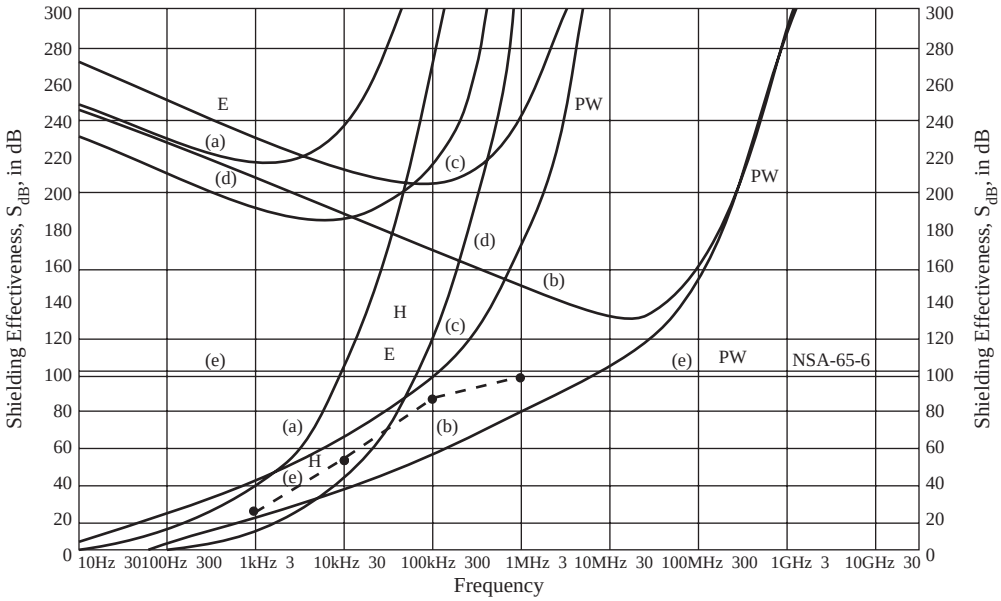


FIGURE 40.4 The shielding effectiveness of common sheet metals, 1 m separation. (a) 26-gage steel; (b) 3-oz. copper foil; (c) 0.030-in. aluminum sheet; (d) 0.003-in. Permalloy; (e) is a common specification for shielded enclosures.

Figure 40.4 illustrates the shielding effectiveness of a variety of common materials versus various thicknesses for a source distance of 1 m. This is the shielding effectiveness of a six-sided enclosure. To be useful, the enclosure must be penetrated for various services or devices. This is illustrated in Fig. 40.5(a) for small enclosures and Fig. 40.5(b) for room-sized enclosures.

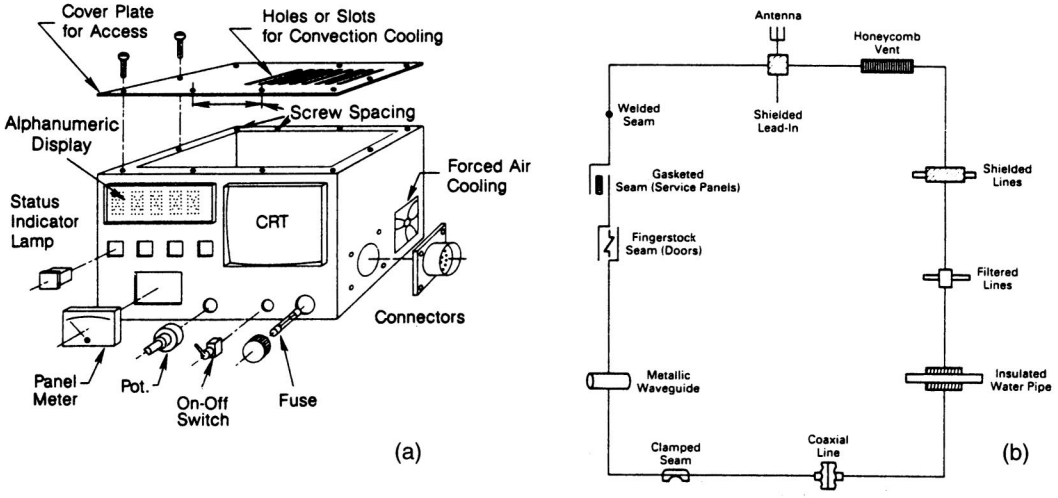


FIGURE 40.5 Penetrations in small (a) and large (b) enclosures.

Shielding Penetrations

Total shielding effectiveness of an enclosure is a function of the basic shield and all of the leakages associated with the penetrations in the enclosure. The latter includes seams, doors, vents, control shafts, piping, filters, windows, screens, and fasteners.

The design of the seams is a function of the type of enclosure and the level and nature of the shielding effectiveness required. For small instruments, computers, and similar equipment, the typical shielding required is on the order of 60 dB for electric and plane wave shielding. EMI gaskets are commonly used to seal the openings in sheet metal construction. In some high-performance applications the shielding is achieved using very tight-fitting machined housings. Examples are IF strips and large dynamic range log amplifier circuits. Various methods of sealing joints are illustrated in Fig. 40.6. EMI gasketing methods are shown in Fig. 40.7. For large room-sized enclosures, the performance requirements typically range from 60 to 120 dB. Conductive EMI shielding tape is used in the 60-dB realm, clamped seams for 80–100 dB, and continuous welded seams for 120-dB performance. These are illustrated in Fig. 40.8.

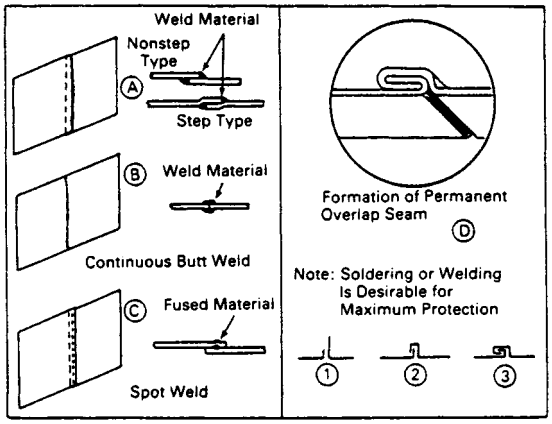


FIGURE 40.6 Methods of sealing enclosure seams.

A good electromagnetic shielded door design must meet a variety of physical and electrical requirements. Figure 40.9 illustrates a number of ways this is accomplished.

For electronic equipment, a variety of penetrations must be made to make the shielded volume functional. These include control shafts, windows, lights, filters, and displays. Careful design is required to maintain the required shielding integrity.

Shield Testing

The most common specification used for shield evaluation is the procedure given in MIL-STD-285. This consists of establishing a reference level without the shield and then enclosing the receiver within the shield and determining the difference. The ratio is the shielding effectiveness. This applies regardless of materials used in the construction of the shield. Care must be taken in evaluating the results since the measured value is a function of a variety of factors, not all of which are definable.

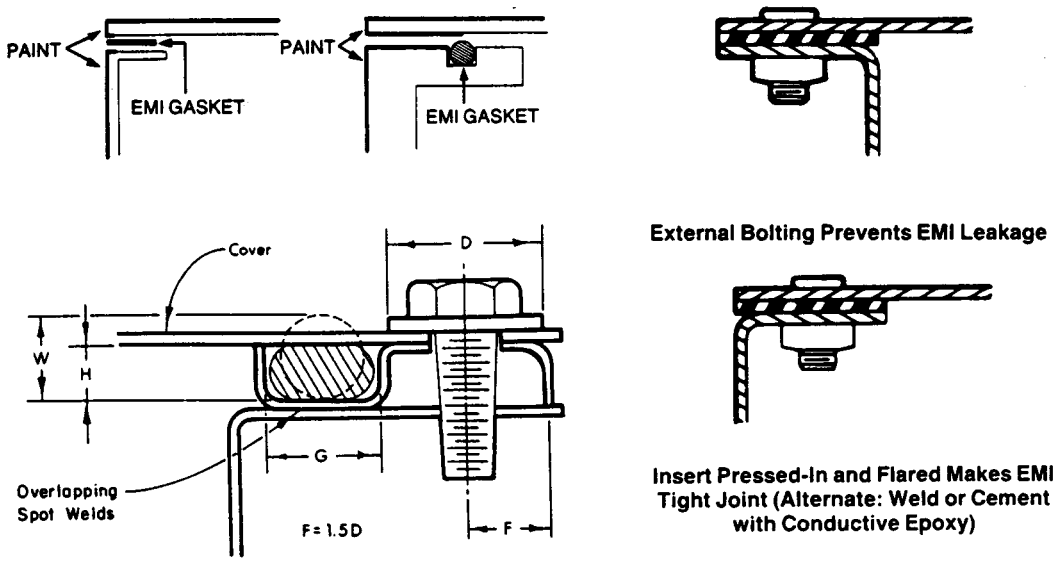


FIGURE 40.7 Methods of constructing gasketed joints.

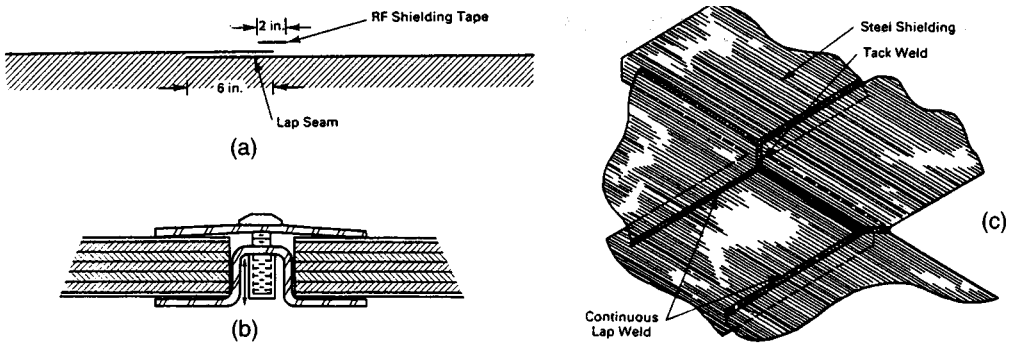


FIGURE 40.8 Most common seams in large enclosures. (a) Foil and shielding tape; (b) clamped; (c) welded.

Summary of Good Shielding Practice

Shielding Effectiveness

- Good conductors, such as copper and aluminum, should be used for electric field shields to obtain high reflection loss. A shielding material thick enough to support itself usually provides good electric shielding at all frequencies.
- Magnetic materials, such as iron and special high-permeability alloys, should be used for magnetic field shields to obtain high absorbtion loss.
- In the plane wave region, the sealing of all apertures is critical to good shielding practice.

Multiple Shields

- Multiple shields are quite useful where high degrees of shielding effectiveness are required.

Shield Seams

- All openings or discontinuities should be addressed in the design process to ensure achievement of the required shielding effectiveness. Shield material should be selected not only from a shielding requirement, but also from electrochemical corrosion and strength considerations.
- Whenever system design permits, use continuously overlapping welded seams. Obtain intimate contact between mating surfaces over as much of the seam as possible.

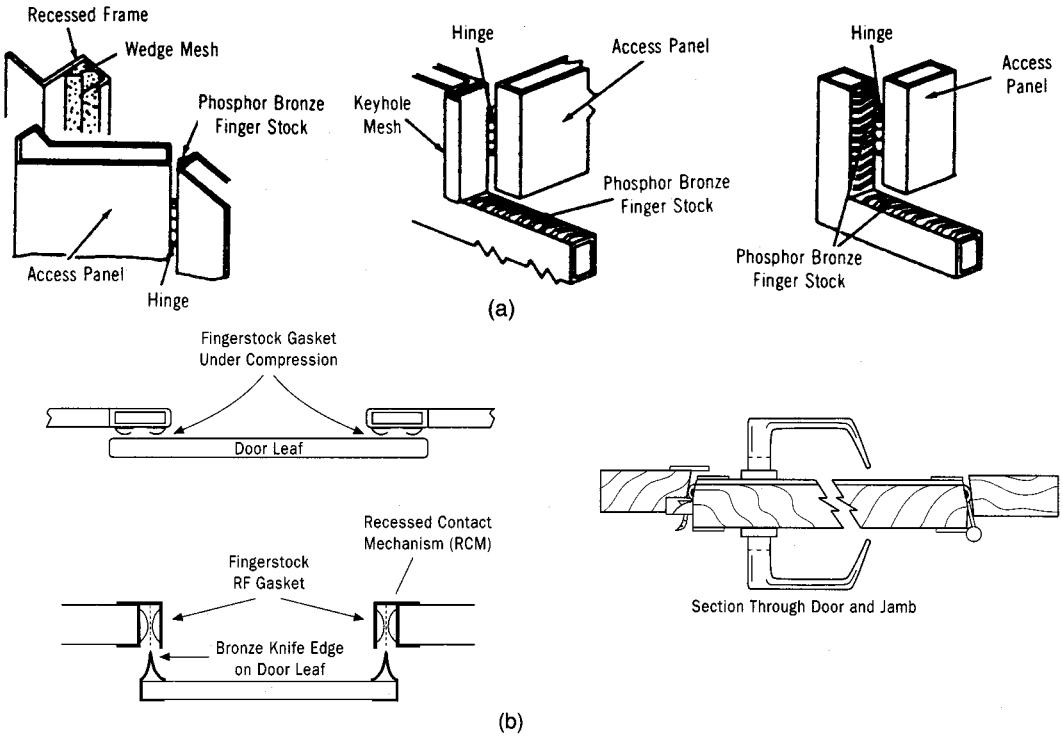


FIGURE 40.9 Methods of sealing seams in RF enclosure small (a) and large (b) doors.

- Surfaces to be mated must be clean and free from nonconducting finishes, unless the bonding process positively and effectively cuts through the finish. When electromagnetic compatibility (**EMC**) and finish specifications conflict, the finishing requirements must be modified.

Case Construction

- Case material should have good shielding properties.
- Seams should be welded or overlapped.
- Panels and cover plates should be attached using conductive gasket material with closely spaced fasteners.
- Mating surfaces should be cleaned just before assembly to ensure good electrical contact and to minimize corrosion.
- A variety of special devices are available for sealing around doors, vents, and windows.
- Internal interference generating circuits must be isolated both electrically and physically. Electrical isolation is achieved by circuit design; physical isolation may be achieved by proper shielding.
- For components external to the case, use EMI boots on toggle switches, EMI rotary shaft seals on rotary shafts, and screening and shielding on meters and other indicator faces.

Cable Shields

- Cabling that penetrates a case should be shielded and the shield should be terminated in a peripheral bond at the point of entry. This peripheral bond should be made to the connector or adaptor shell.

Filtering

An electrical filter is a combination of lumped or distributed circuit elements arranged so that it has a frequency characteristic that passes some frequencies and blocks others.

Filters provide an effective means for the reduction and suppression of electromagnetic interference as they control the spectral content of signal paths. The application of filtering requires careful consideration of an extensive list of factors including insertion loss, impedance, power handling capability, signal distortion, tunability, cost, weight, size, and rejection of undesired signals. Often they are used as stopgap measures, but if suppression techniques are used early in the design process, then the complexity and cost of interference fixes can be minimized. There are many textbooks on filtering, which should be used for specific applications.

The types of filters are classified according to the band of frequencies to be transmitted or attenuated. The basic types illustrated in Fig. 40.10 include low-pass, high-pass, bandpass, and bandstop (reject).

Filters can be composed of lumped, distributed, or dissipative elements; the type used is mainly a function of frequency.

Filtering Guidance

- It is best to filter at the interference source.
- Suppress all spurious signals.
- Design nonsusceptible circuits.
- Ensure that all filter elements interface properly with other EMC elements, i.e., proper mounting of a filter in a shielded enclosure.

Filter Design

Filters using lumped and distributive elements generally are reflective, in that the various component combinations are designed for high series impedance and low shunt impedance in the stopband while providing low series impedance and high shunt impedance in the passband.

The impedance mismatches associated with the use of reflective filters can result in an increase of interference. In such cases, the use of dissipative elements is found to be useful. A broad range of ferrite components are available in the form of beads, tubes, connector shells, and pins. A very effective method of low-pass filtering is to form the ferrite into a coaxial geometry, the properties of which are proportional to the length of the ferrite, as shown in Fig. 40.11.

Application of filtering takes many forms. A common problem is transient suppression as illustrated in Fig. 40.12. All sources of transient interference should be treated at the source.

Power line filtering is recommended to eliminate conducted interference from reaching the powerline and adjacent equipment. Active filtering is very useful in that it can be built in as part of the circuit design and can be effective in passing only the design signals. A variety of noise blankers, cancelers, and limiter circuits are available for active cancellation of interference.

Special Filter Types

A variety of special-purpose filters are used in the design of electronic equipment. Transmitters require a variety of filters to achieve a noise-free output.

Receive preselectors play a useful role in interference rejection. Both distributed (cavity) and lumped element components are used.

IF filters control the selectivity of a receiving system and use a variety of mechanical and electrical filtering components.

Testing

The general requirements for electromagnetic filters are detailed in MIL-F-15733, MIL-F-18327, and MIL-F-25880. Insertion loss is measured in accordance with MIL-STD-220.

Defining Terms

Earth electrode system: A network of electrically interconnected rods, plates, mats, or grids, installed for the purpose of establishing a low-resistance contact with earth. The design objective for resistance to earth of this subsystem should not exceed 10 Ω.

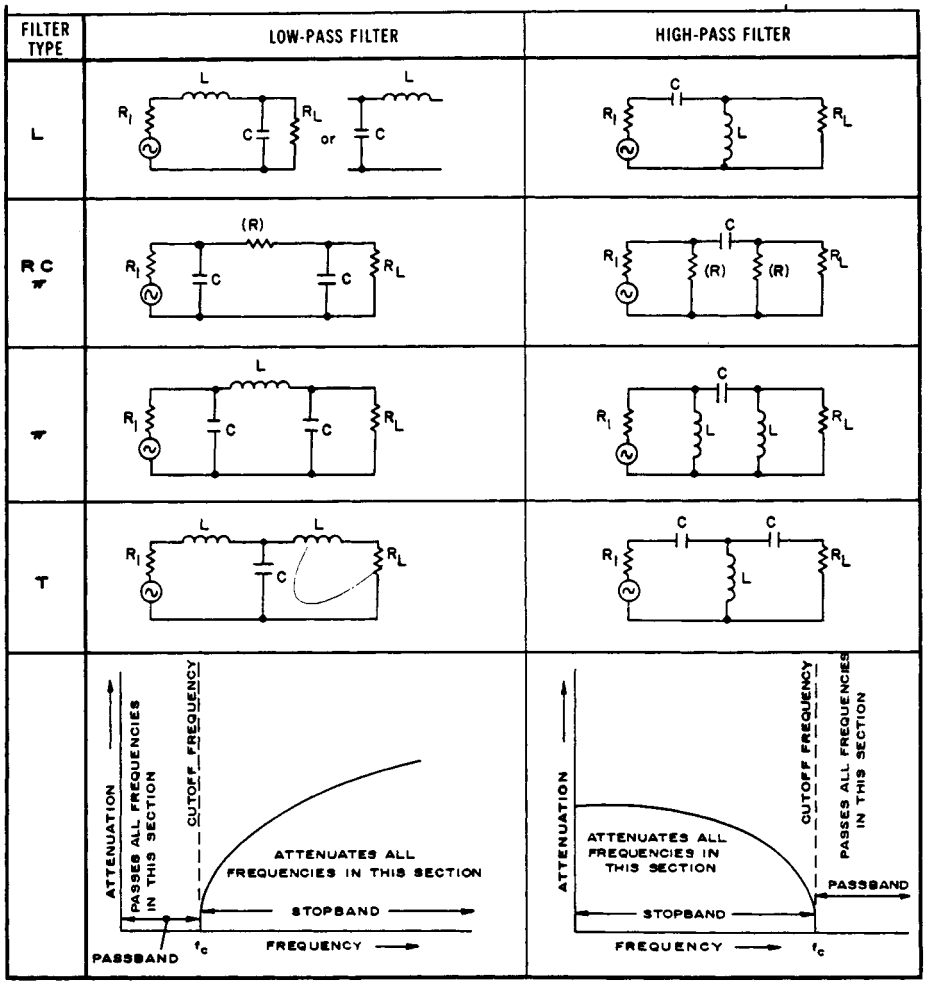


FIGURE 40.10 Filters provide a variety of frequency characteristics.

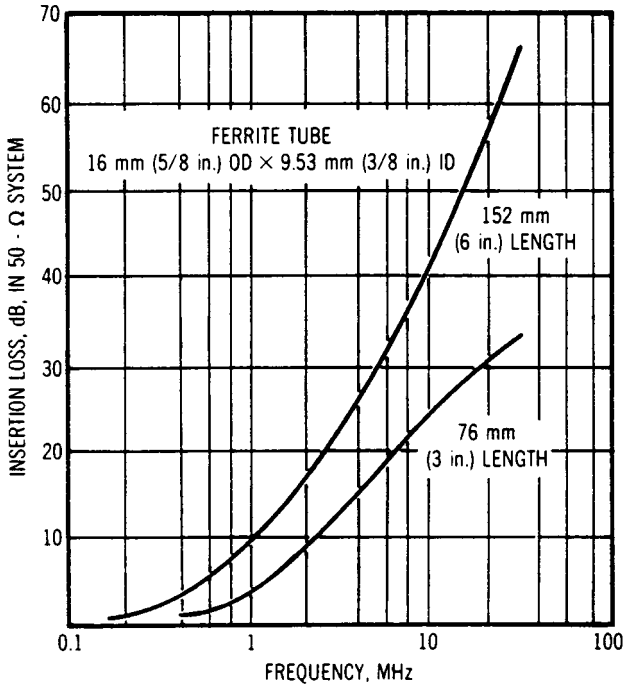


FIGURE 40.11 Ferrite provides a flexible means of achieving a low-pass filter with good high-frequency loss characteristics.

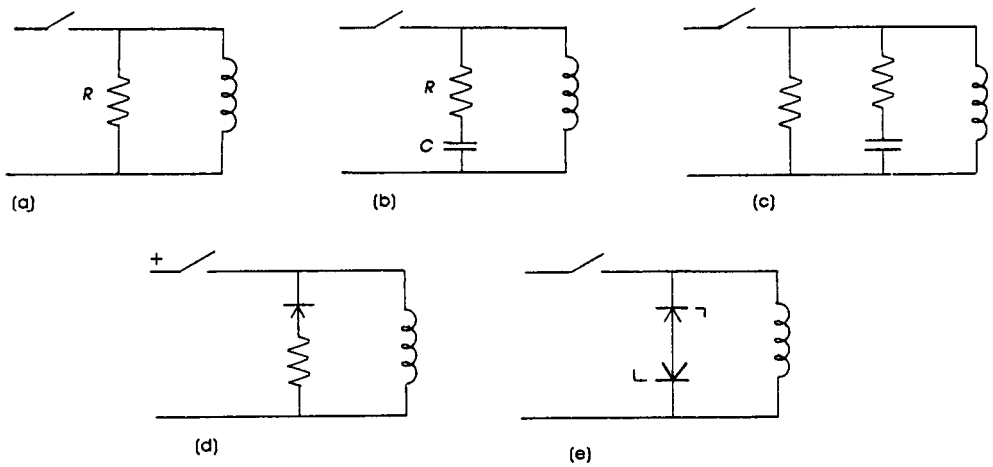


FIGURE 40.12 Transient responses are controlled using simple filters at the source. (a) Resistance damping; (b) capacitance suppression; (c) RC suppression; (d) diode suppression; (e) back-to-back diode suppression.

Electromagnetic compatibility (EMC): The capability of equipment or systems to be operated in their intended operational environment at designed levels of efficiency without causing or receiving degradation owing to unintentional electromagnetic interference. Electromagnetic compatibility is the result of an engineering planning process applied during the life cycle of the equipment. The process involves careful considerations of frequency allocation, design, procurement, production, site selection, installation, operation, and maintenance.

Electromagnetic pulse (EMP): A large impulsive-type electromagnetic wave generated by nuclear or chemical explosions.

Field strength: A general term that means the magnitude of the electric field vector (in volts per meter) or the magnitude of the magnetic field vector (in ampere-turns per meter). As used in the field of EMC/EMI, the term *field strength* shall be applied only to measurements made in the far field and shall be abbreviated as FS. For measurements made in the near field, the term *electric field strength* (EFS) or *magnetic field strength* (MFS) shall be used, according to whether the resultant electric or magnetic field, respectively, is measured.

Penetration: The passage through a partition or wall of an equipment or enclosure by a wire, cable, pipe, or other conductive object.

Radio frequency interference (RFI): Synonymous with *electromagnetic interference.*

Shielding effectiveness: A measure of the reduction or attenuation in the electromagnetic field strength at a point in space caused by the insertion of a shield between the source and that point.

Signal reference subsystem: This subsystem provides the reference points for all signal grounding to control static charges, noise, and interference. It may consist of any one or a combination of the lower frequency network, higher frequency network, or hybrid signal reference network.

TEMPEST: A code word (not an acronym) which encompasses the government/industrial program for controlling the emissions from systems processing classified data. Individual equipment may be *TEMPESTed* or commercial equipment may be placed in shielded enclosures.

Related Topics

10.3 The Ideal Linear-Phase Low-Pass Filter • 10.4 Ideal Linear-Phase Bandpass Filters • 55.3 Dielectric Breakdown

References

AFSC Design Handbook, DH1-4, Electromagnetic Compatibility, 4th ed., U.S. Air Force, Wright-Patterson Air Force Base, Ohio, January 1991.

R. F. Ficchi, Ed., *Practical Design for Electromagnetic Compatibility*, Hayden, 1971.

E. R. Freeman, *Electromagnetic Compatibility Design Guide for Avionics and Related Ground Support Equipment*, Norwood, Mass.: Artech House, 1982.

L. H. Hemming, *Architectural Electromagnetic Shielding Handbook*, New York: IEEE Press, 1991.

B. Keiser, *Principles of Electromagnetic Compatibility*, 3rd ed., Norwood, Mass.: Artech House, 1987.

Y.J. Lubkin, *Filter Systems and Design: Electrical, Microwave, and Digital*, Reading, Mass.: Addison-Wesley, 1970.

MIL-HDBK-419A, Grounding, Bonding, and Shielding of Electronic Equipment and Facilities, U.S. Department of Defense, Washington, D.C., 1990.

R. Morrison, *Grounding and Shielding Techniques in Instrumentation*, New York: John Wiley, 1986.

R. Morrison and W. H. Lewis, *Grounding and Shielding Techniques in Facilities*, New York: John Wiley, 1990.

T. Rikitake, *Magnetic and Electromagnetic Shielding*, Amsterdam: D. Reidel, 1987.

N. O. N. Violetto, *Electromagnetic Compatibility Handbook*, New York: Van Nostrand Reinhold, 1987.

D. R. J. White, *Shielding Design, Methodology and Procedures*, Springfield, Va.: Interference Control Technologies, 1986.

D. R. J. White, *A Handbook on Electromagnetic Shielding Materials and Performance*, Springfield Va.: Interference Control Technologies, 1975.

Further Information

The annual publication *Interference Technology Engineers' Master (Item),* published by R&B Enterprises, West Conshohocken, Pennsylvania, covers all aspects of EMI including an extensive product directory.

The periodical *IEEE Transactions on Electromagnetic Compatibility,* which is published by The Institute of Electrical and Electronics Engineers, Inc., provides theory and practice in the EMI field.

The periodical *EMC Test & Design,* published by the Cardiff Publishing Company, is a good source for practical EMI design information.

The periodical "emf-emi control" published bimonthly by EEC Press, Gainesville, VA, is an excellent source of practical EMI information.

The periodical *Compliance Engineering*, published quarterly by Compliance Engineering, Inc., is a good source for information on EMC regulations and rules.

40.2 Spectrum, Specifications, and Measurement Techniques

Vichate Ungvichian and John M. Roman

Electromagnetic radiation is a form of energy at a particular frequency that can propagate through a medium. This intentionally or unintentionally generated electromagnetic energy is considered as **electromagnetic interference** (EMI) if it degrades the performance of electronic systems. The purposeful generation of electromagnetic energy for communications can be defined as intentionally generated EMI; unintentionally generated EMI can be created, for example, by the electrical signals in a computer and may be radiated into space by way of the interconnecting cables and/or by openings in the device enclosures.

All electrical devices create some form of electromagnetic energy that may potentially interfere with the operation of other electrical devices outside the system (inter-system) or within the system (intra-system). Due to the increasing man-made EMI generated around the globe, allowable limits as well as measurement techniques on RF noise/interference have been set at national and international levels. The Federal Communications Commission and the Military are the two governing bodies in the United States setting standards on EMI, whereas the International Electrotechnical Commission is the ruling body in Europe. These ruling bodies are concerned with only a fraction of the total electromagnetic spectrum.

Electromagnetic Spectrum

The frequency spectrum of electromagnetic energy can span from dc to gamma ray (10^{21} Hz) and beyond. Figure 40.13 shows the typical frequency spectrum chart over a fraction of hertz to 6×10^{22} Hz.

The spectrum for use in electromagnetic compatibility (EMC) purposes covers only from a few hertz (extreme low frequency, ELF) to 40 GHz (microwave bands). ELF has been in use mostly in the area of biological research and ELF communications. On the other side of the spectrum, the electronic devices must function in a hostile environment, in military applications, over the gigahertz frequency range.

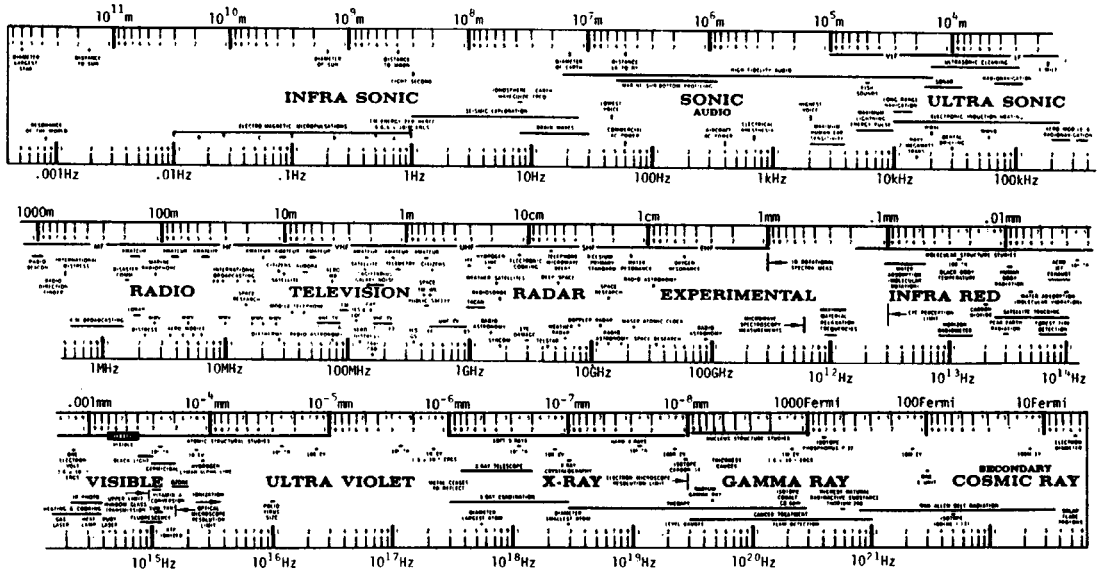


FIGURE 40.13 The frequency spectrum chart. (Contributed by Luther Monell, North America Rockwell Corp.)

TABLE 40.2 Unintentional Radiator Equipment Authorizations

Type of Device	Equipment Authorization Required[a]
TV broadcast receiver	Verification
FM broadcast receiver	Do
CB receiver	Certification
Superregenerative receiver	Do
Scanning receiver	Do
All other receivers subject to Part 15	Notification
TV interface device	Certification
Cable system terminal device	Notification
Stand-alone cable input selector switch	Verification
Class B personal computers and peripherals	Certification
Other Class B digital devices and peripherals	Verification
Class A digital devices and peripherals	Do
External switching power supplies	Do
All other devices	Do

[a]See additional provisions in CFR Part 15.101 and Part 15.103.

Specifications

In the United States, the Federal Communications Commission (FCC) and the military (MIL) are the two regulating bodies governing the EMC standards for commercial and military-based electronic devices and systems. In Europe, each country has its own EMC governing body as well as its own standards. The Verband Deutscher Elektrotechniker (VDE), British Standards Institute (BSI), European Telecommunications Standards Institute (ETSI), and International Special Committee on Radio Interference (CISPR) standards are very few examples of the acceptable standards in European countries such as Germany and Great Britain. Not until January 1, 1996, more than 15 European countries have adopted the IEC 1000 Part 4 Sections 2 to 4 as common EMI/EMC standards with several additional sections being in the final stages of approval.

The United States of America

Federal Communications Commission. The FCC sets limits on the amount of electromagnetic radiation allowed to be emitted from commercial electronic equipment. Any electronic device capable of emitting radio frequency energy by radiation or conduction is defined by the Commission as a radio frequency device and is subject to comply with the standards. The relevant limits and some general measurement techniques, along with equipment authorization procedures, are given in the Code of Federal Regulations (CFR), Title 47 (Telecommunications).

Listed in the CFR, Title 47 are five different types of equipment authorization procedures, namely, type acceptance, type approval, notification, certification, and verification. Restrictions are placed on the marketing and sale of radio frequency devices until the appropriate equipment authorization criteria are met.

Devices and systems that require allocation of the frequency spectrum fall under either type acceptance or type approval equipment authorization procedures. These radio frequency devices usually radiate high powers as in radio or television broadcast transmitters. Radio frequency devices not within the allocated part of the RF spectrum would require either certification or verification equipment authorizations. Some receivers, such as pagers, require notification equipment authorization as well. The requirements for these types of radio frequency devices are listed in Part 15 of the CFR, Title 47.

Part 15 contains three categories of equipment. Incidental radiators (such as dc motors, mechanical light switches, etc.) are not subject to FCC Part 15 emission control requirements. Unintentional radiators are radio frequency devices that intentionally generate radio frequency energy for use within the device, but which are not intended to emit RF energy by radiation or induction. Intentional radiators are radio frequency devices that intentionally radiate radio frequency energy by radiation or by induction.

Unintentional Radiators. There are two different classifications of digital devices listed under unintentional radiators. Class A digital devices are defined as devices that are intended for use in the commercial, industrial, or business environment. Class B digital devices are defined as devices that are intended to be used in a residential

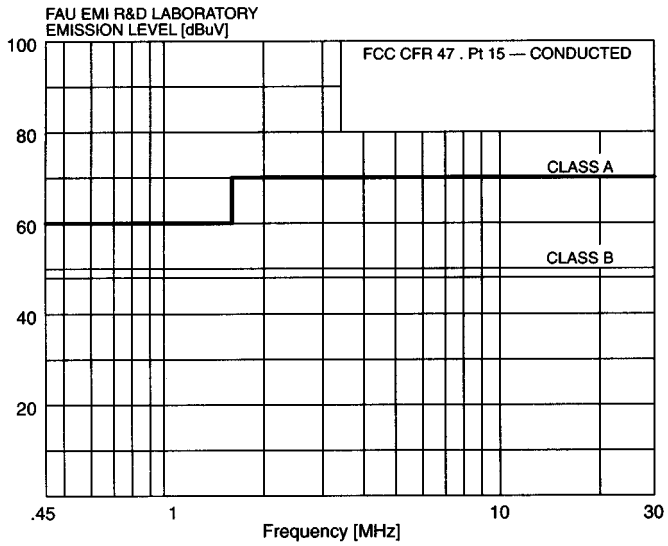


FIGURE 40.14 Class A and Class B conducted emission limits.

environment. Table 40.2 lists the different types of unintentional radiators and their corresponding equipment authorization procedures.

Unintentional Radiator Exempted Devices. There are some unintentional radiators that are listed as exempt devices. These devices are exempted from the technical requirements but are subject to the general requirements of Part 15, Sections 15.5 and 15.29, which state that if the devices cause harmful interference their operation must be ceased until such time that the interference is corrected, and that the device must be made available for inspection upon request by the Commission. It is also recommended (although not required) that these devices meet the technical specifications in Part 15. The exempted devices are as follows:

a. A digital device utilized exclusively in any transportation vehicle, including motor vehicles and aircraft.
b. A digital device used exclusively as an electronic control or power system utilized by a public utility or in an industrial plant. The term public utility includes equipment only to the extent that it is in a dedicated building or in a large room owned or leased by the utility and does not extend to equipment installed in a subscriber facility.
c. A digital device used exclusively as industrial, commercial, or medical test equipment.
d. A digital device utilized exclusively in an appliance, e.g., microwave oven, dishwasher, clothes dryer, air conditioner, etc.
e. Specialized medical digital devices (generally used at the direction of or under the supervision of a licensed health care practitioner) whether used in a patient's home or a health care facility. Nonspecialized medical devices, i.e., devices marketed through retail channels for use by the general public, are not exempted. This exemption also does not apply to digital devices used for recordkeeping or any purpose not directly connected with medical treatment.
f. Digital devices having a power consumption not exceeding 6 nW.
g. Joystick controllers, or similar devices such as a mouse, used with digital devices but which contain only nondigital circuitry or a simple circuit to convert the signal to the format required are viewed as passive add-on devices and are not directly subject to the technical standards or the equipment authorization requirements.
h. Digital devices in which the highest frequency generated and the highest frequency used are less than 1.705 MHz and which do not operate from the ac power lines or contain provisions to operate while connected to the ac power lines.
i. It should be noted that equipment containing more than one device is not exempt from the technical standards in Part 15 unless all of the devices meet the criteria for exemption.

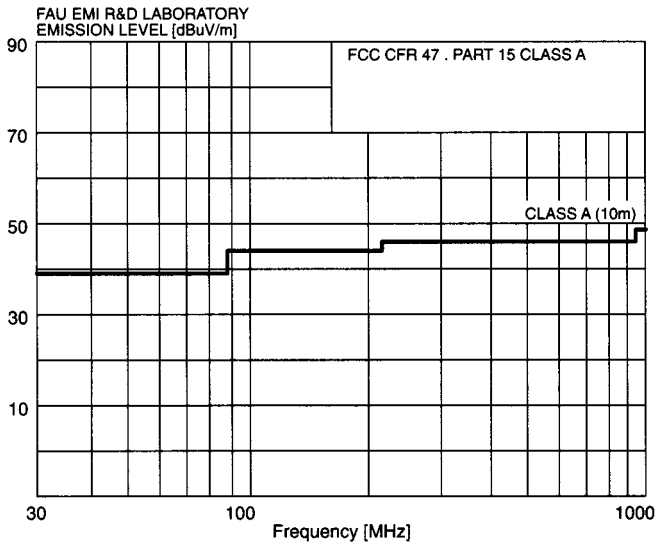


FIGURE 40.15(a) Unintentional (10 m) Class A radiated emission limit.

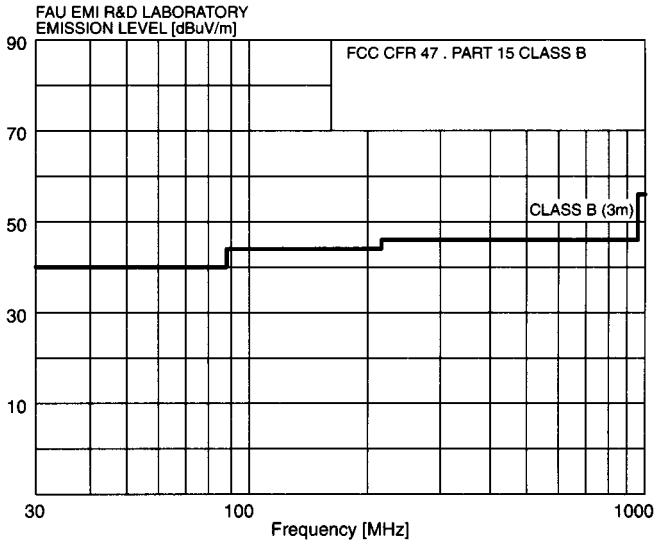


FIGURE 40.15(b) Unintentional (3 m) Class B radiated emission limits.

Unintentional Radiator Conducted Emission Limits. The limits on the interference conducted back into the ac power distribution system are given in Fig. 40.14.

Unintentional Radiator Radiated Emission Limits. Unintentional radiator radiated emission limits for Classes A and B devices are given in Figs. 40.15(a) and (b). It should be noted that a 10-m distance is required for Class A limit whereas 3 m is needed for Class B limit.

Intentional Radiators. There are different limits on devices which intentionally radiate radio frequency energy. The authorization procedure required by the Commission is the same as a certification. In addition to the radiated and conducted emission limits, there are restricted bands of operation in which the devices may not intentionally radiate (spurious emissions are permitted in these bands); these bands are given in Table 40.3.

TABLE 40.3 Restricted Bands of Operation

MHz	MHz	MHz	GHz
0.090–0.110	162.0125–167.17	2310–2390	9.3–9.5
0.49–0.51	167.72–173.2	2483.5–2500	10.6–12.7
2.1735–2.1905	240–285	2655–2900	13.25–13.4
8.362–8.366	322–335.4	3260–3267	14.47–14.5
13.36–13.41	399.9–410	3332–3339	15.35–16.2
25.5–25.67	608–614	3345.8–3358	17.7–21.4
37.5–38.25	960–1240	3600–4400	22.01–23.12
73–75.4	1300–1427	4500–5250	23.6–24.0
108–121.94	1435–1626.5	5350–5460	31.2–31.8
123–138	1660–1710	7250–7750	36.43–36.5
149.9–150.05	1718.8–1722.2	8025–8500	Above 38.6
156.7–156.9	2200–2300	9000–9200	

TABLE 40.4 General Requirement Radiated Emission Limits

Frequency (MHz)	Field Strength (μV/m)	Measurement Distance (m)
0.009–0.490	2400/F (kHz)	300
0.490–1.705	24000/F (kHz)	30
1.705–30.0	30	30
30–88	100*	3
88–216	150*	3
216–960	200*	3
Above 960	500	3

*Except as provided in paragraph (g), fundamental emissions from intentional radiators operating under this Section shall not be located in the frequency bands 54–72 MHz, 76–88 MHz, 174–216 MHz, or 470–806 MHz. However, operation within these frequency bands is permitted under other sections of this part, e.g., §§ 15.231 and 15.241.

Note: All of the limits specified above are measured using a CISPR Quasi-Peak adapter except for the frequency bands 9–90 kHz, 110–490 kHz, and above 1000 MHz; these are specified for an average measurement.

There is also a requirement that the antenna on the device be attached such that it may not be replaced by a different antenna.

Intentional Radiator Conducted Emissions. The conducted emission limits on intentional radiators are the same as for a Class B digital device (see Fig. 40.14).

Intentional Radiator Radiated Emissions. There are general provisions for the amount of radio frequency energy the intentional radiators are allowed to emit. Table 40.4 shows the general requirements for the radiated emission limits.

There are also some additional provisions for operation in specific frequency bands. Listed below are the frequency bands (and their corresponding part in the CFR) that contain additional provisions on the limits of radiated emissions. It is recommended that the CFR, Title 47 be consulted for the limits if the device in question intentionally radiates in the listed bands.

1. Part 15.217: Operation in the band 160–190 kHz.
2. Part 15.219: Operation in the band 510–1705 kHz.
3. Part 15.221: Operation in the band 525–1705 kHz (carrier current systems, am broadcast stations on a college or university).

TABLE 40.5 Appropriated Requirements for Different Platforms and Installations

Requirement	Description
CE101	Conducted emissions, power leads, 30 Hz to 10 kHz
CE102	Conducted emissions, power leads, 10 kHz to 10 MHz
CE106	Conducted emissions, antenna terminal, 10 kHz to 40 GHz
CS101	Conducted susceptibility, power leads, 30 Hz to 50 kHz
CS103	Conducted susceptibility, antenna port, intermodulation, 15 kHz to 10 GHz
CS104	Conducted susceptibility, antenna port, rejection of undesired signals, 30 Hz to 20 GHz
CS105	Conducted susceptibility, antenna port, cross-modulation, 30 Hz to 20 GHz
CS109	Conducted susceptibility, structure current, 60 Hz to 100 kHz
CS114	Conducted susceptibility, bulk cable injection, 10 kHz to 400 MHz
CS115	Conducted susceptibility, bulk cable injection, impulse excitation
CS116	Conducted susceptibility, damped sinusoidal transients, cables and power leads, 10 kHz to 100 MHz
RE101	Radiated emissions, magnetic field, 30 Hz to 100 kHz
RE102	Radiated emissions, electric field, 10 kHz to 18 GHz
RE103	Radiated emissions, antenna spurious and harmonic outputs, 10 kHz to 40 GHz
RS101	Radiated susceptibility, magnetic field, 30 Hz to 100 kHz
RS103	Radiated susceptibility, electric field, 10 kHz to 40 GHz
RS105	Radiated susceptibility, transient electromagnetic field

4. Part 15.223: Operation in the band 1.705–10 MHz.
5. Part 15.225: Operation in the band 13.553–13.567 MHz.
6. Part 15.227: Operation in the band 26.96–27.28 MHz.
7. Part 15.229: Operation in the band 40.66–40.70 MHz.
8. Part 15.231: Periodic operation in the band 40.55–40.70 MHz and above 70 MHz (alarm systems, door openers, remote switches, etc.).
9. Part 15.233: Operations within the bands 43.71–44.49 MHz, 46.60–46.98 MHz, 48.75–49.51 MHz, and 49.66–50.0 MHz (cordless telephones).
10. Part 15.235: Operation within the band 49.82–49.90 MHz.
11. Part 15.237: Operation within the bands 72.0–73.0 MHz, 74.6–74.8 MHz, and 75.2–76.0 MHz.
12. Part 15.239: Operation in the band 88–108 MHz.
13. Part 15.241: Operation in the band 174–216 MHz (biomedical telemetry devices only).
14. Part 15.243: Operation in the band 890–940 MHz (devices that use radio frequency energy to measure the characteristics of materials only).
15. Part 15.245: Operation in the bands 902–928 MHz, 2435–2465 MHz, 5785–5815 MHz, 10500–10550 MHz, and 24075–24175 MHz (field disturbance sensors only, excluding perimeter protection systems).
16. Part 15.247: Operation in the bands 902–928 MHz, 2400–2483.5 MHz, and 5725–5850 MHz (certain frequency hopping and direct sequence spread spectrum intentional radiators).
17. Part 15.249: Operation within the bands 902–928 MHz, 2400–2483.5 MHz, 5725–5875 MHZ, and 24.0–24.25 GHz.
18. Part 15.251: Operation within the bands 2.9–3.26 GHz, 3.267–3.332 GHz, 3.339–3.3458 GHz, and 3.358–3.6 GHz.

Military Standards. The standards and requirements related to the electromagnetic interference and susceptibility for the Military in the United States are described in the MIL-STD-461C document. Electromagnetic interference is defined as the radiated and conducted energy emitted from the device. Electromagnetic susceptibility is defined as the amount of radiated or conducted energy that the device can withstand without degrading its performance.

The standards are broken down into 17 segments defined by the two-letter suffix code and followed by three numbers ranging from 101–999 in the requirement name. The letter codes are conducted emissions (CE), conducted susceptibility (CS), radiated emissions (RE), radiated susceptibility (RS). Table 40.5 is a list and descriptions of the different emission and susceptibility requirements for a particular branch or type of application.

TABLE 40.6 MIL-STD-461D Emission and Susceptibility Requirements

Equipment and Subsystems Installed In, On, or Launched From the Following Platforms or Installations	Requirement Applicability																
	CE101	CE102	CE106	CS101	CS103	CS104	CS105	CS109	CS114	CS115	CS116	RE101	RE102	RE103	RS101	RS103	RS105
Surface ships	A	A	L	A	S	S	S		A		A	A	A	L	A	A	L
Submarines	A	A	L	A	S	S	S	L	A		A	A	A	L	A	A	L
Aircraft, Army, including flight line	A	A	L	A	S	S	S		A	A	L	A	A	L	A	A	L
Aircraft, Navy	L	A	L	A	S	S	S		A	A	A	L	A	L	L	A	L
Aircraft, Air Force		A	L	A	S	S	S		A	A	A		A	L		A	
Space systems, including launch vehicles		A	L	A	S	S	S		A	A	A		A	L		A	
Ground, Army		A	L	A	S	S	S		A	L	L		A	L	L	A	
Ground, Navy		A	L	A	S	S	S		A		A		A	L	L	A	L
Ground, Air Force		A	L	A	S	S	S		A	L	A		A	L		A	

TABLE 40.7 RE102 Applicability and Frequency Band of Testing

Applied for:	Frequency Band
Ground	2 MHz to 18 GHz[a]
Ships, surface	10 kHz to 18 GHz[a]
Submarines	10 kHz to 1 GHz
Aircraft (Army)	10 kHz to 18 GHz[a]
Aircraft (Air Force and Navy)	2 MHz to 18 GHz[a]

[a]If the highest clock frequency of the device is less than 1.8 GHz, replace the upper frequency limit to 1 GHz or 10 times the clock frequency, whichever is greater.

There are different equipment and subsystem classes defined for the various environments in which they are to be installed into. Table 40.6 gives the descriptions of the different classes and applicapability in MIL-STD-461D. If the requirement is applicable, three letters (A, L, S) are assigned in the matrix entry. The letter A means the equipment must meet the particular requirement. An L means consulting more detail of the requirement which may have different limits due to type of equipment and installation environment. An S means depending on the procurement requirement. From Table 40.6 RE102, RS103, CE102, CS101, and CS114, segments are required for all types of equipment and platforms. In this Handbook only RE102 and CE102 limits will be described.

The RE102 encompasses electric field limits as shown in Figure 40.16 a, b, and c. The frequency bands of the testing requirements typically cover from 10 kHz to 18 GHz depending on the clock frequency of the device or types of platforms which is described in Table 40.7. Up to 30 MHz, only the vertical polarization of the electric field will be measured and compared with the limits. Above 30 MHz, both horizontal and vertical field components must be measured and again compared with the limits. The device is in conformance with the RE102 if the electric field intensity in the appropriated frequency band is less than the prescribed limit.

The CE102 requirement is applicable to all power cords (either AC or DC source) including returns utilizing five nominal voltage levels. The limit based on a nominal 28 V or less and the limit for 200 V are depicted in Fig. 40.16 d. For the other operating voltage levels (115, 270, and 440 V), the limit is modified by adding a relaxation factor corresponding to the intended line voltage to the 28-V limit. The 220-V limit is 9 dB less stringent than the 28-V limit.

The European Union (EU)

During the early 1980s, a plan to harmonize the electromagnetic compatibility (EMC) requirements for the European Nations into one sovereign community was introduced. This effort, still growing today, attempts to align commercial, judicial, and financial objectives. There are currently more than 15 countries that participate

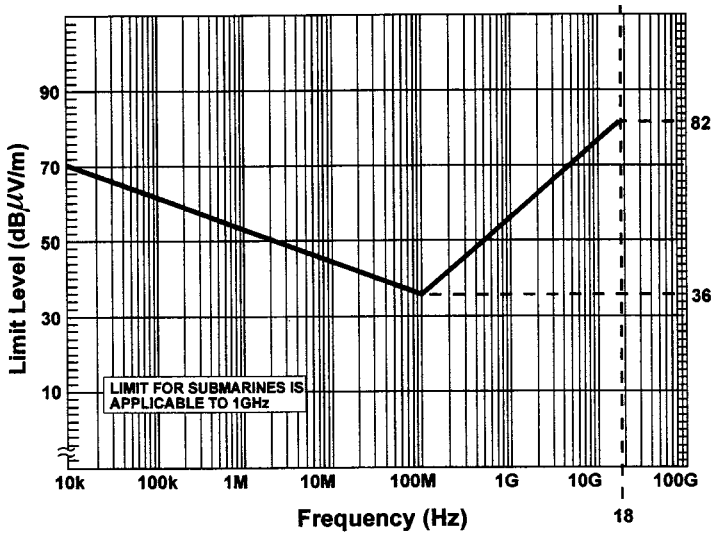


FIGURE 40.16(a) RE102 electric field limit for surface ship and submarine.

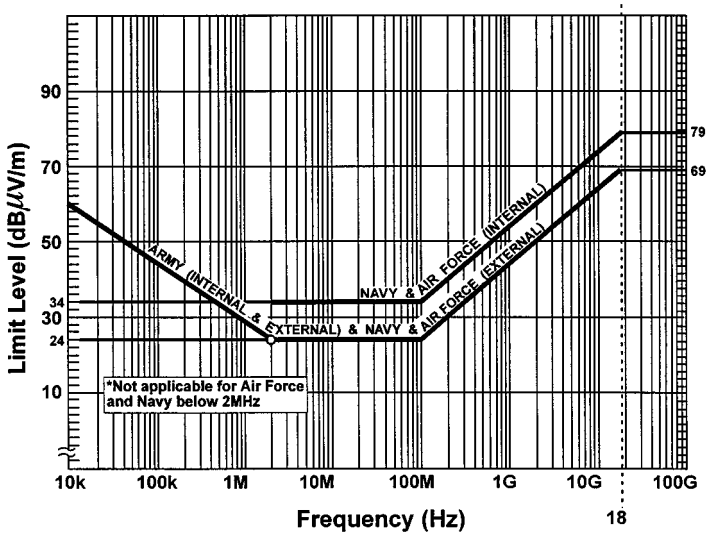


FIGURE 40.16(b) RE102 electric field limit for aircraft and space system.

in the EU requirement for EMI/EMC. Those countries have agreed to the common regulatory requirements placed on commercial products. The agreements, called directives, are listed in the Official Journal of the European Council (OJEC). Different directives for each type of equipment (Telecommunications, Information Technology Equipment [ITE], etc.) spell out the regulatory rules (EMC, Low Voltage, Telecommunications, etc.) required for approval to ship the product into the member countries, which is signified by the application of the conformity mark (CE).

The EU has approved a set of standards called the European Norms (EN). One of the EMC directives contains the International Electrotechnical Commission (IEC) test requirements for individual equipment approval. The IEC series of standards currently called upon by the EMC directive for generic equipment are EN 50081-1 (Generic Emission Standard; Part 1, residential, commercial and light industry) and EN 50082-1 (Generic Immunity Standard; Part 1, residential, commercial, and light industry equipment). These standards reference the IEC documents which list the requirements for measurement of emission and immunity characteristics.

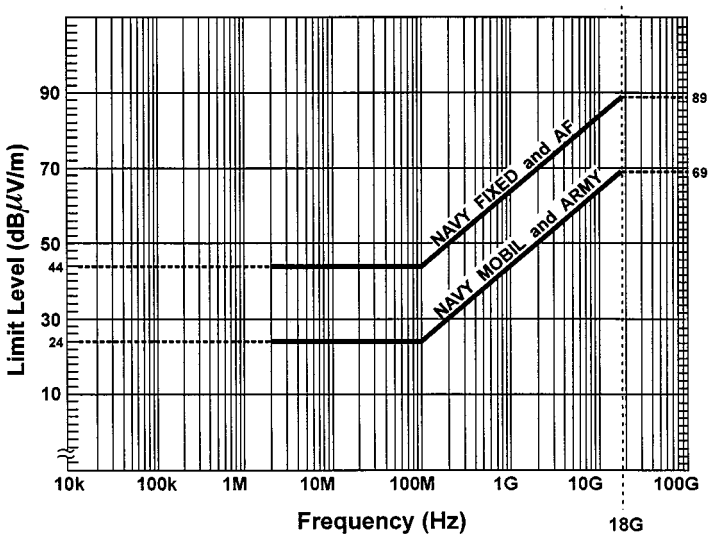


FIGURE 40.16(c) RE102 electric field limit for ground application.

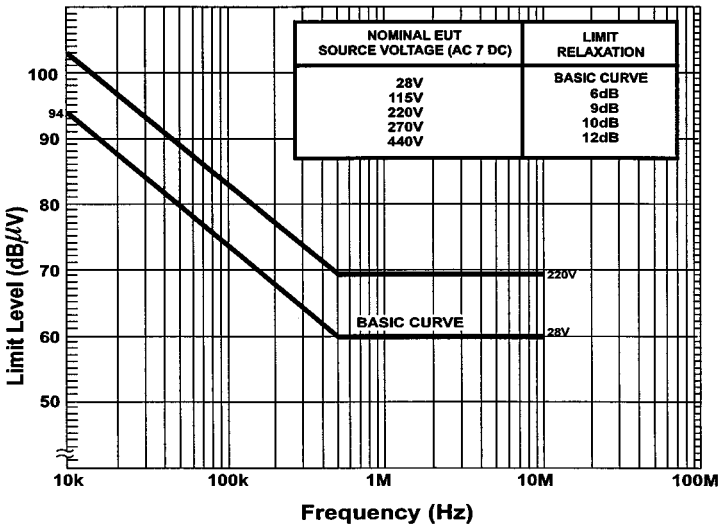


FIGURE 40.16(d) The CE102 limits for 28 V and 220 V nominal operating voltages.

The 1000-4-2 is the Electrostatic Discharge (ESD) test standard; 100-4-3 is the immunity standard; and 100-4-4 is the Electric Fast Transient (EFT) standard. The EN 55022 document contains the limits and test methods for both radiated and conducted emissions of ITE.

EN 55022 and IEC 1000 Series (Part 4)

EN 55022, also known as CISPR Publication 22, involves the limits and methods of measurement of radio interference characteristics of ITE. Currently, the ITE must meet three Parts (standards) of the IEC 1000-4 series. However, the future revision of the IEC 1000-4 series may mandate an additional eight requirements (presently int he process of review). Consult the latest versions of the applicable generic standard for the current requirements.

EN 55022 Document. The EN 55022 procedure describes for measurement of spurious signal strength, in the frequency range 0.15 MHz to 1 GHz, generated by pulsed electrical waveforms either through the power

TABLE 40.8 Limits of Mains Terminal Interference Voltage in the Frequency Range 0.15 to 30 MHz for Class A Equipment

Frequency Range (MHz)	Limits [dB (μV)]	
	Quasi-peak	Average
0.15 to 0.50	79	66
0.50 to 30	73	60

TABLE 40.9 Limits of Mains Terminal Interference Voltage in the Frequency Range 0.15 to 30 MHz for Class B Equipment

Frequency Range (MHz)	Limits [dB (μV)]	
	Quasi-peak	Average
0.15 to 0.50	66 to 56	56 to 46
0.50 to 5	56	46
5 to 30	60	50

The lower limit shall apply at the transition frequencies.
Note: In Table 40.9 the limit decreases linearly with the logarithm of the frequency in the range 0.15 to 0.50 MHz.

TABLE 40.10 Limits of Radiated Interference Field Strength in the Frequency Range 30 MHz to 1 GHz at a Test Distance of 30 m for Class A Equipment

Frequency Range (MHz)	Quasi-peak Limits [dB (μV/m)]
30 to 230	30
230 to 1000	37

TABLE 40.11 Limits of Radiated Interference Field Strength in the Frequency Range 30 MHz to 1 GHz at a Test Distance of 10 m for Class B Equipment

Frequency Range (MHz)	Quasi-peak Limits [dB (μV/m)]
30 to 230	30
230 to 1000	37

The lower limit shall apply at the transition frequencies.

TABLE 40.12(a) Contact Discharge Severity Levels

Level	Test Voltage Contact Discharge (kV)
1	2
2	4
3	6
4	8
x^1	Special

TABLE 40.12(b) Air Discharge Severity Levels

Level	Test Voltage Air Discharge (kV)
1	2
2	4
3	8
4	15
x^1	Special

[1] *x* is an open level. The level is subject to negotiations and has to be specified in the dedicated equipment specification. If higher voltages than those shown are specified, special test equipment may be needed.

main cable or through direct radiation. There are two classes of ITE: Class A and Class B. Class A equipment is usually intended for use in commercial establishments whereas Class B equipment is for domestic use. In order to conform with EN 55022, the measured voltages or field strengths shall meet the following limits.

Tables 40.8 and 40.9 are the limits of mains terminal interference voltage in the frequency range 0.15 to 30 MHz for Class A and Class B, respectively.

Tables 40.10 and 40.11 are the limits of radiated interference field strength in the frequency range 30 MHz to 1 GHz for Class A and Class B equipment.

IEC 1000-4 Section 2. The IEC 1000-4-2 document describes the static electricity discharge (ESD) requirements. At the time of publication, the current limits tabulated herein are utilized; however, the limits are subject for a revision.

Since contact discharge method and air discharge method are acceptable for ESD tests, two limits have been specified and tabulated in Table 40.12(a) and (b).

IEC 1000-4 Section 3. The IEC 1000-4-3 document involves the susceptibility requirements for equipment under test (EUT). The primary concern is the degradation of EUT under the influence of the hand-held transceiver or other sources of radiation in the frequency range 80 to 1000 MHz. Table 40.13 gives the tabulated limits of the susceptibility requirements.

TABLE 40.13 Severity Levels for Susceptibility Requirements

Level	Test Field Strength, (V/m)
1	1
2	3
3	10
x^1	Special

[1] *x* is an open class.

TABLE 40.14 Severity Levels for the Fast Transient/Burst Requirements

	Open Circuit Output Test Voltage ±10% (kV)	
Level	On Power Supply	On I/O (Input/Output) Signal, Data and Control Lines
1	0.5	0.25
2	1	0.5
3	2	1
4	4	2
x[1]	Special	Special

[1] *x* is an open level. The level is subject to negotiation between the user and the manufacturer or is specified by the manufacturer.

There are four classes of severity levels. Class 1 is for low-level electromagnetic radiation environments, such as those typical of local radio/television stations located more than 1 km away and levels typical of low-power transceivers. Class 2 is for moderate electromagnetic radiation environments, such as portable transceivers that are close to the EUT but not closer than 1 m. Class 3 is for severe electromagnetic radiation environments, such as levels typical of high-power transceivers in close proximity to the EUT. Class 4 is an open class for situations involving very severe electromagnetic radiation environments. The manufacturer sets the level relative to the environmental conditions into which the EUT would be installed.

IEC 1000-4 Section 4. The IEC 1000-4-4 document involves electrical fast transient/burst requirements. The purpose of this requirement is to evaluate the performance of the EUT when exposed to switching transients with high repetition frequency which may couple into the power main supply and external communication lines. Table 40.14 gives the test severity levels recommended at the time of publication. The repetition rate of the impulses is 5 kHz with a tolerance of ±20% for Levels 1 through 4, with an exception of 2.5 kHz for Level-4 power supply port.

Measurement Procedures

To determine the emission or susceptibility levels, measurement procedures were established. There are many protocols existing around the world. Each country may adopt its own measurement guidelines. In this section, only the FCC procedure will be described in detail. A list of some other procedures is given below.

FCC

The procedures used by the Commission to determine compliance are as follows:

1. FCC/OET MP-l: FCC Measurements for Determining Compliance of Radio Control and Security Alarm Devices and Associated Receivers.
2. FCC/OET MP-2: Measurement of UHF Noise Figures of TV Receivers.
3. FCC/OET MP-3: FCC Methods of Measurement of Output Signal Level, Output Terminal Conducted Spurious Emissions, Transfer Switch Characteristics, and Radio Noise Emissions from TV Interface Devices.
4. ANSI C63.4, 1992: FCC Procedure for Measuring RF Emissions from Computing Devices.
5. FCC/OET MP-9: FCC Procedure for Measuring Cable Television Switch Isolation.

In addition to the documents listed above, the FCC has outlined some generic measurement characteristics in the CFR, Title 47, Part 15, Sections 15.31 through 15.35.

Listed in the American National Standards Institute (ANSI) Specification C63.4-1992 measurement procedure document are the configuration setups for Class A and B computing devices and peripheral equipment. Some of the pertinent highlights for Class B computing devices are as follows:

- The equipment must be set up in a system configuration which includes the computer controller, a monitor, keyboard, serial device, parallel device, and any other device which may typically be connected to the system.

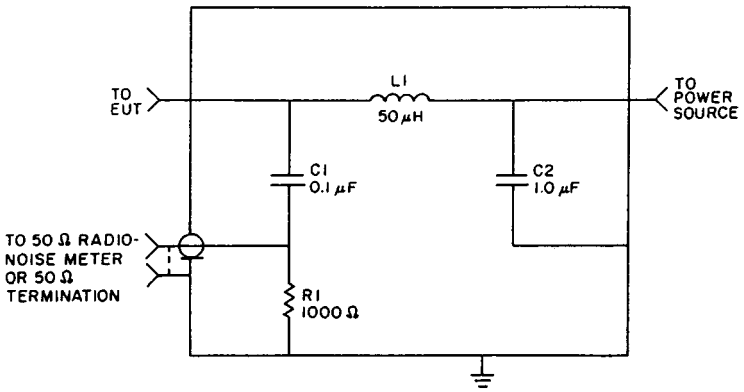


FIGURE 40.17 LISN circuit diagram.

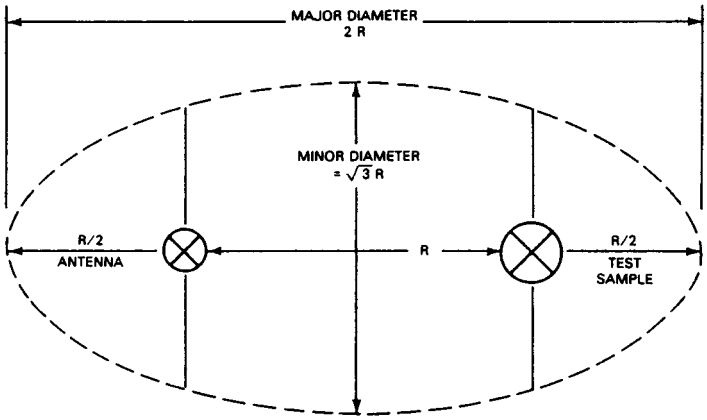


FIGURE 40.18 Minimum obstruction-free area for open field test site.

- The computer controller must be configured with the peripheral cards needed for typical operation (serial card, parallel card, video card, disk controller, memory), along with any other specialized cards defined in the typical setup to be marketed.
- A program to display, print, store, and/or send capital H characters to all of the pertinent devices (inclusive of the drives, CRT, printer, and any other data receiving devices) must be run for the duration of the evaluation.
- Typical cables and power cords are required for the test; the cables are bundled serpentine fashion at the center of the cable in 30- to 40-cm bundle lengths, to an overall length of 1 meter.

Conducted Emission Testing. Measurements are recommended to be performed inside of an RF shielded room in order to eliminate interference from ambient electromagnetic fields. The system units are placed on a nonconducting table 0.8 m high and 10 cm from the rear wall of the RF shielded enclosure. The measurements are performed with a line impedance stabilization network (LISN). This type of network is specifically designed to present a known impedance to the device under test, filter the noise present on the power line, and to match impedances with the measurement receiver. Figure 40.17 shows the typical FCC LISN circuit, applicable for monitoring the conducted noise present on either the phase or neutral line. Data is collected across phase and neutral to ground over the 450-kHz to 30-MHz frequency range and is compared with the aforementioned limits.

Radiated Emission Testing. The radiated emissions are measured at an FCC listed site (either semi-anechoic or open field), which requires a metal ground plane over the floor (typically hardware cloth). The site must satisfy a certain minimum size criteria, depending on the prescribed measurement distance. The accepted criteria is based on the Fresnel ellipse, which is presented in Fig. 40.18. The procedure for listing a site with the FCC includes the submittal of the site attenuation measurements, the site description, and a list of measurement equipment. The qualified site should meet the ±4-dB variation from theoretical values.

The EUT is placed on a nonconducting table 0.8 m above the ground plane floor. The receiving antenna is placed at the prescribed measurement distance (R) from the system (3 m for Class B and 10 m for Class A) and is scanned from 1 to 4 m in height while the EUT is rotated 360 degrees. The maximum emission data (per azimuth, elevation, and antenna orientation) is collected over the appropriate frequency range.

Abbreviations

ANSI: American National Standards Institute
BSI: The British Standards Institute
CFR: Code of Federal Regulations
CISPR: International Special Committee on Radio Interference
EMC: Electromagnetic compatibility
EMI: Electromagnetic interference
EN: European Norms
ESD: Electrostatic discharge
EU: European Union
EUT: Equipment under test
FCC: The Federal Communications Commission
IEC: International Electrotechnical Commission
IEEE: The Institute of Electrical and Electronics Engineers
ITE: Information technology equipment
LISN: Line impedance stabilization network
MIL: The U.S. Military

Defining Terms

Conducted emission: An RF current propagated through an electrical conductor.
Electromagnetic interference: An unwanted electromagnetic signal which may degrade the performance of an electronic device.
Far field: The region where the ratio of the electric to magnetic field is approximately equal to 377 Ω.
Field strength: An amount of electric or magnetic field measured in far-field region and expressed in volts/meter or amps/meter.
Line impedance stabilization network (LISN): A network designed to present a defined impedance at high frequency to a device under test, to filter any existing noise on the power mains, and to provide a 50-Ω impedance to the noise receiver.
Radiated emission: An electromagnetic field propagated through space.

Related Topics

16.1 Spectral Analysis • 73.2 Noise

References

Air Force Systems Command Electromagnetic Compatibility Handbook, 3rd ed., January 5, 1975.
CEI International Standard, IEC 801-1 to 801-4, 2nd ed., 1991.
Code of Federal Regulations, Title 47, Telecommunications, Part 15, October, 1995.

Electromagnetic Interference and Compatibility Handbook, vol. 1, Donald White Consultants, Inc., 1971.
Military Standard 461C, Requirements for the Control of Electromagnetic Interference Emissions and Susceptibility, January, 1993.
C.R. Paul, *Introduction to Electromagnetic Compatibility,* New York: John Wiley & Sons, 1992.

Further Information

The aforementioned measurement procedures used by the FCC are available from the Government Printing Office, Washington, D.C., 20402. The ANSI C63.4 document is available from the IEEE, 345 East 47th Street, New York, NY, 10017-2394.

The procedures utilized for the measurements performed to military EMC specifications are given in the following documents which are available from the Naval Publications and Forms Center, NPODS, 700 Robbins Ave., Philadelphia, Pa., 19111-2394.

MIL-STD-285 Attenuation Measurements for Enclosure, EM Shielding, for Electronic Test Purposes, Method of
MIL-STD-462D Electromagnetic Emission and Susceptibility, Measurement of Electromagnetic Interference Characteristics.
MIL-STD-463 Definitions and System of Units, Electromagnetic Interference and Electromagnetic Compatibility
MIL-STD-1377 Effectiveness of Cable, Connector and Weapon Enclosure Shielding and Filters in Precluding Hazards of Electromagnetic Radiation to Ordnance, Measurement of

The EC procedures and standards listed below are available from the Bureau Central de la Commission Electrotechnique Internationale 3, rue de Varembe, Geneve, Suisse.

EN 50081-1 Electromagnetic Compatibility—Generic Emission Standard; Part 1: Residential, Commercial and Light Industry.
EN 50082-1 Electromagnetic Compatibility—Generic Immunity Standard; Part 1: Residential, Commercial and Light Industry
EN55022 Limits and Methods of Measurement of Radio Interference Characteristics of Information Technology Equipment
IEC 1000-4-2 Electrostatic Discharge Immunity Test
IEC 1000-4-3 Radiated, Radio-Frequency, Electromagnetic Field Immunity Test
IEC 1000-4-4 Electrical Fast Transient/Burst Immunity Test

40.3 Lightning

Martin A. Uman and Marcos Rubinstein

An understanding of lightning and of the electric and magnetic fields produced by lightning is critical to an understanding of lightning-induced effects on electronic and electric power systems. This section begins with an overview of the terminology and physics of lightning. Then, statistics on lightning occurrence are given. Next the characteristics of the electric and the magnetic fields resulting from lightning charges and currents are examined, and the models used to describe that relationship are discussed. The section ends with a discussion of the coupling of the electric and magnetic fields from lightning to overhead wires.

Terminology and Physics

Lightning is a transient, high-current electric spark whose length is measured in kilometers. Lightning discharges can occur within a cloud, between clouds, from cloud to air, and from cloud to ground. All discharges except the latter are known as cloud discharges. The usual cloud-to-ground lightning is initiated in the cloud, lasts about half a second, and lowers to ground some 20 to 30 Coulombs of negative cloud charge. A less frequent type of cloud-to-ground discharge, accounting for less than 10% of all cloud-to-ground lightning, also begins

in the cloud but lowers positive cloud charge. An even less frequent type of cloud-to-ground lightning is initiated in an upward direction from tall man-made structures such as TV towers or tall geographical features such as mountaintops. A complete lightning discharge of any type is called a **flash**. The usual negative cloud-to-ground lightning flash starts in the cloud when a so-called **preliminary breakdown**, a particular type of electric discharge in the cloud, occurs. This process is followed by a discharge, termed the **stepped leader**, that propagates towards the ground in a series of luminous steps tens of meters in length. In progressing toward the ground, the negatively charged stepped leader branches in a downward direction. When one or more leader branches approach within a hundred meters or so of the ground, after 10 to 20 ms of stepped leader travel at an average speed of 10^5 to 10^6 m/s, the electric field at the ground (or at objects on the ground) increases above the critical breakdown field of the surrounding air and one or more upward-going discharges is initiated, starting the **attachment process.** After traveling a few tens of meters, one of the upward-going discharges, which is essentially at ground potential, contacts the tip of one branch of the stepped leader, which is at a high negative potential, probably some tens of megavolts. From that point, ground potential propagates upward, discharging to ground some or all of the negative charge previously deposited along the channel by the stepped leader. This upward propagating potential discontinuity is called the **return stroke**. Its front is a region of high electric field that causes increased ionization, current, temperature, and pressure as it travels the 5-km or more length of the leader channel. That trip is made in about 100 μs at an initial speed of the order of one third to one half the speed of light, the speed decreasing with height. The current at ground associated with the negative first return stroke has a peak of typically 35 kA achieved in a few microseconds, has a maximum current derivative of about 10^{11} A/s and falls to half of peak value in some tens of microseconds. The cessation of the first return stroke current may or may not end the flash. If more cloud charge is made available to the first stroke channel by in-cloud discharges, another leader-return stroke sequence may ensue, typically after tens of milliseconds. Preceding and initiating a subsequent return stroke is a continuous leader lowering negative charge, called a **dart leader**. The dart leader typically propagates down the residual channel of the previous stroke, generally ignoring the first stroke branches, although in about 50% of cloud-to-ground flashes there is at least one dart leader which transforms to a stepped leader on the downward trip, creating a new path to ground. There are typically three or four leader-return stroke sequences per negative cloud-to-ground flash, but ten or more is not uncommon.

Of the many different processes that occur during the various phases of a negative cloud-to-ground lightning (e.g., the in-cloud K processes, in-cloud J processes, and cloud to ground M components that occur between strokes and after the final stroke and are not discussed here), the electric and magnetic fields associated with the return stroke described above generally are the largest and hence the most significant in inducing unwanted voltages in electronic and electric power systems. This is the case because the currents in all other lightning processes are generally smaller than return stroke currents and the ground strike point of the return stroke can be much closer to objects on the ground than are in-cloud discharges. Cloud discharges exhibit currents similar to those of the in-cloud processes occurring in ground discharges and hence produce similar relatively small fields at or near ground level.

Positive flashes to ground, those initiated in the cloud and lowering positive charge to earth, generally contain only one return stroke, which is preceded by a "pulsating" leader rather than the stepped leader characteristically preceding negative first strokes and is generally followed by a period of continuous current flow. Positive flashes contain a greater percentage of very large return stroke currents, in the 100- to 300-kA range, than do negative flashes. Positive flashes may represent half of all flashes to ground in winter storms, which produce few total flashes, and typically represent 1 to 20% of the overall flashes in summer storms, that percentage increasing with increasing latitude.

Lightning Occurrence Statistics

Lightning flash density is defined as the number of lightning flashes per unit time per unit area and is usually measured in units of lightning flashes, either cloud or cloud-to-ground or both, per square kilometer per year. The two most common techniques for directly measuring flash density are (1) the use of so-called flash counters, relatively crude devices which trigger on electric fields above a value of the order of 1 kV/m in a frequency band centered in the hundreds of hertz to kilohertz range, of which two models are extensively used, the CIGRE

10-kHz and the CIGRE 500-Hz and (2) the use of networks of wideband magnetic direction finders, networks of wideband time-of-arrival detectors, and networks combining the two technologies, such networks now covering the U.S., Canada, Japan, Korea, Taiwan, most of Europe, and parts of many other countries. The average flash density varies considerably with geographical location, generally increasing with decreasing latitude. Typical ground flash densities are 1 to 5 $km^{-2}\ yr^{-1}$, with the world's highest being 30 to 50 $km^{-2}\ yr^{-1}$. Significant variations in flash density are observed with changes in local meteorological conditions within distances of the order of 10 km, for example, perpendicular to and inland from the Florida coastline. A ground flash density map of the U.S. for 1989, obtained from the U.S. National Lightning Detection Network of 114 wideband magnetic direction finders, is given by Orville [1991]. Flash densities in the U.S. are maximum in Florida with 10 to 15 $km^{-2}\ yr^{-1}$ and minimum along portions of the Pacific coast which has essentially no lightning.

An extensively measured parameter used to describe lightning activity worldwide is the thunderday or isokeraunic level, $T_{D,}$ the number of days per year that thunder is heard at a given location. This parameter has been recorded by weather station observers worldwide for many decades, whereas the accurate direct measurement of flash density has been possible only recently. Commonly used relations to convert thunderday level to ground flash density N_g are of the form

$$N_g = aT_D^b \quad \text{km}^{-2}\ \text{yr}^{-1} \tag{40.6}$$

where the value of a is near, and usually less than, 0.1 and the value of b is near, and usually greater than, 1. It should be noted that Eq. (40.6) is relatively inaccurate in that the data to which it is a fit is highly variable. The literature contains more than ten different values of a and b determined by different investigators.

Finally, from both worldwide thunderday and earth-orbiting satellite measurements, it has been estimated that there are about 100 total flashes, cloud and cloud-to-ground, per second over the whole earth. This number corresponds to an average global total flash density of 6 $km^{-2}\ yr^{-1}$.

Electric and Magnetic Fields

For the usual negative return stroke, measurements of the vertical component of the electric field and the two horizontal components of the magnetic field at ground level using wideband systems with upper frequency 3-dB points in the 1- to 20-MHz range are well documented in the literature. Measured vertical electric field and horizontal magnetic field waveshapes are shown in Fig. 40.19. Sketches of typical electric and magnetic fields are given in Fig. 40.20 for lightning in the 1- to 5-km range and in Fig. 40.21 for lightning at 10, 15, 50, and 200 km. Measured vertical and measured horizontal electric fields near ground are shown in Fig. 40.22. The mean value of the initial peak vertical electric field, normalized to 100 km by assuming an inverse distance dependence, is about 7 V/m for negative first strokes and about 4 V/m for negative subsequent strokes.

The return stroke vertical electric field rise to peak is comprised of two distinguishable parts, evident in Fig. 40.19: a slow front immediately followed by a fast transition to peak. For first strokes the slow front has a duration of a few microseconds and rises to typically half the peak amplitude, while for subsequent strokes the same slow front lasts less than 1 μs and rises only to typically 20% of the peak. The mean 10–90% fast transition time is about 200 ns regardless of

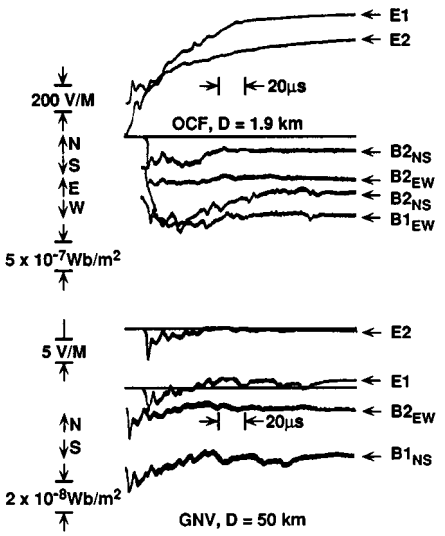


FIGURE 40.19 Simultaneously measured return stroke vertical electric field (E) and two horizontal magnetic flux densities (B_{EW} and B_{NS}) as observed about 2 and 50 km from a two-stroke flash, the first stroke being designated "1", the second "2". (*Source:* Adapted from Y.T. Lin et al., *J. Geophys. Res.*, vol. 84, pp. 6307–6314, 1979. With permission.)

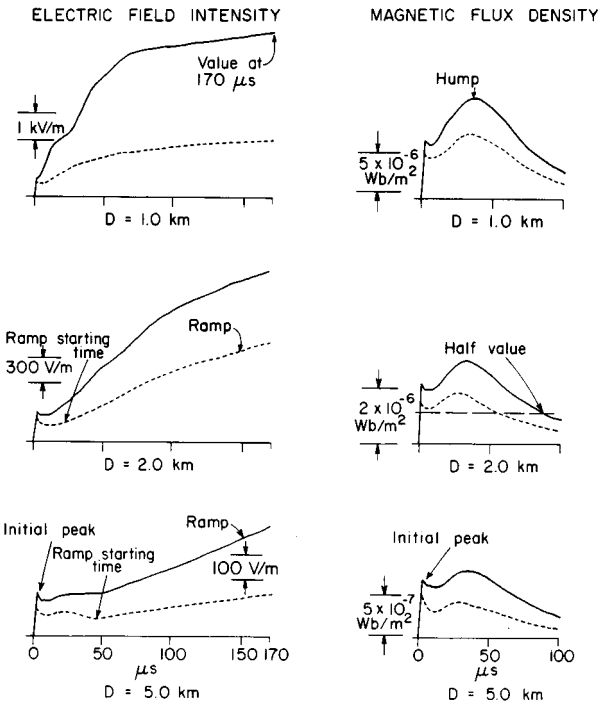


FIGURE 40.20 Drawings of typical return stroke electric fields and magnetic flux densities at 1, 2, and 5 km with definition of pertinent characteristic features. Solid lines represent first strokes, dotted subsequent strokes. (*Source:* Adapted from Y.T. Lin et al., *J. Geophys. Res.*, vol. 84, pp. 6307–6314, 1979. With permission.)

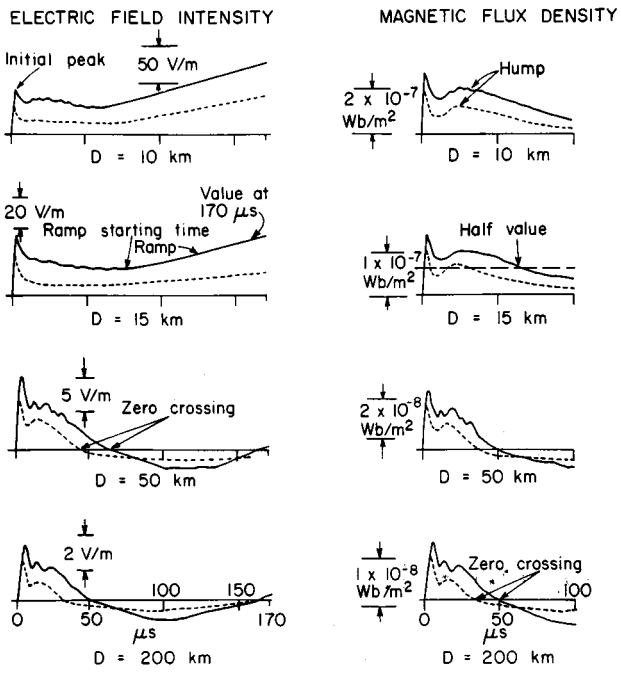


FIGURE 40.21 Drawings of typical return stroke electric fields and magnetic flux densities at 10, 15, 50, and 200 km; a continuation of Fig. 40.24. (*Source:* Adapted from Y.T. Lin et al., *J. Geophys. Res.*, vol. 84, pp. 6307–6314, 1979. With permission.)

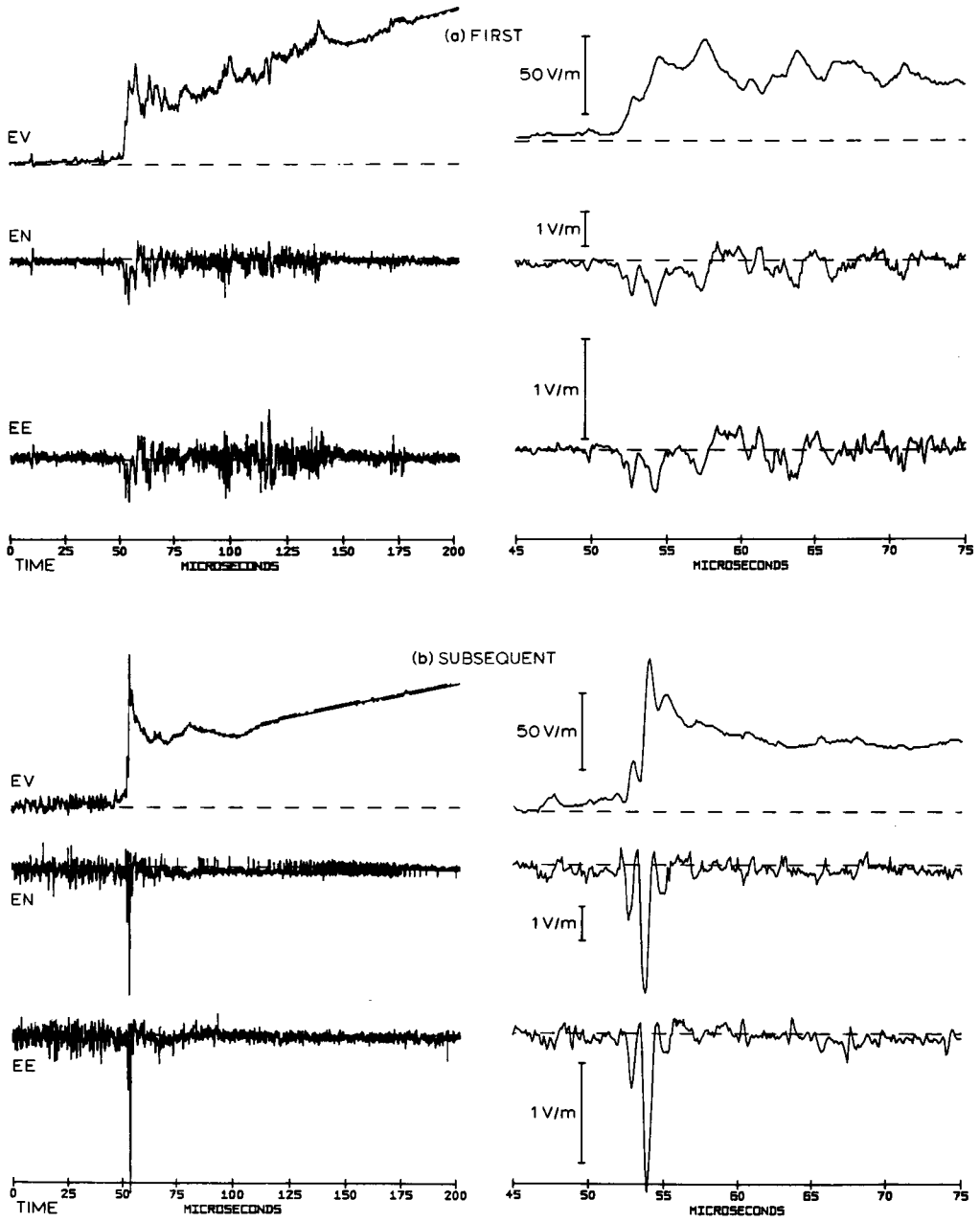


FIGURE 40.22 Measured horizontal electric field components (EN and EE) and vertical electric field (EV) one meter above ground for a first stroke (a) and a subsequent stroke (b) at a distance of 7 km presented on two time scales. (*Source:* Adapted from E.M. Thomson et al., *J. Geophys. Res.*, vol. 93, pp. 2429–2441, 1988. With permission.)

stroke order for strokes observed over saltwater where there is minimal distortion of the waveform due to propagation. The waveforms in Fig. 40.19 have suffered distortion in propagating over land.

After the initial field peak, the waveshapes of the vertical electric field and the horizontal magnetic field for close lightning exhibit a valley followed by a hump in the case of the magnetic field and by a ramp in the case of the electric field, as is evident from Figs. 40.19 through 40.22. Relative to the amplitude of the initial peak, the hump and the ramp decrease with increasing distance of the return stroke. For distances of 25 km or greater, the ramp in the electric field is no longer significant, and for distances of 50 km or more and for times of the

order of 100 μs, the waveshapes of the electric and magnetic fields are nearly identical, exhibiting a zero crossing and polarity reversal at some tens of microseconds.

For positive return strokes, there are more very large peak currents at the channel base, in the 100-kA range, than for return strokes lowering negative charge to ground, although the median value for both positives and negatives is not much different [Berger et al., 1975]. This observation is supported by measurements of the initial peak magnetic field from positive and negative return strokes made with magnetic direction-finding networks worldwide, where various investigations have found the mean peak positive field to be about twice the mean peak negative.

The horizontal component of the electric field has not been as well studied or characterized as the vertical. For the case of a finite-conducting earth and lightning beyond a few kilometers, Thomson et al. [1988] give wideband measurements of the three perpendicular components of the electric field about 1 m above ground level. An example is shown in Fig. 40.22. The horizontal field waveshapes are more impulsive and vary on a faster time scale than their associated vertical electric field waveshapes. In fact, the horizontal field appears to be a crude derivative of the vertical. The peak amplitudes of the horizontal electric fields are on the order of 30 times smaller than those of the vertical fields for ground conductivities of the order of 10^{-2} mho/m, this ratio being roughly proportional to the square root of the ground conductivity. The horizontal field, although considerably smaller for distant lightning, can be as important as the vertical electric field in inducing voltages on an overhead horizontal wire because of the greater horizontal extent of the wire relative to its height above ground, a fact well established by recent research, whereas in the earlier literature on power line coupling, for example, only the vertical field was considered to be important.

The so-called wavetilt formula, given in Eq. (40.7), models the ratio, in the frequency domain, of the horizontal to vertical electric field of a plane wave at the surface of an earth of conductivity σ and permittivity $\epsilon_r\epsilon_0$ for the case of grazing incidence and is certainly applicable to lightning return strokes occurring beyond a few kilometers, probably beyond a few hundred meters.

$$\frac{E_H(\omega)}{E_V(\omega)} = \frac{1}{\sqrt{\epsilon_r + \left(\dfrac{\sigma}{j\omega\epsilon_0}\right)}} \tag{40.7}$$

To the best of our knowledge, no horizontal or vertical electric fields very close to natural lightning, at distances from tens to hundreds of meters, necessary to the understanding of the voltages induced by very close lightning, have been published, although such close fields have been calculated by Diendorfer [1990] and by Rubinstein et al. [1990] using different return stroke models. These two sets of calculated fields are to be considered model-dependent estimates. Although there is disagreement between the two studies as to the waveshape of fields and in how appreciable is the influence of a finite ground conductivity at small distances, both studies yield horizontal field amplitudes at the height of a typical power distribution line comparable to the amplitude of the vertical field. Note that no return stroke model used to date (see next section) takes proper account of the attachment process referred to earlier and hence probably none accurately models the fields at very early times. Further, the leader fields preceding the return stroke field change are not taken into account in the existing models, although such fields at very close range are clearly important since it is the leader charge near ground that the return stroke discharges to ground, and hence the leader and return stroke electrostatic field changes should be of equivalent magnitude very close to the ground strike point (Rubinstein et al., 1995).

Modeling of the Return Stroke

General

A number of return stroke current models are found in the literature from which, if the current at the channel base is specified (e.g., from measurement) along with the model parameters, the channel current can be calculated as a function of height and time: the Bruce–Golde (BG) model, the transmission line (TL) model, the modified transmission line (MTL) model, the traveling current source (TCS) model, the Lin–Uman–Standler (LUS) model, the Diendorfer–Uman (DU) model, and the modified DU model. Two assumptions are

common to all of these models: that the lightning channel is perfectly straight and vertical and that the ground is a perfect conductor. Once the channel currents are determined as a function of height and time, the remote electric and magnetic fields can be calculated from Eqs. (40.8) through (40.14)

$$\overline{E} = \overline{E}_{\text{ele}} + \overline{E}_{\text{ind}} + \overline{E}_{\text{rad}} \tag{40.8}$$

$$\overline{E}_{\text{ele}} = \frac{1}{4\pi\epsilon_0}\int_{-h}^{h}\left\{\frac{2\cos\theta'\hat{a}_R + \sin\theta'\hat{a}_{\theta'}}{R^3}\int_0^t i\left(|z'|, \tau - \frac{R}{c}\right)dt\right\}dz \tag{40.9}$$

$$\overline{E}_{\text{ind}} = \frac{1}{4\pi\epsilon_0}\int_{-h}^{h}\frac{2\cos\theta'\hat{a}_R + \sin\theta'\hat{a}_{\theta'}}{cR^2} i\left(|z'|, t - \frac{R}{c}\right)dz' \tag{40.10}$$

$$\overline{E}_{\text{rad}} = \frac{1}{4\pi\epsilon_0}\int_{-h}^{h}\frac{1}{c^2R}\frac{\partial i\left(|z'|, t - \frac{R}{c}\right)}{\partial t}\hat{a}_{\theta'}dz' \tag{40.11}$$

$$\overline{B} = \overline{B}_{\text{ind}} + \overline{B}_{\text{rad}} \tag{40.12}$$

$$\overline{B}_{\text{ind}} = \frac{\mu_0}{4\pi}\int_{-h}^{h}\frac{\sin\theta'}{R^2} i\left(|z'|, t - \frac{R}{c}\right)\hat{a}_{\phi'}dz' \tag{40.13}$$

$$\overline{B}_{\text{rad}} = \frac{\mu_0}{4\pi}\int_{-h}^{h}\frac{\sin\theta'}{cR}\frac{\partial i\left(|z'|, t - \frac{R}{c}\right)}{\partial t}\hat{a}_{\phi'}dz' \tag{40.14}$$

where $i(z',t)$ is the current along the channel obtained from one of the return stroke current models mentioned above, and the geometry by which the above equations are to be interpreted is shown in Fig. 40.23. Note that the spatial integral includes the image current below the perfectly conducting ground plane so as to take account of reflections from the earth's surface. The three terms on the right-hand side of Eq. (40.8) [expanded in Eqs. (40.9) through (40.11)] are called, from left to right, the electrostatic, induction, and radiation terms. Similarly, the two terms on the right-hand side of Eq. (40.12) [expanded in Eqs. (40.13) and (40.14)] are termed the induction and radiation terms.

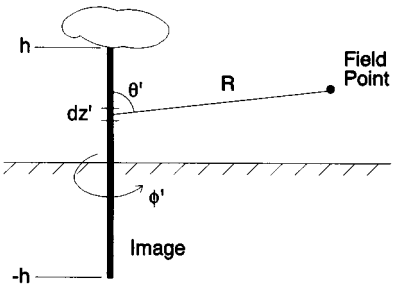


FIGURE 40.23 The geometry for the calculation of the fields using Eqs. (40.8) through (40.14).

For large distances to the lightning channel, the radiation part of the electric and magnetic fields is dominant due to its $1/R$ dependence (as compared to the $1/R^2$ and $1/R^3$ dependencies of the induction and electrostatic terms, respectively). By a similar argument, for close distances, the dominant terms will be the electrostatic term in the case of the electric field and the induction term for the magnetic field. It can be readily shown from Eqs. (40.8) through (40.14) and the preceding discussion that for any individual lightning return stroke model, the waveshapes of the vertical electric field and the horizontal magnetic field are almost identical for great distances, and this fact is also evident in the experimental data (see Fig. 40.21). Moreover, it can be shown that for great distances, the ratio of the electric field intensity E to the magnetic flux density B is the speed of light c.

A brief examination of return stroke current models follows. We discuss here only the transmission line (TL) model and the model of Diendorfer and Uman [1990] along with its modified version. More details of all models are found in Nucci et al. [1990], Diendorfer and Uman [1990], and Thottappillil et al. [1991], including fields calculated from the various models.

The Transmission Line Model

In the transmission line model it is assumed that the current waveform at the ground travels undistorted up the lightning channel at a constant speed v. Mathematically, this current is represented by

$$i(z',t) = i(0,t - z'/v) \qquad z' < vt \tag{40.15}$$

$$i(z',t) = 0 \qquad z' > vt$$

No charge is removed by the transmission line return stroke current along the channel since the charge entering the bottom of any section of the channel leaves the top when the current reaches it. All the charge is therefore transferred from the bottom to the top of the channel, an unrealistic situation given our knowledge of lightning physics.

Willett et al. [1989] have presented return stroke current, field, and speed data from artificially initiated (by firing small rockets trailing grounded wires) lightning in an attempt to validate the TL model. Using these data, Rakov et al. [1992] have shown, at least for subsequent strokes in artifically initiated lightning, that return stroke peak current can be derived from return stroke peak field by the expression $I = 1.5 - 0.037DE$ where the peak current I is in kA and is negative, the distance D is in km, and the peak electric field E is in V/m and is positive. Several investigators have published lightning peak current statistics derived from the magnetic radiation fields recorded by networks of magnetic direction finders by making use of the transmission line model. These studies are discussed by Rakov et al. [1992].

The Diendorfer–Uman (DU) Model and a Modification of It

The DU model [Diendorfer and Uman, 1990] is a physically reasonable model that can predict the salient features of the measured lightning electric and magnetic fields. Given the return stroke current at ground level, the channel current above ground is assumed to discharge the leader by way of two independent processes: (1) the discharge of the highly ionized core of the leader channel, termed the breakdown discharge process, with a time constant of 1 μs or less, and (2) the discharge of the corona envelope with a larger time constant. In both cases, the discharge at a height z' starts when the return stroke front, assumed to travel up at a constant speed v, arrives at z'. The liberated currents are assumed to flow to the ground at the speed of light.

For a current at ground $i(0,t)$, Diendorfer and Uman [1990] show that the current as a function of height and time is

$$i(z',t) = i(0,t_m) - i(0,z'/v*)\ \exp(-t_e/\tau) \tag{40.16}$$

where $t_m = (t + z'/v)$, $t_e = (t - z'/v)$, $v* = v\,(1 + v/c)$ and τ is the discharge time constant.

The Diendorfer–Uman model described above assumes that the return stroke propagates up the lightning channel at a constant speed and that the current from activated sections of the channel travels to ground at the speed of light. An analytical generalization of the DU model which allows for the return stroke speed and the downward current speed to be arbitrary functions of height has been presented by Thottappillil et al.[1991].

Lightning-Overhead Wire Interactions

General

Lightning interactions with overhead wires such as power distribution lines are a major source of electromagnetic compatibility problems, resulting in inferior power quality, power outages, and damaged electronics. Only a small fraction of all the cloud-to-ground lightning flashes directly strike overhead lines, making induced overvoltages a significant source of power disturbances. This section begins with a discussion of the appropriate transmission line equations. Then, examples of measured lightning-induced voltages on overhead lines as well as calculated voltages are presented.

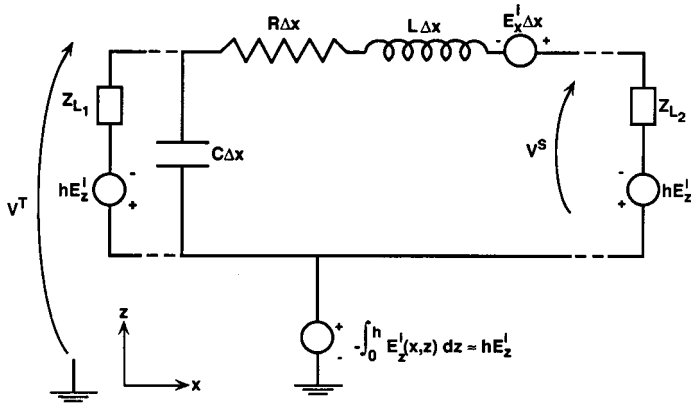


FIGURE 40.24 Equivalent circuit model obtained from Eqs. (40.17) through (40.19).

Transmission Line Equations

The transmission line equations for a nonuniform electromagnetic field impinging on a system of horizontal wires have been derived in the time domain by Agrawal et al. [1980], who adapted the theory to the case of wires above an imperfectly conducting ground. The main advantage of a time domain model over an equivalent frequency domain model is its applicability to cases of time varying and nonlinear loads and its ability to account for multiple reflections on a line with two or more discontinuities. On the other hand, with a frequency domain model it is intrinsically easier to handle frequency-dependent parameters such as the ground impedance.

The derivation of the time domain coupling equations is conceptually simple: Maxwell's equations are first integrated over closed cylindrical surfaces and along closed rectangular paths. The resulting integral equations, which are in terms of electric and magnetic fields, are then recast in terms of voltages and currents. One version of the transmission line equations, due to Agrawal et al. [1980], follows:

$$\frac{\partial V^s(x,t)}{\partial t} + Z_g * I(x,t) + L\frac{\partial I(x,t)}{\partial t} = E^i_x(x, z = h, t) \tag{40.17}$$

$$\frac{\partial I(x,t)}{\partial x} + C\frac{\partial V^s(x,t)}{\partial t} = 0 \tag{40.18}$$

$$V^t = V^i + V^s = -\int_0^h E^i_z(x,z,t)\,dz + V^s \tag{40.19}$$

where the superscript *s* identifies the "scattered" quantities, the superscript *i* identifies the "incident" quantities, the superscript *t* identifies the total, measurable quantities, and the asterisk is the convolution operator.

In these equations, the only source along the horizontal portion of the line is the horizontal component of the incident electric field. At the line terminations, the boundary condition and the termination current, *I*, are used to determine the end voltage. At those vertically oriented terminations, the vertical electric fields drive currents through the terminations into the line. The total voltage, $V^t(x,t)$, at the line terminations must equal $I_T * Z_T$ at all times, where Z_T is the termination impedance. Equations (40.17) through (40.19) can be represented by the circuit model in Fig. 40.24.

Two basic assumptions are used to arrive at Eqs. (40.17) through (40.19): (1) The response of the power line (scattered voltages and currents) to the impinging EM wave (incident field) is quasi-TEM (i.e., the scattered fields can be approximated as transverse electromagnetic). This allows us to define a "static" voltage along the line and to relate the line current and the scattered magnetic flux by an inductance, as well as the line scattered voltage and charge by a capacitance. (2) The transverse dimensions of the line system are small compared to the minimum wavelength, λ_{min}, of the excitation wave, and the height of the line is much larger than the diameter of the wire.

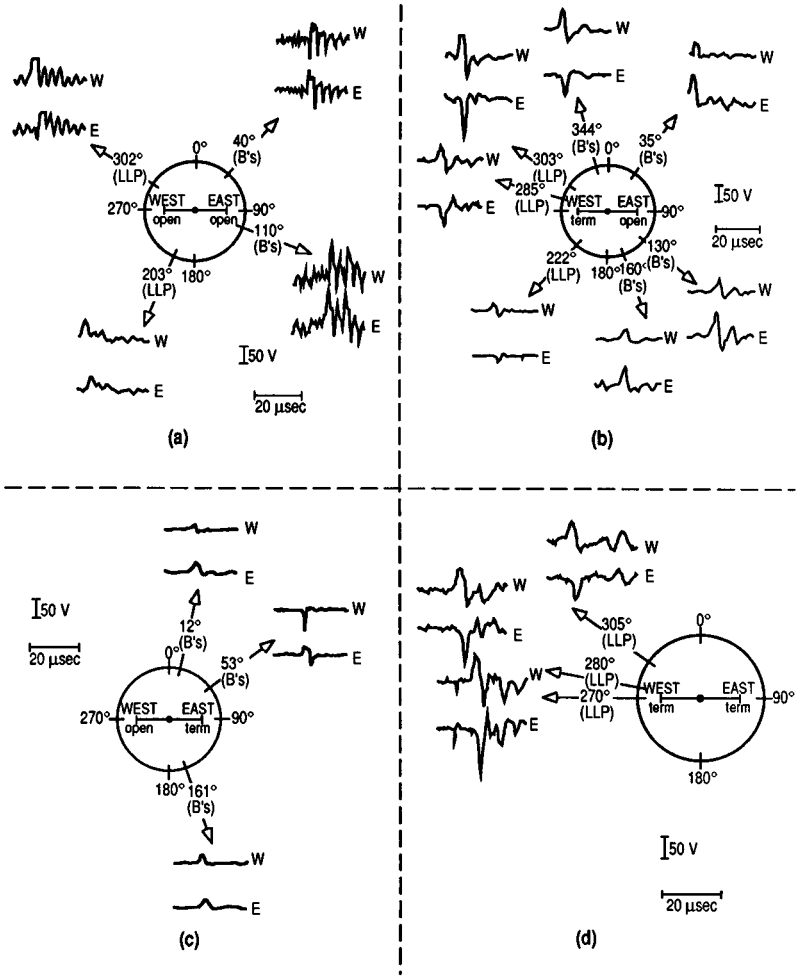


FIGURE 40.25 Examples of simultaneously measured lightning-induced voltages at the east end (E) and west end (W) of a 450-m line. Each line end is either open or terminated in its characteristic impedance, as noted. Directions to the lightning are determined from the ratio of the horizontal magnetic flux densities (*B*'s) or from a commercial lightning location system (LLP). (*Source:* Adapted from N. Georgiadis et al., "Lightning-induced voltages at both ends of a 450-meter distribution line," *IEEE Trans. EMC*, vol. 34, pp. 451–460, 1992. ©1992 IEEE. With permission.)

For other formulations of the overhead wire coupling equations, written in terms of field variables different from those used by Agrawal et al. [1980], see Rachidi [1993], Cooray [1994], and Nucci and Rachidi [1995].

Measured and Calculated Lightning-Induced Voltages on Overhead Wires

Several experiments have been carried out to test the coupling theory [e.g., Georgiadis et al., 1992; Rubinstein et al., 1994; Barker et al. 1996]. The basic strategy is the same in each experiment: to measure the lightning electric and magnetic fields in the vicinity of an instrumented overhead line while simultaneously measuring the voltages induced on the line, the measured fields then being used as inputs to a computer program written to solve Eqs. (40.17) through (40.19) and the computer-calculated voltage waveforms being compared with the measured voltage waveforms. The following discussion illustrates the types of voltage waveforms induced on overhead wires by lightning beyond a few kilometers and the degree of agreement that has been obtained in the coupling-model calculations. Examples of voltages induced on a 450-m overhead line about 10 m above the ground are shown in Fig. 40.25. Each line end was either terminated in its characteristic impedance or

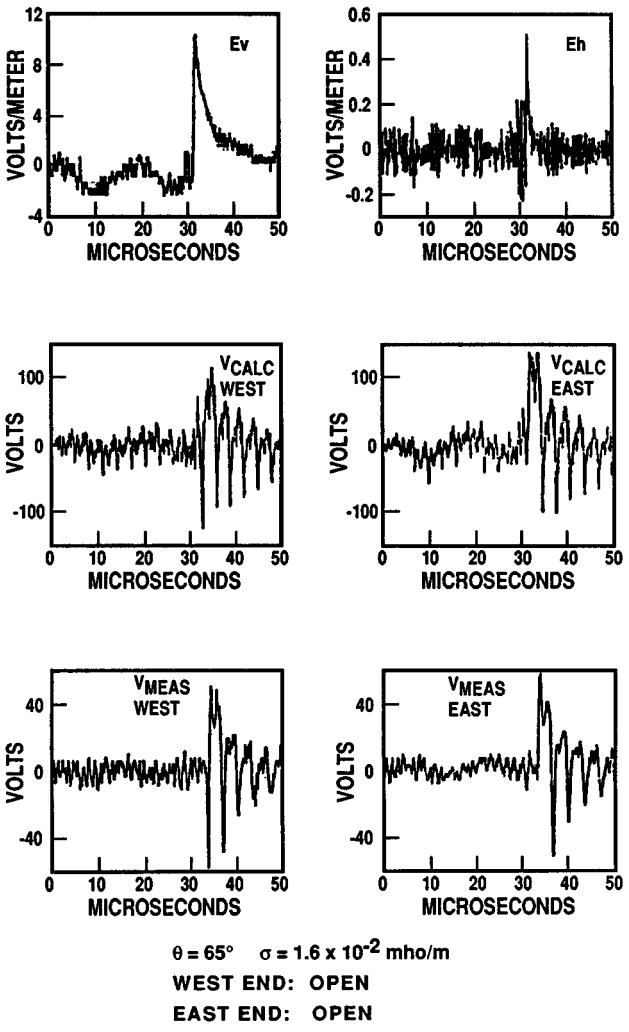


FIGURE 40.26 Measured and calculated voltages at the east and west ends of a 450-m line for both line ends open. Although the direction to the lightning as determined from the ratio of the two magnetic flux density (B) components was 40°, as shown in Fig. 40.29(a), the best calculated fit to the data was found for 65° as shown, the angular error apparently being caused by variation in the magnitudes of the magnetic flux density components due to nearby conductors as determined from comparing azimuths computed from the B's and from a commercial lightning location system (LLP). (*Source:* Adapted from N. Georgiadis et al., "Lightning-induced voltages at both ends of a 450-meter distribution line," *IEEE Trans. EMC*, vol. 34, pp. 451–460, 1992. ©1992 IEEE. With permission.)

open-circuited (four different cases), and voltages were measured simultaneously at each end. Figures 40.26 and 40.27 contain specific examples of measured and calculated voltage waveforms at each line end as well as the measured vertical electric field and calculated horizontal electric field via Eq. (40.7). It is clear from Figs. 40.25 through 40.27 that the induced voltage polarities and waveshapes are strongly dependent on the angle to the lightning and on the line end terminations. It is apparent also from Georgiadis et al. [1992] that while measured and calculated voltage waveshapes are in good agreement, the measured voltage amplitudes are, on average, a factor of three smaller than calculated voltages. This amplitude discrepancy remains unexplained but is probably due to the fact that the fields reaching the power line were shielded by trees along the line whereas the fields measured were in an open area and hence were unshielded.

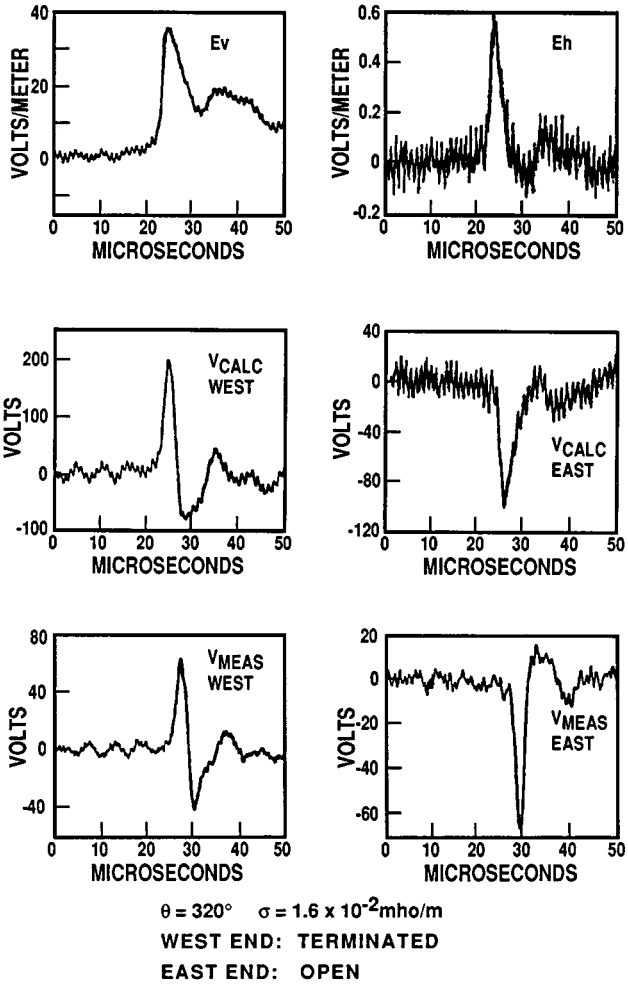


FIGURE 40.27 Measured and calculated voltages at the east and west ends of a 450-m line with the west end terminated and the east end open. The azimuth was determined from LLP data and is shown in Fig. 40.29(b). (*Source:* Adapted from N. Georgiadis et al., "Lightning-induced voltages at both ends of a 450-meter distribution line," *IEEE Trans. EMC*, vol. 34, pp. 451–460, 1992. ©1992 IEEE. With permission.)

Defining Terms

Attachment process: A process that occurs when one or more stepped leader branches approach within a hundred meters or so of the ground and the electric field at the ground increases above the critical breakdown field of the surrounding air. At that time one or more upward-going discharges is initiated. After traveling a few tens of meters, one of the upward discharges, which is essentially at ground potential, contracts the tip of one branch of the stepped leader, which is at a high potential, completing the leader path to ground.

Dart leader: A continuously moving leader lowering charge preceding a return stroke subsequent to the first. A dart leader typically propagates down the residual channel of the previous stroke.

Flash: A complete lightning discharge of any type.

Preliminary breakdown: An electrical discharge in the cloud that initiates a cloud-to-ground flash.

Return stroke: The upward propagating high-current, bright, potential discontinuity following the leader that discharges to the ground some or all of the charge previously deposited along the channel by the leader.

Stepped leader: A discharge following the preliminary breakdown that propagates from cloud towards the ground in a series of intermittant luminous steps with an average speed of 10^5 to 10^6 m/s. Negatively charged leaders clearly step, while positively charged leaders are more pulsating than stepped.

Related Topic

33.1 Maxwell Equations

References

A. K. Agrawal, H. J. Price, and S. H. Gurbaxani, "Transient response of multiconductor transmission lines excited by a non-uniform electromagnetic field," *IEEE Trans. EMC*, vol. EMC-22, pp. 119–129, 1980.

P. Barker, T. Short, A. Eybert-Berard, and J. Berlandis, "Induced voltage measurements on an experimental distribution line during nearby rocket triggered lightning flashes," *IEEE Trans. Pow. Delivery,* Vol. 11, pp. 980–995, 1996.

K. Berger, R. B. Anderson, and H. Kroninger, "Parameters of lightning flashes," *Electra*, vol. 80, pp. 23–37, 1975.

V. Cooray, "Calculating lightning-induced voltages in power lines: a comparison of two coupling models," *IEEE Trans. EMC,* vol. 36, pp. 170–182, 1994.

G. Diendorfer, "Induced voltage on an overhead line due to nearby lightning," *IEEE Trans. Electromag. Comp.*, vol. 32, pp. 292–299, 1990.

G. Diendorfer and M. A. Uman, "An improved return stroke model with specified channel-base current," *J. Geophys. Res.*, vol. 95, pp. 13,621–13,644, 1990.

N. Georgiadis, M. Rubinstein, M. A. Uman, P. J. Medelius, and E. M. Thomson, "Lightning-induced voltages at both ends of a 450-meter distribution line," *IEEE Trans. EMC*, vol. 34, pp. 451–460, 1992.

Y. T. Lin, M. A. Uman, J. A. Tiller, R. D. Brantley, W. H. Beasley, E. P. Krider, and C. D. Weidman, "Characterization of lightning return stroke electric and magnetic fields from simultaneous two-station measurements," *J. Geophys. Res.*, vol. 84, pp. 6307–6314, 1979.

C. A. Nucci, G. Diendorfer, M. A. Uman, F. Rachidi, M. Ianoz, and C. Mazzetti, "Lightning return stroke current models with specified channel-base current: A review and comparison," *J. Geophys. Res.*, vol. 95, pp. 20,395–20,408, 1990.

C. A. Nucci and F. Rachidi, "On the contribution of the electromagnetic field components in field-to-transmission line interaction," *IEEE Trans. EMC,* vol. 37, pp. 505–508, 1995.

R. E. Orville, "Annual summary—Lightning ground flash density in the contiguous United States—1989," *Monthly Weather Review*, vol. 119, pp. 573–577, 1991.

F. Rachidi, "Formulation of the field-to-transmission line coupling equations in terms of magnetic excitation field," *IEEE Trans. EMC,* col. 35, pp. 404–407, 1993.

V. A. Rakov, R. Thottappillil, and M. A. Uman, "On the empirical formula of Willett, et al., relating lightning return stroke peak current and peak electric field," *J. Geophys. Res.,* vol. 97, pp. 11,527–11,533, 1992.

M. Rubinstein, M. A. Uman, E. M. Thomson, and P. J. Medelius, "Voltages induced on a test distribution line by artificially initiated lightning at close range: Measurement and theory," in Proceedings of the 20th International Conference on Lightning Protection, Interlaken, Switzerland, September 24–28, 1990.

M. Rubinstein, M. A. Uman, P. J. Medelius, and E. M. Thomson, "Measurements of the voltage induced on an overhead power line 20 m from triggered lightning," *IEEE Trans. EMC,* vol. 36, pp. 134–140, 1994.

M. Rubinstein, F. Rachidi, M. A. Uman, R. Thottappillil, V. A. Rakov, and C. A. Nucci, "Characterization of vertical electric fields 500 m and 30 m from triggered lightning, *J. Geophys. Res.,* vol. 100, pp. 8863–8872, 1995.

E. M. Thomson, P. Medelius, M. Rubinstein, M. A. Uman, J. Johnson, and J. Stone, "Horizontal electric fields from lightning return strokes," *J. Geophys. Res.*, vol. 93, pp. 2429–2441, 1988.

R. Thottappillil, D. K. McLain, G. Diendorfer, and M. A. Uman, "Extension of the Diendorfer–Uman lightning return stroke model to the case of a variable upward return stroke speed and a variable downward discharge current speed," *J. Geophys. Res.*, vol. 96, pp. 17,143–17,150, 1991.

J. E. Willett, J. C. Bailey, V. P. Idone, A. Eybert-Berard, and L. Barret, "Submicrosecond intercomparison of radiation fields and currents in triggered lightning return strokes based on the transmission-line model," *J. Geophys. Res.*, vol. 94, pp. 13,275–13,286, 1989.

Further Information

For more details on the material presented here, see *The Lightning Discharge* (Academic Press, San Diego, 1987) by M. A. Uman and the review article "Natural and Artificially Initiated Lightning" (*Science*, vol. 246, 457–464, 1989) by M. A. Uman and E. P. Krider. For the most recent information on return stroke properties and references to previous work, see "Some Properties of Negative Cloud to Ground Lightning vs. Stroke Order" (*J. Geophys. Res.*, vol. 95, 5447–5453, 1990), by V. A. Rakov and M. A. Uman, and "Lightning Subsequent Stroke Electric Field Peak Greater than the First Stroke Peak and Multiple Ground Terminations" (*J. Geophys. Res.*, vol. 97, 7503–7509, 1992), by R. Thottappillil, V. A. Rakov, M. A. Uman, W. H. Beasley, M. J. Master, and D. V. Shelukhin.

For more information on lightning properties derived from networks of wideband magnetic direction finders, see "Cloud to Ground Lightning Flash Characteristics from June 1984 through May 1985" (*J. Geophys. Res.*, vol. 92, 5640–5644, 1992), by R. E. Orville, R. A. Weisman, R. B. Pyle, R. W. Henderson, and R. E. Orville, Jr., and "Calibration of a Magnetic Direction Finding Network Using Measured Triggered Lightning Return Stroke Peak Currents" (*J. Geophys. Res.*, vol. 96, 17,135–17,142, 1991), by R. E. Orville.

Triggered Lightning

The photograph shows lightning that was artificially initiated or "triggered" from a natural thunderstorm at the University of Florida's International Center for Lightning Research and Testing at Camp Blanding Army National Guard Base in Florida. Triggered lightning is presently being used to study the close electromagnetic environment of lightning as well as to determine lightning's effects on communication and power systems. (Photo courtesy of the University of Florida, Department of Electrical and Computer Engineering.)

Belcher, M.L., Nessmith, J.T., Wiltse, J.C. “Radar”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

41

Radar

Melvin L. Belcher
Georgia Tech Research Institute

Josh T. Nessmith
Georgia Tech Research Institute

James C. Wiltse
Georgia Tech Research Institute

41.1 Pulse Radar
Overview of Pulsed Radars • Critical Subsystem Design and Technology • Radar Performance Prediction • Radar Waveforms • Detection and Search • Estimation and Tracking

41.2 Continuous Wave Radar
CW Doppler Radar • FM/CW Radar • Interrupted Frequency-Modulated CW (IFM/CW) • Applications • Summary Comments

41.1 Pulse Radar

Melvin L. Belcher and Josh T. Nessmith

Overview of Pulsed Radars

Basic Concept of Pulse Radar Operation

The basic operation of a pulse radar is depicted in Fig. 41.1. The radar transmits a pulse of RF energy and then receives returns (reflections) from desired and undesired targets. Desired targets may include space, airborne, and sea- and/or surface-based vehicles. They can also include the earth's surface and the atmosphere, depending on the application. Undesired targets are termed *clutter*. Clutter sources include the ground, natural and man-made objects, sea, atmospheric phenomena, and birds. Short-range/low-altitude radar operation is often constrained by clutter since the multitude of undesired returns masks returns from targets of interest such as aircraft.

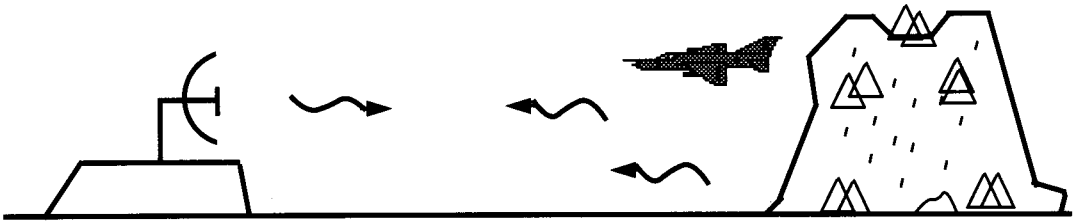
The range, azimuth angle, elevation angle, and range rate can be directly measured from a return to estimate target position and velocity. Signature data can be extracted by measuring the amplitude, phase, and polarization of the return.

Pulse radar affords a great deal of design and operational flexibility. Pulse duration and pulse rate can be tailored to specific applications to provide optimal performance. Modern computer-controlled multiple-function radars exploit this capability by choosing the best waveform from a repertoire for a given operational mode and interference environment automatically.

Radar Applications

The breadth of pulse radar applications is summarized in Table 41.1. Radar applications can be grouped into search, track, and signature measurement applications. Search radars are used for tracking but have relatively large range and angle errors. The search functions favor broad beam-widths and low bandwidths in order to efficiently search over a large spatial volume. As indicated in Table 41.1, search is preferably performed in the lower frequency bands. The antenna pattern is narrow in azimuth and has a cosecant pattern in elevation to provide acceptable coverage from the horizon to the zenith.

Tracking radars are typically characterized by a narrow beamwidth and moderate bandwidth in order to provide accurate range and angle measurements on a given target. The antenna pattern is a pencil beam with approximately the same dimensions in azimuth and elevation. Track is usually conducted at the higher frequency bands in order to minimize the beamwidth for a given antenna aperture area. After each return from a target



$$\text{Target Range} = \frac{\text{Two-Way-Time-Delay} \cdot \text{Speed-of-Light}}{}$$

FIGURE 41.1 Pulse radar.

TABLE 41.1 Radar Bands

Band	Frequency Range	Principal Applications
HF	3–30 MHz	Over-the-horizon radar
VHF	30–300 MHz	Long-range search
UHF	300–1000 MHz	Long-range surveillance
L	1000–2000 MHz	Long-range surveillance
S	2000–4000 MHz	Surveillance
		Long-range weather characterization
		Terminal air traffic control
C	4000–8000 MHz	Fire control
		Instrumentation tracking
X	8–12 GHz	Fire control
		Air-to-air missile seeker
		Marine radar
		Airborne weather characterization
Ku	12–18 GHz	Short-range fire control
		Remote sensing
Ka	27–40 GHz	Remote sensing
		Weapon guidance
V	40–75 GHz	Remote sensing
		Weapon guidance
W	75–110 GHz	Remote sensing
		Weapon guidance

is received, the range and angle are measured and input into a track filter. Track filtering smooths the data to refine the estimate of target position and velocity. It also predicts the target's flight path to provide range gating and antenna pointing control to the radar system.

Signature measurement applications include remote sensing of the environment as well as the measurement of target characteristics. In some applications, synthetic aperture radar (SAR) imaging is conducted from aircraft or satellites to characterize land usage over broad areas. Moving targets that present changing aspect to the radar can be imaged from airborne or ground-based radars via inverse synthetic aperture radar (ISAR) techniques. As defined in the subsection "Resolution and Accuracy," cross-range resolution improves with increasing antenna extent. SAR/ISAR effectively substitutes an extended observation interval over which coherent returns are collected from different target aspect angles for a large antenna structure that would not be physically realizable in many instances.

In general, characterization performance improves with increasing frequency because of the associated improvement in range, range rate, and cross-range resolution. However, phenomenological characterization to support environmental remote sensing may require data collected across a broad swath of frequencies.

A multiple-function **phased array** radar generally integrates these functions to some degree. Its design is usually driven by the track function. Its operational frequency is generally a compromise between the lower

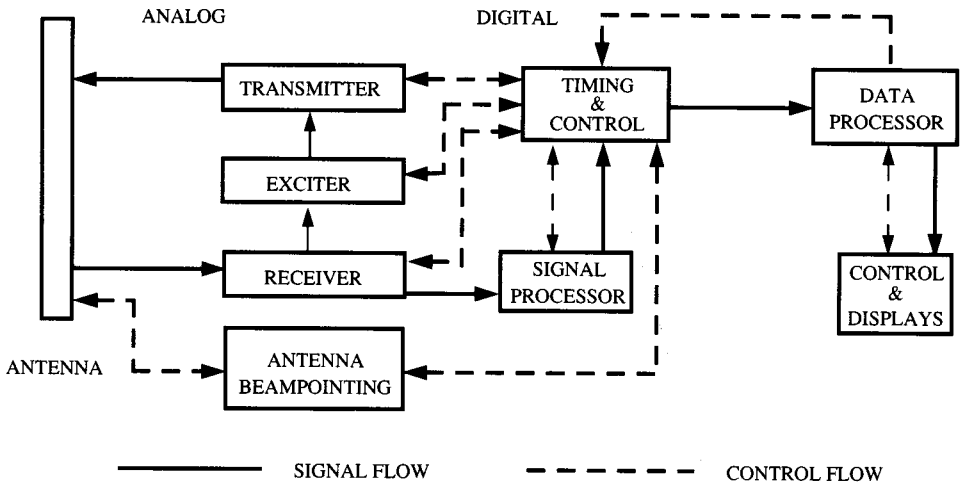


FIGURE 41.2 Radar system architecture.

frequency of the search radar and the higher frequency desired for the tracking radar. The degree of signature measurement implemented to support such functions as noncooperative target identification depends on the resolution capability of the radar as well as the operational user requirements. Multiple-function radar design represents a compromise among these different requirements. However, implementation constraints, multiple-target handling requirements, and reaction time requirements often dictate the use of phased array radar systems integrating search, track, and characterization functions.

Critical Subsystem Design and Technology

The major subsystems making up a pulse radar system are depicted in Fig. 41.2. The associated interaction between function and technology is summarized in this subsection.

Antenna

The radar antenna function is to first provide spatial directivity to the transmitted EM wave and then to intercept the scattering of that wave from a target. Most radar antennas may be categorized as mechanically scanning or electronically scanning. Mechanically scanned reflector antennas are used in applications where rapid beam scanning is not required. Electronic scanning antennas include phased arrays and frequency scanned antennas. Phased array beams can be steered to any point in their field-of-view, typically within 10 to 100 μs, depending on the latency of the beam steering subsystem and the switching time of the phase shifters. Phased arrays are desirable in multiple function radars since they can interleave search operations with multiple target tracks.

There is a Fourier transform relationship between the antenna illumination function and the far-field antenna pattern. Hence, tapering the illumination to concentrate power near the center of the antenna suppresses sidelobes while reducing the effective antenna aperture area. The phase and amplitude control of the antenna illumination determines the achievable sidelobe suppression and angle measurement accuracy.

Perturbations in the illumination due to the mechanical and electrical sources distort the illumination function and constrain performance in these areas. Mechanical illumination error sources include antenna shape deformation due to sag and thermal effects as well as manufacturing defects. Electrical illumination error is of particular concern in phased arrays where sources include beam steering computational error and phase shifter quantization. Control of both the mechanical and electrical perturbation errors is the key to both low sidelobes and highly accurate angle measurements. Control denotes that either tolerances are closely held and maintained or that there must be some means for monitoring and correction. Phased arrays are attractive for low sidelobe applications since they can provide element-level phase and amplitude control.

TABLE 41.2 Pulse Radar Transmitter Technology

Technology	Mode of Operation	Maximum Frequency (GHz)	Demonstrated Peak/ Average Power (kW)	Typical Gain	Typical Bandwidth
Thermionic					
Magnetron	Oscillator	95	1 MW/500 W @ X-band	n/a	Fixed–10%
Helix traveling wave tube (TWT)	Amplifier	95	4 kW/400 W @ X-band	40–60 dB	Octave/multioctave
Ring-loop TWT	Amplifier	18	8 kW/200 W @ X-band	40–60 dB	5–15%
Coupled-cavity TWT	Amplifier	95	100 kW/25 kW @ X-band	40–60 dB	5–15%
Extended interaction oscillator (EIO)	Oscillator	220	1 kW/10 W @ 95 GHz	n/a	0.2% (elec.) 4% (mech.)
Extended interaction Klystron (EIK)	Amplifier	140	1 kW/10 W @ 95 GHz	40–50 dB	0.5–1%
Klystron	Amplifier	35	50 kW/5 kW @ X-band	30–60 dB	0.1–2% (inst.) 1–10% (mech.)
Crossed-field amplifier (CFA)	Amplifier	18	500 kW/1 kW @ X-band	10–20 dB	5–15%
Solid state					
Silicon BJT	Amplifier	5	300 W/30 W @1 GHz	5–10 dB	10–25%
GaAs FET	Amplifier	30	15 W/5 W @ X-band	5–10 dB	5–20%
Impatt diode	Oscillator	140	30 W/10 W @ X-band	n/a	Fixed–5%

Source: Tracy V. Wallace, Georgia Tech Research Institute, Atlanta, Georgia.

Transmitter

The transmitter function is to amplify waveforms to a power level sufficient for target detection and estimation. There is a general trend away from tube-based transmitters toward solid-state transmitters. In particular, solid-state transmit/receive modules appear attractive for constructing phased array radar systems. In this case, each radiating element is driven by a module that contains a solid-state transmitter, phase shifter, low-noise amplifier, and associated control components. Active arrays built from such modules appear to offer significant reliability advantages over radar systems driven from a single transmitter. However, microwave tube technology continues to offer substantial advantages in power output over solid-state technology. Transmitter technologies are summarized in Table 41.2.

Receiver and Exciter

This subsystem contains the precision timing and frequency reference source or sources used to derive the master oscillator and local oscillator reference frequencies. These reference frequencies are used to downconvert received signals in a multiple-stage superheterodyne architecture to accommodate signal amplification and interference rejection. The receiver front end is typically protected from overload during transmission through the combination of a circulator and a transmit/receive switch.

The exciter generates the waveforms for subsequent transmission. As in signal processing, the trend is toward programmable digital signal synthesis because of the associated flexibility and performance stability.

Signal and Data Processing

Digital processing is generally divided between two processing subsystems, i.e., signals and data, according to the algorithm structure and throughput demands. Signal processing includes pulse compression, Doppler filtering, and detection threshold estimation and testing. Data processing includes track filtering, user interface support, and such specialized functions as electronic counter-counter measures (ECCM) and built-in test (BIT), as well as the resource management process required to control the radar system.

The signal processor is often optimized to perform the repetitive complex multiply-and-add operations associated with the fast Fourier transform (FFT). FFT processing is used for implementing **pulse compression** via fast convolution and for Doppler filtering. Fast convolution consists of taking the FFT of the digitized receiver output, multiplying it by the stored FFT of the desired filter function, and then taking the inverse FFT

of the resulting product. Fast convolution results in significant computational saving over performing the time-domain convolution of returns with the filter function corresponding to the matched filter. The signal processor output can be characterized in terms of range gates and Doppler filters corresponding approximately to the range and Doppler resolution, respectively.

In contrast, the radar data processor typically consists of a general-purpose computer with a real-time operating system. Fielded radar data processors range from microcomputers to mainframe computers, depending on the requirements of the radar system. Data processor software and hardware requirements are significantly mitigated by off loading timing and control functions to specialized hardware. This timing and control subsystem typically functions as the two-way interface between the data processor and the other radar subsystems. The increasing inclusion of BIT (built-in-test) and built-in calibration capability in timing and control subsystem designs promises to result in significant improvement in fielded system performance.

Radar Performance Prediction

Radar Line-of-Sight

With the exception of over-the-horizon (OTH) radar systems, which exploit either sky-wave bounce or ground-wave propagation modes and sporadic ducting effects at higher frequencies, surface and airborne platform radar operation is limited to the refraction-constrained line of sight. Atmospheric refraction effects can be closely approximated by setting the earth's radius to 4/3 its nominal value in estimating horizon-limited range. The resulting line-of-sight range is depicted in Fig. 41.3 for a surface-based radar, an airborne surveillance radar, and a space-based radar.

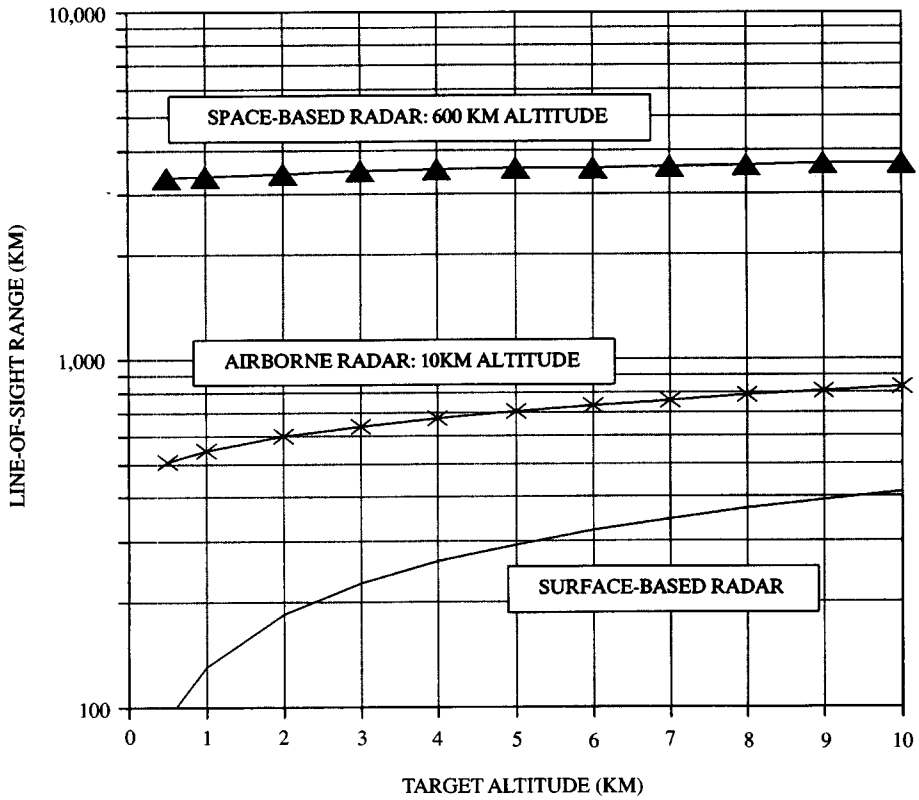


FIGURE 41.3 Maximum line-of-sight range for surface-based radar, an airborne surveillance radar, and a space-based radar.

As evident in the plot, airborne and space-based surveillance radar systems offer significant advantages in the detection of low-altitude targets that would otherwise be masked by earth curvature and terrain features from surface-based radars. However, efficient clutter rejection techniques must be used in order to detect targets since surface clutter returns will be present at almost all ranges of interest.

Radar Range Equation

The radar range equation is commonly used to estimate radar system performance, given that line-of-sight conditions are satisfied. This formulation essentially computes the signal-to-noise ratio (S/N) at the output of the radar signal processor. In turn, S/N is used to provide estimates of radar detection and position measurement performance as described in the subsections "Detection and Search" and "Estimation and Tracking." S/N can be calculated in terms of the number of pulses coherently integrated over a single coherent processing interval (CPI) using the radar range equation such that

$$S/N = \frac{PDAT_p N_p \sigma}{(4\pi)^2 R^4 L_t L_{rn} L_{sp} k T_s} \tag{41.1}$$

where P is peak transmitter power output, D is directivity of the transmit antenna, A is effective aperture area of the receive antenna in meters squared, T_p is pulse duration, σ is **radar cross section** in square meters, N_p is the number of coherently integrated pulses within the coherent processing interval, R is range to target in meters, L_t is system ohmic and nonohmic transmit losses, L_{rn} is system nonohmic receive losses, L_{sp} is signal processing losses, k is Boltzmann's constant (1.38×10^{-23} K), and T_s is system noise temperature, including receive ohmic losses (kelvin).

At X-band and above it may also be necessary to include propagation loss due to atmospheric absorption [Blake, 1986]. This form of the radar range equation is applicable to radar systems using pulse compression or pulse Doppler waveforms as well as the unmodulated single-pulse case. In many applications, average power is a better measure of system performance than peak power since it indicates the S/N improvement achievable with pulse integration over a given interval of time. Hence, the radar range equation can be modified such that

$$S/N = \frac{P_a DAT_c \sigma}{(4\pi)^2 R^4 L_t L_{rn} L_{sp} k T_s} \tag{41.2}$$

where P_a is average transmitter power and T_c is coherent processing interval (CPI).

The portion of time over which the transmitter is in operation is referred to as the radar duty cycle. The average transmitter power is the product of duty cycle and peak transmitter power. Duty cycle ranges from less than 1% for typical **noncoherent** pulse radars to somewhat less than 50% for high pulse repetition frequency (PRF) pulse Doppler radar systems. High PRF systems are sometimes referred to as interrupted continuous wave (ICW) systems because they operate essentially as a CW radar system with transmitter and receiver alternately turned on and off.

The CPI is the period over which returns are collected for **coherent** processing functions such as integration and Doppler filtering. The CPI can be estimated as the product of the number of coherently integrated pulses and the interval between pulses. Noncoherent integration is less efficient and alters the statistical character of the signal and interference.

Antenna Directivity and Aperture Area

The directivity of the antenna is

$$D = \frac{4\pi A \eta}{\lambda^2} \tag{41.3}$$

TABLE 41.3 Median Target RCS (m^2)

Carrier Frequency, GHz	1–2	3	5	10	17
Aircraft (nose/tail avg.)					
Small propeller	2	3	2.5		
Small jet (Lear)	1	1.5	1	1.2	
T38-twin jet, F5	2	2–3	2	1–2/6	
T39-Sabreliner	2.5		10/8	9	
F4, large fighter	5–8/5	4–20/10	4	4	
737, DC9, MD80	10	10	10	10	10
727, 707, DC8-type	22–40/15	40	30	30	
DC-10-type, 747	70	70	70	70	
Ryan drone				2/1	
Standing man (180 lb)	0.3	0.5	0.6	0.7	0.7
Automobiles	100	100	100	100	100
Ships-incoming ($\times 10^4$ m^2)					
4K tons	1.6	2.3	3.0	4.0	5.4
16K tons	13	18	24	32	43
Birds					
Sea birds	0.002	0.001–0.004	0.004		
Sparrow, starling, etc.	0.001	0.001	0.001	0.001	0.001

Slash marks indicate different set.

where η is aperture efficiency and λ is radar carrier wavelength. Aperture inefficiency is due to the antenna illumination factor.

The common form of the radar range equation uses power gain rather than directivity. Antenna gain is equal to the directivity divided by the antenna losses. In the design and analysis of modern radars, directivity is a more convenient measure of performance because it permits designs with distributed active elements, such as solid-state phased arrays, to be assessed to permit direct comparison with passive antenna systems. Beamwidth and directivity are inversely related; a highly directive antenna will have a narrow beamwidth. For typical design parameters,

$$D = \frac{10^7}{\theta_{az}\theta_{el}} \tag{41.4}$$

where θ_{az} and θ_{el} are the radar azimuth and elevation beamwidths, respectively, in milliradians.

Radar Cross Section

In practice, the *radar cross section* (RCS) of a realistic target must be considered a random variable with an associated correlation interval. Targets are composed of multiple interacting scatters so that the composite return varies in magnitude with the constructive and destructive interference of the contributing returns. The target RCS is typically estimated as the mean or median of the target RCS distribution. The associated correlation interval indicates the rate at which the target RCS varies over time. RCS fluctuation degrades target detection performance at moderate to high probability of detection.

The median RCS of typical targets is given in Table 41.3. The composite RCS measured by a radar system may be composed of multiple individual targets in the case of closely spaced targets such as a bird flock.

Loss and System Temperature Estimation

Sources of S/N loss include ohmic and nonohmic (mismatch) loss in the antenna and other radio frequency components, propagation effects, signal processing deviations from matched filter operation, detection thresholding, and search losses. Scan loss in phased array radars is due to the combined effects of the decrease in projected antenna area and element mismatch with increasing scan angle.

TABLE 41.4 Typical Microwave Loss and System Temperature Budgets

	Mechanically Scanned Reflector Antenna	Electronically Scanned Slotted Array	Solid-State Phased Array
Nominal losses			
Transmit loss, L_t (dB)	1	1.5	0.5
Nonohmic receiver loss, L_r (dB)	0.5	0.5	0.1
Signal processing loss, L_{sp} (dB)	1.4	1.4	1.4
Scan loss (dB)	N/A	N/A	30 log [cos (scan angle)]
Search losses, L_{DS}			
Beam shape (dB)	3	3	3
Range gate straddle (dB)	0.5	0.5	0.5
Doppler filter straddle (dB)	0.5	0.5	0.5
Detection thresholding (dB)	1	1	1
System noise temperature (kelvin)	500	600	400

Search operations impose additional losses due to target position uncertainty. Because the target position is unknown before detection, the beam, range gate, and Doppler filter will not be centered on the target return. Hence, straddling loss will occur as the target effectively straddles adjacent resolution cells in range and Doppler. Beamshape loss is a consequence of the radar beam not being pointed directly at the target so that there is a loss in both transmit and receive antenna gain. In addition, detection threshold loss associated with radar system adaptation to interference must be included [Nathanson, 1991]).

System noise temperature estimation corresponds to assessing the system thermal noise floor referenced to the antenna output. Assuming the receiver hardware is at ambient temperature, the system noise temperature can be estimated as

$$T_s = T_a + 290\ (L_{ro}\ F - 1) \tag{41.5}$$

where T_a is the antenna noise temperature, L_{ro} is receive ohmic losses, and F is the receiver noise figure.

In phased array radars, the thermodynamic temperature of the antenna receive beam-former may be significantly higher than ambient, so a more complete analysis is required. The antenna noise temperature is determined by the external noise received by the antenna from solar, atmospheric, earth surface, and other sources.

Table 41.4 provides typical loss and noise temperature budgets for several major radar classes. In general, loss increases with the complexity of the radar hardware between the transmitter/receiver and the antenna radiator. Reflector antennas and active phased arrays impose relatively low loss, while passive array antennas impose relatively high loss.

Resolution and Accuracy

The fundamental resolution capabilities of a radar system are summarized in Table 41.5. In general, there is a trade-off between mainlobe resolution corresponding to the nominal range, Doppler, and angle resolution, and effective dynamic range corresponding to suppression of sidelobe components. This is evident in the use of weighting to suppress Doppler sidebands and angle sidelobes at the expense of broadening the mainlobe and S/N loss.

Cross range denotes either of the two dimensions orthogonal to the radar line of sight. Cross-range resolution in real-aperture antenna systems is closely approximated by the product of target range and radar beamwidth in radians. Attainment of the nominal ISAR/SAR cross-range resolution generally requires complex signal processing to generate a focused image, including correction for scatterer change in range over the CPI.

The best accuracy performance occurs for the case of thermal noise-limited error. The resulting accuracy is the resolution of the radar divided by the square root of the S/N and an appropriate monopulse or interpolation factor. In this formulation, the single-pulse S/N has been multiplied by the number of pulses integrated within the CPI as indicated in Eqs. (41.1) and (41.2).

TABLE 41.5 Resolution and Accuracy

Dimension	Nominal Resolution	Noise-Limited Accuracy
Angle	$\frac{\alpha\lambda}{d}$	$\frac{\alpha\lambda}{dK_m\sqrt{2S/N}}$
Range	$\frac{\alpha C}{2B}$	$\frac{\alpha C}{2BK_i\sqrt{2S/N}}$
Doppler	$\frac{\alpha}{\text{CPI}}$	$\frac{\alpha}{\text{CPI}K_i\sqrt{2S/N}}$
SAR/ISAR	$\frac{\alpha\lambda}{2\Delta\theta}$	$\frac{\alpha\lambda}{2\Delta\theta K_i\sqrt{2S/N}}$

α, taper broadening factor, typically ranging from 0.89 (unweighted) to 1.3 (Hamming); d, antenna extent in azimuth/elevation; B, waveform bandwidth; K_m, monopulse slope factor, typically on the order of 1.5; K_i, interpolation factor, typically on the order of 1.8; $\Delta\theta$, line-of-sight rotation of target relative to radar over CPI.

In practice, accuracy is also constrained by environmental effects, target characteristics, and instrumentation error as well as the available S/N. Environmental effects include multipath and refraction. Target glint is characterized by an apparent wandering of the target position because of coherent interference effects associated with the composite return from the individual scattering centers on the target. Instrumentation error is minimized with alignment and calibration but may significantly constrain track filter performance as a result of the relatively long correlation interval of some error sources.

Radar Range Equation for Search and Track

The radar range equation can be modified to directly address performance in the two primary radar missions: search and track.

Search performance is basically determined by the capability of the radar system to detect a target of specific RCS at a given maximum detection range while scanning a given solid angle extent within a specified period of time. S/N can be set equal to the minimum value required for a given detection performance, $S/N|r$, while R can be set to the maximum required target detection range, $R_{\max}$. Manipulation of the radar range equation results in the following expression:

$$\frac{P_a A}{L_t L_r L_{sp} L_{os} T_s} \geq \left(\frac{S}{N}\right)_r \frac{R_{\max}^4 \Omega}{\sigma\ T_{fs}} \cdot 16\text{k} \tag{41.6}$$

where Ω is the solid angle over which search must be performed (steradians), T_{fs} is the time allowed to search Ω by operational requirements, and L_{os} is the composite incremental loss associated with search.

The left-hand side of the equation contains radar design parameters, while the right-hand side is determined by target characteristics and operational requirements. The right-hand side of the equation is evaluated to determine radar requirements. The left-hand side of the equation is evaluated to determine if the radar design meets the requirements.

The track radar range equation is conditioned on noise-limited angle accuracy as this measure stresses radar capabilities significantly more than range accuracy in almost all cases of interest. The operational requirement is to maintain a given data rate track providing a specified single-measurement angle accuracy for a given number of targets with specified RCS and range. Antenna beamwidth, which is proportional to the radar carrier wavelength divided by antenna extent, impacts track performance since the degree of S/N required for a given measurement accuracy decreases as the beamwidth decreases. Track performance requirements can be bounded as

TABLE 41.6 Selected Waveform Characteristics

	Comments	Time Bandwidth Product	Range Sidelobes (dB)	*S/N* Loss (dB)	Range/Doppler Coupling	ECM/EMI Robustness
Unmodulated	No pulse compression	~1	Not applicable	0	No	Poor
Linear frequency modulation	Linearly swept over bandwidth	>10	Unweighted: –13.5 Weighted: >–40[a]	0 0.7–1.4	Yes	Poor
Nonlinear FM	Multiple variants	Waveform specific	Waveform specific	0	Waveform specific	Fair
Barker	*N*-bit biphase	≤ 13 (N)	–20 log(N)	0	No	Fair
LRS	*N*-bit biphase	~N; >64/pulse[a]	~–10 log (N)	0	No	Good
Frank	*N*-bit polyphase (N = integer^2)	~N	~–10 log ($\pi^2 N$)	0	Limited	Good
Frequency coding	*N* subpulses noncoincidental in time and frequency	~N^2	Waveform specific • Periodic • Pseudorandom	 0.7–1.40 0	Waveform specific	Good

[a]Constraint due to typical technology limitations rather than fundamental waveform characteristics.

$$\frac{P_a A^3}{\lambda^4 L_t L_r L_{sp} T_s} k_m^2 \eta^2 \geq 5k \frac{rN_t R^4}{\sigma \sigma_\theta^2} \tag{41.7}$$

where r is the single-target track rate, N_t is the number of targets under track in different beams, σ_θ is the required angle accuracy standard deviation (radians), and σ is the RCS. In general, a phased array radar antenna is required to support multiple target tracking when $N_t > 1$.

Incremental search losses are suppressed during single-target-per-beam tracking. The beam is pointed as closely as possible to the target to suppress beamshape loss. The tracking loop centers the range gate and Doppler filter on the return. Detection thresholding loss is minimal since the track range window is small.

Radar Waveforms

Pulse Compression

Typical pulse radar waveforms are summarized in Table 41.6. In most cases, the signal processor is designed to closely approximate a matched filter. As indicated in Table 41.5, the range and Doppler resolution of any match-filtered waveform are inversely proportional to the waveform bandwidth and duration, respectively. Pulse compression, using modulated waveforms, is attractive since *S/N* is proportional to pulse duration rather than bandwidth in matched filter implementations. Ideally, the intrapulse modulation is chosen to attain adequate range resolution and range sidelobe suppression performance while the pulse duration is chosen to provide the required sensitivity. Pulse compression waveforms are characterized as having a time bandwidth product (TBP) significantly greater than unity, in contrast to an unmodulated pulse, which has a TBP of approximately unity.

Pulse Repetition Frequency

The radar system pulse repetition frequency (PRF) determines its ability to unambiguously measure target range and range rate in a single CPI as well as determining the inherent clutter rejection capabilities of the radar system. In order to obtain an unambiguous measurement of target range, the interval between radar pulses (1/PRF) must be greater than the time required for a single pulse to propagate to a target at a given range and back. The maximum unambiguous range is then given by $C/(2 \cdot \text{PRF})$ where C is the velocity of electromagnetic propagation.

Returns from moving targets and clutter sources are offset from the radar carrier frequency by the associated Doppler frequency. As a function of range rate, $R\cdot$, the Doppler frequency, f_D, is given by $2R\cdot/\lambda$. A coherent pulse train samples the returns' Doppler modulation at the PRF. Most radar systems employ parallel sampling in the in-phase and quadrature baseband channels so that the effective sampling rate is twice the PRF. The target's return is folded in frequency if the PRF is less than the target Doppler.

Clutter returns are primarily from stationary or near-stationary surfaces such as terrain. In contrast, targets of interest often have a significant range rate relative to the radar clutter. Doppler filtering can suppress returns from clutter. With the exception of frequency ambiguity, the Doppler filtering techniques used to implement pulse Doppler filtering are quite similar to those described for CW radar in Section 41.2. Ambiguous measurements can be resolved over multiple CPIs by using a sequence of slightly different PRFs and correlating detections among the CPIs [Morris, 1988].

Detection and Search

Detection processing consists of comparing the amplitude of each range gate/Doppler filter output with a threshold. A detection is reported if the amplitude exceeds that threshold. A false alarm occurs when noise or other interference produces an output of sufficient magnitude to exceed the detection threshold. As the detection threshold is decreased, both the detection probability and the false alarm probability increase. S/N must be increased to enhance detection probability while maintaining a constant false alarm probability.

As noted in the subsection "Radar Cross Section," RCS fluctuation effects must be considered in assessing detection performance. The Swerling models which use chi-square probability density functions (PDFs) of 2 and 4 degrees of freedom (DOF) are commonly used for this purpose [Nathanson, 1991]. The Swerling 1 and 2 models are based on the 2 DOF PDF and can be derived by modeling the target as an ensemble of independent scatterers of comparable magnitude. This model is considered representative of complex targets such as aircraft. The Swerling 3 and 4 models use the 4 DOF PDF and correspond to a target with a single dominant scatterer and an ensemble of lesser scatterers. Missiles are sometimes represented by Swerling 2 and 4 models. The Swerling 1 and 3 models presuppose slow fluctuation such that the target RCS is constant from pulse to pulse within a scan. In contrast, the RCS of Swerling 2 and 4 targets is modeled as independent on a pulse to pulse basis.

Single-pulse detection probabilities for nonfluctuating, Swerling 1/2, and Swerling 3/4 targets are depicted in Fig. 41.4. This curve is based on a typical false alarm number corresponding approximately to a false alarm probability of 10^{-6}. The difference in S/N required for a given detection probability for a fluctuating target relative to the nonfluctuating case is termed the fluctuation loss.

The detection curves presented here and in most other references presuppose noise-limited operation. In many cases, the composite interference present at the radar system output will be dominated by clutter returns or electromagnetic interference such as that imposed by hostile electronic countermeasures. The standard textbook detection curves cannot be applied in these situations unless the composite interference is statistically similar to thermal noise with a Gaussian PDF and a white power spectral density. The presence of non-Gaussian interference is generally characterized by an elevated false alarm probability. Adaptive detection threshold estimation techniques are often required to search for targets in environments characterized by such interference.

Estimation and Tracking

Measurement Error Sources

Radars measure target range and angle position and, potentially, Doppler frequency. Angle measurement performance is emphasized here since the corresponding cross-range error dominates range error for most practical applications. Target returns are generally smoothed in a tracking filter, but tracking performance is largely determined by the measurement accuracy of the subject radar system. Radar measurement error can be characterized as indicated in Table 41.7.

The radar design and the alignment and calibration process development must consider the characteristics and interaction of these error components. Integration of automated techniques to support alignment and

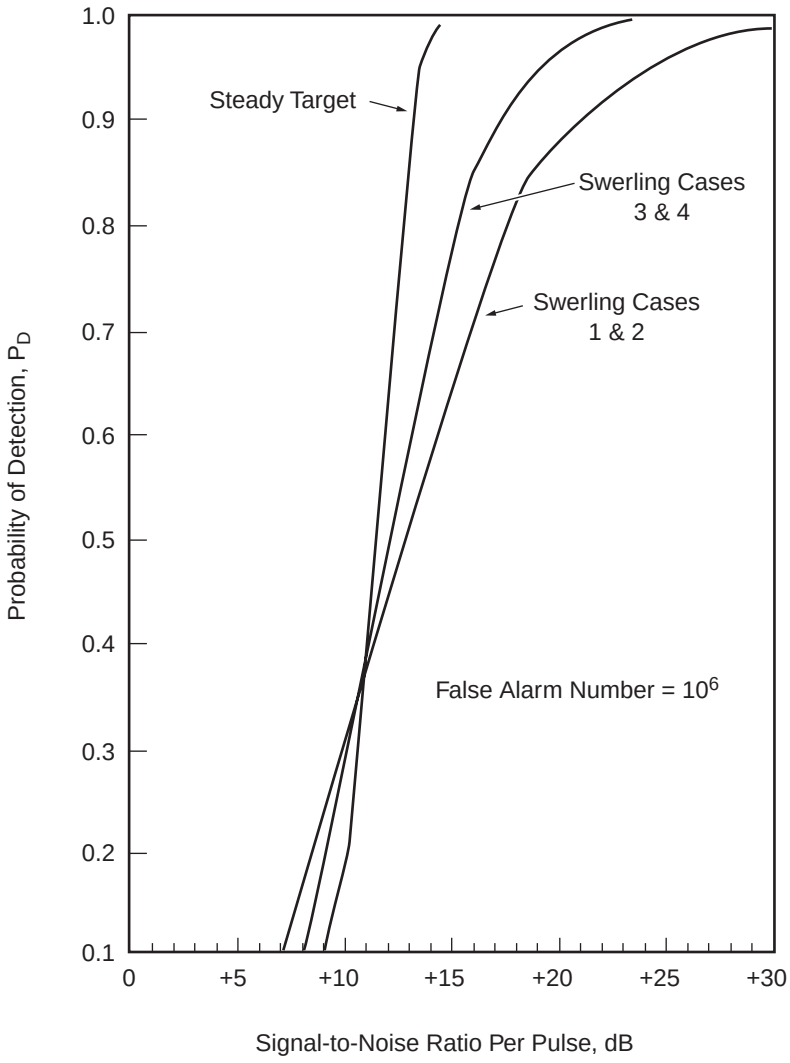


FIGURE 41.4 Detection probabilities for various target fluctuation models.

TABLE 41.7 Radar Measurement Error

Random errors	Those errors that cannot be predicted except on a statistical basis. The magnitude of the random error can be termed the *precision* and is an indication of the repeatability of a measurement.
Bias errors	A systematic error whether due to instrumentation or propagation conditions. A nonzero mean value of a random error.
Systematic error	An error whose quantity can be measured and reduced by calibration.
Residual systematic error	Those errors remaining after measurement and calibration. A function of the systematic and random errors in the calibration process.
Accuracy	The magnitude of the rms value of the residual systematic and random errors.

calibration is an area of strong effort in modern radar design that can lead to significant performance improvement in fielded systems.

As indicated previously, angle measurement generally is the limiting factor in measurement accuracy. Target azimuth and elevation position is primarily measured by a monopulse technique in modern radars though early systems used sequential lobing and conical scanning. Specialized monopulse tracking radars utilizing

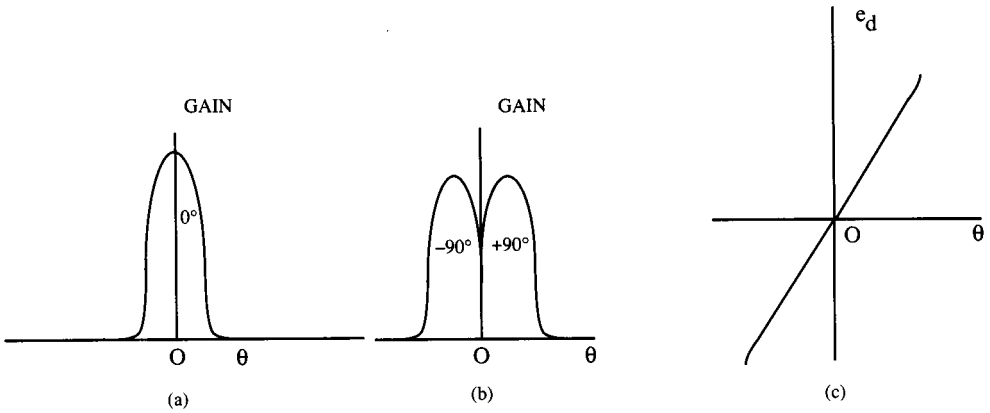


FIGURE 41.5 Monopulse beam patterns and difference voltage: (a) sum (Σ); (b) difference (Δ); (c) difference voltage.

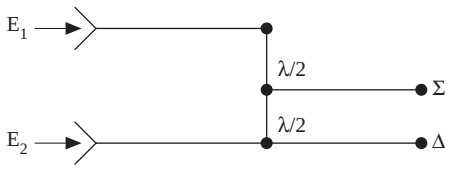


FIGURE 41.6 Monopulse comparator.

reflectors have achieved instrumentation and *S/N* angle residual systematic error as low as 50 μrad. Phased array antennas have achieved a random error of less than 60 μrad, but the composite systematic residual errors remain to be measured. The limitations are primarily in the tolerance on the phase and amplitude of the antenna illumination function.

Figure 41.5 shows the monopulse beam patterns. The first is the received sum pattern that is generated by a feed that provides the energy from the reflector or phased array antenna through two ports in equal amounts and summed in phase in a monopulse comparator shown in Fig. 41.6. The second is the difference pattern generated by providing the energy through the same two ports in equal amounts but taken out with a phase difference of π radians, giving a null at the center. A target located at the center of the same beam would receive a strong signal from the sum pattern with which the target could be detected and ranged. The received difference pattern would produce a null return, indicating the target was at the center of the beam. If the target were off the null, the signal output or difference voltage would be almost linear proportional to the distance off the center (off-axis), as shown in the figure. This output of the monopulse processor is the real part of the dot product of the complex sums and the difference signals divided by the absolute magnitude of the sum signal squared, i.e.,

$$e_d = \mathrm{Re}\left[\frac{\Sigma \cdot \Delta}{|\Sigma|^2}\right] \tag{41.8}$$

The random instrumentation measurement errors in the angle estimator are caused by phase and amplitude errors of the antenna illumination function. In reflector systems, such errors occur because of the position of the feedhorn, differences in electrical length between the feed and the monopulse comparator, mechanical precision of the reflector, and its mechanical rotation. In phased array radars, these errors are a function of the phase shifters, time delay units, and combiners between the antenna elements and the monopulse comparator as well as the precision of the array. Although these errors are random, they may have correlation intervals considerably longer than the white noise considered in the thermal-noise random error and may depend upon

the flight path of the target. For a target headed radially from or toward the radar, the correlation period of angle-measurement instrumental errors is essentially the tracking period. For crossing targets, the correlation interval may be pulse to pulse.

As in the estimate of range, the propagation effects of refraction and multipath also enter into the tracking error. The bias error in range and elevation angle by refraction can be estimated as

$$\Delta R = 0.007\ N_s \text{ cosecant } E_o \text{ (meters)} \tag{41.9}$$

$$\Delta E_o = N_s \cot E_o \text{ (μrad)}$$

where N_s is the surface refractivity and E_o is the elevation angle [Barton and Ward, 1984].

One can calculate the average error in multipath. However, one cannot correct for it as in refraction since the direction of the error cannot be known in advance unless there are controlled conditions such as in a carefully controlled experiment. Hence, the general approach is to design the antenna sidelobes to be as low as feasible and accept the multipath error that occurs when tracking close to the horizon. There has been considerable research to find means to reduce the impact, including using very wide bandwidths to separate the direct path from the multipath return.

Tracking Filter Performance

Target tracking based on processing returns from multiple CPIs generally provides a target position and velocity estimate of greater accuracy than the single-CPI measurement accuracy delineated in Table 41.5. In principle, the error variance of the estimated target position with the target moving at a constant velocity is approximately $4/n \cdot \sigma_m^2$ where n is the number of independent measurements processed by the track filter and σ_m is the single measurement accuracy. In practice, the variance reduction factor afforded by a track filter is often limited to about an order of magnitude because of the reasons summarized in the following paragraphs.

Track filtering generally provides smoothing and prediction of target position and velocity via a recursive prediction-correction process. The filter predicts the target's position at the time of the next measurement based on the current smoothed estimates of position, velocity, and possibly acceleration. The subsequent difference between the measured position at this time and the predicted position is used to update the smoothed estimates. The update process incorporates a weighting vector that determines the relative significance given the track filter prediction versus the new measurement in updating the smoothed estimate.

Target model fidelity and adaptivity are fundamental issues in track filter mechanization. Independent one-dimensional tracking loops may be implemented to control pulse-to-pulse range gate positioning and antenna pointing. The performance of one-dimensional polynomial algorithms, such as the alpha-beta filter, to track targets from one pulse to the next and provide modest smoothing is generally adequate. However, one-dimensional closed-loop tracking ignores knowledge of the equations of motion governing the target so that their smoothing and long-term prediction performance is relatively poor for targets with known equations of motion. In addition, simple one-dimensional tracking-loop filters do not incorporate any adaptivity or measure of estimation quality.

Kalman filtering addresses these shortcomings at the cost of significantly greater computational complexity. Target equations of motion are modeled explicitly such that the position, velocity, and potentially higher-order derivatives of each measurement dimension are estimated by the track filter as a state vector. The error associated with the estimated state vector is modeled via a covariance matrix that is also updated with each iteration of the track filter. The covariance matrix determines the weight vector used to update the smoothed state vector in order to incorporate such factors as measurement S/N and dynamic target maneuvering.

Smoothing performance is constrained by the degree of *a priori* knowledge of the target's kinematic motion characteristics. For example, Kalman filtering can achieve significantly better error reduction against ballistic or orbital targets than against maneuvering aircraft. In the former case the equations of motion are explicitly known, while the latter case imposes motion model error because of the presence of unpredictable pilot or guidance system commands. Similar considerations apply to the fidelity of the track filter's model of radar measurement error. Failure to consider the impact of correlated measurement errors may result in underestimating track error when designing the system.

Defining Terms

Coherent: Integration where magnitude and phase of received signals are preserved in summation.

Noncoherent: Integration where only the magnitude of received signals is summed.

Phased array: Antenna composed of an aperture of individual radiating elements. Beam scanning is implemented by imposing a phase taper across the aperture to collimate signals received from a given angle of arrival.

Pulse compression: The processing of a wideband, coded signal pulse, of initially long time duration and low-range resolution, to result in an output pulse of time duration corresponding to the reciprocal of the bandwidth.

Radar cross section (RCS): A measure of the reflective strength of a radar target; usually represented by the symbol σ, measured in square meters, and defined as 4π times the ratio of the power per unit solid angle scattered in a specified direction of the power unit area in a plane wave incident on the scatterer from a specified direction.

Related Topics

35.1 Maxwell Equations • 69.1 Modulation and Demodulation

References

D.K. Barton and H.R. Ward, *Handbook of Radar Measurement,* Dedham, Mass.: Artech, 1984.
L.V. Blake, *Radar Range-Performance Analysis,* Dedham, Mass.: Artech, 1986.
J.L. Eaves and E.K. Reedy, Eds., *Principles of Modern Radar,* New York: Van Nostrand, 1987.
G.V. Morris, *Airborne Pulsed Doppler Radar,* Dedham, Mass.: Artech, 1988.
F.E. Nathanson, *Radar Design Principles,* 2nd ed., New York: McGraw-Hill, 1991.

Further Information

M.I. Skolnik, Ed., *Radar Handbook,* 2nd ed., New York: McGraw-Hill, 1990.
IEEE Standard Radar Definitions, IEEE Standard 686-1990, April 20, 1990.

41.2 Continuous Wave Radar

James C. Wiltse

Continuous wave (CW) radar employs a transmitter which is on all or most of the time. Unmodulated CW radar is very simple and is able to detect the **Doppler-frequency shift** in the return signal from a target which has a component of motion toward or away from the transmitter. While such a radar cannot measure range, it is used widely in applications such as police radars, motion detectors, burglar alarms, proximity fuzes for projectiles or missiles, illuminators for semiactive missile guidance systems (such as the Hawk surface-to-air missile), and scatterometers (used to measure the scattering properties of targets or clutter such as terrain surfaces) [Nathanson, 1991; Saunders, 1990; Ulaby and Elachi, 1990].

Modulated versions include frequency-modulated (FM/CW), interrupted frequency-modulated (IFM/CW), and phase-modulated. Typical waveforms are indicated in Fig. 41.7. Such systems are used in altimeters, Doppler navigators, proximity fuzes, over-the-horizon radar, and active seekers for terminal guidance of air-to-surface missiles. The term *continuous* is often used to indicate a relatively long waveform (as contrasted to pulse radar using short pulses) or a radar with a high duty cycle (for instance, 50% or greater, as contrasted with the typical duty cycle of less than 1% for the usual pulse radar). As an example of a long waveform, planetary radars may transmit for up to 10 hours and are thus considered to be CW [Freiley et al., 1992]. Another example is interrupted CW (or **pulse-Doppler**) radar, where the transmitter is pulsed at a high rate for 10 to 60% of the total time [Nathanson, 1991]. All of these modulated CW radars are able to measure range.

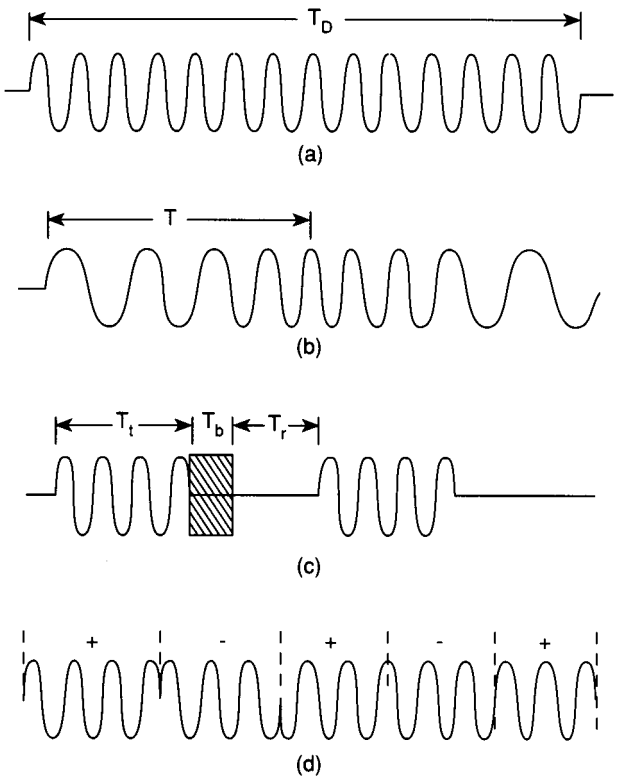


FIGURE 41.7 Waveforms for the general class of CW radar: (a) continuous sine wave CW; (b) frequency modulated CW; (c) interrupted CW; (d) binary phase-coded CW.

The first portion of this section discusses concepts, principles of operation, and limitations. The latter portion describes various applications. In general, CW radars have several potential advantages over pulse radars. Advantages include simplicity and the facts that the transmitter leakage is used as the local oscillator, transmitter spectral spread is minimal (not true for wide-deviation FM/CW), and peak power is the same as (or only a little greater than) the average power. This latter situation means that the radar is less detectable by intercepting equipment.

The largest disadvantage for CW radars is the need to provide antenna isolation (reduce spillover) so that the transmitted signal does not interfere with the receiver. In a pulse radar, the transmitter is off before the receiver is enabled (by means of a duplexer and/or receiver-protector switch). Isolation is frequently obtained in the CW case by employing two antennas, one for transmit and one for reception. When this is done, there is also a reduction of close-in clutter return from rain or terrain. A second disadvantage is the existence of noise sidebands on the transmitter signal which reduce sensitivity because the Doppler frequencies are relatively close to the carrier. This is considered in more detail below.

CW Doppler Radar

If a sine wave signal were transmitted, the return from a moving target would be Doppler-shifted in frequency by an amount given by the following equation:

$$f_d = \frac{2v_r f_T}{c} = \text{Doppler frequency} \tag{41.10}$$

where f_T = transmitted frequency; c = velocity of propagation, 3×10^8 m/s; and v_r = radial component of velocity between radar and target.

Radio Engineering during World War II

No single event had a greater effect on electrical engineering than the Second World War. The years from 1939 to 1945 saw a radical change in the field of electrical engineering as it was transformed from a specialty with well-defined applications, primarily in power and communications, into the source for the most powerful and pervasive technologies of the 20th century.

In the heat of war, radio engineering was transformed into electronics. Electronics became a technology to harness the most advanced and subtle knowledge of the very parts of matter itself, manipulating electrons and electromagnetic waves in an effort not simply to communicate, but to detect, control, and even as some saw it, think. The tremendous pressures of wartime development forged a new relationship between engineers and physical scientists. More and more the realms and tasks of both overlapped and advances in electronics made use of the latest findings, theories, and techniques of physicists and chemists, while scientific discovery came to rely progressively more on the instrumentation created by engineers. This merging of science and technology was one of the war's greatest legacies and has continued to shape our times.

The enormous demands that the war put on the world also marked the indispensable and strategic place of electric power. Electric power rose to the status of necessity, not only increasing the general industrial consumption of power, but also highlighting specialized uses of electricity, such as the production of aluminum and explosives, that were critical to the pursuit of the war. In Europe, the targeting of power plants and dams by both allied and axis bombers provided proof of electricity's central place in modern warfare.

The postwar years were ones of growth and change, accompanied by tensions and conflicts both within the engineering community and in society at large. The war was, again, followed by unprecedented prosperity, but at this time it was in a world where the dangers and possible consequences of international conflict were distressingly obvious. The efforts of engineers were, therefore, divided between the creation of a consumer society, powered by electricity and tuned by electronics, and the demands of national and international security. Alongside this division was another division of the engineering community. The split between the AIEE and the IRE became less and less justifiable and in the coming decades, this problem was solved, as engineers everywhere recognized their common interests. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

TABLE 41.8 Doppler Frequencies for Several Transmitted Frequencies and Various Relative Speeds

Microwave Frequency—f_T	Relative Speed			
	1 m/s	300 m/s	1 mph	600 mph
3 GHz	20 Hz	6 kHz	8.9 Hz	5.4 kHz
10 GHz	67 Hz	20 kHz	30 Hz	17.9 kHz
35 GHz	233 Hz	70 kHz	104 Hz	63 kHz
95 GHz	633 Hz	190 kHz	283 Hz	170 kHz

Using Eq. (41.10) the Doppler frequencies have been calculated for several speeds and are given in Table 41.8.

As may be seen, the Doppler frequencies at 10 GHz (X-band) range from 30 Hz to about 18 kHz for a speed range between 1 and 600 mph. The spectral width of these Doppler frequencies will depend on target fluctuation and acceleration, antenna scanning effects, frequency variation in oscillators or components (for example, due

to microphonism from vibrations), but most significantly, by the spectrum of the transmitter, which inevitably will have noise sidebands that extend much higher than these Doppler frequencies, probably by orders of magnitude. At higher microwave frequencies the Doppler frequencies are also higher and more widely spread. In addition, the spectra of higher frequency transmitters are also wider, and, in fact, the transmitter noise-sideband problem is usually worse at higher frequencies, particularly at millimeter wavelengths (i.e., above 30 GHz). These characteristics may necessitate frequency stabilization or phase locking of transmitters to improve the spectra.

Simplified block diagrams for CW Doppler radars are shown in Fig. 41.8. The transmitter is a single-frequency source, and leakage (or coupling) of a small amount of transmitter power serves as a local oscillator signal in the mixer. The transmitted signal will produce a Doppler-shifted return from a moving target. In the case of scatterometer measurements, where, for example, terrain reflectivity is to be measured, the relative motion may be produced by moving the radar (perhaps on a vehicle) with respect to the stationary target [Wiltse et al., 1957]. The return signal is collected by the antenna and then also fed to the mixer. After mixing with the transmitter leakage, a difference frequency will be produced which is the Doppler shift. As indicated in Table 41.8, this difference is apt to range from low audio to over 100 kHz, depending on relative speeds and choice of microwave frequency. The Doppler amplifier and filters are chosen based on the information to be obtained, and this determines the amplifier bandwidth and gain, as well as the filter bandwidth and spacing. The transmitter leakage may include reflections from the antenna and/or nearby clutter in front of the antenna, as well as mutual coupling between antennas in the two-antenna case.

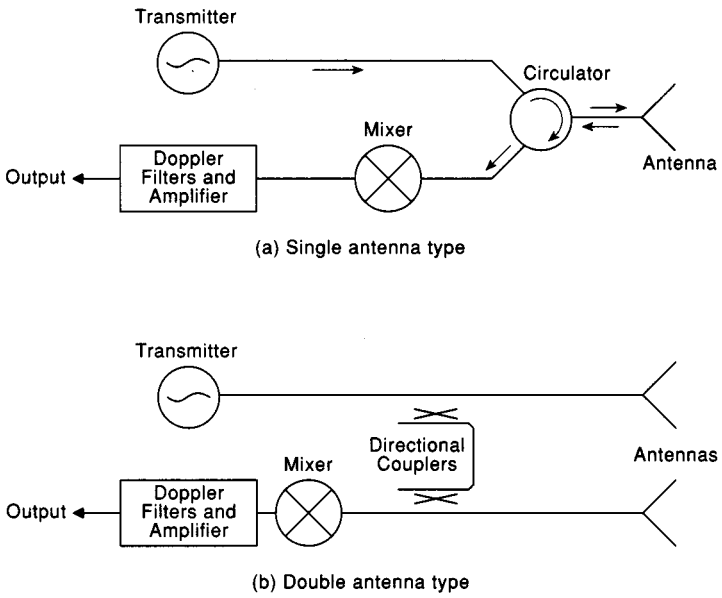


FIGURE 41.8 Block diagrams of CW-Doppler radar systems: (a) single antenna type; (b) double antenna type.

The detection range for such a radar can be obtained from the following [Nathanson, 1991]:

$$R^4 = \frac{\overline{P}_T G_T L_T A_e L_R L_p L_a L_s \delta_T}{(4\pi)^2 k T_s b (S/N)} \tag{41.11}$$

where R = the detection range of the desired target.
$-P_T$ = the average power during the pulse.
G_T = the transmit power gain of the antenna with respect to an omnidirectional radiator.

L_T = the losses between the transmitter output and free space including power dividers, waveguide or coax, radomes, and any other losses not included in A_e.

A_e = the effective aperture of the antenna, which is equal to the projected area in the direction of the target times the efficiency.

L_R = the receive antenna losses defined in a manner similar to the transmit losses.

L_p = the beam shape and scanning and pattern factor losses.

L_a = the two-way-pattern propagation losses of the medium; often expressed as $\exp(-2\propto R)$, where $\propto$ is the attenuation constant of the medium and the factor 2 is for a two-way path.

L_s = signal-processing losses that occur for virtually every waveform and implementation.

δ_T = the radar cross-sectional area of the object that is being detected.

k = Boltzmann's constant (1.38×10^{-23} W-s/K).

T_s = system noise temperature.

b = Doppler filter or *speedgate* bandwidth.

S/N = signal-to-noise ratio.

$S_{\min}$ = the minimum detectable target-signal power that, with a given probability of success, the radar can be said to *detect*, *acquire*, or *track* in the presence of its own thermal noise or some external interference. Since all these factors (including the target return itself) are generally noiselike, the criterion for a detection can be described only by some form of probability distribution with an associated probability of detection P_D and a probability that, in the absence of a target signal, one or more noise or interference samples will be mistaken for the target of interest.

While the Doppler filter should be a matched filter, it usually is wider because it must include the target spectral width. There is usually some compensation for the loss in detectability by the use of postdetection filtering or integration. The S/N ratio for a CW radar must be at least 6 dB, compared with the value of 13 dB required with pulse radars when detecting steady targets [Nathanson, 1991, p. 449].

The Doppler system discussed above has a maximum detection range based on signal strength and other factors, but it cannot measure range. The rate of change in signal strength as a function of range has sometimes been used in fuzes to estimate range closure and firing point, but this is a relative measure.

FM/CW Radar

The most common technique for determining target range is the use of frequency modulation. Typical modulation waveforms include sinusoidal, linear sawtooth, or triangular, as illustrated in Fig. 41.9. For a linear sawtooth, a frequency increasing with time may be transmitted. Upon being reflected from a stationary point target, the same linear frequency change is reflected back to the receiver, except it has a time delay which is related to the range to the target. The time is $T = (2R)/c$, where R is the range. The received signal is mixed with the transmit signal, and the difference or beat frequency (F_b) is obtained. (The sum frequency is much higher and is rejected by filtering.) For a stationary target this is given by

$$F_b = \frac{4R}{c} \cdot \Delta F \cdot F_m \tag{41.12}$$

where ΔF = frequency deviation and F_m = modulation rate.

The beat frequency is constant except near the turn-around region of the sawtooth, but, of course, it is different for targets at different ranges. (If it is desired to have a constant intermediate frequency for different ranges, which is a convenience in receiver design, then the modulation rate or the frequency deviation must be adjusted.) Multiple targets at a variety of ranges will produce multiple-frequency outputs from the mixer and frequently are handled in the receiver by using multiple range-bin filters.

If the target is moving with a component of velocity toward (or away) from the radar, then there will be a Doppler frequency component added to (or subtracted from) the difference frequency (F_b), and the Doppler will be slightly higher at the upper end of the sweep range than at the lower end. This will introduce an

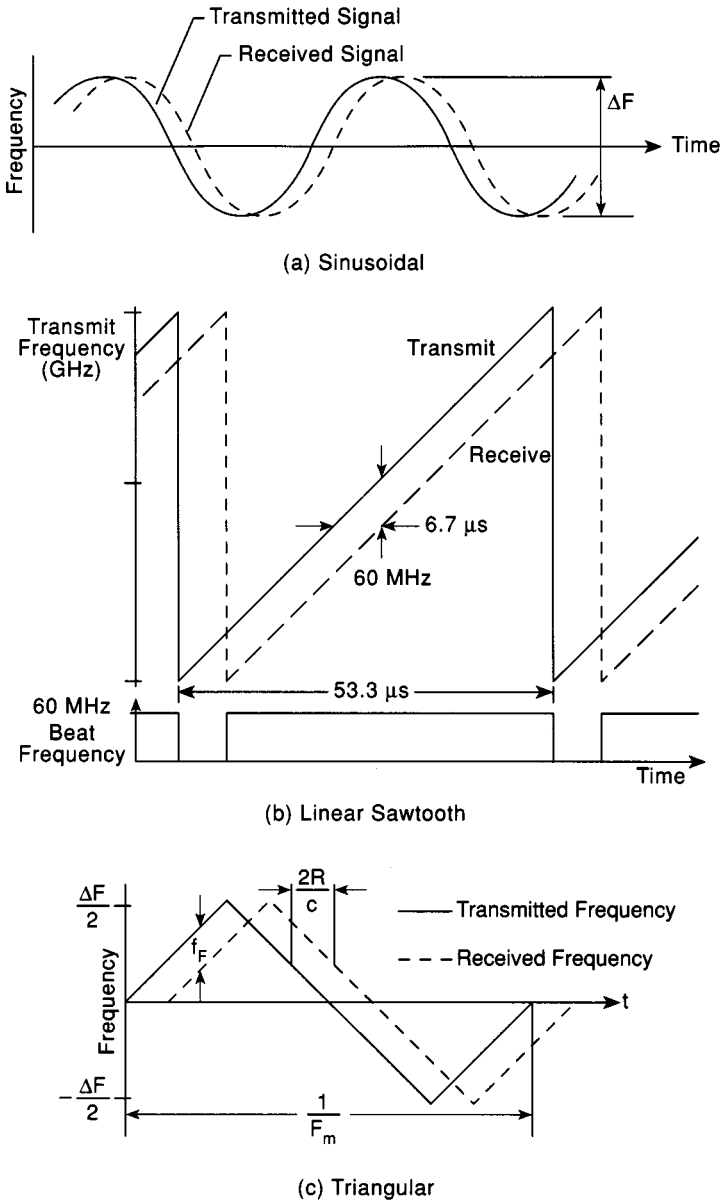


FIGURE 41.9 Frequency vs. time waveforms for FM/CW radar: (a) sinusoidal, (b) linear sawtooth, (c) triangular modulations.

uncertainty or ambiguity in the measurement of range, which may or may not be significant, depending on the parameters chosen and the application. For example, if the Doppler frequency is low (as in an altimeter) and/or the difference frequency is high, the error in range measurement may be tolerable. For the symmetrical triangular waveform, a Doppler less than F_b averages out, since it is higher on one-half of a cycle and lower on the other half. With a sawtooth modulation, only a decrease or increase is noted, since the frequencies produced in the transient during a rapid flyback are out of the receiver passband. Exact analyses of triangular, sawtooth, dual triangular, dual sawtooth, and combinations of these with noise have been carried out by Tozzi [1972]. Specific design parameters are given later in this chapter for an application utilizing sawtooth modulation in a **missile terminal guidance seeker**.

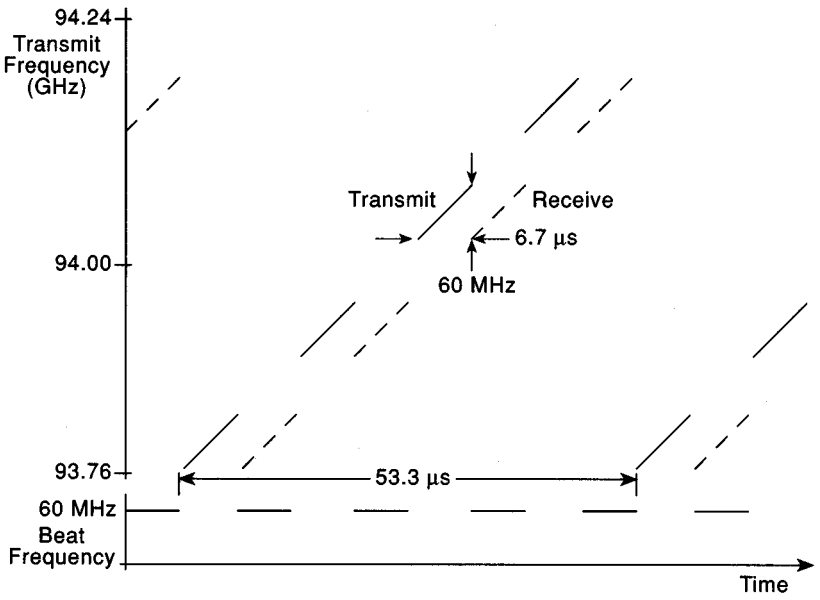


FIGURE 41.10 Interrupted FM/CW waveform. (*Source:* S.O. Piper, "MMW seekers," in *Principles and Applications of Millimeter Wave Radar*, N. Currie and C. E. Brown, Eds., Norwood, Mass.: Artech House, 1987, p. 683. With permission.)

For the case of sinusoidal frequency modulation the spectrum consists of a series of lines spaced away from the carrier by the modulating frequency or its harmonics. The amplitudes of the carrier and these sidebands are proportional to the values of the Bessel functions of the first kind (J_n, $n = 0, \ldots 1, \ldots 2, \ldots 3, \ldots$), whose argument is a function of the modulating frequency and range. By choosing a particular modulating frequency, the values of the Bessel functions and thus the characteristics of the spectral components can be influenced. For instance, the signal variation with range at selected ranges can be optimized, which is important in fuzes. A short-range dependence that produces a rapid increase in signal, greater than that corresponding to the normal range variation, is beneficial in producing well-defined firing signals. This can be accomplished by proper choice of modulating frequency and filtering to obtain the signal spectral components corresponding to the appropriate order of the Bessel function. In a similar fashion, spillover and/or reflections from close-in objects can be reduced by filtering to pass only certain harmonics of the modulating frequency (F_m). Receiving only frequencies near $3F_m$ results in considerable spillover rejection, but at a penalty of 4 to 10 dB in signal-to-noise [Nathanson, 1991].

For the sinusoidal modulation case, Doppler frequency contributions complicate the analysis considerably. For details of this analysis the reader is referred to Saunders [1990] or Nathanson [1991].

Interrupted Frequency-Modulated CW (IFM/CW)

To improve isolation during reception, the IFM/CW format involves preventing transmission for a portion of the time during the frequency change. Thus, there are frequency gaps, or interruptions, as illustrated in Fig. 41.10. This shows a case where the transmit time equals the round-trip propagation time, followed by an equal time for reception. This duty factor of 0.5 for the waveform reduces the average transmitted power by 3 dB relative to using an interrupted transmitter. However, the improvement in the isolation should reduce the system noise by more than 3 dB, thus improving the signal-to-noise ratio [Piper, 1987]. For operation at short range, Piper points out that a high-speed switch is required [1987]. He also points out that the ratio of frequency deviation to beat frequency should be an even integer and that the minimum ratio is typically 6, which produces an out-of-band loss of 0.8 dB.

IFM/CW may be compared with pulse compression radar if both use a wide bandwidth. Pulse compression employs a "long" pulse (i.e., relatively long for a pulse radar) with a large frequency deviation or "chirp." A long pulse is often used when a transmitter is peak-power limited, because the longer pulse produces more

energy and gives more range to targets. The frequency deviation is controlled in a predetermined way (frequently a linear sweep) so that a matched filter can be used in the receiver. The large time-bandwidth product permits the received pulse to be compressed in time to a short pulse in order to make an accurate range measurement. A linear-sawtooth IFM/CW having similar pulse length, frequency deviation, and pulse repetition rate would thus appear similar, although arrived at from different points of view.

Applications

Space does not permit giving a full description of the many applications mentioned at the beginning of this chapter, but several will be discussed.

Radar Proximity Fuzes

Projectiles or missiles designed to be aimed at ships or surface land targets often need a height-of-burst (HOB) sensor (or target detection device) to fire or fuze the warhead at a height of a few meters. There are two primary generic methods of sensing or measuring height to generate the warhead fire signal. The most obvious, and potentially the most accurate, is to measure target round trip propagation delay employing conventional radar ranging techniques. The second method employs a simple CW Doppler radar or variation thereof, with loop gain calibrated in a manner that permits sensing the desired burst height by measurement of target return signal amplitude and/or rate of change. Often the mission requirements do not justify the complexity and cost of the radar ranging approach. Viable candidates are thus narrowed down to variations on the CW doppler fuze.

In its simplest form, the CW Doppler fuze consists of a fractional watt RF oscillator, homodyne detector, Doppler amplifier, Doppler envelope detector, and threshold circuit. When the Doppler envelope amplitude derived from the returned signal reaches the preset threshold, a fire signal is generated. The height at which the fire signal occurs depends on the radar loop gain, threshold level, and target reflectivity. Fuze gain is designed to produce the desired height of burst under nominal trajectory angle and target reflectivity conditions, which may have large fluctuations due to glint effects, and deviations from the desired height due to antenna gain variations with angle, target reflectivity, and fuze gain tolerances are accepted. A loop gain change of 6 dB (2 to 1 in voltage), whether due to a change in target reflection coefficient, antenna gain, or whatever, will result in a 2 to 1 HOB change.

HOB sensitivity to loop gain factors can be reduced by utilizing the slope of the increasing return signal, or so-called rate-of-rise. Deriving HOB solely from the rate-of-rise has the disadvantage of rendering the fuze sensitive to fluctuating signal levels such as might result from a scintillating target. The use of logarithmic amplifiers decreases the HOB sensitivity to the reflectivity range. An early (excessively high) fire signal can occur if the slope of the signal fluctuations equals the rate-of-rise threshold of the fuze. In practice a compromise is generally made in which Doppler envelope amplitude and rate-of-rise contribute in some proportion of HOB.

Another method sometimes employed to reduce HOB sensitivity to fuze loop gain factors and angle of fall is the use of FM sinusoidal modulation of suitable deviation to produce a range correlation function comprising the zero order of a Bessel function of the first kind. The subject of sinusoidal modulation is quite complex, but has been treated in detail by Saunders [1990, pp. 1422–1446 and 144.41]. The most important aspects of fuze design have to do with practical problems such as low cost, small size, ability to stand very high-g accelerations, long life in storage, and countermeasures susceptibility.

Police Radars

Down-the-road police radars, which are of the CW Doppler type, operate at 10.525, 24.150, or in the 33.4 to 36.0 GHz range, frequencies approved in the United States by the Federal Communications Commission. Half-power beamwidths are typically in the 0.21 to 0.31 radian range. The sensitivity is usually good enough to provide a range exceeding 800 meters. Target size has a dynamic range of 30 dB (from smallest cars or motorcycles to large trucks). This means that a large target can be seen well outside the antenna 3-dB point at a range exceeding the range of a smaller target near the center of the beam. Thus there can be uncertainty about which vehicle is the target. Fisher [1992] has given a discussion of a number of the limitations of these systems, but in spite of these factors probably a hundred thousand have been built.

The designs typically have three amplifier gains for detection of short, medium, or maximum range targets, plus a squelch circuit so that sudden spurious signals will not be counted. The Doppler signal is integrated and this direct current provides a speed readout. Provision is made for calibration to assure the accuracy of the readings.

TABLE 41.9 Parameters for Two Commercial Altimeters

Modulation Frequency	Frequency Deviation	Prime Power	Weight (pounds)	Radiated Power
Bendix ALA-52A	150 Hz	130 MHz	30 W	11*
Collins ALT-55	100 kHz	100 MHz	8	350 mW

*Not including antenna and indicator.

Altimeters

A very detailed discussion of FM/CW altimeters has been given by Saunders [1990, pp. 14.34–14.36], in which he has described modern commercial products built by Bendix and Collins. The parameters will be summarized below and if more information is needed, the reader may want to turn to other references [Saunders, 1990; Bendix Corp., 1982; and Maoz et al., 1991]. In his material, Saunders gives a general overview of modern altimeters, all of which use wide-deviation FM at a low modulation frequency. He discusses the limitations on narrowing the antenna pattern, which must be wide enough to accommodate attitude changes of the aircraft. Triangular modulation is used, since for this waveform the Doppler averages out, and dual antennas are employed. There may be a step error or quantization in height (which could be a problem at low altitudes), due to the limitation of counting zero crossings. A difference of one zero crossing (i.e., 1/2 Hz) corresponds to 3/4 meter for a frequency deviation of 100 MHz. Irregularities are not often seen, however, since meter response is slow. Also, if terrain is rough, there will be actual physical altitude fluctuations. Table 41.9 shows some of the altimeters' parameters. These altimeters are not acceptable for military aircraft, because their relatively wide-open front ends make them potentially vulnerable to electronic countermeasures. A French design has some advantages in this respect by using a variable frequency deviation, a difference frequency that is essentially constant with altitude, and a narrowband front-end amplifier [Saunders, 1990].

Doppler Navigators

These systems are mainly sinusoidally modulated FM/CW radars employing four separate downward looking beams aimed at about 15 degrees off the vertical. Because commercial airlines have shifted to nonradar forms of navigation, these units are designed principally for helicopters. Saunders [1990] cites a particular example of a commercial unit operating at 13.3 GHz, employing a Gunn oscillator as the transmitter, with an output power of 50 mW, and utilizing a 30-kHz modulation frequency. A single microstrip antenna is used. A low-altitude equipment (below 15,000 feet), the unit weighs less than 12 pounds. A second unit cited has an output power of 300 mW, dual antennas, dual modulating frequencies, and an altitude capability of 40,000 feet.

Millimeter-Wave Seeker for Terminal Guidance Missile

Terminal guidance for short-range (less than 2 km) air-to-surface missiles has seen extensive development in the last decade. Targets such as tanks are frequently immersed in a clutter background which may give a radar return that is comparable to that of the target. To reduce the clutter return in the antenna footprint, the antenna beamwidth is reduced by going to millimeter wavelengths. For a variety of reasons the choice is usually a frequency near 35 or 90 GHz. Antenna beamwidth is inversely proportional to frequency, so in order to get a reduced beamwidth we would normally choose 90 GHz; however, more deleterious effects at 90 GHz due to atmospheric absorption and scattering can modify that choice. In spite of small beamwidths, the clutter is a significant problem, and in most cases signal-to-clutter is a more limiting condition than signal-to-noise in determining range performance. Piper [1987] has done an excellent job of analyzing the situation for 35- and 90-GHz pulse radar seekers and comparing those with a 90-GHz FM/CW seeker. His FM/CW results will be summarized below.

In his approach to the problem, Piper gives a summary of the advantages and disadvantages of a pulse system compared to the FM/CW approach. Most of these have already been covered in earlier sections, but one difficulty for the FM/CW can be emphasized again. That is the need for a highly linear sweep, and, because of the desire for the wide bandwidth, this requirement is accentuated. The wide bandwidth is desired in order to average the clutter return and to smooth the glint effects. In particular, glint occurs from a complex target because of

the vector addition of coherent signals scattered back to the receiver from various reflecting surfaces. At some angles the vectors may add in phase (constructively) and at others they may cancel, and the effect is specifically dependent on wavelength. For a narrowband system, glint may provide a very large signal change over a small variation of angle, but, of course, at another wavelength it would be different. Thus, very wide bandwidth is desirable from this smoothing point of view, and typical numbers used in millimeter-wave radars are in the 450- to 650-MHz range. Piper chose 480 MHz.

Another tradeoff involves the choice of FM waveform. Here the use of a triangular waveform is undesirable because the Doppler frequency averages out and Doppler compensation is then required. Thus the sawtooth version is chosen, but because of the large frequency deviation desired, the difficulty of linearizing the frequency sweep is made greater. In fact many components must be extremely wideband, and this generally increases cost and may adversely affect performance. On the other hand, the difference frequency (F_b) and/or the intermediate frequency (F_{IF}) will be higher and thus further from the carrier, so the phase noise will be lower. After discussing the other tradeoffs, Piper chose 60 MHz for the beat frequency.

With a linear FM/CW waveform, the inverse of the frequency deviation provides the theoretical time resolution, which is 2.1 ns for 480 MHz (or range resolution of 0.3 meter). For an RF sweep linearity of 300 kHz, the range resolution is actually 5 meters at the 1000-meter nominal search range. (The system has a mechanically scanned antenna.) An average transmitting power of 25 mW was chosen, which was equal to the average power of the 5-W peak IMPATT assumed for the pulse system. The antenna diameter was 15 cm. For a target radar cross section of 20 m^2 and assumed weather conditions, the signal-to-clutter and signal-to-noise ratios were calculated and plotted for ranges out to 2 km and for clear weather or 4 mm per hour rainfall. The results show that for 1 km range the target-to-clutter ratios are higher for the FM/CW case than the pulse system in clear weather or in rain, and target-to-clutter is the determining factor.

Summary Comments

From this brief review it is clear that there are many uses for CW radars, and various types (such as fuzes) have been produced in large quantities. Because of their relative simplicity, today there are continuing trends toward the use of digital processing and integrated circuits. In fact, this is exemplified in articles describing FM/CW radars built on single microwave integrated circuit chips [Maoz et al., 1991; Chang et al., 1995].

Defining Terms

Doppler-frequency shift: The observed frequency change between the transmitted and received signal produced by motion along a line between the transmitter/receiver and the target. The frequency increases if the two are closing and decreases if they are receding.

Missile terminal guidance seeker: Located in the nose of a missile, a small radar with short-range capability which scans the area ahead of the missile and guides it during the terminal phase toward a target such as a tank.

Pulse Doppler: A coherent radar, usually having high pulse repetition rate and duty cycle and capable of measuring the Doppler frequency from a moving target. Has good clutter suppression and thus can see a moving target in spite of background reflections.

Related Topic

69.1 Modulation and Demodulation

References

Bendix Corporation, Service Manual for ALA-52A Altimeter; Design Summary for the ALA-52A, Bendix Corporation, Ft. Lauderdale, Fla., May 1982.

K.W. Chang, H. Wang, G. Shreve, J.G. Harrison, M. Core, A. Paxton, M. Yu, C.H. Chen, and G.S. Dow, "Forward-looking automotive radar using a W-band single-chip transceiver," *IEEE Transactions on Microwave Theory and Techniques,* vol. 43, pp. 1659–1668, July, 1995.

Collins (Rockwell International), ALT-55 Radio Altimeter System; Instruction Book, Cedar Rapids, Iowa, October 1984.

P.D. Fisher, "Improving on police radar," *IEEE Spectrum,* vol. 29, pp. 38–43, July 1992.

A.J. Freiley, B.L. Conroy, D.J. Hoppe, and A.M. Bhanji, "Design concepts of a 1-MW CW X-band transmit/receive system for planetary radar," *IEEE Transactions on Microwave Theory and Techniques,* vol. 40, pp. 1047–1055, June 1992.

B. Maoz, L.R. Reynolds, A. Oki, and M. Kumar, "FM-CW radar on a single GaAs/AlGaAs HBT MMIC chip," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 3–6, June 1991.

F.E. Nathanson, *Radar Design Principles,* New York: McGraw-Hill, 1991, pp. 448–467.

S.O. Piper, "MMW seekers,"in *Principles and Applications of Millimeter Wave Radar*, N. C. Currie and C. E. Brown, Eds., Norwood, Mass.: Artech House, 1987, chap. 14.

W.K. Saunders, "CW and FM radar," in *Radar Handbook*, M.I. Skolnik, Ed., New York: McGraw-Hill, 1990, chap. 14.

L.M. Tozzi, "Resolution in frequency-modulated radars," Ph.D. thesis, University of Maryland, College Park, 1972.

F.T. Ulaby and C. Elachi, *Radar Polarimetry for Geoscience Applications*, Norwood, Mass.: Artech House, 1990, pp. 193–200.

J.C. Wiltse, S.P. Schlesinger, and C.M. Johnson, "Back-scattering characteristics of the sea in the region from 10 to 50 GHz," *Proceedings of the IRE*, vol. 45, pp. 220–228, February 1957.

Further Information

For a general treatment, including analysis of clutter effects, Nathanson's [1991] book is very good and generally easy to read. For extensive detail and specific numbers in various actual cases, Saunders [1990] gives good coverage. The treatment of millimeter-wave seekers by Piper [1987] is excellent, both comprehensive and easy to read.

Agbo, S.O., Cherin, A.H, Tariyal, B.K. “Lightwave”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

42

Lightwave

Samuel O. Agbo
California Polytechnic State University

Allen H. Cherin
Lucent Technologies

Basant K. Tariyal
Lucent Technologies

42.1 Lightwave Waveguides
Ray Theory • Wave Equation for Dielectric Materials • Modes in Slab Waveguides • Fields in Cylindrical Fibers • Modes in Step-Index Fibers • Modes in Graded-Index Fibers • Attenuation • Dispersion and Pulse Spreading

42.2 Optical Fibers and Cables
Introduction • Classification of OpticalFibers and Attractive Features • Fiber Transmission Characteristics • Optical Fiber Cable Manufacturing

42.1 Lightwave Waveguides

Samuel O. Agbo

Lightwave waveguides fall into two broad categories: dielectric slab waveguides and optical fibers. As illustrated in Fig. 42.1, slab waveguides generally consist of a middle layer (the film) of **refractive index** n_1 and lower and upper layers of refractive indices n_2 and n_3, respectively.

Optical fibers are slender glass or plastic cylinders with annular cross sections. The core has a refractive index, n_1, which is greater than the refractive index, n_2, of the annular region (the cladding). Light propagation is confined to the core by total internal reflection, even when the fiber is bent into curves and loops. Optical fibers fall into two main categories: step-index and graded-index (GRIN) fibers. For step-index fibers, the refractive index is constant within the core. For GRIN fibers, the refractive index is a function of radius r given by

$$n(r) = \begin{cases} n_1\left[1 - 2\Delta\left(\dfrac{r}{a}\right)^{\alpha}\right]^{1/2} & ; r < a \\ n_1(1-2\Delta)^{1/2} = n_2 ; & a < r \end{cases} \tag{42.1}$$

In Eq. (42.1), Δ is the **relative refractive index difference**, a is the core radius, and α defines the type of graded-index profile. For triangular, parabolic, and step-index profiles, α is, respectively, 1, 2, and ∞. Figure 42.2 shows the raypaths in step-index and graded-index fibers and the cylindrical coordinate system used in the analysis of lightwave propagation through fibers. Because rays propagating within the core in a GRIN fiber undergo progressive refraction, the raypaths are curved (sinusoidal in the case of parabolic profile).

Ray Theory

Consider Fig 42.3, which shows possible raypaths for light coupled from air (refractive index n_0) into the film of a slab waveguide or the core of a step-index fiber. At each interface, the transmitted raypath is governed by Snell's law. As θ_0 (the acceptance angle from air into the waveguide) decreases, the angle of incidence θ_i increases

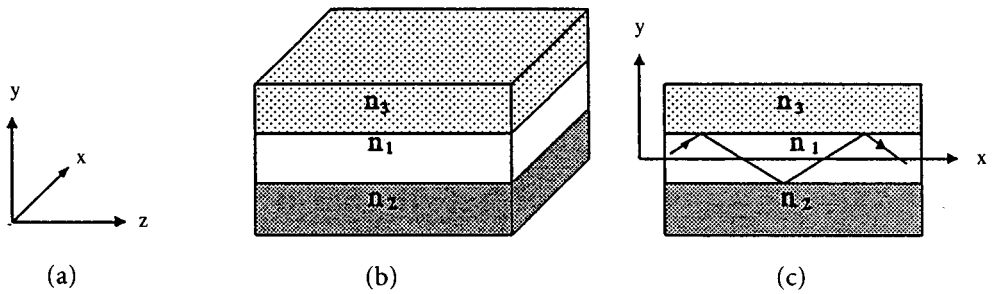


FIGURE 42.1 Dielectric slab waveguide: (a) the Cartesian coordinates used in analysis of slab waveguides; (b) the slab waveguide; (c) light guiding in a slab waveguide.

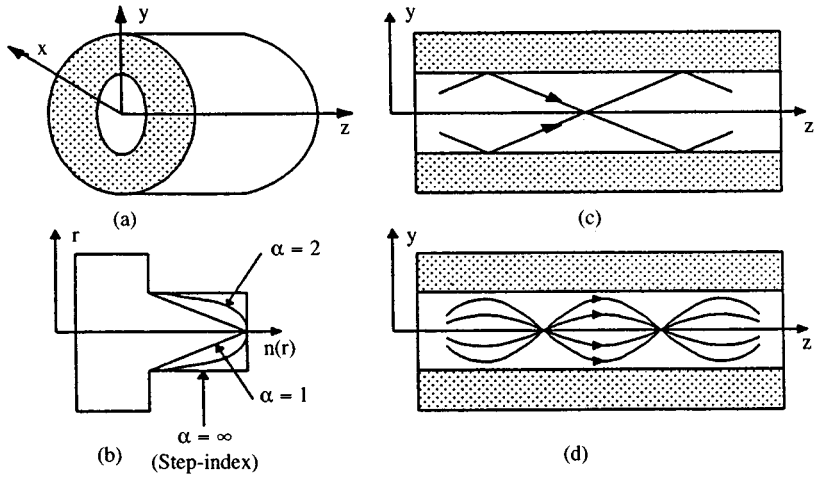


FIGURE 42.2 The optical fiber: (a) the cylindrical coordinate system used in analysis of optical fibers; (b) some graded-index profiles; (c) raypaths in step-index fiber; (d) raypaths in graded-index fiber.

until it equals the critical angle, θ_c, making θ_0 equal to the maximum acceptance angle, θ_a. According to ray theory, all rays with acceptance angles less than θ_a propagate in the waveguide by total internal reflections. Hence, the numerical aperture (NA) for the waveguide, a measure of its light-gathering ability, is given by

$$NA = n_0 \sin \theta a = n_1 \sin\left(\frac{\pi}{2} - \theta_c\right) \tag{42.2}$$

By Snell's law, $\sin \theta_c = n_2/n_1$. Hence,

$$NA = \left[n_1^2 - n_2^2\right]^{1/2} \tag{42.3}$$

For step-index fibers, the preceding analysis applies to **meridional rays**. Skew (nonmeridional) rays have larger maximum acceptance angles, θ_{as}, given by

$$\sin \theta_{as} = \frac{NA}{\cos \gamma} \tag{42.4}$$

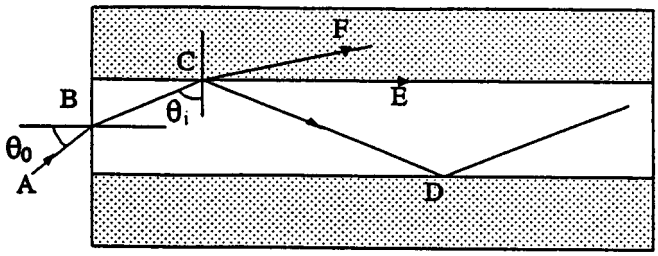


FIGURE 42.3 Possible raypaths for light coupled from air into a slab waveguide or a step-index fiber.

where NA is the numerical aperture for meridional rays and γ is the angle between the core radius and the projection of the ray onto a plane normal to the fiber axis.

Wave Equation for Dielectric Materials

Only certain discrete angles, instead of all acceptance angles less than the maximum acceptance angle, lead to guided propagation in lightwave waveguides. Hence, ray theory is inadequate, and wave theory is necessary, for analysis of light propagation in optical waveguides.

For lightwave propagation in an unbounded dielectric medium, the assumption of a linear, homogeneous, charge-free, and nonconducting medium is appropriate. Assuming also sinusoidal time dependence of the fields, the applicable Maxwell's equations are

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{42.5a}$$

$$\nabla \times \mathbf{H} = j\omega\epsilon\mathbf{E} \tag{42.5b}$$

$$\nabla \times \mathbf{E} = 0 \tag{42.5c}$$

$$\nabla \times \mathbf{H} = 0 \tag{42.5d}$$

The resulting wave equations are

$$\nabla^2\mathbf{E} - \gamma^2\mathbf{E} = 0 \tag{42.6a}$$

$$\nabla^2\mathbf{H} - \gamma^2\mathbf{H} = 0 \tag{42.6b}$$

where

$$\gamma^2 = \omega^2\mu\epsilon = (j\kappa)^2 \tag{42.7}$$

and

$$\kappa = n\kappa_0 = \omega\sqrt{\mu\epsilon} = \frac{\omega}{\nu} \tag{42.8}$$

In Eq. (42.8) κ is the phase propagation constant and n is the refractive index for the medium, while κ_0 is the phase propagation constant for free space. The velocity of propagation in the medium is $\nu = 1/\sqrt{\mu\epsilon}$.

Modes in Slab Waveguides

Consider a plane wave polarized in the y direction and propagating in z direction in an unbounded dielectric medium in the Cartesian coordinates. The vector wave equations (42.6) lead to the scalar equations:

$$\frac{\partial^2 E_y}{\partial z} - \partial^2 E_y = 0 \tag{42.9a}$$

$$\frac{\partial^2 H_x}{\partial z} - \partial^2 H_x = 0 \tag{42.9b}$$

The solutions are

$$E_y = A e^{j(\omega t - \kappa z)} \tag{42.10a}$$

$$H_x = \frac{-E_y}{\eta} = \frac{A}{\eta} e^{j(\omega t - \kappa z)} \tag{42.10b}$$

where A is a constant and $\eta = \sqrt{\mu\epsilon}$ is the intrinsic impedance of the medium.

Because the film is bounded by the upper and lower layers, the rays follow the zigzag paths as shown in Fig. 42.3. The upward and downward traveling waves interfere to create a standing wave pattern. Within the film, the fields transverse to the z axis, which have even and odd symmetry about the x axis, are given, respectively, by

$$E_y = A \cos(hy) e^{j(\omega t - \beta z)} \tag{42.11a}$$

$$E_y = A \sin(hy) e^{j(\omega t - \beta z)} \tag{42.11b}$$

where β and h are the components of κ parallel to and normal to the z axis, respectively. The fields in the upper and lower layers are evanescent fields decaying rapidly with attenuation factors α_3 and α_2, respectively, and are given by

$$E_y = A_3 e^{-\alpha_3\left(y - \frac{d}{2}\right)} e^{j(\omega t - \beta z)} \tag{42.12a}$$

$$E_y = A_2 e^{-\alpha_2\left(y + \frac{d}{2}\right)} e^{j(\omega t - \beta z)} \tag{42.12b}$$

Only waves with raypaths for which the total phase change for a complete (up and down) zigzag path is an integral multiple of 2π undergo constructive interference, resulting in guided modes. Waves with raypaths not satisfying this mode condition interfere destructively and die out rapidly. In terms of a raypath with an angle of incidence $\theta_i = \theta$ in Fig. 42.3, the mode conditions [Haus, 1984] for fields transverse to the z axis and with even and odd symmetry about the x axis are given, respectively, by

$$\tan\left(\frac{hd}{2}\right) = \frac{1}{n_1 \cos\theta}\left[n_1^2 \sin^2\theta - n_2^2\right]^{1/2} \tag{42.13a}$$

$$\tan\left(\frac{hd}{2} - \frac{\pi}{2}\right) = \frac{1}{n_1 \cos\theta}\left[n_1^2 \sin^2\theta - n_2^2\right]^{1/2} \tag{42.13b}$$

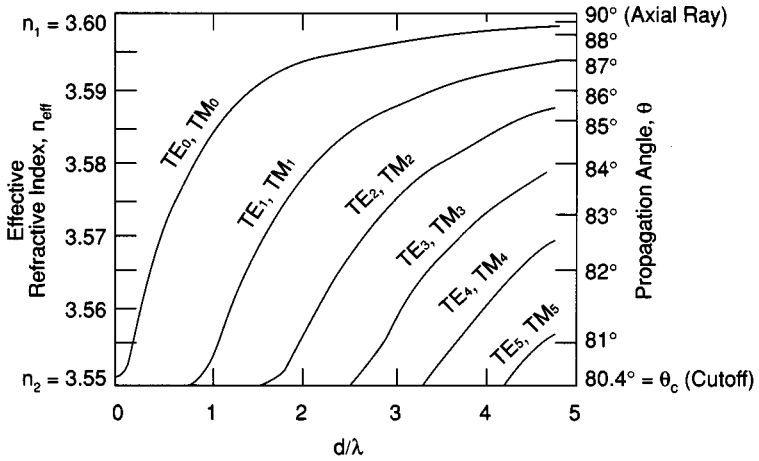


FIGURE 42.4 Mode chart for the symmetric slab waveguide with $n_1 = 3.6$, $n_2 = 3.55$.

where $h = \kappa \cos\theta = (2\pi n_1/\lambda)\cos\theta$ and λ is the free space wavelength.

Equations (42.13a) and (42.13b) are transcendental, have multiple solutions, and are better solved graphically. Let $(d/\lambda)_0$ denote the smallest value of d/λ, the film thickness normalized with respect to the wavelength, satisfying Eqs. (42.13a) and (42.13b). Other solutions for both even and odd modes are given by

$$\left(\frac{d}{\lambda}\right)_m = \left(\frac{d}{\lambda}\right)_0 + \frac{m}{2n_1 \cos\theta} \tag{42.14}$$

where m is a nonnegative integer denoting the order of the mode.

Figure 42.4 [Palais, 1992] shows a **mode chart** for a symmetrical slab waveguide obtained by solving Eqs. (42.13a) and (42.13b). For the TE_m modes, the E field is transverse to the direction (z) of propagation, while the H field lies in a plane parallel to the z axis. For the TM_m modes, the reverse is the case. The highest-order mode that can propagate has a value m given by the integer part of

$$m = \frac{2d}{\lambda}\left[n_1^2 - n_2^2\right]^{1/2} \tag{42.15}$$

To obtain a single-mode waveguide, d/λ should be smaller than the value required for $m = 1$, so that only the $m = 0$ mode is supported. To obtain a multimode waveguide, d/λ should be large enough to support many modes.

Shown in Fig. 42.5 are transverse mode patterns for the electric field in a symmetrical slab waveguide. These are graphical illustrations of the fields given by Eqs. (42.11) and (42.12). Note that, for TE_m, the field has m zeros in the film, and the evanescent field penetrates more deeply into the upper and lower layers for high-order modes.

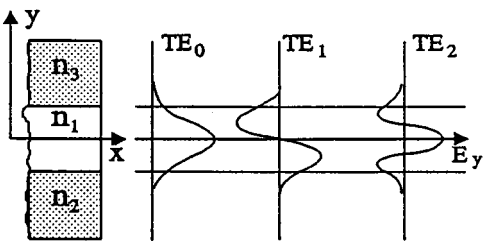


FIGURE 42.5 Transverse mode field patterns in the symmetric slab waveguide.

For asymmetric slab waveguides, the equations and their solutions are more complex than those for symmetric slab waveguides. Shown in Fig. 42.6 [Palais, 1992] is the mode chart for the asymmetric slab waveguide. Note that the TE_m and TM_m modes in this case have different

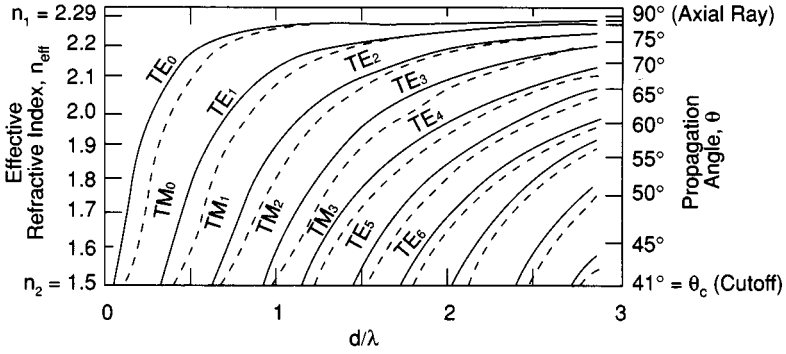


FIGURE 42.6 Mode chart for the asymmetric slab waveguide with $n_1 = 2.29$, $n_2 = 1.5$, and $n_3 = 1.0$.

propagation constants and do not overlap. By contrast, for the symmetric case, TE_m and TM_m modes are degenerate, having the same propagation constant and forming effectively one mode for each value of m.

Figure 42.7 shows typical mode patterns in the asymmetric slab waveguide. Note that the asymmetry causes the evanescent fields to have unequal amplitudes at the two boundaries and to decay at different rates in the two outer layers.

The preceding analysis of slab waveguides is in many ways similar to, and constitutes a good introduction to, the more complex analysis of cylindrical (optical) fibers. Unlike slab waveguides, cylindrical waveguides are bounded in two dimensions rather than one. Consequently, skew rays exist in optical fibers, in addition to the meridional rays found in slab waveguides. In addition to transverse modes similar to those found in slab waveguides, the skew rays give rise to hybrid modes in optical fibers.

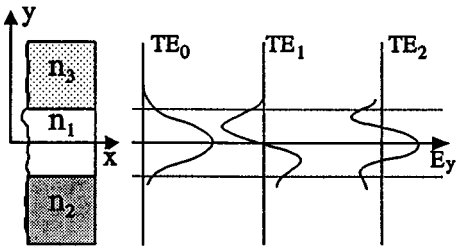


FIGURE 42.7 Transverse mode field patterns in the asymmetric slab waveguide.

Fields in Cylindrical Fibers

Let ψ represent E_z or H_z and β be the component of κ in z direction. In the cylindrical coordinates of Fig. 42.2, with wave propagation along the z axis, the wave equations (42.6) correspond to the scalar equation

$$\frac{\partial^2\psi}{\partial r^2} + \frac{1}{r}\frac{\partial\psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\psi}{\partial\Phi^2} + (\kappa^2 - \beta^2)\psi = 0 \tag{42.16}$$

The general solution to the preceding equation is

$$\psi(r) = C_1 J_\ell(\mathrm{hr}) + C_2 Y_\ell(\mathrm{hr}); \quad \kappa^2 > \beta^2 \tag{42.17a}$$

$$\psi(r) = C_1 I_\ell(\mathrm{qr}) + C_2 K_\ell(\mathrm{qr}); \quad \kappa^2 < \beta^2 \tag{42.17b}$$

In Eqs. (42.17) and (42.17b), J_ℓ and Y_ℓ are Bessel functions of the first kind and second kind, respectively, of order ℓ; I_ℓ and K_ℓ are modified Bessel functions of the first kind and second kind, respectively, of order ℓ; C_1 and C_2 are constants; $h^2 = \kappa^2 - \beta^2$ and $q^2 = \beta^2 - \kappa^2$.

E_z and H_z in a fiber core are given by Eq. (42.17a) or (42.17b), depending on the sign of $\kappa^2 - \beta^2$. For guided propagation in the core, this sign is negative to ensure that the field is evanescent in the cladding. One of the

coefficients vanishes because of asymptotic behavior of the respective Bessel functions in the core or cladding. Thus, with A_1 and A_2 as arbitrary constants, the fields in the core and cladding are given, respectively, by

$$\psi(r) = A_1 J_\ell(\mathrm{h}r) \tag{42.18a}$$

$$\psi(r) = A_2 \kappa_\ell(\mathrm{h}r) \tag{42.18b}$$

Because of the cylindrical symmetry,

$$\psi(r,t) = \psi(r,\phi)\, e^{j(\omega t - \beta z)} \tag{42.19}$$

Thus, the usual approach is to solve for E_z and H_z and then express E_r, E_ϕ, H_r, and H_ϕ in terms of E_z and H_z.

Modes in Step-Index Fibers

Derivation of the exact modal field relations for optical fibers is complex. Fortunately, fibers used in optical communication satisfy the weekly guiding approximation in which the relative index difference, ∇, is much less than unity. In this approximation, application of the requirement for continuity of transverse and tangential electric field components at the core-cladding interface (at $r = a$) to Eqs. (42.18a) and (42.18b) results in the following eigenvalue equation [Snyder, 1969]:

$$haJ_{\ell\pm1}\frac{(ha)}{J_\ell(ha)} = \pm\frac{qa\kappa_{\ell\pm1}(qa)}{\kappa_\ell(qa)} \tag{42.20}$$

Let the normalized frequency V be defined as

$$V = a(q^2 + h^2)^{1/2} = a\kappa_0\left(n_1^2 - n_2^2\right)^{1/2} = \frac{2\pi}{\lambda}a(NA) \tag{42.21}$$

Solving Eq. (42.20) allows β to be calculated as a function of V. Guided modes propagating within the core correspond to $n_2\kappa_0 \le \beta \le n_1\kappa$. The normalized frequency V corresponding to $\beta = n_1\kappa$ is the cut-off frequency for the mode.

As with planar waveguides, *TE* ($E_z = 0$) and *TM* ($H_z = 0$) modes corresponding to meridional rays exist in the fiber. They are denoted by *EH* or *HE* modes, depending on which component, *E* or *H*, is stronger in the plane transverse to the direction or propagation. Because the cylindrical fiber is bounded in two dimensions rather than one, two integers, ℓ and m, are needed to specify the modes, unlike one integer, m, required for planar waveguides. The exact modes, $\mathrm{TE}_{\ell\mathrm{m}}$, $\mathrm{TM}_{\ell\mathrm{m}}$, $\mathrm{EH}_{\ell\mathrm{m}}$, and $\mathrm{HE}_{\ell\mathrm{m}}$, may be given by two linearly polarized modes, $\mathrm{LP}_{\ell\mathrm{m}}$. The subscript ℓ is now such that $\mathrm{LP}_{\ell\mathrm{m}}$ corresponds to $\mathrm{HE}_{\ell+1,\mathrm{m}}$, $\mathrm{EH}_{\ell-1,\mathrm{m}}$, $\mathrm{TE}_{\ell-1,\mathrm{m}}$, and $\mathrm{TM}_{\ell-1,\mathrm{m}}$. In general, there are 2ℓ field maxima around the fiber core circumference and m field maxima along a radius vector. Figure 42.8 illustrates the correspondence between the exact modes and the LP modes and their field configurations for the three lowest LP modes.

Figure 42.9 gives the mode chart for step-index fiber on a plot of the refractive index, β/κ_0, against the normalized frequency. Note that for a single-mode (LP_{01} or HE_{11}) fiber, $V < 2.405$. The number of modes supported as a function of V is given by

$$N = \frac{V^2}{2} \tag{42.22}$$

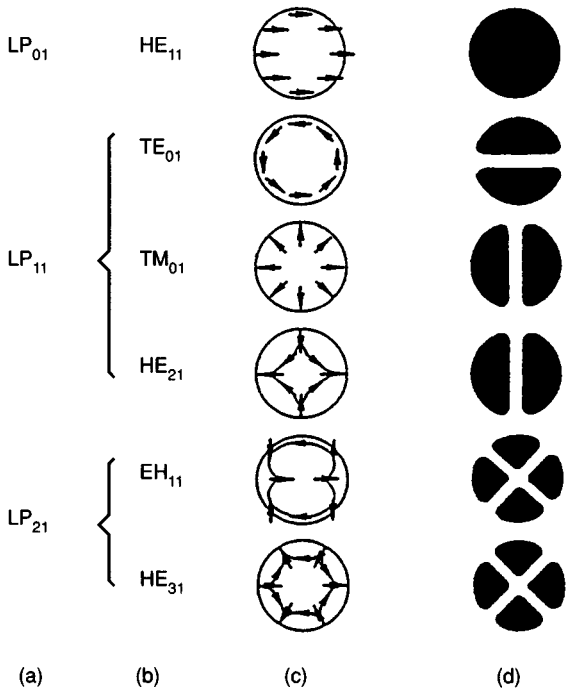


FIGURE 42.8 Transverse electric field patterns and field intensity distributions for the three lowest LP modes in a step-index fiber: (a) mode designations; (b) electric field patterns; (c) intensity distribution. (*Source:* J. M. Senior, *Optical Fiber Communications: Principles and Practice*, Englewood Cliffs, N.J.: Prentice-Hall, 1985, p. 36. With permission.)

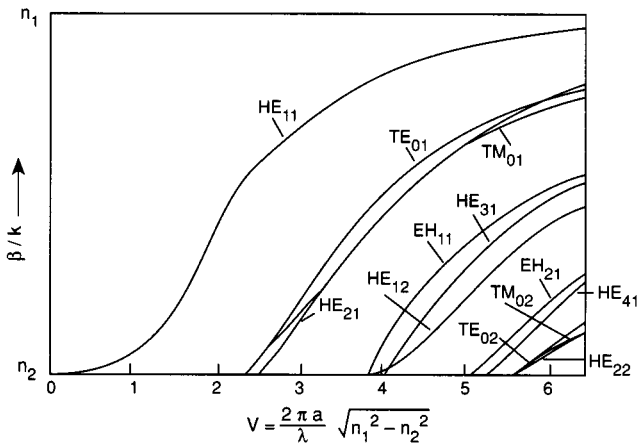


FIGURE 42.9 Mode chart for step-index fibers: $b = (\beta/\kappa_0 - n_2)/(n_1 - n_2)$ is the normalized propagation constant. (*Source:* D. B. Keck, *Fundamentals of Optical Fiber Communications,* M. K. Barnoski, Ed., New York: Academic Press, 1981, p. 13. With permission.)

Modes in Graded-Index Fibers

A rigorous modal analysis for optical fibers based on the solution of Maxwell's equations is possible only for step-index fiber. For graded-index fibers, approximate methods are used. The most widely used approximation is the WKB (Wenzel, Kramers, and Brillouin) method [Marcuse, 1982]. This method gives good modal solutions

Miniature Radar

An inexpensive miniaturized radar system developed at Lawrence Livermore National Labs (LLNL) may become the most successful technology ever privatized by a federal lab, with a potential market for the product estimated at between $100 million and $150 million.

The micropower impulse radar was developed by engineer Tom McEwan as part of a device designed to measure the one billion pulses of light emitted from LLNL's Nova laser in a single second. The system he developed is the size of a cigarette box and consists of about $10 worth of parts. The same measurement had been made previously using $40,000 worth of equipment.

Titan Technologies of Edmonton, AL, Canada, was the first to bring to market a product using the technology when they introduced storage-tank fluid sensors incorporating the system. The new radar allowed Titan to reduce its devices from the size of an apple crate to the size of a softball, and to sell them for one-third the cost of a comparable device. The Federal Highway Administration is preparing to use the radar for highway inspections and the Army Corps of Engineers has contracted with LLNL to use the system for search and rescue radar. Other applications include a monitoring device to check the heartbeats of infants to guard against Sudden Infant Death Syndrome (SIDS), robot guide sensors, automatic on/off switches for bathroom hand dryers, hand-held tools, automobile back-up warning systems, and home security.

AERES, a San Jose-based company, has developed a new approach to ground-penetrating radar using impulse radar. The first application of the technology was an airborne system for detecting underground bunkers. The design can be altered to provide high depth capability for large targets, or high resolution for smaller targets near the surface. This supports requirements in land mine searches and explosive ordinance disposal for the military. AERAS has developed both aircraft and ground-based systems designed for civilian applications as well as military. Underground utility mapping, such as locating pipes and cables; highway and bridge under-surface inspection; and geological and archeological surveying are examples of the possible civilian applications. (Reprinted with permission from *NASA Tech Briefs,* 20(10), 24, 1996.)

for graded-index fiber with arbitrary profiles, when the refractive index does not change appreciably over distances comparable to the guided wavelength [Yariv, 1991]. In this method, the transverse components of the fields are expressed as

$$E_t = \psi(r)\, e^{j\ell\phi}\ \ e^{j(\omega t - \beta z)} \tag{42.23}$$

$$H_t = \frac{\beta}{\omega\mu} E_t \tag{42.24}$$

In Eq. (42.23), ℓ is an integer. Equation (42.16), the scalar wave equation in cylindrical coordinates can now be written with $\kappa = n(r)\,\kappa_0$ as

$$\left[\frac{d^2}{dr^2} + \frac{1}{2}\frac{d}{dr} + p^2(r)\right]\psi(r) = 0 \tag{42.25}$$

where

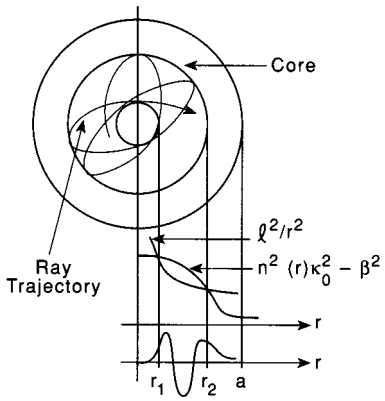


FIGURE 42.10 End view of a skew ray in a graded-index fiber, its graphical solution in the WKB method, and the resulting field that is oscillatory between r_1 and r_2 and evanescent outside that region.

$$p^2(r) = n^2(r)\kappa_0^2 - \frac{\ell^2}{r^2} - \beta^2 \tag{42.26}$$

Let r_1 and r_2 be roots of $p^2(r) = 0$ such that $r_1 < r_2$. A ray propagating in the core does not necessarily reach the core-cladding interface or the fiber axis. In general, it is confined to an annular cylinder bounded by the two caustic surfaces defined by r_1 and r_2. As illustrated in Fig. 42.10, the field is oscillatory within this annular cylinder and evanescent outside it. The fields obtained as solutions to Eq. (42.25) are

$$\psi(r) = \frac{A}{[rp(r)]^{1/2}} \exp\left[-\int_r^{r_1} |p(r)|\, dr\right]; r < r_1 \tag{42.27a}$$

$$\psi(r) = \frac{B}{[rp(r)]^{1/2}} \sin\left[\int_{r_1}^{r} p(r)dr + \frac{\pi}{4}\right]; r_1 < r \tag{42.27b}$$

$$\psi(r) = \frac{C}{[rp(r)]^{1/2}} \sin\left[\int_{r}^{r_2} p(r)dr + \frac{\pi}{4}\right]; r < r_2 \tag{42.27c}$$

$$\psi(r) = \frac{D}{[rp(r)]^{1/2}} \exp\left[-\int_{r_2}^{r} |p(r)|\, dr\right]; r_2 < r \tag{42.27d}$$

Equations (42.27b) and (42.27c) represent fields in the same region. Equating them leads to the mode condition:

$$\int_{r_1}^{r_2} \left[n^2(r)\kappa_0^2 - \frac{\ell^2}{r^2} - \beta^2\right]^{1/2} dr = (2m + 1)\frac{\pi}{2} \tag{42.28}$$

In Eq. (42.28) ℓ and m are the integers denoting the modes. A closed analytical solution of this equation for β is possible only for a few simple graded-index profiles. For other cases, numerical or approximate methods

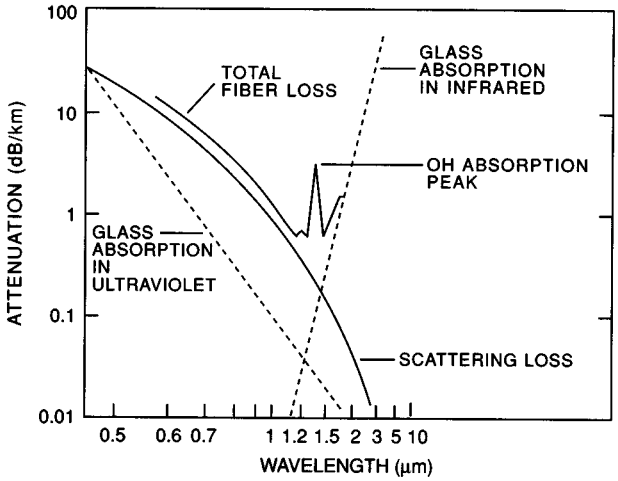


FIGURE 42.11 Attenuation of a germanium-doped low-loss silica glass fiber. (*Source:* H. Osanai et al., "Effects of dopants on transmission loss of low-OH content optical fibers," *Electronic Letters*, vol. 12, no. 21, p. 550, 1976. With permission.)

are used. It can be shown [Marcuse, 1982] that for fiber of graded index profile α, the number of modes supported N_g, and the normalized frequency V, (and hence the core radius) for single mode operations are given, respectively, by

$$N_g = \left(\frac{\alpha}{2+\alpha}\right)\left(\frac{V^2}{2}\right) \tag{42.29}$$

$$V = 2.405\left(\frac{2+\alpha}{\alpha}\right)^{\frac{1}{2}} \tag{42.30}$$

For parabolic (a = 2) index profile Eq. (40.29) give $N_g = \frac{V^2}{4}$, which is half the corresponding number of modes for step index fiber, and Eq. (40.30) gives $V \leq 2.405\sqrt{2}$. Thus, compared with step index fiber, graded index fiber will have larger core radios for single mode operation, and for the same core radius, will support a fewer number of modes.

Attenuation

The assumption of a nonconducting medium for dielectric waveguides led to solutions to the wave equation with no attenuation component. In practice, various mechanisms give rise to losses in lightwave waveguides. These mechanisms contribute a loss factor of $e^{-\alpha z}$ to Eq. (42.10) and comparable field expressions, where α is the attenuation coefficient. The attenuation due to these mechanisms and the resulting total attenuation as a function of wavelength is shown in Fig. 42.11 [Osanai et al., 1976]. Note that the range of wavelengths (0.8 to 1.6 μm) in which communication fibers are usually operated corresponds to a region of low overall attenuation. Brief discussions follow of the mechanisms responsible for the various types of attenuation shown in Fig. 42.11.

Intrinsic Absorption

Intrinsic absorption is a natural property of glass. In the ultraviolet region, it is due to strong electronic and molecular transition bands. In the infrared region, it is caused by thermal vibration of chemical bonds.

Extrinsic Absorption

Extrinsic absorption is caused by metal (Cu, Fe, Ni, Mn, Co, V, Cr) ion impurities and hydroxyl (OH) ion impurity. Metal ion absorption involves electron transition from lower to higher energy states. OH absorption

is caused by thermal vibration of the hydroxyl ion. Extrinsic absorption is strong in the range of normal fiber operation. Thus, it is important that impurity level be limited.

Rayleigh Scattering

Rayleigh scattering is caused by localized variations in refractive index in the dielectric medium, which are small relative to the optic wavelength. It is strong in the ultraviolet region. It increases with decreasing wavelength, being proportional to λ^{-4}. It contributes a loss factor of $\exp(-\alpha_R z)$. The Rayleigh scattering coefficient, α_R, is given by

$$\alpha_R = \left(\frac{8\pi^3}{3\lambda^4}\right)(\delta n^2)^2 \delta V \tag{42.31}$$

where δn^2 is the mean-square fluctuation in refractive index and V is the volume associated with this index difference.

Mie Scattering

Mie scattering is caused by inhomogeneities in the medium, with dimensions comparable to the guided wavelength. It is independent of wavelength.

Dispersion and Pulse Spreading

Dispersion refers to the variation of velocity with frequency or wavelength. Dispersion causes pulse spreading, but other nonwavelength-dependent mechanisms also contribute to pulse spreading in optical waveguides. The mechanisms responsible for pulse spreading in optical waveguides include material dispersion, waveguide dispersion, and multimode pulse spreading.

Material Dispersion

In material dispersion, the velocity variation is caused by some property of the medium. In glass, it is caused by the wavelength dependence of refractive index. For a given pulse, the resulting pulse spread per unit length is the difference between the travel times of the slowest and fastest wavelengths in the pulse. It is given by

$$\Delta\tau = \frac{-\lambda}{c} n''\Delta\lambda = -M\Delta\lambda \tag{42.32}$$

In Eq. 42.32, n'' is the second derivative of the refractive index with respect to λ, $M = (\lambda/c)n''$ is the material dispersion, and $\Delta\lambda$ is the **linewidth** of the pulse. Figure 42.12 shows the wavelength dependence of material dispersion [Wemple, 1979]. Note that for silica, zero dispersion occurs around 1.3 μm, and material dispersion is small in the wavelength range of small fiber attenuation.

Waveguide Dispersion

The effective refractive index for any mode varies with wavelength for a fixed film thickness, for a slab waveguide, or a fixed core radius, for an optical fiber. This variation causes pulse spreading, which is termed waveguide dispersion. The resulting pulse spread is given by

$$\Delta\tau = \frac{-\lambda}{c} n''_{\text{eff}}\Delta\lambda = -M_G\Delta\lambda \tag{42.33}$$

where $M_G = (\lambda/c)n''_{\text{eff}}$ is the waveguide dispersion.

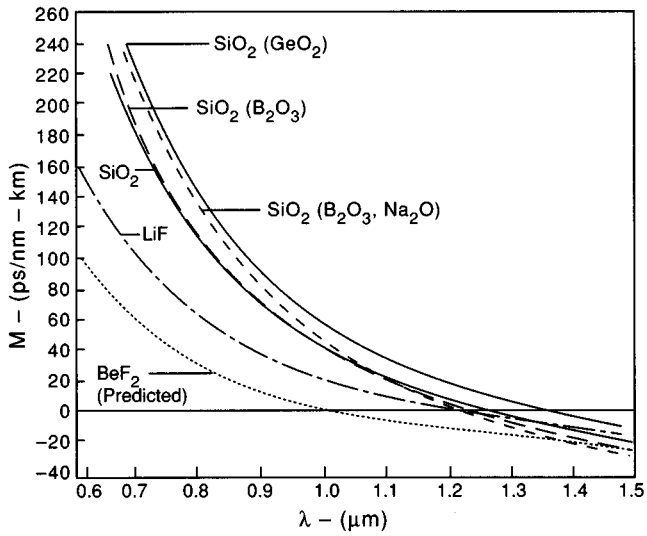


FIGURE 42.12 Material dispersion as a function of wavelength for silica and several solids. (*Source:* S.H. Wemple, "Material dispersion in optical fibers," *Applied Optics*, vol. 18, no. 1, p. 33, 1979. With permission.)

Polarization Mode Dispersion

The HE_{11} propagating in a single mode fiber actually consists of two orthogonally polarized modes, but the two modes have the same effective refractive index and propagation velocity except in birefringent fibers. Birefringent fibers have asymmetric cores or asymmetric refractive index distribution in the core, which result in different refractive indices and group velocities for the orthogonally polarized modes. The different group velocities result in a group delay of one mode relative to the other, known as polarization mode dispersion. Birefringent fibers are polarization preserving and are required for several applications, including coherent optical detection and fiber optic gyroscopes. In high birefringence fibers, polarization dispersion can exceed 1 ns/km. However, in low birefringence fibers, polarization mode dispersion is negligible relative to other pulse spreading mechanisms [Payne et al., 1982].

Multimode Pulse Spreading

In a multimode waveguide, different modes travel different path lengths. This results in different travel times and, hence, in pulse spreading. Because this pulse spreading is not wavelength dependent, it is not usually referred to as dispersion. Multimode pulse spreads are given, respectively, for a slab waveguide, a step-index fiber, and a parabolic graded-index fiber by the following equations:

$$\Delta\tau_{\text{mod}} = \frac{n_1(n_1 - n_2)}{cn_2} \quad \text{(slab waveguide)} \tag{42.34}$$

$$\Delta\tau_{\text{mod}} = \frac{n_1\Delta}{c} \quad \text{(step-index fiber)} \tag{42.35}$$

$$\Delta\tau_{\text{mod}} = \frac{n_1\Delta^2}{8c} \quad \text{(GRIN fiber)} \tag{42.36}$$

Total Pulse Spread

Total pulse spread is the overall effect of material dispersion, waveguide dispersion, and multimode pulse spread. It is given by

$$\Delta\tau_T^2 = \Delta\tau_{\text{mod}}^2 + \Delta\tau_{\text{dis}}^2 \tag{42.37}$$

where

$$\Delta\tau_{\text{dis}} = \text{total dispersion} = -(M + M_G)\Delta\lambda$$

In a multimode waveguide, multimode pulse spread dominates, and dispersion can often be ignored. In a single-mode waveguide, only material and waveguide dispersion exist; material dispersion dominates, and waveguide dispersion can often be ignored.

Total pulse spread imposes an upper limit on the bandwidth of an optical fiber. This upper limit is equal to $1/(2\Delta\tau_T)Hz$.

Defining Terms

Linewidth: The range of wavelengths emitted by a source or present in a pulse.

Meridional ray: A ray that is contained in a plane passing through the fiber axis.

Mode chart: A graphical illustration of the variation of effective refractive index (or, equivalently, propagation angle θ) with normalized thickness d/λ for a slab waveguide or normalized frequency V for an optical fiber.

Refractive index: The ratio of the velocity of light in free space to the velocity of light in a given medium.

Relative refractive index difference: The ratio $(n_1^2 - n_2^2)/2n_1^2 \approx (n_1 - n_2)/n_1$, where $n_1 > n_2$ and n_1 and n_2 are refractive indices.

Related Topics

31.1 Lasers • 37.2 Waveguides

References

H.A. Haus, *Waves and Fields in Optoelectronics*, Englewood Cliffs, N.J.: Prentice-Hall, 1984.

D.B. Keck, "Optical fiber waveguides," in *Fundamentals of Optical Fiber Communications*, 2nd ed., M. K. Barnoski, Ed., New York: Academic Press, 1981.

D. Marcuse, *Light Transmission Optics*, 2nd ed., New York: Van Nostrand Reinhold, 1982.

H. Osanai et al., "Effects of dopants on transmission loss of low-OH-content optical fibers," *Electronic Letters*, vol. 12, no. 21, 1976.

J.C. Palais, *Fiber Optic Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1992.

D.N. Payne, A.J. Barlow, and J.J.R. Hansen, "Development of low-and-high birefringence optical fibers," *IEEE J. Quantum Electron.*, QE-18 no. 4, pp. 477–487, 1982.

J.M. Senior, *Optical Fiber Communications: Principles and Practice*, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

J.M. Snyder, "Asymptotic expressions for eigenfunctions and eigenvalues of a dielectric or optical waveguide," *Trans. IEEE Microwave Theory Tech.*, vol. MTT-17, pp. 1130–1138, 1969.

S.H. Wemple, "Material dispersion in optical fibers," *Applied Optics*, vol. 18, no. 1, p. 33, 1979.

A. Yariv, *Optical Electronics*, 4th ed., Philadelphia: Saunders College Publishing, 1991.

Further Information

IEEE Journal of Lightwave Technology, a bimonthly publication of the IEEE, New York.

IEEE Lightwave Telecommunications Systems, a quarterly magazine of the IEEE, New York.

Applied Optics, a biweekly publication of the Optical Society of America, 2010 Massachusetts Avenue NW, Washington, DC 20036.

D. Macruse, *Theory of Optical Waveguides*, 2nd ed., Boston: Academica Press, 1991.

42.2 Optical Fibers and Cables[1]

Allen H. Cherin and Basant K. Tariyal

Communications using light as a signal carrier and optical fibers as transmission media are termed optical fiber communications. The applications of optical fiber communications have increased at a rapid rate, since the first commercial installation of a fiber-optic system in 1977. Today every major telecommunication company is spending millions of dollars on optical fiber communication systems. In an optical fiber communication system voice, video, or data are converted to a coded pulse stream of light using a suitable light source. This pulse stream is carried by optical fibers to a regenerating or receiving station. At the final receiving station the light pulses are converted to electric signals, decoded, and converted into the form of the original information. Optical fiber communications arc currently used for telecommunications, data communications, military applications, industrial controls, medical applications, and CATV.

Introduction

Since ancient times humans have used light as a vehicle to carry information. Lanterns on ships and smoke signals or flashing mirrors on land are early examples of uses of how humans used light to communicate. It was just over a hundred years ago that Alexander Graham Bell (1880) transmitted a telephone signal a distance greater than 200 m using light as the signal carrier. Bell called his invention a "photophone" and obtained a patent for it. Bell, however, wisely gave up the photophone in favor of the electric telephone. Photophone at the time of its invention could not be exploited commercially because of two basic drawbacks: (1) the lack of a reliable light source and (2) the lack of a dependable transmission medium.

The invention of the laser in 1960 gave a new impetus to the idea of lightwave communications (as scientists realized the potential of the dazzling information-carrying capacity of these lasers). Much research was undertaken by different laboratories around the world during the early 1960s on optical devices and transmission media. The transmission media, however, remained the main problem, until K.C. Kao and G.A. Hockham in 1966 proposed that glass fibers with a sufficiently high-purity **core** surrounded by a lower refractive index **cladding** could be used for transmitting light over long distances. At the time, available glasses had losses of several thousand decibels per kilometer. In 1970, Robert Maurer of Corning Glass Works was able to produce a fiber with a loss of 20 dB/km. Tremendous progress in the production of low-loss optical fibers has been made since then in the various laboratories in the United States, Japan, and Europe, and today optical fiber communication is one of the fastest growing industries. Optical fiber communication is being used to transmit voice, video, and data over long distance as well as within a local network.

Fiber optics appears to be the future method of choice for many communications applications. The biggest advantage of a lightwave system is its tremendous information-carrying capacity. There are already systems that can carry several thousand simultaneous conversations over a pair of optical fibers thinner than human hair. In addition to this extremely high capacity, the lightguide cables are light weight, they are immune to electromagnetic interference, and they are potentially very inexpensive.

A lightwave communication system (Fig. 42.13) consists of a transmitter, a transmission medium, and a receiver. The transmitter takes the coded electronic signal (voice, video, or data) and converts it to the light signal, which is then carried by the transmission medium (an optical fiber cable) to either a **repeater** or the receiver. At the receiving end the signal is detected, converted to electrical pulses, and decoded to the proper output. This article provides a brief overview of the different components used in an optical fiber system, along with examples of various applications of optical fiber systems.

Classification of Optical Fibers and Attractive Features

Fibers that are used for optical communication are waveguides made of transparent dielectrics whose function is to guide light over long distances. An optical fiber consists of an inner cylinder of glass called the core,

[1]This section, including all illustrations, is modified from A. H. Cherin and B. K. Tariyal, "Optical fiber communication," in *Encyclopedia of Telecommunications*, R. A. Meyers, Ed., San Diego: Academic Press, 1988. With permission.

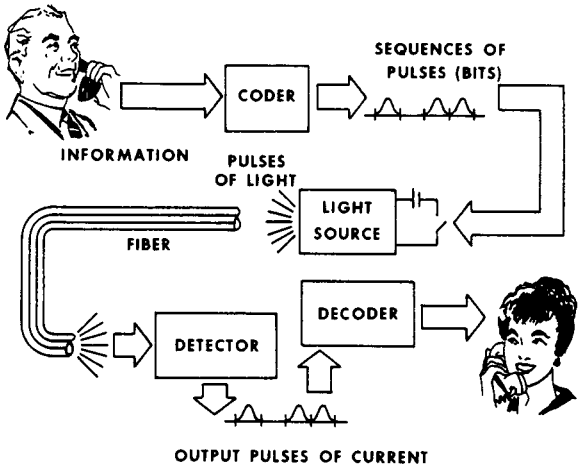


FIGURE 42.13 Schematic diagram of a lightwave communications system.

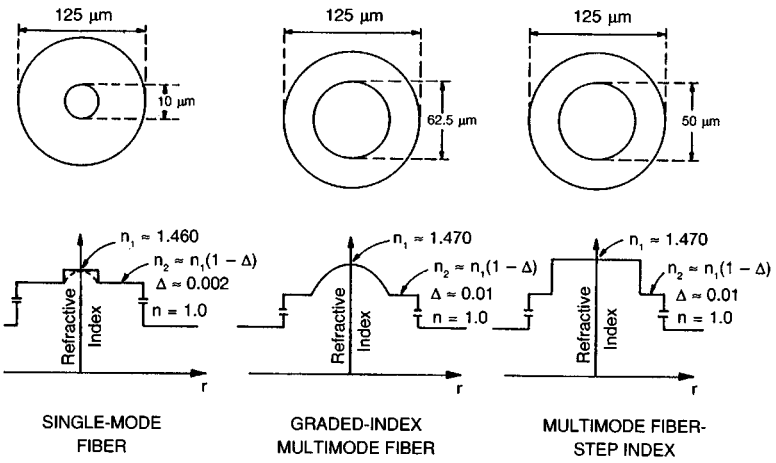


FIGURE 42.14 Geometry of single-mode and multimode fibers.

surrounded by a cylindrical shell of glass of lower refractive index, called the cladding. Optical fibers (lightguides) may be classified in terms of the refractive index profile of the core and whether one **mode** (single-mode fiber) or many modes (multimode fiber) are propagating in the guide (Fig. 42.14). If the core, which is typically made of a high-silica-content glass or a multicomponent glass, has a uniform refractive index n_1, it is called a *step-index fiber.* If the core has a nonuniform refractive index that gradually decreases from the center toward the core-cladding interface, the fiber is called a *graded-index fiber.* The cladding surrounding the core has a uniform refractive index n_2 that is slightly lower than the refractive index of the core region. The cladding of the fiber is made of a high-silica-content glass or a multicomponent glass. Figure 42.14 shows the dimensions and refractive indexes for commonly used telecommunication fibers. Figure 42.15 enumerates some of the advantages, constraints, and applications of the different types of fibers. In general, when the transmission medium must have a very high **bandwidth**—for example, in an undersea or long-distance terrestrial system—a single-mode fiber is used. For intermediate system bandwidth requirements between 200 MHz-km and 2 GHz-km, such as found in local-area networks, either a single-mode or graded-index multimode fiber would be the choice. For applications such as short data links where lower bandwidth requirements are placed on the transmission medium, either a graded-index or a step-index multimode fiber may be used.

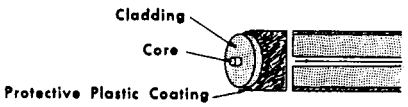
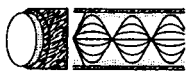
	Single-mode Fiber	Graded-index Multimode Fiber	Step-index Multimode Fiber
			
Source	Laser preferred	Laser or LED	Laser or LED
Bandwidth	Very very large >2 GHz-km	Very large 150 MHz to 2 GHz-km	Large <200 MHz-km
Example of application	Submarine cable system	LANS	Data links

FIGURE 42.15 Applications and characteristics of fiber types.

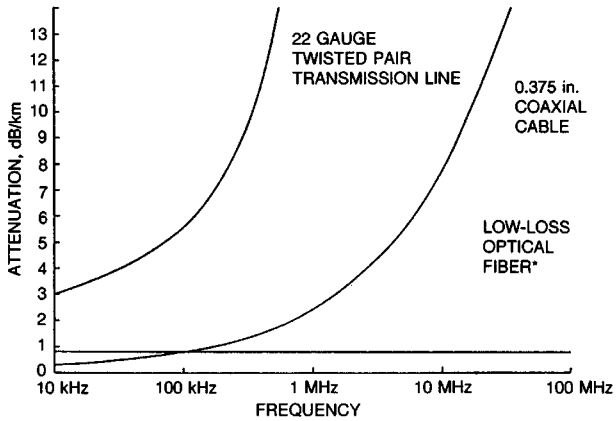


FIGURE 42.16 Attenuation versus frequency for three different transmission media. Asterisk indicates fiber loss at a carrier wavelength of 1.3 μm.

Because or their low loss and wide bandwidth capabilities, optical fibers have the potential for being used wherever twisted wire pairs or coaxial cables are used as the transmission medium in a communication system. If an engineer were interested in choosing a transmission medium for a given transmission objective, he or she would tabulate the required and desired features of alternate technologies that may be available for use in the applications. With that process in mind, a summary of the attractive features and the advantages of optical fiber transmission will be given. Some of these advantages include (a) low loss and high bandwidth; (b) small size and bending radius; (c) nonconductive, nonradiative, and noninductive; (d) light weight; and (e) providing natural growth capability.

To appreciate the low loss and wide bandwidth capabilities of optical fibers, consider the curves of signal **attenuation** versus frequency for three different transmission media shown in Fig. 42.16. Optical fibers have a "flat" transfer function well beyond 100 MHz. When compared with wire pairs of coaxial cables, optical fibers have far less loss for signal frequencies above a few megahertz. This is an important characteristic that strongly influences system economics, since it allows the system designer to increase the distance between regenerators (amplifiers) in a communication system.

The small size, small bending radius (a few centimeters), and light weight of optical fibers and cables are very important where space is at a premium, such as in aircraft, on ships, and in crowded ducts under city streets.

Because optical fibers are dielectric waveguides, they avoid many problems such as radiative interference, ground loops, and, when installed in a cable without metal, lightning-induced damage that exists in other transmission media.

Finally, the engineer using optical fibers has a great deal of flexibility. He or she can install an optical fiber cable and use it initially in a low-capacity (low-bit-rate) system. As the system needs grow, the engineer can take advantage of the broadband capabilities of optical fibers and convert to a high-capacity (high-bit-rate) system by simply changing the terminal electronics.

Fiber Transmission Characteristics

The proper design and operation of an optical communication system using optical fibers as the transmission medium requires a knowledge of the transmission characteristics of the optical sources, fibers, and interconnection devices (connectors, couplers, and splices) used to join lengths of fibers together. The transmission criteria that affect the choice of the fiber type used in a system are signal attenuation, information transmission capacity (bandwidth), and source coupling and interconnection efficiency. Signal attenuation is due to a number of loss mechanisms within the fiber, as shown in Table 42.1, and due to the losses occurring in splices and connectors. The information transmission capacity of a fiber is limited by **dispersion**, a phenomenon that causes light that is originally concentrated into a short pulse to spread out into a broader pulse as it travels along an optical fiber. Source and interconnection efficiency depends on the fiber's core diameter and its **numerical aperture**, a measure of the angle over which light is accepted in the fiber. Absorption and scattering of light traveling through a fiber leads to signal attenuation, the rate of which is measured in decibels per kilometer (dB/km). As can be seen in Fig. 42.17, for both multimode and single-mode fibers, attenuation depends strongly on wavelength. The decrease in scattering losses with increasing wavelength is offset by an increase in material absorption such that attenuation is lowest near 1.55 μm (1550 nm).

TABLE 42.1 Loss Mechanisms

Intrinsic material absorption loss
Ultraviolet absorption tail
Infrared absorption tail
Absorption loss due to impurity ions
Rayleigh scattering loss
Waveguide scattering loss
Microbending loss

The measured values given in Table 42.2 are probably close to the lower bounds for the attenuation of optical fibers. In addition to intrinsic fiber losses, extrinsic loss mechanisms, such as absorption due to impurity ions, and **microbending** loss due to jacketing and cabling can add loss to a fiber.

The bandwidth or information-carrying capacity of a fiber is inversely related to its total dispersion. The total dispersion in a fiber is a combination of three components: intermodal dispersion (modal delay distortion), material dispersion, and waveguide dispersion.

Intermodal dispersion occurs in multimode fibers because rays associated with different modes travel different effective distances through the optical fiber. This causes light in the different modes to spread out temporally as it travels along the fiber. Modal delay distortion can severely limit the bandwidth of a step-index multimode fiber to the order of 20 MHz-km. To reduce modal delay distortion in multimode fibers, the core is carefully doped to create a graded (approximately parabolic shaped) refractive index profile. By carefully designing this index profile, the group velocities of the propagating modes are nearly equalized. Bandwidths of 1.0 GHz-km are readily attainable in commercially available graded-index multimode fibers. The most effective way of eliminating intermodal dispersion is to use a single-mode fiber. Since only one mode propagates in a single-mode fiber, modal delay distortion between modes does not exist and very high bandwidths are possible. The bandwidth of a single-mode fiber, as mentioned previously, is limited by the combination of material and waveguide dispersion. As shown in Fig. 42.18, both material and waveguide dispersion are dependent on wavelength.

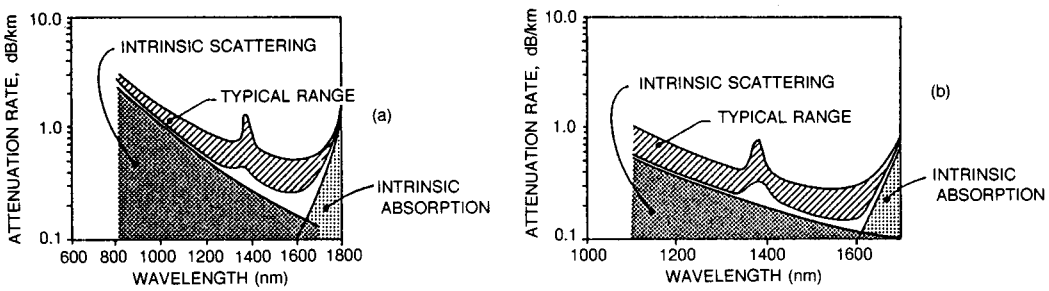


FIGURE 42.17 Spectral attenuation rate. (a) Graded-index multimode fibers. (b) Single-mode fibers.

TABLE 42.2 Best Attenuation Results (dB/km) in Ge-P-SiO_2 Core Fibers

Wavelength (nm)	$\Delta \approx 0.2\%$ (Single-mode Fibers)	$\Delta \approx 1.0\%$ (Graded-index Multimode Fibers)
850	2.1	2.20
1300	0.27	0.44
1500	0.16	0.23

Material dispersion is caused by the variation of the refractive index of the glass with wavelength and the spectral width of the system source. Waveguide dispersion occurs because light travels in both the core and cladding of a single-mode fiber at an effective velocity between that of the core and cladding materials. The waveguide dispersion arises because the effective velocity, the waveguide dispersion, changes with wavelength. The amount of waveguide dispersion depends on the design of the waveguide structure as well as on the fiber material. Both material and waveguide dispersion are measured in picoseconds (of pulse spreading) per nanometer (of source spectral width) per kilometer (of fiber length), reflecting both the increases in magnitude in source linewidth and the increase in dispersion with fiber length.

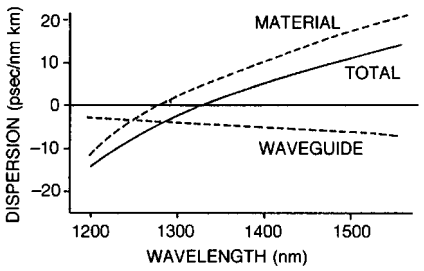


FIGURE 42.18 Single-mode step-index dispersion curve.

Material and waveguide dispersion can have different signs and effectively cancel each other's dispersive effect on the total dispersion in a single-mode fiber. In conventional germanium-doped silica fibers, the "zero-dispersion" wavelength at which the waveguide and material dispersion effects cancel each other out occurs near 1.30 μm. The zero-dispersion wavelength can be shifted to 1.55 μm, or the low-dispersion characteristics of a fiber can be broadened by modifying the refractive index profile shape of a single-mode fiber. This profile shape modification alters the waveguide dispersion characteristics of the fiber and changes the wavelength region in which waveguide and material dispersion effects cancel each other. Figure 42.19 illustrates the profile shapes of "conventional," "dispersion-shifted," and "dispersion-flattened" single-mode fibers. Single-mode fibers operating in their zero-dispersion region with system sources of finite spectral width do not have infinite bandwidth but have bandwidths that are high enough to satisfy all current high-capacity system requirements.

Optical Fiber Cable Manufacturing

Optical fiber cables should have low loss and high bandwidth and should maintain these characteristics while in service in extreme environments. In addition, they should be strong enough to survive the stresses encountered during manufacture, installation, and service in a hostile environment. The manufacturing process used

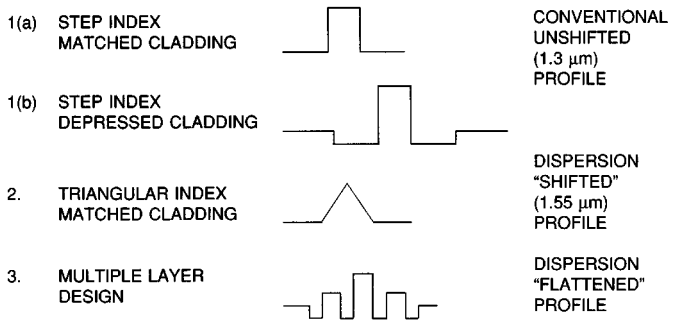


FIGURE 42.19 Single-mode refractive index profiles.

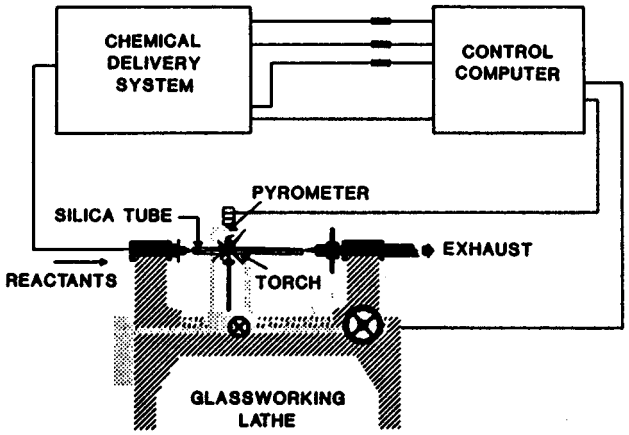


FIGURE 42.20 Schematic diagram of the MCVD process.

to fabricate the optical fiber cables can be divided into four steps: (1) preform fabrication, (2) fiber drawing and coating, (3) fiber measurement, and (4) fiber packaging.

Preform Fabrication

The first step in the fabrication of optical fiber is the creation of a glass preform. A preform is a large blank of glass several millimeters in diameter and several centimeters in length. The preform has all the desired properties (e.g., geometrical ratios and chemical composition) necessary to yield a high-quality fiber. The preform is subsequently drawn into a multi-kilometer-long hair-thin fiber. Four different preform manufacturing processes are currently in commercial use.

The most widely used process is the modified **chemical vapor deposition** (MCVD) process invented at the AT&T Bell Laboratories. Outside vapor deposition process (OVD) is used by Corning Glass Works and some of its joint ventures in Europe. Vapor axial deposition (VAD) process is the process used most widely in Japan. Philips, in Eindhoven, Netherlands, uses a low-temperature plasma chemical vapor deposition (PCVD) process.

In addition to the above four major processes, other processes are under development in different laboratories. Plasma MCVD is under development at Bell Laboratories, hybrid OVD-VAD processes are being developed in Japan, and Sol-Gel processes are being developed in several laboratories. The first four processes are the established commercial processes and are producing fiber economically. The new processes are aimed at greatly increasing the manufacturing productivity of preforms, and thereby reducing their cost.

All the above processes produce high-silica fibers using different dopants, such as germanium, phosphorus, and fluorine. These dopants modify the refractive index of silica, enabling the production of the proper core refractive index profile. Purity of the reactants and the control of the refractive index profile are crucial to the low loss and high bandwidth of the fiber.

MCVD Process. In the MCVD process (Fig. 42.20), a fused-silica tube of extremely high purity and dimensional uniformity is cleaned in an acid solution and degreased. The clean tube is mounted on a glass working lathe. A mixture of reactants is passed from one end of the tube and exhaust gases are taken out at the other end while the tube is being rotated. A torch travels along the length of the tube in the direction of the reactant flow. The reactants include ultra-high-purity oxygen and a combination of one or more of the halides and oxyhalides ($SiCl_4$, $GeCl_4$, $POCl_3$, BCl_3, BBr_3, SiF_4, CCl_4, CCl_2F_2, Cl_2, SF_6, and $SOCl_2$).

The halides react with the oxygen in the temperature range of 1300–1600°C to form oxide particles, which are driven to the wall of the tube and subsequently consolidated into a glassy layer as the hottest part of the flame passes over. After the completion of one pass, the torch travels back and the next pass is begun. Depending on the type of fiber (i.e., multimode or single-mode), a **barrier layer** or a cladding consisting of many thin layers is first deposited on the inside surface of the tube. The compositions may include B_2O_3-P_2O_5-SiO_2 or F-P_2O_5-SiO_2 for barrier layers and SiO_2, F-SiO_2, F-P_2O_5-SiO_2, or F-GeO_2-SiO_2-P_2O_5 for cladding layers. After the required number of barrier or cladding layers has been deposited the core is deposited. The core compositions

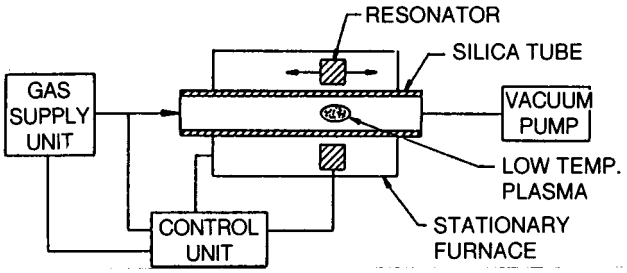


FIGURE 42.21 Schematic diagram of the PCVD process.

depend on whether the fiber is single-mode, multimode, step-index, or multimode graded-index. In the case of graded-index multimode fibers, the dopant level changes with every layer, to provide a refractive index profile that yields the maximum bandwidth.

After the deposition is complete, the reactant flow is stopped except for a small flow of oxygen, and the temperature is raised by reducing the torch speed and increasing the flows of oxygen and hydrogen through the torch. Usually the exhaust end of the tube is closed first and a small positive pressure is maintained inside the deposited tube while the torch travels backward. The higher temperatures cause the glass viscosity to decrease, and the surface tension causes the tube to contract inward. The complete collapse of the tube into a solid preform is achieved in several passes. The speed of the collapse, the rotation of the tube, the temperature of collapse, and the positive pressure of oxygen inside the tube are all accurately controlled to predetermined values in order to produce a straight and bubble-free preform with minimum ovality. The complete preform is then taken off the lathe. After an inspection to assure that the preform is free of defects, the preform is ready to be drawn into a thin fiber.

The control of the refractive index profile along the cross section of the deposited portion of the preform is achieved through a vapor delivery system. In this system, liquids are vaporized by passing a carrier gas (pure O_2) through the bubblers, made of fused silica. Accurate flows are achieved with precision flow controllers that maintain accurate carrier gas flows and extremely accurate temperatures within the bubblers. Microprocessors are used to automate the complete deposition process, including the torch travel and composition changes throughout the process. Impurities are reduced to very low levels by starting with pure chemicals, and there is further reducing of the impurities with in-house purification of these chemicals. Ultra-pure oxygen and a completely sealed vapor-delivery system are used to avoid chemical contamination. Transition-metal ion impurities of well below 1 ppb and OH^- ion impurities of less than 1 ppm are typically maintained to produce high-quality fiber.

The PCVD Process. The PCVD process (Fig. 42.21) also uses a starting tube, and the deposition takes place inside the tube. Here, however, the tube is either stationary or oscillating and the pressure is kept at 10–15 torr. Reactants are fed inside the tube, and the reaction is accomplished by a traveling microwave plasma inside the tube. The entire tube is maintained at approximately 1200°C. The plasma causes the heterogeneous depositions of glass on the tube wall, and the deposition efficiency is very high. After the required depositions of the cladding and core are complete, the tube is taken out and collapsed on a separate equipment. Extreme care is required to prevent impurities from getting into the tube during the transport and collapse procedure. The PCVD process has the advantages of high efficiency, no tube distortion because of the lower temperature, and very accurate profile control because of the large number of layers deposited in a short time. However, going to higher rates of flow presents some difficulties, because of a need to maintain the low pressure.

The PMCVD Process. The PMCVD is an enhancement of the MCVD process. Very high rates of deposition (up to 10 g/min, compared to 2 g/min for MCVD) are achieved by using larger diameter tubes and an RF plasma for reaction (Fig. 42.22). Because of the very high temperature of the plasma, water cooling is essential. An oxyhydrogen torch follows the plasma and sinters the deposition. The high rates of deposition are achieved because of very high thermal gradients from the center of the tube to the wall and the resulting high thermophoretic driving force. The PMCVD process is still in the development stage and has not been commercialized.

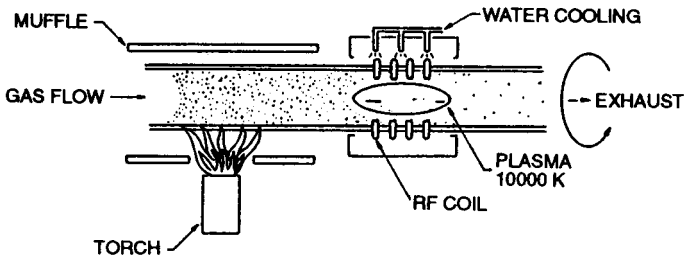


FIGURE 42.22 Schematic diagram of the PMCVD process.

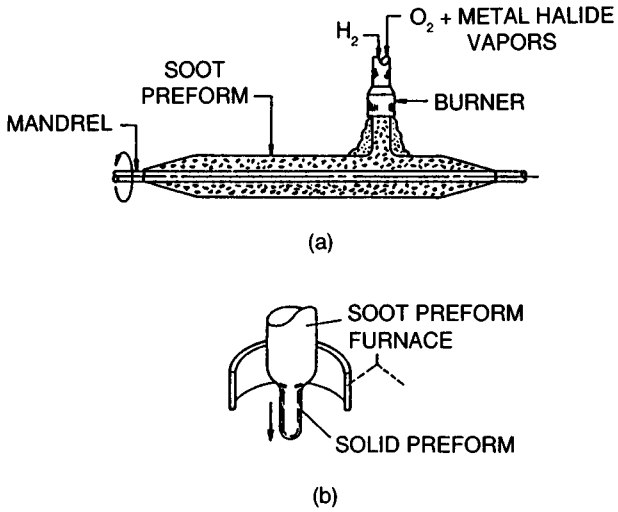


FIGURE 42.23 Schematic diagram of the outside vapor deposition. (a) Soot deposition. (b) Consolidation.

The OVD Process. The OVD process does not use a starting tube; instead, a stream of soot particles of desired composition is deposited on a bait rod (Fig. 42.23). The soot particles are produced by the reaction of reactants in a fuel gas-oxygen flame. A cylindrical porous soot preform is built layer by layer. After the deposition of the core and cladding is complete, the bait rod is removed. The porous preform is then sintered and dried in a furnace at 1400–1600°C to form a clear bubble-free preform under a controlled environment. The central hold left by the blank may or may not be closed, depending on the type of preform. The preform is now ready for inspection and drawing.

The VAD Process. The process is very similar to the OVD process. However, the soot deposition is done axially instead of radially. The soot is deposited at the end of a starting silica-glass rod (Fig. 42.24). A special torch using several annular holes is used to direct a stream of soot at the deposition surface. The reactant vapors, hydrogen gas, argon gas, and oxygen gas flow through different annular openings. Normally the core is deposited and the rotating speed is gradually withdrawn as the deposition proceeds at the end. The index profile is controlled by the composition of the gases flowing through the torch and the temperature distribution at the deposition surface. The porous preform is consolidated and dehydrated as it passes through a carbon-ring furnace in a

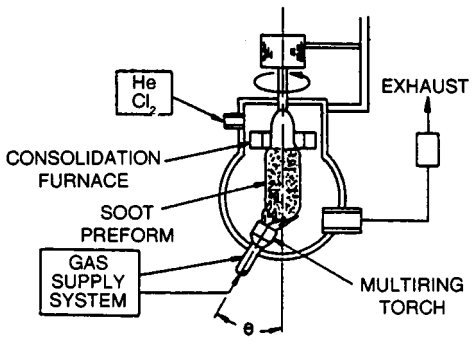


FIGURE 42.24 Schematic diagram of the vapor axial deposition.

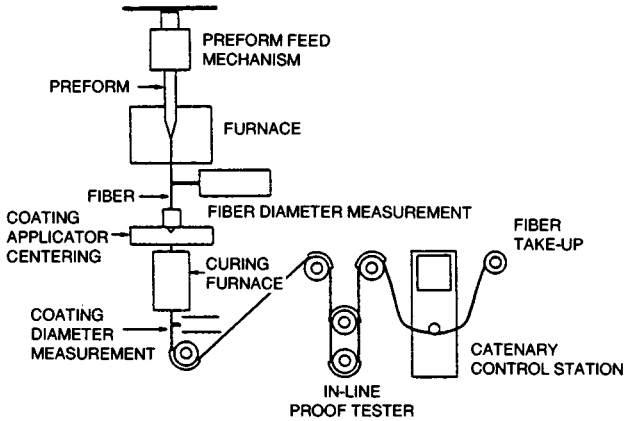


FIGURE 42.25 The fiber drawing process.

controlled environment. $SOCl_2$ and Cl are used to dehydrate the preform. Because of the axial deposition, this process is semicontinuous and is capable of producing very large preforms.

Fiber Drawing

After a preform has been inspected for various defects such as bubbles, ovality, and straightness, it is taken to a fiber drawing station. A large-scale fiber drawing process must repeatedly maintain the optical quality of the preform and produce a dimensionally uniform fiber with high strength.

Draw Process. During fiber drawing, the inspected preform is lowered into a hot zone at a certain feed rate V_p, and the fiber is pulled from the softened neck-down region (Fig. 42.25) at a rate V_f. At steady state,

$$\pi D_p^2 V_p/4 = \pi D_f^2 V_f/4 \qquad (42.38)$$

where D_p and D_f are the preform and fiber diameters, respectively. Therefore,

$$V_f = (D_p^2/D_f^2) V_p \qquad (42.39)$$

A draw machine, therefore, consists of a preform feed mechanism, a heat source, a pulling device, a coating device, and a control system to accurately maintain the fiber diameter and the furnace temperature.

Heat Source. The heat source should provide sufficient energy to soften the glass for pulling the fiber without causing excessive tension and without creating turbulence in the neck-down region. A proper heat source will yield a fiber with uniform diameter and high strength. Oxyhydrogen torches, CO_2 lasers, resistance furnaces, and induction furnaces have been used to draw fibers. An oxyhydrogen torch, although a clean source of heat, suffers from turbulence due to flame. A CO_2 laser is too expensive a heat source to be considered for the large-scale manufacture of fibers. Graphite resistance furnaces and zirconia induction furnaces are the most widely used heat sources for fiber drawing. In the graphite resistance furnace, a graphite resistive element produces the required heat. Because graphite reacts with oxygen at high temperatures, an inert environment (e.g., carbon) is maintained inside the furnace. The zirconia induction furnace does not require inert environment. It is extremely important that the furnace environment be clean in order to produce high-strength fibers. A zirconia induction furnace, when properly designed and used, has produced very-high-strength long-length fibers (over 2.0 GPa) in lengths of several kilometers.

Mechanical Systems. An accurate preform feed mechanism and drive capstan form the basis of fiber speed control. The mechanism allows the preform to be fed at a constant speed into the hot zone, while maintaining the preform at the center of the furnace opening at the top. A centering device is used to position preforms that are not perfectly straight. The preform is usually held with a collet-type chuck mounted in a vertically movable carriage, which is driven by a lead screw. A precision stainless-steel drive capstan is mounted on the

shaft of a high-performance dc servomotor. The fiber is taken up on a proper-diameter spool. The fiber is wound on the spool at close to zero tension with the help of a catenary control. In some cases fiber is proof-tested in-line before it is wound on a spool. The proof stress can be set at different levels depending on the application for which the fiber is being manufactured.

Fiber Coating System. The glass fiber coming out of the furnace has a highly polished pristine surface and the theoretical strength of such a fiber is in the range of 15–20 GPa. Strengths in the range of 4.5–5.5 GPa are routinely measured on short fiber lengths. To preserve this high strength, polymeric coatings are applied immediately after the drawing. The coating must be applied without damaging the fiber, it must solidify before reaching the capstan, and it should not cause microbending loss. To satisfy all these requirements, usually two layers of coatings are applied: a soft inner coating adjacent to the fiber to avoid microbending loss and a hard outer coating to resist abrasion. The coatings are a combination of ultraviolet- (UV) curable acrylates, UV-curable silicones, hot melts, heat-curable silicones, and nylons. When dual coatings are applied, the coated fiber diameter is typically 235–250 μm. The nylon-jacketed fiber typically used in Japan has an outside diameter of 900 μm. All coating materials are usually filtered to remove particles that may damage the fiber. Coatings are usually applied by passing the fiber through a coating cup and then curing the coating before the fiber is taken up by the capstan. The method of application, the coating material, the temperature, and the draw speed affect the proper application or a well-centered, bubble-free coating.

Fiber drawing facilities are usually located in a clean room where the air is maintained at class 10,000. The region of the preform and fiber from the coating cup to the top of the preform is maintained at class 100 or better. A class 100 environment means that there are no more than 100 particles of size greater than 0.5 μm in 1 ft^3 of air. A clean environment, proper centering of the preform in the furnace and fiber in the coating cup, and proper alignment of the whole draw tower ensure a scratch-free fiber of a very high tensile strength. A control unit regulates the draw speed, preform feed speed, preform centering, fiber diameter, furnace temperature, and draw tension.

The coated fiber wound on a spool is next taken to the fiber measurement area to assure proper quality control.

Proof Testing of Fibers. Mechanical failure is one of the major concerns in the reliability of optical fibers. Fiber drawn in kilometer lengths must be strong enough to survive all of the short- and long-term stresses that it will encounter during the manufacture, installation, and long service life. Glass is an ideal elastic isotropic solid and does not contain dislocations. Hence, the strength is determined mainly by inclusions and surface flaws. Although extreme care is taken to avoid inhomogeneities and surface flaws during fiber manufacture, they cannot be completely eliminated. Since surface flaws can result from various causes, they are statistical in nature and it is very difficult to predict the long-length strength of glass fibers. To guarantee a minimum fiber strength, proof testing has been adopted as a manufacturing step. Proof testing can be done in-line immediately after the drawing and coating or off-line before the fiber is stored.

In proof testing, the entire length of the fiber is subjected to a properly controlled proof stress. The proof stress is based on the stresses likely to be encountered by the fiber during manufacture, storage, installation, and service. The fibers that survive the proof test are stored for further packaging into cables.

Proof testing not only guarantees that the fiber will survive short-term stresses but also guarantees that the fiber will survive a lower residual stress that it may be subjected to during its long service life. It is well known that glass, when used in a humid environment, can fail under a long-term stress well below its instantaneous strength. This phenomenon is called static fatigue. Several models have been proposed to quantitatively describe the relationship between residual stress and the life of optical fibers. Use is made of the most conservative of these models, and the proof stress is determined by a consideration of the maximum possible residual stress in service and the required service life.

Fiber Packaging

In order to efficiently use one or more fibers, they need to be packaged so that they can be handled, transported, and installed without damage. Optical fibers can be used in a variety of applications, and hence the way they are packaged or cabled will also vary. There are numerous cable designs that are used by different cable manufacturers. All these designs, however, must meet certain criteria. A primary consideration in a cable design

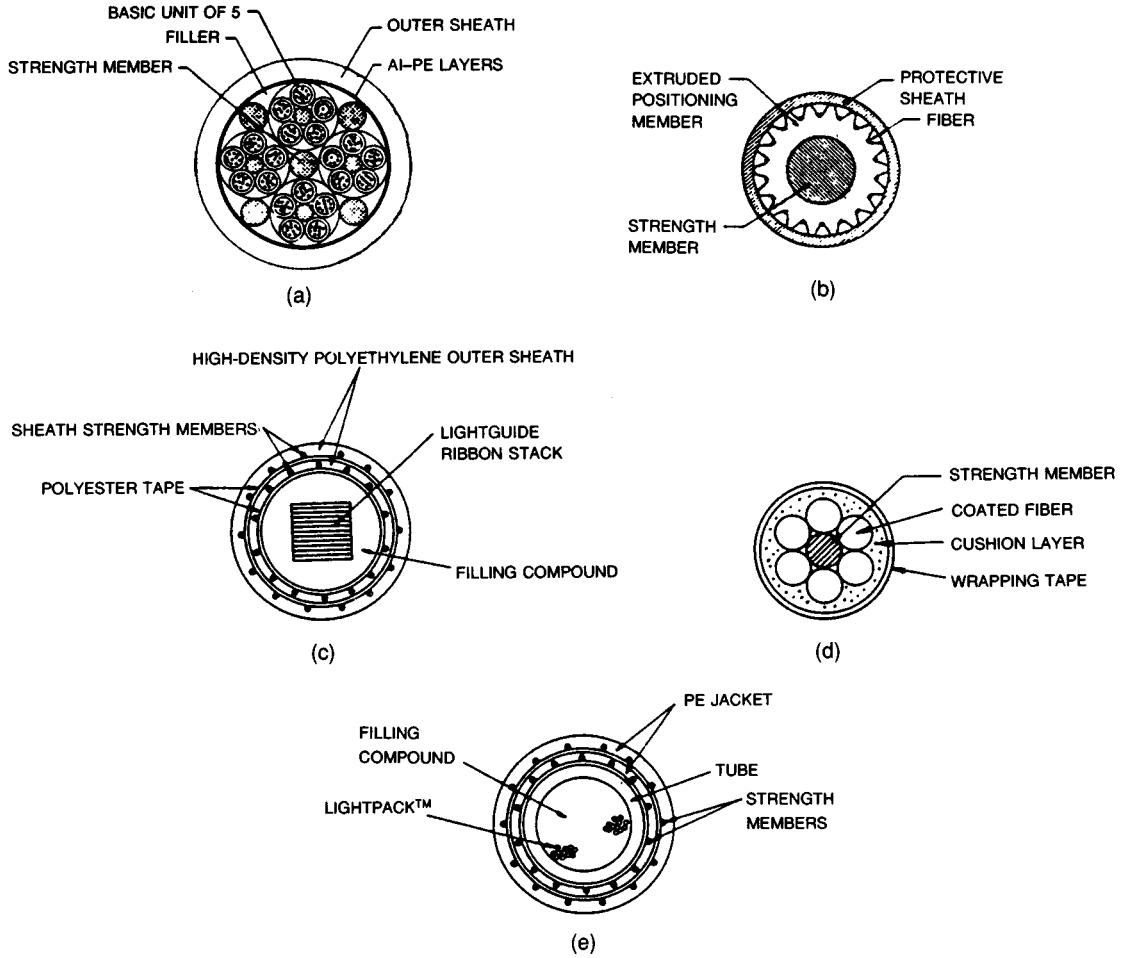


FIGURE 42.26 Fiber cable designs. (a) Loose tube design. (b) Slotted design. (c) Ribbon design. (d) Stranded unit. (e) Lightpack™ Cable design.

is to assure that the fibers in the cables maintain their optical properties (attenuation and dispersion) during their service life under different environmental conditions. The design, therefore, must minimize microbending effects. This usually means letting the fiber take a minimum energy position at all times in the cable structure. Proper selection of cabling materials so as to minimize differential thermal expansion or contraction during temperature extremes is important in minimizing microbending loss. The cable structure must be such that the fibers carry a load well below the proof-test level at all times, and especially while using conventional installation equipment. The cables must provide adequate protection to the fibers under all adverse environmental conditions during their entire service life, which may be as long as 40 years. Finally, the cable designs should be cost effective and easily connectorized or spliced.

Five different types (Fig. 42.26) of basic cable designs are currently in use: (a) loose tube, (b) fluted, (c) ribbon, (d) stranded, and (e) Lightpack Cable. The loose tube design was pioneered by Siemens in Germany. Up to 10 fibers are enclosed in a loose tube, which is filled with a soft filling compound. Since the fibers are relatively free to take the minimum energy configuration, the microbending losses are avoided. Several of these buffered loose tube units are stranded around a central glass-resin support member. Aramid yarns are stranded on the cable core to provide strength members (for pulling through ducts), with a final polyethylene sheath on the outside. The stranding lay length and pitch radius are calculated to permit tensile strain on the cable up to the rated force and to permit cooling down to the rated low temperature without affecting the fiber attenuation.

In the fluted designs, fibers are laid in the grooves of plastic central members and are relatively free to move. The shape and size of the grooves vary with the design. The grooved core may also contain a central strength member. A sheath is formed over the grooved core, and this essentially forms a unit. Several units may then be stranded around a central strength member to form a cable core of desired size, over which different types of sheaths are formed. Fluted designs have been pioneered in France and Canada.

The ribbon design was invented at AT&T Bell Laboratories and consists of a linear array of 12 fibers sandwiched between two polyester tapes with pressure-sensitive adhesive on the fiber side. The spacing and the back tension on the fibers is accurately maintained. The ribbons are typically 2.5 mm in width. Up to 12 ribbons can be stacked to give a cable core consisting of 144 fibers. The core is twisted to some lay length and enclosed in a polyethylene tube. Several combinations of protective plastic and metallic layers along with metallic or nonmetallic strength members are then applied around the core to give the final cable its required mechanical and environmental characteristics needed for use in specified conditions. The ribbon design offers the most efficient and economic packaging of fibers for high-fiber-count cables. It also lends the cable to preconnectorization and makes it extremely convenient for installation and splicing.

The tight-bound stranded designs were pioneered by Japanese and are used in the United States for several applications. In this design, several coated fibers are stranded around a central support member. The central support member may also serve as a strength member, and it may be metallic or nonmetallic. The stranded unit can have up to 18 fibers. The unit is contained within a plastic tube filled with a water-blocking compound. The final cable consists of several of these units stranded around a central member and protected on the outside with various sheath combinations.

The Lightpack Cable design, pioneered by AT&T, is one of the simplest designs. Several fibers are held together with a binder to form a unit. One or more units are laid inside a large tube, which is filled with a water-blocking compound. This design has the advantage of the loose tube design in that the fibers are free of strain, but is more compact. The tube-containing units can then be projected with various sheath options and strength members to provide the final cable.

The final step in cabling is the sheathing operation. After the fibers have been made into identifiable units, one or more of the units (as discussed earlier) form a core which is then covered with a combination of sheathing layers. The number and combination of the sheathing layers depend on the intended use. Typically, a polyethylene sheath is extruded over the filled cable core. In a typical cross-ply design, metallic or nonmetallic strength members are applied over the first sheath layer, followed by another polyethylene sheath, over which another layer of strength members is applied. The direction of lay of the two layers of the strength members is opposite to each other. A final sheath is applied and the cable is ready for the final inspection, preconnectorization, and shipment. Metallic vapor barriers and lightning- and rodent-protection sheath options are also available. Further armoring is applied to cables made for submarine application.

In addition to the above cable designs, there are numerous other cable designs used for specific applications, such as fire-resistant cables, military tactical cables, cables for missile guidance systems, cables for field communications established by air-drop operations, air deployment cables, and cables for industrial controls. All these applications have unique requirements, such as ruggedness, low loss, and repeaterless spans, and the cable designs are accordingly selected. However, all these cable designs still rely on the basic unit designs discussed above.

Defining Terms

Attenuation: Decrease of average optical power as light travels along the length or an optical fiber.

Bandwidth: Measure of the information-carrying capacity of the fiber. The greater the bandwidth, the greater the information-carrying capacity.

Barrier layer: Layer of deposited glass adjacent to the inner tube surface to create a barrier against OH diffusion.

Chemical vapor deposition: Process in which products of a heterogeneous gas-liquid or gas-solid reaction are deposited on the surface of a substrate.

Cladding: Low refractive index material that surrounds the fiber core.

Core: Central portion of a fiber through which light is transmitted.

Cut-off wavelength: Wavelength greater than which a particular mode ceases to be a bound mode.

Dispersion: Cause of distortion of the signal due to different propagation characteristics of different modes, leading to bandwidth limitations.

Graded-index profile: Any refractive index profile that varies with radius in the core.

Microbending: Sharp curvatures involving local fiber axis displacements of a few micrometers and spatial wavelengths of a few millimeters. Microbending causes significant losses.

Mode: Permitted electromagnetic field pattern within an optical fiber.

Numerical aperture: Acceptance angle of the fiber.

Optical repeater: Optoelectric device that receives a signal and amplifies it and retransmits it. In digital systems the signal is regenerated.

Related Topics

31.3 Circuits • 71.1 Lightwave Technology for Video Transmission

References

M.K. Barnoski, Ed., *Fundamentals of Optical Fiber Communications*, New York: Academic Press, 1976.

B. Bendow and S. M. Shashanka. Eds., *Fiber Optics: Advances in Research and Development*, New York: Plenum Press, 1979.

A.H. Cherin, *Introduction to Optical Fibers*, New York: McGraw-Hill, 1983.

T. Li, Ed., *Optical Fiber Communications*, New York: Academic Press, 1985.

J.E. Midwinter, *Optical Fibers for Transmission*, New York: Wiley, 1979.

S.E. Miller and A.G. Chynoweth, Eds., *Optical Fiber Telecommunications*, New York: Academic Press, 1979.

Y. Suematsu and I. Ken-ichi, *Introduction to Optical Fiber Communication*, New York: Wiley, 1982.

Bahl, I.J. “Solid State Circuits”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

43

Solid State Circuits

43.1 Introduction
43.2 Amplifiers
43.3 Oscillators
43.4 Multipliers
43.5 Mixers
43.6 Control Circuits
43.7 Summary and Future Trends

I. J. Bahl
ITT Gallium Arsenide Technology Center

43.1 Introduction

Over the past two decades, microwave active circuits have evolved from individual solid state transistors and passive elements housed in conventional waveguides and/or coaxial lines to fully integrated planar assemblies, including active and passive components and interconnections, generically referred to as a microwave integrated circuit (MIC). The **hybrid microwave integrated circuit** (HMIC) consists of an interconnect pattern and distributed circuit components printed on a suitable substrate, with active and lumped circuit components (in packaged or chip form) attached individually to the printed interconnect circuit by the use of soldering and wire bonding techniques. The solid state active elements are either silicon or gallium arsenide (or other III–V compound) devices. More recently, the solid state **monolithic microwave integrated circuit** (MMIC) approach has become commonplace. In MMICs, all interconnections and components, both active and passive, are fabricated simultaneously on a semi-insulating semiconductor substrate (usually gallium arsenide, GaAs) using deposition and etching processes, thereby eliminating discrete components and wire bond interconnects. The term MMIC is used for circuits operating in the millimeter wave (30–300 GHz) region of the frequency spectrum as well as the microwave (1–30 GHz) region. Major advantages of MMICs include low cost, small size, low weight, circuit design flexibility, broadband performance, elimination of circuit tweaking, high-volume manufacturing capability, package simplification, improved reproducibility, improved reliability, and multifunction performance on a single chip.

Microwave circuits use two types of active devices: two-terminal devices, referred to as diodes, such as Schottky, Gunn, tunnel, impact avalanche and transit time (IMPATT), varactor, and **PIN**, and three-terminal devices, referred to as transistors, such as bipolar junction transistor (BJT), metal semiconductor field effect transistor (MESFET), high electron mobility transistor (HEMT), heterostructure FET (HFET), and heterojunction bipolar transistor (HBT). Microwave circuits using these devices include **amplifiers**, **oscillators**, **multipliers**, **mixers**, **switches**, **phase shifters**, **attenuators**, modulators, and many others used for receiver or transmitter applications covering microwave and millimeter wave frequency bands. New devices, microwave **computer-aided design (CAD)** tools, and automated testing have played a significant role in the advancement of these circuits during the past decade. The theory and performance of most of these circuits have been well documented [Kollberg, 1984; Bhartia and Bahl, 1984; Pucel, 1985; Maas, 1986; Bahl and Bhartia, 1988; Goyal, 1989; Ali et al., 1989; Chang, 1990; Vendelin et al., 1990; Ali and Gupta, 1991; Chang, 1994]. Solid state circuits are extensively used in such applications as radar, communication, navigation, electronic warfare (EW), smart weapons,

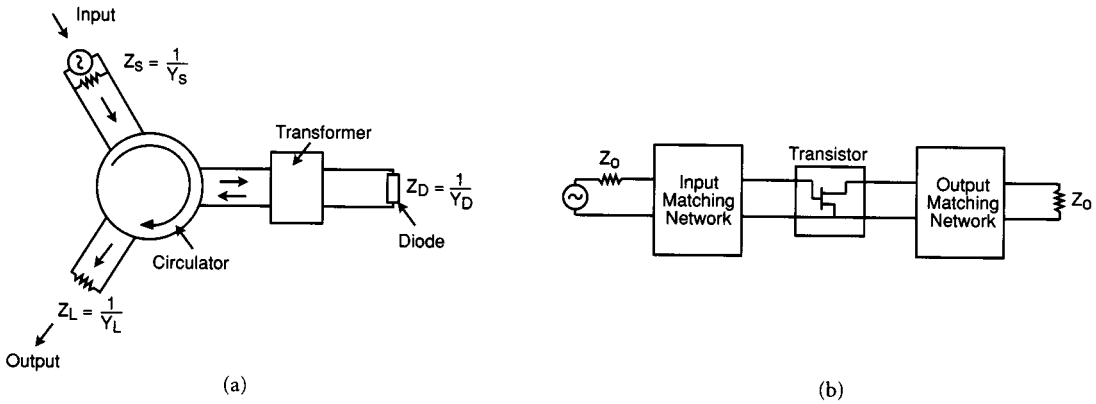


FIGURE 43.1 Amplifier circuits configurations. (a) Two-terminal negative resistance type requires a circulator to isolate the input and output ports. (b) Three-terminal transistor type requires input and output matching networks.

consumer electronics, and microwave instruments and equipment. This section will briefly describe the performance status of amplifiers, oscillators, multipliers, mixers, and microwave control circuits.

43.2 Amplifiers

Amplifier circuits have received maximum attention in solid state circuits development. The two-terminal device amplifiers, such as parametric, tunnel, Gunn, and IMPATT, are normally called *reflection-type circuits,* or *negative resistance amplifiers.* A diagram for these amplifiers is shown in Fig. 43.1(a). Parametric amplifiers are narrowband (<10%) and have very good noise figure. Tunnel-diode amplifiers are high-gain, low-noise figure, and low-power circuits. Octave **bandwidth** of such amplifiers is possible. The performance of Gunn-diode amplifiers is quite similar to tunnel-diode amplifiers. IMPATT-diode amplifiers are high power and high efficiency. They are moderately noisy, and bandwidths up to an octave are possible.

The basic circuit configuration for three-terminal device amplifiers is shown in Fig. 43.1(b). Several different types of amplifiers developed using transistors are low noise, power, high linearity, broadband, high efficiency, logarithmic, limiting, transimpedance, and variable gain. The silicon bipolar transistor performs very well up to about 4 GHz, with reliable performance, high power, high gain, and low cost. The GaAs MESFETs perform better than the bipolar transistors above 4 GHz. They are broadband, have a wide dynamic range, are highly reliable, and are low cost. Both low-noise and medium-power MESFET amplifiers are available. They compete with uncooled parametric amplifiers as well as moderate-power IMPATTs. HEMTs find a niche in low-noise and high-frequency applications. The **noise figure** of HEMT amplifiers is better than that of uncooled parametric amplifiers up to 100 GHz, as shown in Fig. 43.2.

Various techniques are used to realize small signal or low-power broadband amplifiers. Five of them are shown in Fig. 43.3. The distributed approach provides the unique capability of excellent gain-bandwidth product, low **VSWR (voltage standing wave ratio)**, and moderately low noise figure. This technique has been successfully used in monolithic ultrabroadband amplifiers. The performance of such amplifiers using various transistor devices is given in Table 43.1.

The performance of solid state power amplifiers is shown in Fig. 43.4. Currently, IMPATT and Gunn diodes provide maximum power above 10 GHz, whereas bipolar junction transistor and MESFET technologies offer the most promise to generate higher power levels below 10 GHz. In particular, IMPATT devices have been operated over the complete millimeter wave band and have shown good continuous wave (CW) and pulsed power efficiency and reliability.

During the past decade significant progress has been made in monolithic power amplifiers operating over both the narrowband and broadband [Williams and Bahl, 1992; Tserng and Saunier, 1991]. Power levels as high as 12 W from a single MMIC chip at C-band with 60% power-added efficiency (PAE) have been demonstrated. A 6-W MMIC chip has been developed at X-band. A 2-W power output has been obtained at 30 GHz.

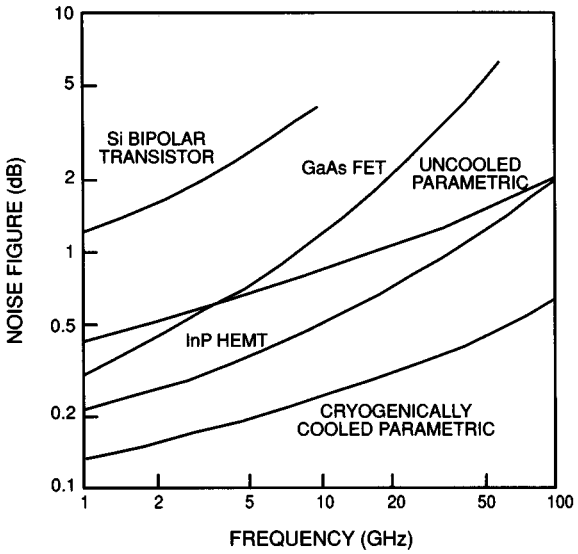


FIGURE 43.2 Comparison of noise performance of various solid state amplifiers; the InP HEMT LNA, which is also compatible with MMIC technology, is a clear choice for receiver applications where cryogenic cooling is precluded. (*Source:* D. Willems and I. Bahl, "Advances in Monolithic Microwave and Millimeter Wave Integrated Circuits," *IEEE Int. Circuits and Systems Symp. Digest,* pp. 783–786. © 1992 IEEE. With permission.)

In the high- efficiency area, a C-band MMIC amplifier with 70% PAE, 8-dB gain, and 1.7-W power output has been demonstrated. For broadband amplifiers having an octave or more bandwidth, MMIC technology has been exclusively used and is quite promising. Figure 43.5 depicts power performance for single-chip MMIC amplifiers spanning microwave and millimeter wave frequencies. The state of the art in high efficiency and broadband power MMIC amplifiers is summarized in Tables 43.2 and 43.3, respectively. Note that the high-efficiency examples included in Table 43.2 all exceed 40% PAE.

43.3 Oscillators

Solid state oscillators represent the basic microwave energy source and have the advantages of light weight and small size compared with microwave tubes. As shown in Fig. 43.6, a typical microwave oscillator consists of a MESFET as an active device (a diode can also be used) and a passive frequency-determining resonant element, such as a microstrip, surface acoustic wave (SAW), cavity resonator, or dielectric resonator for fixed tuned oscillators and a varactor or a yttrium iron garnet (YIG) sphere for tunable oscillators. These oscillators have the capability of temperature stabilization and phase locking. Dielectric resonator oscillators provide stable operation from 1 to 100 GHz as fixed frequency sources. In addition to their good frequency stability, they are simple in design, have high efficiency, and are compatible with MMIC technology. Gunn and IMPATT oscillators provide higher power levels and cover microwave and millimeter wave bands. The transistor oscillators using MESFETs, HEMTS, and HBTs provide highly cost-effective, miniature, reliable, and low-noise sources for use up to the millimeter wave frequency range, while BJT oscillators reach only 20 GHz. Compared to a GaAs MESFET oscillator, a BJT or a HBT oscillator typically has 6 to 10 dB lower phase noise very close to the carrier. Figure 43.7 shows the performance of various solid state oscillators. Higher power levels for oscillators are obtained by connecting high-power amplifiers at the output of medium-power oscillators.

43.4 Multipliers

Microwave frequency multipliers are used to generate microwave power at levels above those obtainable with fundamental frequency oscillators. Several different nonlinear phenomena can be used to achieve frequency

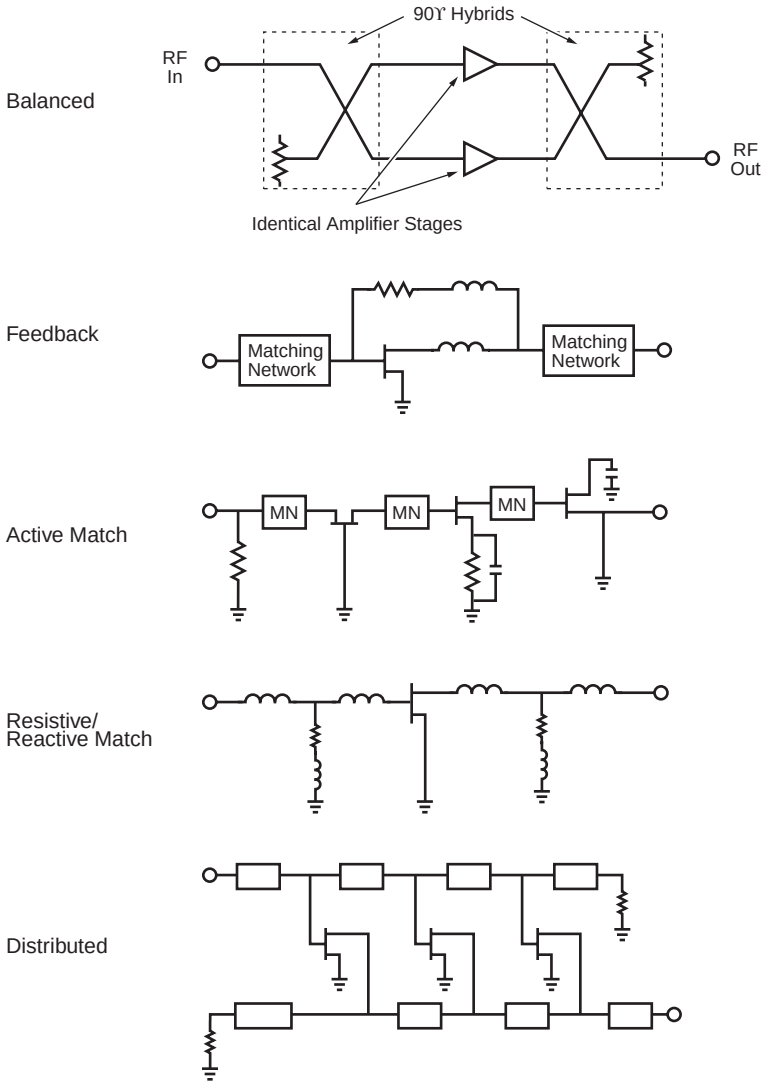


FIGURE 43.3 Broadband amplifier configurations. *Balanced* has low noise figure and better cascadability, *feedback* has small size, *active match* is more suitable for monolithic approach, and *distributed* is good for multioctave bandwidths.

multiplication, e.g., nonlinear reactance in varactors and step-recovery diodes and nonlinear resistance in Schottky barrier diodes and three-terminal devices (BJT, MESFET, HEMT, HBT).

Varactor multipliers offer the best frequency multiplier performance. Varactor multipliers (pulsed) have achieved power output in excess of 100 and 10 W at 4 and 10 GHz, respectively [Bahl and Bhartia, 1988]. Table 43.4 shows the best performance measured in the millimeter wave range and above.

43.5 Mixers

Mixers convert (heterodyne) the input frequency to a new frequency, where filtering and/or gain is easier to implement, in contrast to detectors, which are used to provide an output signal that contains the amplitude or amplitude variation information of the input signal. A mixer is basically a multiplier, which requires two

TABLE 43.1 Broadband Single-Chip Distributed MMIC Amplifier Performance

Frequency Range (GHz)	Gain (dB)	Noise Figure (dB)	Device Used
0.5–26.5	6	5.2	0.32 μm GaAs HEMT
0.5–50	6	—	0.32 μm GaAs HEMT
2–18	9	5.7	0.5 μm dual gate FET
2–20	9.5	3.5	0.2 μm GaAs HEMT
2–24	6	—	2 μm SABM GaAs HBT
5–40	9	4.0	0.25 μm GaAs HEMT
5–60	8	—	0.25 μm GaAs HEMT
5–100	5	—	0.1 μm InP HEMT
6–18	10.5	—	0.4 μm GaAs MESFET
9–70	3.5	7.0	0.2 μm GaAs PHEMT

SABM, self-aligned base ohmic metal; PHEMT, pseudomorphic HEMT.

Source: D. Willems and I. Bahl, "Advances in Monolithic Microwave and Millimeter Wave Integrated Circuits," *IEEE Int. Circuits and Systems Symp. Digest*, pp. 783–786. © 1992 IEEE. With permission.

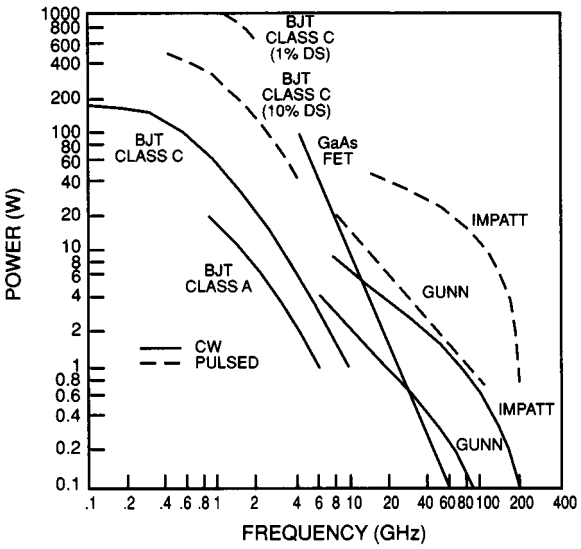


FIGURE 43.4 Power performance of microwave power amplifiers.

signals and uses any solid state device that exhibits nonlinear properties. Mixing is achieved by applying an RF and a high-power local oscillator signal to a nonlinear element, which can be a diode or a transistor.

As illustrated in Fig. 43.8, there are many types of mixers: one diode (single ended), two diodes (balanced or antiparallel), four diodes (double balanced), and eight diodes (double-double balanced). Mixers can also be realized using the nonlinearities associated with transistors that provide conversion gain. The most commonly used mixer configuration in the microwave frequency band is the double-balanced mixer, which has better isolation between the ports and better spurious response. However, the single and balanced mixers place lower power requirements on the local oscillator and have lower conversion loss.

Subharmonic mixing (where the local oscillator frequency is approximately half that needed in conventional mixers) has been extensively used at millimeter wave frequencies. This technique is quite useful when reliable stable local oscillators are either unavailable or prohibitively expensive at high frequencies. Figure 43.9 gives the performance of millimeter wave mixers.

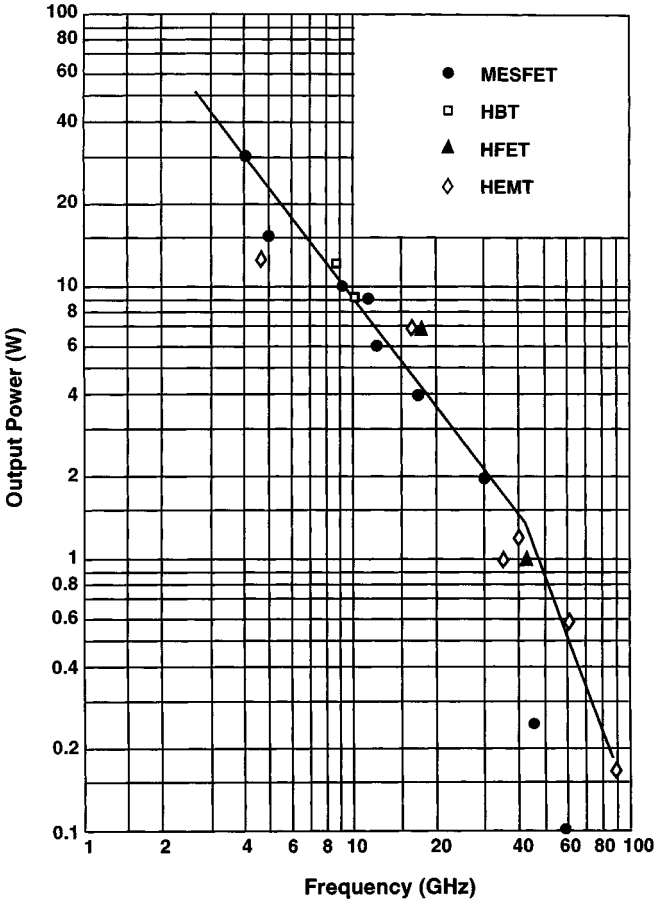


FIGURE 43.5 Performance status of single-chip power MMIC amplifiers using MESFET, HFET, HEMT, and HBT technologies.

TABLE 43.2 Single-Chip High-Efficiency Power MMIC Performance

Frequency (GHz)	No. of Stages	P_O (W)	PAE (%)	Gain (dB)
5.2	1	12.0*	60	9
5.5	1	1.7	70	8
8.5	2	3.2	52	—
10.0	1	5.0	48	7
10.0	1	6.0	44	6
11.5	2	3.0	42	12

Source: D. Willems and I. Bahl, "Advances in Monolithic Microwave and Millimeter Wave Integrated Circuits," *IEEE Int. Circuits and Systems Symp. Digest,* pp. 783–786. © 1992 IEEE. With permission.

*W.L. Pribble and E.L. Griffin, "An ion-implanted 13W C-band MMIC with 60% peak power added efficiency," *IEEE 1996 Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest,* pp. 25–28.

TABLE 43.3 Single-Chip Broadband Power MMIC Performance

Frequency (GHz)	Configuration	No. of Stages	Gain (dB)	P_O (W)	PAE(%)
1.5–9.0	Reactive match	2	5	0.5	14
2.0–8.0	Distributed	1	5	1.0	—
2.0–20.0	Distributed	1	4	0.8	15
3.5–8.0	Reactive match	2	10	2.0	20
6–17	Distributed/reactive	4	16	0.8	11
6–20	Distributed	1	11	0.25	—
7–10.5	Reactive match	2	12.5	3.0	35
7.7–12.2	Reactive match	2	8.0	3.0	14
12–16	Reactive match	3	18	1.8	18
14–33	Distributed	1	4	0.1	—

Source: D. Willems and I. Bahl, "Advances in Monolithic Microwave and Millimeter Wave Integrated Circuits," *IEEE Int. Circuits and Systems Symp. Digest,* pp. 783–786. © 1992 IEEE. With permission.

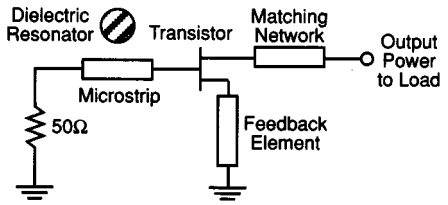


FIGURE 43.6 Basic configuration of a dielectric resonator oscillator. The feedback element is used to make the active device unstable, the matching network allows transfer of maximum power to the load, and the dielectric resonator provides frequency stability.

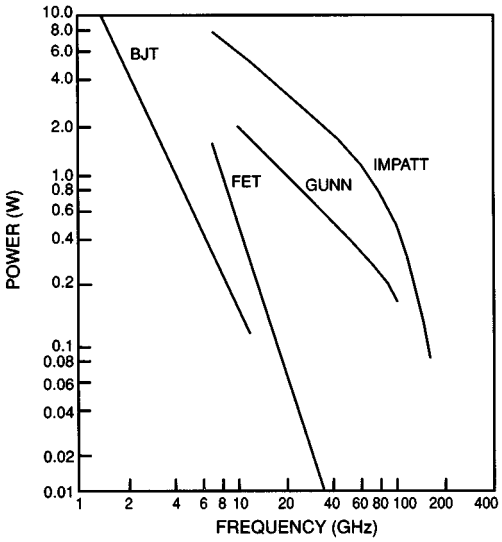


FIGURE 43.7 Maximum CW power obtained from solid state microwave oscillators.

TABLE 43.4 Summary of State-of-the-Art Performance for Millimeter Wave Frequency Multipliers

Mount Type	Tunable Output Operating Band (GHz)	Minimum Output		Maximum Output			Maximum Pump Power (mW)	Notes[a]
		Effic. (%)	Power (mW)	Effic. (%)	Power (mW)	Freq. (GHz)		
Doubler	180–120	9.5	18	14.0	26.6	188 and 105	190	2, 3, 9
	180–120	10.7	16	15.5	23.2	100	150	1, 2, 3
	180–120	10	7	16	11	104	70	1, 4, 3
	100	—	—	25	20	100	80	6, 4
	110–170	10	8	15	12.0	120	80	1, 2, 3
	140–150	10	8	22	17.6	145	80	1, 2, 3, 5
	190–260	10	8	27	21.5	215	80	1, 2, 3
	200	—	—	19	18	200	150	6, 4
	400	—	—	8.5	10.44	300	5.1	1, 2, 3, 7
	500–600	7	0.7	—	—	—	10	1, 2, 8
Tripler	85–115	4	1.2	8	2.4	106	28	1, 2, 8
	96–120	1.8	1.8	8.2	8.2	110	100	1, 2, 3
	105	—	—	25	18	105	72	6, 4
	200–290	2.5	2.0	7.5	6	225	80	1, 2, 3
	190–240	1	0.3	10	3	230	30	1, 2, 8
	260–350	1.8	1.5	3.75	3.0	340	80	2, 3, 6
	300	—	—	2	2	300	100	6, 4
	450	—	—	1	0.079	450	6.3	1, 2, 3, 7
× 6 balanced doubler/tripler	310–350	0.3	0.6	0.4	0.75	345	190	1, 2, 3, 6, 9

[a] 1, Crossed waveguide mount; 2, tuning and bias optimized at each operating frequency; 3, microstrip low-pass filter; 4, fixed tuning and bias; 5, narrowbanded version of NRAO 110- to 170-GHz doubler; 6, quasi-optical mount; 7, limited pump power available; 8, coaxial low-pass filter; 9, two-diode balanced cross guide mounts.

43.6 Control Circuits

Control components are widely used in communication, radar, EW, instrument, and other systems for controlling the signal flow or to adjust the phase and amplitude of the signal [Bahl and Bhartia, 1988; Chang, 1990; Sharma, 1989; Sokolov, 1991]. PIN diodes and MESFETs are extensively used in HMICs and MMICs, respectively, for microwave control circuits, such as switches, phase shifters, attenuators, and limiters. PIN diode circuits have low loss and can handle higher power levels than do MESFET components; conversely, the latter have great flexibility in the design of integrated subsystems, consume negligible power, and are low cost.

Figure 43.10 shows various control configurations being developed using PIN and MESFET devices. Either device can be used in these circuits.

The most commonly used configuration for microwave switches is the single-pole double throw (SPDT) as shown in Fig. 43.10(a), which requires a minimum of two switching devices (diodes or transistors). Table 43.5 provides typical performance for broadband SPDT switches developed using GaAs MESFET monolithic technology. Table 43.5 also summarizes performance for phase shifters and attenuators, which are described briefly below.

There are four main types of solid state digitally controlled phase shifters: switched line, reflection, loaded line, and low-pass/high-pass, as shown in Fig. 43.10(b). The switched-line and low-pass/high-pass configurations, which are most suitable for broadband applications and compact size, are not suitable for analog operation. Reflection and loaded-line phase shifters are inherently narrowband; however, the loaded-line small bit phase shifters, 22.5 degrees or less, can be designed to have up to an octave bandwidth. Phase shifters using the vector-modulator concept have also been developed in monolithic form.

Voltage-controlled variable attenuators are important control elements and are widely used for automatic gain control circuits. They are indispensable for temperature compensation of gain variation in broadband

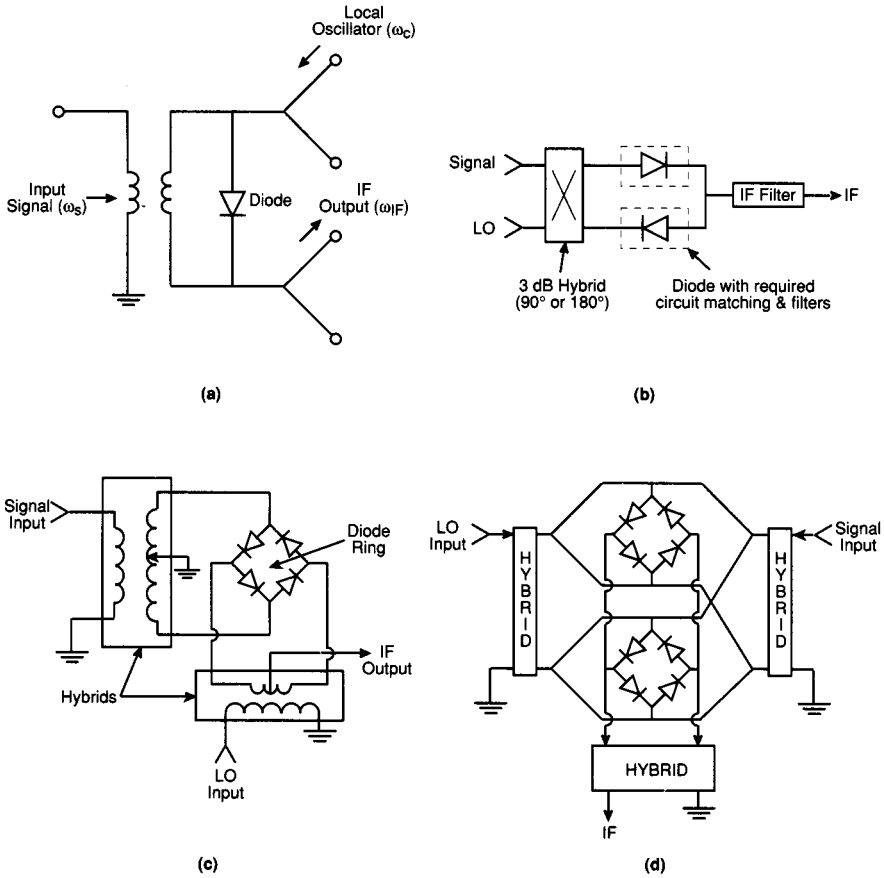


FIGURE 43.8 Basic mixer configurations: (a) single ended, (b) balanced, (c) double balanced, and (d) double-double balanced.

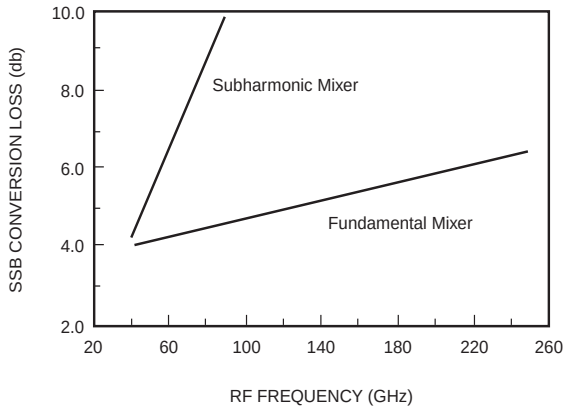


FIGURE 43.9 Single-sideband (SSB) conversion loss of millimeter wave mixers. Subharmonic type mixers have higher conversion loss but are generally less expensive.

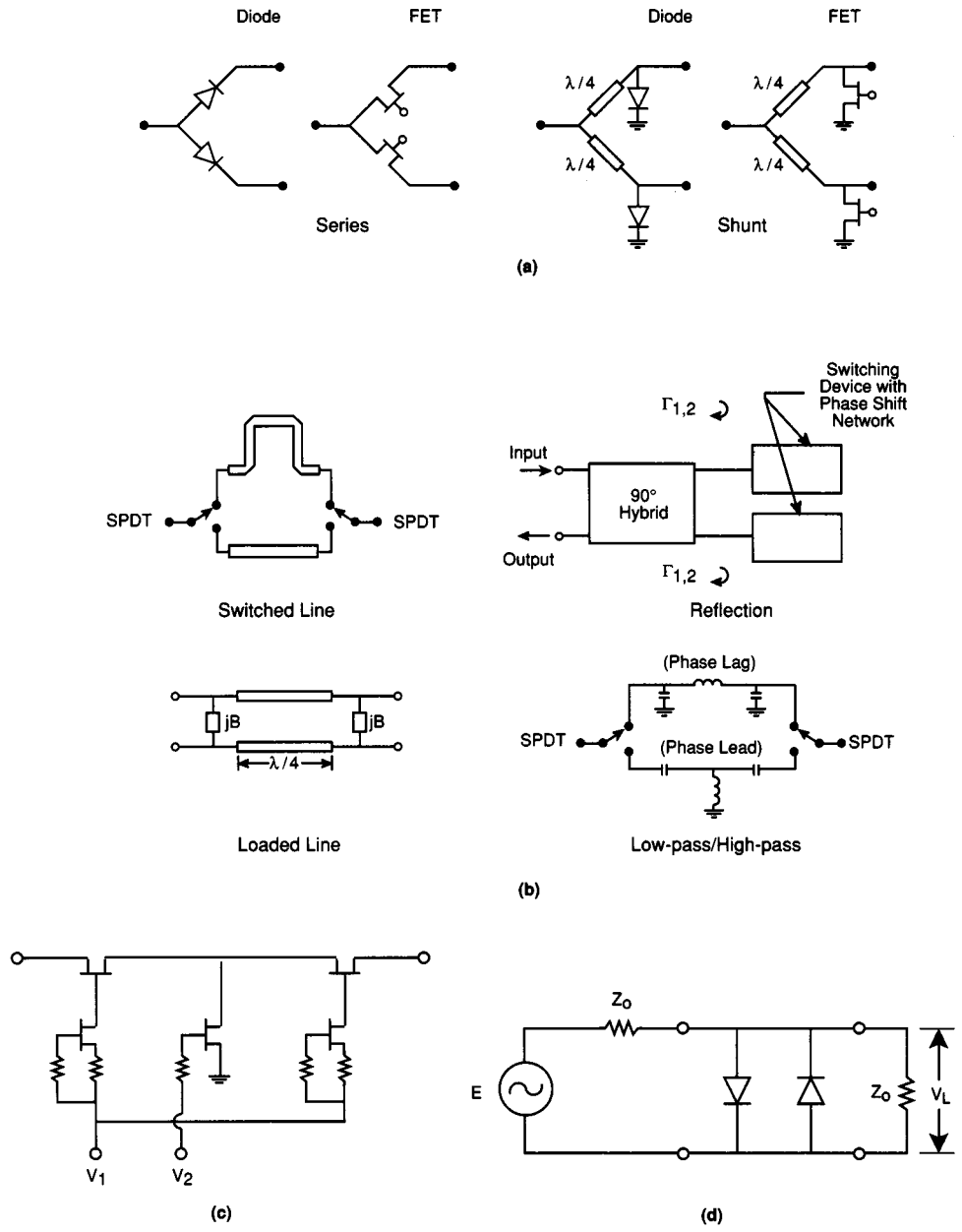


FIGURE 43.10 Microwave control circuits: (a) SPDT switch configurations, (b) basic phase shifter types, (c) schematic of a MESFET attenuator, and (d) basic limiter circuit using two rectifying diodes.

amplifiers. Both PIN diodes and GaAs MESFET devices are used for variable attenuators. Figure 43.10(c) shows a variable attenuator configuration using MESFETs in the passive mode. Apart from the use of MESFETs in the passive mode, active MESFET amplifier circuits may also be used for variable attenuation circuits. **Dual-gate MESFET** amplifiers with controlled voltage applied to the second gate are ideal for this purpose and provide a lower noise figure than the passive attenuators.

A ***limiter***, whose basic configuration is shown in Fig. 43.10(d), is an important control component used at microwave frequencies. An ideal limiter has no attenuation when low power is incident upon it but has an attenuation that increases with increasing power (above a threshold level) to maintain constant output power. Limiters are also used to protect the receivers from other nearby radar transmitters. Schottky and PIN diodes

TABLE 43.5 Performance of GaAs FET Control Monolithic ICs

Circuit Function	Frequency (GHz)	Maximum Insertion Loss (dB)	Minimum Return Loss (dB)	RMS Error (deg.)	Minimum Isolation (dB)	Minimum Dynamic Range (dB)	Maximum Power (dBm)	Minimum Switching Speed (ns)	Size (mm2)
Broadband SPDT switch	DC-20	12.0	17	—	24	—	25	0.4	1.27 × 1.27
	DC-40	13.0	17	—	23	—	25	0.8	0.84 × 1.27
2 × 2 Switch matrix	DC-20	13.5	10	—	30	—	—	—	2.03 × 1.78
6-Bit phase shifter	5–6	16.0	10	±2.0	—	—	—	5.0	9.43 × 4.20
3-Bit phase shifter	6–18	11.5	13	±8.5	—	—	—	—	2.2 × 1.25
Variable attenuator	DC-50	14.2	19	—	—	30	—	1.0	1.52 × 0.65
Multibit attenuator	DC-20	15.0	12	—	—	75	26	—	2.55 × 1.57

are commonly used to realize these components. Because they are used at the front ends of receivers, low loss is one of the basic requirements.

43.7 Summary and Future Trends

In this section we have briefly described the performance of microwave and millimeter wave solid state circuits. Currently the discrete silicon bipolar transistors, IMPATT diodes, and GaAs MESFETs are most widely used in solid state circuits and bipolar transistors, and IMPATT diodes are still the most powerful solid state sources below S-band and 10 to 300 GHz, respectively. The MESFET is the workhorse for microwave circuits up to 40 GHz. The HEMT applications are similar to the MESFET, including amplifiers, oscillators, frequency multipliers, mixers, and control circuits, and above 50 GHz, this device is exclusively used for power and low-noise amplifiers. The HBTs applications include high-efficiency amplifiers, ultrabroadband dc amplifiers, low-noise oscillators, frequency multipliers, and mixers.

During the past decade, GaAs monolithic technology has made tremendous progress in producing single and multifunction microwave and millimeter wave circuits using MESFETs, HEMTs, and HBTs. This technology is exerting a profound impact in producing low-cost and high-volume solid state circuits. Since monolithic circuits have the advantage of bandwidth and reproducibility over discrete devices because of wire bond elimination, the growth in millimeter wave solid state circuits will be based on this technology, and a cost reduction by a factor greater than 10 is expected in the near future. Wherever possible, the two-terminal devices such as IMPATT and Gunn diodes will be replaced by the transistors. Future trends in solid state circuits also include their optical control, tuning, and stabilization.

Defining Terms

Amplifier: Active two-port device with signal of higher amplitude than the input signal while retaining the essential signal characteristics of the input signal.

Attenuator: Two-port device with output signal of lower amplitude than the input signal while retaining the essential signal characteristics of the input signal.

Bandwidth (BW): A measure of the frequency range over which the circuit performs to specified parameters such as gain, noise figure, power output, etc.

Computer Aided Design (CAD): A design tool that constitutes circuit simulators and optimization programs/algorithms to aid the design of microwave circuits to meet the specified performance goals.

Dual-Gate MESFET: This device is similar in operation to MESFET but possesses two gates: an RF gate and a control gate, instead of a single RF gate.

Gain: The ratio of the output signal to the input signal of an amplifier.

Hybrid microwave integrated circuit (HMIC): A planar assembly that combines different circuit functions formed by strip or microstrip transmission lines printed onto a dielectric substrate and incorporating discrete semiconductor solid state devices and passive distributed or lumped circuit elements, interconnected with wire bonds.

Limiter: A circuit in which the output power is prevented from exceeding a predetermined value.

Mixer: A three-port device in which the output signal is a harmonic of the add mixture of an input signal and a local oscillator signal. An up converter is a mixer in which the output signal frequency is above the input signal. A down converter is a mixer in which the output signal frequency is below the input signal.

Monolithic microwave integrated circuit (MMIC): An MMIC is formed by fabricating all active and passive circuit elements or components and interconnections onto or into the surface of a semi-insulating semiconducting substrate by deposition and etching schemes such as epitaxy, ion implantation, sputtering, evaporation, and/or diffusion, and utilizing photolithographic processes for pattern definition, thus eliminating the need for internal wire bond interconnects.

Multiplier: A two-port device in which the output signal is a harmonic of the input signal.

Noise Figure: The noise figure of any linear two-port circuit is defined as the signal-to-room temperature thermal noise ratio at the input divided by the signal-to-noise ratio at the output.

Oscillator: An active one-port device which produces a nominally frequency stable constant amplitude signal.

PIN Diode: A two-port semiconductor device in which a p doped contact is isolated from an *n*-doped contact by an intrinsic region forming an anisotropic junction.

Phase Shifter: A circuit that provides a shift in the phase of the output signal with respect to a reference value.

Return Loss: Ratio of reflected power to input power of a signal port.

Switch: A circuit designed to close or open one or more transmission paths for the microwave signals.

Voltage Standing Wave Ratio (VSWR): Ratio of maximum voltage amplitude to the minimum voltage amplitude at the specified port.

Related Topics

25.1 Integrated Circuit Technology • 25.3 Application-Specific Integrated Circuits • 37.2 Waveguides

References

E. L. Kollberg (Ed.), Microwave and Millimeter-Wave Mixers, New York: IEEE Press, 1984.

P. Bhartia and I. J. Bahl, *Millimeter Wave Engineering and Applications,* New York: John Wiley, 1984.

R. A. Pucel (Ed.), *Monolithic Microwave Integrated Circuits,* New York: IEEE Press, 1985.

S. A. Maas, *Microwave Mixers,* Norwood, Mass.: Artech House, 1986.

I. J. Bahl and P. Bhartia, *Microwave Solid State Circuit Design,* New York: John Wiley, 1988.

R. Goyal (Ed.), *Monolithic Microwave Integrated Circuits: Technology and Design,* Norwood, Mass.: Artech House, 1989.

F. Ali, I. Bahl, and A. Gupta (Eds.), *Microwave and Millimeter-Wave Heterostructure Transistors and Their Applications,* Norwood, Mass.: Artech House, 1989.

K. Chang (Ed.), *Handbook of Microwave and Optical Components,* vol. 2, New York: John Wiley, 1990.

G. D. Vendelin, A. M. Pavio, and U. L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques,* New York: John Wiley, 1990.

F. Ali and A. Gupta (Eds.), *HEMTs and HBTs: Devices, Fabrication and Circuits,* Norwood, Mass.: Artech House, 1991.

K. Chang, *Microwave Solid-State Circuits and Applications,* New York: John Wiley, 1994.

D. Willems and I. Bahl, "Advances in monolithic microwave and millimeter wave integrated circuits," *IEEE Int. Circuits and System Symp. Digest,* pp. 783–786, 1992.

H. Q. Tserng and P. Saunier, "Advances in power MMIC amplifier technology in space communications," *Proc. SPIE-Monolithic Microwave Integrated Circuits for Sensors, Radar, and Communications Systems,* pp. 74–85, 1991.

A. K. Sharma, "Solid-state control devices: State of the art," *Microwave Journal,* 1989 State of the Art Reference, pp. 95–112, Sept. 1989.

V. Sokolov, "Phase shifters technology assessment: Prospects and applications," *Proc. SPIE-Monolithic Microwave Integrated Circuits for Sensors, Radar and Communications Systems,* vol. 1475, pp. 228–332, 1991.

D. Fisher and I. Bahl, *Gallium Arsenide IC Applications Handbook,* San Diego: Academic Press, 1995.

Further Information

The monthly journal *IEEE Transactions on Microwave Theory and Techniques* routinely publishes articles on the design and performance of solid state circuits. Special issues published in July 1982, January and February 1983, March 1984, and September 1989 exclusively deal with this topic.

IEEEMicrowave and Millimeter-Wave Monolith Circuits Symposium Digests, published every year since 1982, include comprehensive information on the design and performance of monolithic microwave and millimeter-wave solid state circuits.

Books included in the references of this chapter discuss thoroughly the design, circuit implementation, and performance of solid state circuits.

Trowbridge, C.W. “Three-Dimensional Analysis”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

44

Three-Dimensional Analysis

44.1 Introduction
44.2 The Field Equations
Low Frequency Fields • Statics Limit
44.3 Numerical Methods
Finite Elements • Edge Elements • Integral Methods
44.4 Modern Design Environment

C. W. Trowbridge
Vector Fields, Inc.

44.1 Introduction

The three-dimensional analysis of electromagnetic fields requires the use of numerical techniques exploiting the best available computer systems. The well-found laboratory will have at its disposal a range of machines that allow interactive data processing with access to software that provides geometric modeling tools and has sufficient central processing unit (CPU) power to solve the large (>10,000) set of algebraic equations involved, see Section 44.4.

44.2 The Field Equations

The classical equations governing the physical behavior of electromagnetic fields over the frequency range dc to light are **Maxwell's equations**. These equations relate the magnetic flux density (**B**), the electric field intensity (**E**), the magnetic field intensity (**H**), and the electric field displacement (**D**) with the electric charge density (ρ) and electric current density (J). The field vectors are not independent since they are further related by the material constitutive properties: $B = \mu H$, $D = \epsilon E$, and $J = \sigma E$ where μ, ϵ, and σ are the material permeability, permittivity, and conductivity, respectively. In practice these quantities may often be field dependent, and furthermore, some materials will exhibit both anisotropic and hysteretic effects. It should be strongly stated that accurate knowledge of the material properties is one of the most important factors in obtaining reliable simulations.

Because the flux density vector satisfies a zero divergence condition (*div* $\mathbf{B} = 0$), it can be expressed in terms of a magnetic vector potential **A**, i.e., $\mathbf{B} = curl\ \mathbf{A}$, and it follows from Faraday's law that $\mathbf{E} = -(\partial A/\partial t + \nabla V)$, where V is the electric scalar potential. Neither A nor V is completely defined since the gradient of an arbitrary scalar function can be added to A and the time derivative of the same function can be subtracted from V without affecting the physical quantities E and B. These changes to A and V are the gauge transformations, and uniqueness is usually ensured by specifying the divergence of A and sufficient boundary conditions. If $\nabla \cdot A = -(\mu\sigma V + \mu\epsilon\partial V/\partial t)$ (Lorentz gauge) is selected, then the field equations in terms of A and V are:

$$\nabla \times \frac{1}{\mu} \nabla \times \mathbf{A} + \sigma \frac{\partial \mathbf{A}}{\partial t} + \epsilon \frac{\partial^2 \mathbf{A}}{\partial t^2} = \nabla \left[\frac{1}{\mu} \nabla \cdot \mathbf{A} \right]$$
$$\mu\epsilon \frac{\partial^2 V}{\partial t^2} + \mu\sigma \frac{\partial V}{\partial t} = \nabla \cdot \nabla V \tag{44.1}$$

where σ has been assumed piecewise constant. This choice of gauge decouples the vector potential from the scalar potential. For the important class of two-dimensional problems there will be only one component of A parallel to the excitation current density. For fields involving time, at least two types can be distinguished: the time harmonic (ac) case in which the fields are periodic at a given frequency ω, i.e., $A = A_o \exp(j\omega t)$, and the transient case in which the time dependence is arbitrary.

Low Frequency Fields

In the important class of problems belonging to the low frequency limit, i.e., eddy current effects at power frequencies, the second derivative terms with respect to time (wave terms) in Eq. (44.1) vanish. This approximation is valid if the dimensions of the material regions are small compared with the wavelength of the prescribed fields. In such circumstances the displacement current term in Maxwell's equations is small compared to the free current density and there will be no radiation [Stratton, 1941]. In this case, while a full vector field solution is necessary in the conducting regions, in free space regions, where $\sigma = 0$ and curl $H = J_s$ Eqs. (44.1) can be replaced by $\nabla^2\psi = 0$, where ψ is a scalar potential defined by $H = -\nabla\psi$. The scalar and vector field regions are coupled together by the standard interface conditions of continuity of normal flux (B) and tangential field (H).

Statics Limit

In the statics limit (dc) the time-dependent terms in Eq. (44.1) vanish, and the field can be described entirely by the Poisson equation in terms of a single component scalar potential, which will be economic from the numerical point of view. In this case the defining equation is

$$\nabla \cdot \mu\nabla\phi = \nabla \cdot \mu H_s \tag{44.2}$$

where ϕ is known as the reduced magnetic scalar potential with $H = H_s - \nabla\phi$, and H_s the source field given by the Biot Savart law. Some care is needed in solving Eq. (44.2) numerically, in practice, as H_s will often be calculated to a higher accuracy than ϕ. For instance, in regions with high permeability (e.g., ferromagnetic materials), the total field intensity H tends to a small quantity which can lead to significant errors due to cancellation between grad ϕ and H_s, depending upon how the computations are carried out. One approach that avoids this difficulty is to use the total scalar potential ψ in regions that have zero current density [Simkin and Trowbridge, 1979], i.e., where $H = -\nabla\psi$ and H_c is the coercive field for the material ψ satisfies

$$\nabla \cdot \mu\nabla\psi = \nabla \cdot \mu H_c \tag{44.3}$$

Again, the two regions are coupled together by the standard interface condition that results, in this case, in a potential "jump" obtained by integrating the tangential continuity condition, i.e.,

$$\phi = \psi + \int_0^\Gamma \mathbf{H}_s d\Gamma \tag{44.4}$$

where Γ is the contour delineating the two regions that must not intersect a current-carrying region; otherwise the definition of ψ will be violated.

44.3 Numerical Methods

Numerical solutions for the field equations are now routine for a large number of problems encountered in magnet design; these include, for example, two-dimensional models taking into account nonlinear, anisotropic, and even hysteretic effects. Their use for complete three-dimensional models is not so widespread because of the escalating computer resources needed as the problem size increases. Nevertheless, 3-D solutions for nonlinear statics devices are regularly obtained in industry, and time-dependent solutions are rapidly becoming more cost effective as computer hardware architectures develop.

Finite Elements

This increasing use of computer-based solutions has come about largely because of the generality of the finite element method (FEM). In this method, the problem space is subdivided (discretized) into finite regions (elements) over which the solution is assumed to follow a simple local approximating trial function (shape functions). In the simplest situation, a particular element could be a plane hexahedra defined by its eight vertices or nodes and a solution of Eq. (44.3) may be approximated by

$$\psi \approx u = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 z + \alpha_5 xy + \alpha_6 yz + \alpha_7 zx + \alpha_8 xyz = \sum N_i U_i \quad (44.5)$$

Because a hexahedra has eight nodes it is natural to select a bilinear trial function with eight parameters; see Fig. 44.1 for other examples. The functions N_i are called the local shape functions and the parameters U_i are the solution values at the nodes. The finite elements can be integrated into a numerical model for the whole problem space either by (a) the variational method in which the total energy of the system is expressed in terms of the finite element trial functions and then minimized to determine the best solution or (b) the weighted residual method in which the formal error (residual), arising by substituting the trial functions into the defining equation, is weighted by a suitably chosen function and then integrated over the problem domain. The best fit for the trial function parameters can then be obtained by equating the integral to zero. Both methods lead to a set of algebraic equations and are equivalent if the weighting functions are chosen to be the trial functions (Galerkin's method [Zienkiewicz, 1990]). At the element level, the residual R_i is given by

$$R_i = \left[\int_{\text{elem}} \nabla N_i \mu \nabla N_j d\Omega \right] U_j + \int_{\text{elem}} N_i Q d\Omega \quad (44.6)$$

where Q (RHS) denotes the sources of Eqs. (44.2) or (44.3). The integrals can be readily evaluated and assembled for the whole problem by superposition, taking account of the boundary conditions and removing the redundancy at shared nodes. At interelement boundaries in a region of particular potential [reduced Eq. (44.2) or total Eq. (44.3)] the solution is forced to be continuous, but the normal flux (i.e., $\mu \partial U/\partial n$) will only be continuous in a weak sense, that is to say the discontinuity will depend upon the mesh refinement.

The FEM provides a systematic technique for replacing the continuum field equations with a set of discrete algebraic equations that can be solved by standard procedures. In Fig. 44.2 a typical field map is shown for a permanent magnet machine modeled by a computer simulator that can take into account nonlinearity and permanently magnetized materials. Although hysteresis effects can be included, the computational resources required can be prohibitive because of the vector nature of magnetization. The magnetic material must be characterized by a large number of measurements to take account of the minor loops, and from these the convolution integrals necessary to obtain the constitutive relationships can be evaluated [Mayergoyz, 1990]. These characteristics must then be followed through time; this can be implemented by solving at a discrete set of time points, given the initial conditions in the material.

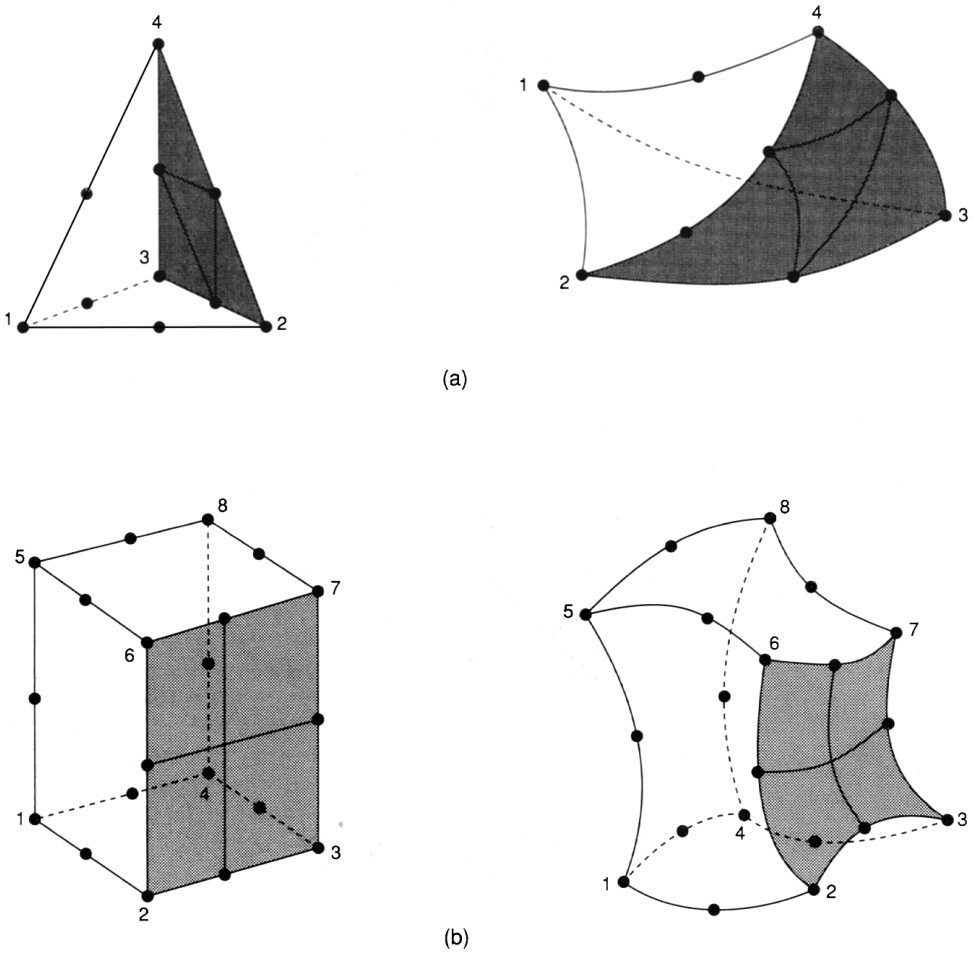


FIGURE 44.1 Three-dimensional second-order Isoparametric finite elements. (a) *Left*, master tetrahedron, 10 nodes in local space (ξ, η, ζ); right, actual tetrahedron, 10 nodes in global space (x, y, z). (b) Left, master hexahedron, 20 nodes in local space (ξ, η, ζ); right, actual hexahedron, 10 nodes in global space (x, y, z).

Although the FEM is widely used by industry for electromagnetic problems covering the entire frequency range, there are many situations where special methods are more effective. This is particularly the case for high-frequency problems, e.g., millimeter and microwave integrated circuit structures where integral equation techniques and such procedures as transmission line modeling (TLM), spectral domain approach, method of lines, and wire grid methods are often preferred [Itoh, 1989] (see Chapter 43).

Edge Elements

Using potentials and nodal finite elements (see Fig. 44.1) rather than field components directly has the advantage that difficulties arising from field discontinuities at material interfaces can be avoided. However, if the element basis functions [see Eq. (44.5)] are expressed in terms of the field (**H**, say) constrained along an element edge, then tangential field continuity is enforced [Bossavit, 1988]. The field equations [Eq. (44.1)] in terms of the field intensity for the low frequency limit reduce to

$$\nabla \times \nabla \times \mathbf{H} + \sigma \frac{\partial(\mu \mathbf{H})}{\partial t} = 0 \tag{44.7}$$

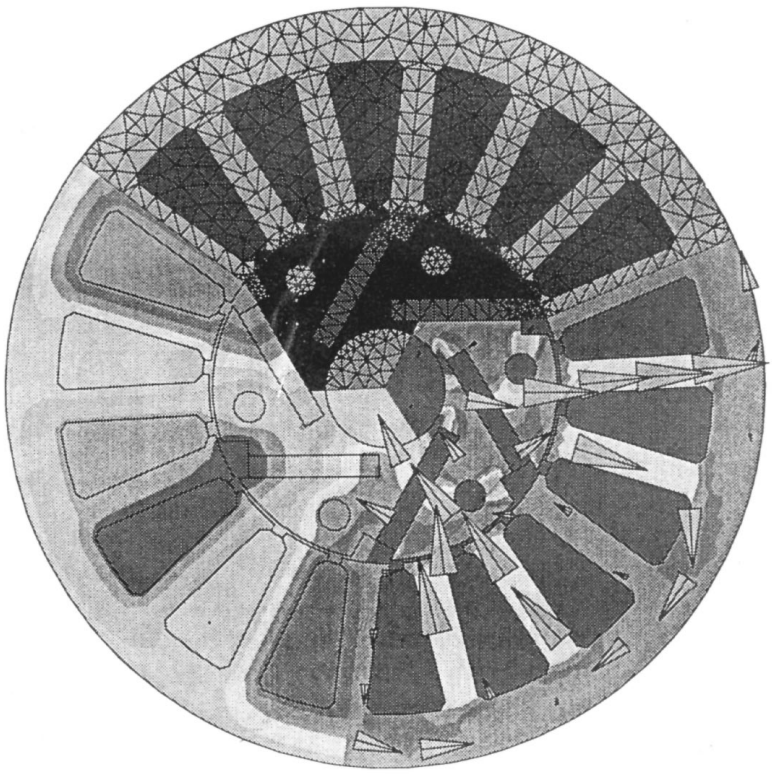


FIGURE 44.2 Permanent magnet motor.

and a suitable edge variable basis function form for solving this equation by finite elements using tetrahedral elements is

$$\mathbf{h}(\mathbf{r}) = \mathbf{a} + \mathbf{b} \times \mathbf{r} \tag{44.8}$$

where **r** is the position vector and **a** and **b**, respectively, are vectors dependent on the geometry of the element. The basis function expansion is given by

$$\mathbf{H} = \sum \mathrm{h_e(r)H_e} \tag{44.9}$$

where $\mathbf{h}_e$ is the vector basis function for edge **e**, and $\mathbf{H}_e$ is the value of the field along an element edge (see Fig. 44.3). The functions, Eqs. (44.8) and (44.9), have the property of being divergence free, and most important they ensure that the tangential component of **H** is continuous while allowing for the possibility of a discontinuity in the normal component. In nonconducting regions where the field can be economically modeled by a scalar potential, standard nodal elements can be used. At the interface the edge elements couple exactly with the nodal elements.

Integral Methods

An alternative procedure is to solve the field equations in their integral form, see also Chapter 43. For example, in magnetostatics, the magnetization vector **M** given by $\mathbf{M} = (\mu - 1)\mathrm{H}$ can be used instead of H to derive an integral equation over all ferromagnetic domains of the problem, i.e.,

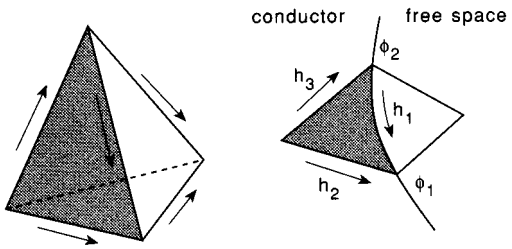


FIGURE 44.3 Edge variables for a tetrahedron element, $\mathbf{h}_1 = (\Phi_2 - \Phi_1)/l$.

$$\mathbf{M}(\mathbf{r}) = (\mu - 1)\left[\mathbf{H}_s(\mathbf{r}') - \frac{1}{4\pi}\nabla\int_{\Omega}\mathbf{M}\cdot\nabla\left(\frac{1}{R}\right)d\Omega\right] \tag{44.10}$$

where R is the distance between the source and field point, respectively. For problems with linear materials Eq. (44.10) reduces to integrations over the bounding surfaces of materials in terms of the magnetic scalar potential, i.e.,

$$4\pi\phi = -\int_{\Gamma}\left(\frac{1}{R}\frac{\partial\phi}{\partial n} - \phi\frac{\partial 1/R}{\partial n}\right)d\Gamma \tag{44.11}$$

Equation (44.11) can be solved numerically by the boundary element method (BEM) in which the active surfaces are discretized into elements. The advantages of integral formulations compared to the standard differential approach using finite elements are (a) only active regions need to be discretized, (b) the far field boundary condition is automatically taken into account, and (c) the fields recovered from the solution are usually very smooth. Unfortunately, the computational costs rapidly escalate as the problem size increases because of the complexity of the system coefficients and because the resulting matrix is fully populated, whereas in the differential approach the coefficients are simple and the matrix is sparse, allowing the exploitation of fast equation solution methods.

44.4 Modern Design Environment

The most common system used in software packages is one in which the pre-processor includes data input, model building, and mesh (element) generation. Although fully automated meshing is now a practical possibility it needs to be combined with error estimation in order to allow the generation of optimal meshes. This approach is now common for 2-D systems and is available in many 3-D systems. Figure 44.4 illustrates a field simulation environment in which the solution processor includes an adaptive mesh generator controlled by *a posteriori* error estimation. This avoids the costly and essentially heuristic task of mesh generation which, in the past, had to be performed by the designer. Furthermore a modern system should have solid modeling capabilities driven by parametric data allowing the user to specify the appropriate engineering quantities, e.g., in the case of a solid cylinder the radius and length are all that is needed to specify the geometry at some predefined location. The system should also be supported by a database which is compliant with evolving standards such as STEP (Standard for the Exchange of Product data-ISO 10303 [Owen, 1993]) thus allowing data communication between other systems.

The environment illustrated in Fig. 44.4 also shows tools for automatic optimization that are now becoming feasible in industrial design applications. Both deterministic and stochastic methods for minimizing constrained objective functions of the design space have been developed for electromagnetic applications (For a review see Russenschuck, 1996). It must be emphasized, however, that the use of optimizing methods is only part of the total problem of design. For example, the process of automatic synthesis based on design rules and engineering

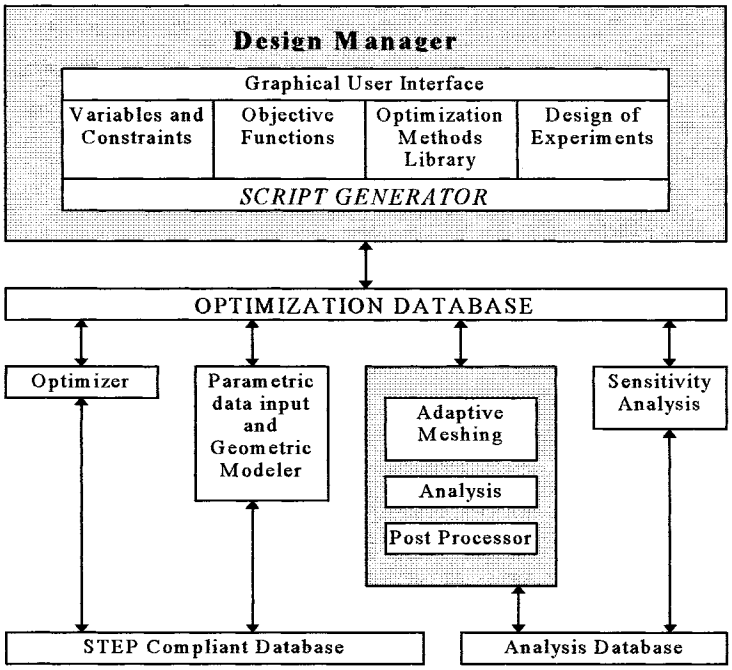


FIGURE 44.4 Electromagnetic design environment.

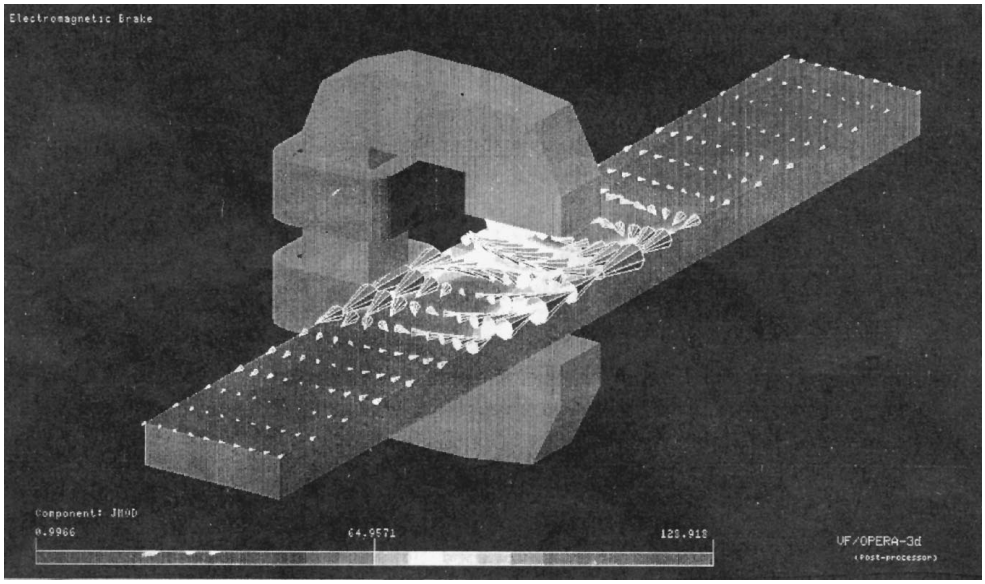


FIGURE 44.5 Moving copper strip, retarded by an electromagnet.

knowledge may provide engineers with a complementary methodology to assist in the creativity that is the essence of design (see Lowther, 1996). An example of a finite element solution displayed on a modern workstation for a three-dimensional eddy current problem with a moving conductor retarded by a *c-core* electromagnet is shown in Fig. 44.5. The post-processor indicates the magnitude and direction of the induced eddy

currents in the moving copper strip by the solid *cones* and the magnetic field by the gray scaled contours. This modest problem was modeled with 5,664 degrees of freedom (No. of equations) and needed 198 cp seconds running on a workstation rated at SPECfp92 96.5. However most industrial problems will require many more degrees of freedom, and typically a non-linear magnetostatic problem with 200,000 equations needed 75 mintues and 25 Mbytes of RAM on the same machine. The resources needed for transient non-linear problems will be far greater.

Defining Terms

Biot Savart law: $$\mathrm{H}_s = \frac{1}{4\pi}\int_{\Omega} \mathrm{J}_s \times \nabla \frac{1}{R} d\Omega$$

where **R** is the distance from the source point to the field point.

Interface conditions:
$$(\mathbf{B}_2 - \mathbf{B}_1) \cdot \mathbf{n} = \mathbf{0}$$
$$(\mathbf{D}_2 - \mathbf{D}_1) \cdot \mathbf{n} = \omega$$
$$(\mathbf{H}_2 - \mathbf{H}_1) \times \mathbf{n} = \mathbf{K}$$
$$(\mathbf{E}_2 - \mathbf{E}_1) \times \mathbf{n} = \mathbf{0}$$

where **K** and ω are the surface current and charge densities, respectively.

Maxwell's equations:
$$\nabla \cdot \mathrm{D} = \rho \quad \text{(Gauss's law)}$$
$$\nabla \cdot \mathrm{B} = 0$$
$$\nabla \times \mathrm{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad \text{(Faraday's law)}$$
$$\nabla \times \mathrm{H} = \mathrm{J} + \frac{\partial \mathbf{D}}{\partial t} \quad \text{(Ampere's law + displacement current)}$$

Related Topics

35.1 Maxwell Equations • 45.1 Introduction • 45.3 Analytical Issues in Developing a Computer Model

References

A. Bossavit, "Rationale for 'edge elements' in 3-D field computation," *IEEE Trans. on Magnetics*, vol. 24, no. 1, p. 74, 1988.

I. Tasuo (Ed.), *Numerical Techniques for Microwave and Millimeter-Wave Passive Structures*, New York: John Wiley, 1989.

D. A. Lowther, "Knowledge-based and numerical optimization techniques for the design of electromagnetic devices," *IJ Num. Mod.*, vol. 9, no. 1,2, pp. 35–44, 1996.

I. Mayergoyz, *Mathematical Models of Hysteresis*, New York: Springer-Verlag, 1990.

J. Owen, *STEP An Introduction,* Information Geometers Ltd, 47 Sockers Avenue, Winchester, UK, 1993.

S. Russenschuck, "Synthesis, inverse problems and optimization in computational electromagnetics," *IJ Num. Mod.*, vol. 9, no. 1,2, pp. 45–58, 1996.

P. P. Silvester and R. L. Ferrari, *Finite Elements for Electrical Engineers*, 2nd ed., Cambridge: Cambridge University Press, 1990.

J. Simkin and C. W. Trowbridge, "On the use of the total scalar potential in the numerical solution of field problems in electromagnetics," *IJNME*, vol. 14, p. 423, 1979.
J. A. Stratton, *Electromagnetic Theory*, New York: McGraw Hill, 1941.
O. C. Zienkiewicz, *The Finite Element Method*, 3rd ed., New York: McGraw Hill, 1990.

Further Information

Conferences on Computation of Electromagnetic Fields *Compumag* Proceedings:

Oxford, UK, 1976 (Ed. J. Simkin) Rutherford Appleton Laboratory, Chilton, Oxon, UK.
Grenoble, France, 1979 (Ed. J. C. Sabonnadiere) ERA 524 CNRS, Grenoble, France.
Chicago, USA, 1981 (Ed. L. Turner) *IEEE Trans. Mag.* 18 (2) 1982.
Genoa, Italy, 1983 (Ed. G. Molinari) *IEEE Trans. Mag.* 19 (6) 1983.
Fort Collins, USA, 1985 (Ed. W. Lord) *IEEE Trans. Mag.* 21 (6) 1985.
Graz, Austria, 1987 (Ed. K. Richter) *IEEE Trans. Mag.* 24 (1) 1988.
Tokyo, Japan, 1989 (Ed. K. Miya) *IEEE Trans. Mag.* 26 (2) 1990.
Sorrento, Italy, 1991 (Ed. R. Martone) *IEEE Trans. Mag.* 28 (2) 1992.
Miami, USA, 1993 (Ed. D. A. Lowther) *IEEE Trans. Mag.* 30 (5) 1994.
Berlin, Germany, 1995 (Ed. O Biro) *IIEEE Trans. Mag.* To be published May 1996.

Conference on Electromagnetic Field Computation, *CEFC* Proceedings

Washington, USA, 1988 (Ed. I. Mayergoyz), *IEEE Trans. Mag.* 25 (4), 1989.
Toronto, Canada, 1990 (Ed. J. Lavers), *IEEE Trans. Mag.* 27 (5), 1991.
Claremont, USA, 1992 (Ed. S. R. Hoole), *IEEE Trans. Mag.*
Aix-les-Bains, France, 1994 (Ed. J. C. Sabonnadiere), *IEEE Trans. Mag.* 31 (3), 1995.

Special Issue on Computational Magnetics (Eds. J. Sykulski and P. Silvester)
Int. Jour. Of Num. Mod. J. Wiley, vol. 9, no. 1 & 2, 1996.

Miller, E.K. “Computational Electromagnetics”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

45

Computational Electromagnetics

45.1 Introduction

45.2 Background Discussion
Modeling as a Transfer Function • Some Issues Involved in Developing a Computer Model

45.3 Analytical Issues in Developing a Computer Model
Selection of Solution Domain • Selection of Field Propagator

45.4 Numerical Issues in Developing a Computer Model
Sampling Functions • The Method of Moments

45.5 Some Practical Considerations
Integral Equation Modeling • Differential Equation Modeling • Discussion • Sampling Requirements

45.6 Ways of Decreasing Computer Time

45.7 Validation, Error Checking, and Error Analysis
Modeling Uncertainties • Validation and Error Checking

45.8 Concluding Remarks

E.K. Miller
Los Alamos National Laboratory

45.1 Introduction

The continuing growth of computing resources is changing how we think about, formulate, solve, and interpret problems. In electromagnetics as elsewhere, computational techniques are complementing the more traditional approaches of measurement and analysis to vastly broaden the breadth and depth of problems that are now quantifiable. Computational electromagnetics (CEM) may be broadly defined as that branch of electromagnetics that intrinsically and routinely involves using a digital computer to obtain numerical results. With the evolutionary development of CEM during the past 20-plus years, the third tool of computational methods has been added to the two classical tools of experimental observation and mathematical analysis.

This discussion reviews some of the basic issues involved in CEM and includes only the detail needed to illustrate the central ideas involved. The underlying principles that unify the various modeling approaches used in electromagnetics are emphasized while avoiding most of the specifics that make them different. Listed throughout are representative, but not exhaustive, numbers of references that deal with various specialty aspects of CEM. For readers interested in broader, more general expositions, the well-known book on the moment method by Harrington [1968]; the books edited by Mittra [1973, 1975], Uslenghi [1978], and Strait [1980]; the monographs by Stutzman and Thiele [1981], Popovic, et al. [1982], Moore and Pizer [1984], and Wang [1991]; and an IEEE Press reprint volume on the topic edited by Miller et al. [1991] are recommended, as is the article by Miller [1988] from which this material is excerpted.

This chapter is excerpted from E. K. Miller, "A selective survey of computational electromagnetics," *IEEE Trans. Antennas Propagat.*, vol. AP-36, pp. 1281–1305, ©1988 IEEE.

45.2 Background Discussion

Electromagnetics is the scientific discipline that deals with electric and magnetic sources and the fields these sources produce in specified environments. Maxwell's equations provide the starting point for the study of electromagnetic problems, together with certain principles and theorems such as superposition, reciprocity, equivalence, induction, duality, linearity, and uniqueness, derived therefrom [Stratton, 1941; Harrington, 1961]. While a variety of specialized problems can be identified, a common ingredient of essentially all of them is that of establishing a quantitative relationship between a cause (forcing function or input) and its effect (the response or output), a relationship which we refer to as a **field propagator**, the computational characteristics of which are determined by the mathematical form used to describe it.

Modeling as a Transfer Function

The foregoing relationship may be viewed as a generalized transfer function (see Fig. 45.1) in which two basic problem types become apparent. For the analysis or the direct problem, the input is known and the transfer function is derivable from the problem specification, with the output or response to be determined. For the case of the synthesis or inverse problem, two problem classes may be identified. The easier synthesis problem involves finding the input, given the output and transfer function, an example of which is that of determining the source voltages that produce an observed pattern for a known antenna array. The more difficult synthesis problem itself separates into two problems. One is that of finding the transfer function, given the input and output, an example of which is that of finding a source distribution that produces a given far field. The other and still more difficult is that of finding the object geometry that produces an observed scattered field from a known exciting field. The latter problem is the most difficult of the three synthesis problems to solve because it is intrinsically transcendental and nonlinear.

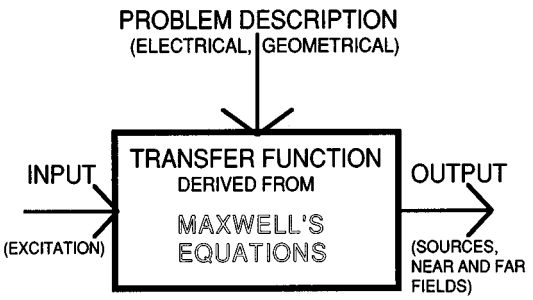


FIGURE 45.1 The electromagnetic transfer function relates the input, output, and problem.

Electromagnetic propagators are derived from a particular solution of Maxwell's equations, as the cause mentioned above normally involves some specified or known excitation whose effect is to induce some to-be-determined response (e.g., a radar cross section, antenna radiation pattern). It therefore follows that the essence of electromagnetics is the study and determination of field propagators to thereby obtain an input–output transfer function for the problem of interest, and it follows that this is also the goal of CEM.

Some Issues Involved in Developing a Computer Model

We briefly consider here a classification of model types, the steps involved in developing a **computer model**, the desirable attributes of a computer model, and finally the role of approximation throughout the modeling process.

Classification of Model Types

It is convenient to classify solution techniques for electromagnetic modeling in terms of the field propagator that might be used, the anticipated application, and the problem type for which the model is intended to be used, as is outlined in Table 45.1. Selection of a field propagator in the form, for example, of the Maxwell curl equations, a Green's function, modal or spectral expansions, or an optical description is a necessary first step in developing a solution to any electromagnetic problem.

Development of a Computer Model

Development of a computer model in electromagnetics or literally any other disciplinary activity can be decomposed into a small number of basic, generic steps. These steps might be described by different names but

TABLE 45.1 Classification of Model Types in CEM

Field Propagator	Description Based on
Integral operator	Green's function for infinite medium or special boundaries
Differential operator	Maxwell curl equations or their integral counterparts
Modal expansions	Solutions of Maxwell's equations in a particular coordinate system and expansion
Optical description	Rays and diffraction coefficients
Application	**Requires**
Radiation	Determining the originating sources of a field and patterns they produce
Propagation	Obtaining the fields distant from a known source
Scattering	Determining the perturbing effects of medium inhomogeneities
Problem type	**Characterized by**
Solution domain	Time or frequency
Solution space	Configuration or wave number
Dimensionality	1D, 2D, 3D
Electrical properties of medium and/or boundary	Dielectric, lossy, perfectly conducting, anisotropic, inhomogeneous, nonlinear, bianisotropic
Boundary geometry	Linear, curved, segmented, compound, arbitrary

TABLE 45.2 Steps in Developing a Computer Model

Step	Activity
Conceptualization	Encapsulating observation and analysis in terms of elementary physical principles and their mathematical descriptions
Formulation	Fleshing out of the elementary description into a more complete, formally solved, mathematical representation
Numerical implementation	Transforming into a computer algorithm using various numerical techniques
Computation	Obtaining quantitative results
Validation	Determining the numerical and physical credibility of the computed results

would include at a minimum those outlined in Table 45.2. Note that by its nature, validation is an open-ended process that cumulatively can absorb more effort than all the other steps together. The primary focus of the following discussion is on the issue of numerical implementation.

Desirable Attributes of a Computer Model

A computer model must have some minimum set of basic properties to be useful. From the long list of attributes that might be desired, we consider: (1) accuracy, (2) efficiency, and (3) utility the three most important as summarized in Table 45.3. Accuracy is put foremost because results of insufficient or unknown accuracy have uncertain value and may even be harmful. On the other hand, a code that produces accurate results but at unacceptable cost will have hardly any more value. Finally, a code's applicability in terms of the depth and breadth of the problems for which it can be used determines its utility.

The Role of Approximation

As approximation is an intrinsic part of each step involved in developing a computer model, we summarize some of the more commonly used approximations in Table 45.4. We note that the distinction between an approximation at the conceptualization step and during the formulation is somewhat arbitrary, but choose to use the former category for those approximations that occur before the formulation itself.

TABLE 45.3 Desirable Attributes in a Computer Model

Attribute	Description
Accuracy	The quantitative degree to which the computed results conform to the mathematical and physical reality being modeled. Accuracy, preferably of known and, better yet, selectable value, is the single most important model attribute. It is determined by the physical modeling error (ε_p) and numerical modeling error (ε_N).
Efficiency	The relative cost of obtaining the *needed* results. It is determined by the *human* effort required to develop the computer input and interpret the output and by the associated *computer* cost of running the model.
Utility	The applicability of the computer model in terms of problem size and complexity. Utility also relates to ease of use, reliability of results obtained, etc.

TABLE 45.4 Representative Approximations that Arise in Model Development

Approximation	Implementation/Implications
Conceptualization	
Physical optics	Surface sources given by tangential components of incident field, with fields subsequently propagated via a Green's function. Best for backscatter and main-lobe region of reflector antennas, from resonance region ($ka > 1$) and up in frequency.
Physical theory of diffraction	Combines aspects of physical optics and geometrical theory of diffraction, primarily via use of edge-current corrections to utilize best features of each.
Geometrical theory diffraction	Fields propagated via a divergence factor with amplitude obtained from diffraction coefficient. Generally applicable for $ka > 2$–5. Can involve complicated ray tracing.
Geometrical optics	Ray tracing without diffraction. Improves with increasing frequency.
Compensation theorem	Solution obtained in terms of perturbation from a reference, known solution.
Born–Rytov	Approach used for low-contrast, penetrable objects where sources are estimated from incident field.
Rayleigh	Fields at surface of object represented in terms of only outward propagating components in a modal expansion.
Formulation	
Surface impedance	Reduces number of field quantities by assuming an impedance relation between tangential E and H at surface of penetrable object. May be used in connection with physical optics.
Thin-wire	Reduces surface integral on thin, wirelike object to a line integral by ignoring circumferential current and circumferential variation of longitudinal current, which is represented as a filament. Generally limited to $ka < 1$ where a is the wire radius.
Numerical Implementation	
$\partial f/\partial x \rightarrow (f_+ - f_-)/(x_+ - x_-)$ $\int f(x)dx \rightarrow \sum f(x_i)\Delta x_i$	Differentiation and integration of continuous functions represented in terms of analytic operations on sampled approximations, for which polynomial or trigonometric functions are often used. Inherently a discretizing operation, for which typically $\Delta x < \lambda/2\pi$ for acceptable accuracy.
Computation	
Deviation of numerical model from physical reality	Affects solution accuracy and relatability to physical problem in ways that are difficult to predict and quantify.
Nonconverged solution	Discretized solutions usually converge globally in proportion to $\exp(-AN_x)$ with A determined by the problem. At least two solutions using different numbers of unknowns N_x are needed to estimate A.

45.3 Analytical Issues in Developing a Computer Model

Attention here is limited primarily to propagators that use either the Maxwell curl equations or source integrals which employ a Green's function, although for completeness we briefly discuss modal and optical techniques as well.

Selection of Solution Domain

Either the **integral equation** (IE) or differential equation (DE) propagator can be formulated in the time domain, where time is treated as an independent variable, or in the frequency domain, where the harmonic

time variation $\exp(j\omega t)$ is assumed. Whatever propagator and domain are chosen, the analytically formal solution can be numerically quantified via the **method of moments** (MoM) [Harrington, 1968], leading ultimately to a linear system of equations as a result of developing a discretized and sampled approximation to the continuous (generally) physical reality being modeled. Developing the approach that may be best suited to a particular problem involves making trade-offs among a variety of choices throughout the analytical formulation and numerical implementation, some aspects of which are now considered.

Selection of Field Propagator

We briefly discuss and compare the characteristics of the various propagator-based models in terms of their development and applicability.

Integral Equation Model

The basic starting point for developing an IE model in electromagnetics is selection of a Green's function appropriate for the problem class of interest. While there are a variety of Green's functions from which to choose, a typical starting point for most IE MoM models is that for an infinite medium. One of the more straightforward is based on the scalar Green's function and Green's theorem. This leads to the Kirchhoff integrals [Stratton, 1941, p. 464 *et seq.*], from which the fields in a given contiguous volume of space can be written in terms of integrals over the surfaces that bound it and volume integrals over those sources located within it.

Analytical manipulation of a source integral that incorporates the selected Green's function as part of its kernel function then follows, with the specific details depending on the particular formulation being used. Perhaps the simplest is that of boundary-condition matching wherein the behavior required of the electric and/or magnetic fields at specified surfaces that define the problem geometry is explicitly imposed. Alternative formulations, for example, the Rayleigh–Ritz variational method and Rumsey's reaction concept, might be used instead, but as pointed out by Harrington [in Miller et al., 1991], from the viewpoint of a numerical implementation any of these approaches lead to formally equivalent models.

This analytical formulation leads to an integral operator, whose kernel can include differential operators as well, which acts on the unknown source or field. Although it would be more accurate to refer to this as an integrodifferential equation, it is usually called simply an integral equation. Two general kinds of integral equations are obtained. In the frequency domain, representative forms for a perfect electric conductor are

$$\mathbf{n} \times \mathbf{E}^{\text{inc}}(\mathbf{r}) = \frac{1}{4\pi}\mathbf{n} \times \int_S \{j\omega\mu[\mathbf{n}' \times \mathbf{H}(\mathbf{r}')]\varphi(\mathbf{r}, \mathbf{r}') - [\mathbf{n}' \cdot \mathbf{E}(\mathbf{r}, \mathbf{r}')\nabla'\varphi(\mathbf{r}, \mathbf{r}')\}ds'; \quad \mathbf{r} \in S \tag{45.1a}$$

$$\mathbf{n} \times \mathbf{H}(\mathbf{r}) = 2\mathbf{n} \times \mathbf{H}^{\text{inc}}(\mathbf{r}) + \frac{1}{2\pi}\mathbf{n} \times \int_S [\mathbf{n}' \times \mathbf{H}(\mathbf{r}')] \times \nabla'\varphi(\mathbf{r}, \mathbf{r}')\}\text{ds}'; \; \mathbf{r} \in S \tag{45.1b}$$

where **E** and **H** are the electric and magnetic fields, respectively, $\mathbf{r}, \mathbf{r}'$ are the spatial coordinate of the observation and source points, the superscript inc denotes incident-field quantities, and $\varphi(\mathbf{r},\mathbf{r}') = \exp[-jk|\mathrm{r} - \mathrm{r}'|]/|\mathrm{r} - \mathrm{r}'|$ is the free-space Green's function. These equations are known respectively as Fredholm integral equations of the first and second kinds, differing by whether the unknown appears only under the integral or outside it as well [Poggio and Miller in Mittra, 1973].

Differential-Equation Model

A DE MoM model, being based on the defining Maxwell's equations, requires intrinsically less analytical manipulation than does derivation of an IE model. Numerical implementation of a DE model, however, can differ significantly from that used for an IE formulation in a number of ways for several reasons:

TABLE 45.5 Comparison of IE- and DE-Field Propagators and Their Numerical Treatment

	Differential Form	Integral Form
Field propagator	Maxwell curl equations	Green's function
Boundary treatment		
At infinity (radiation condition)	Local or global "lookback" to approximate outward propagating wave	Green's function
On object	Appropriate field values specified on mesh boundaries to obtain stairstep, piecewise linear, or other approximation to the boundary	Appropriate field values specified on object contour which can in principle be a general, curvilinear surface, although this possibility seems to be seldom used
Sampling requirements		
No. of space samples	$N_x \propto (L/\Delta L)^D$	$N_x \propto (L/\Delta L)^{D-1}$
No. of time steps	$N_t \propto (L/\Delta L) \approx cT/\delta t$	$N_t \propto (L/\Delta L) \approx cT/\delta t$
No. of excitations (right-hand sides)	$N_{\text{rhs}} \propto (L/\Delta L)$	$N_{\text{rhs}} \propto (L/\Delta L)$
Linear system L is problem size D is no. of problem dimensions (1, 2, 3) T is observation time ΔL is spatial resolution δt is time resolution	Sparse, but larger	Dense, but smaller. In this comparison, note that we assume the IE permits a sampling of order one less than the problem dimension, i.e., inhomogeneous problems are excluded.
Dependence of solution time on highest-order term in ($L/\Delta L$)		
Frequency domain	$T_\omega \propto N_x^{2(D-1)/D+1} = (L/\Delta L)^{3D-2}$	$T_\omega \propto N_x^3 = (L/\Delta L)^{3(D-1)}$
Time domain		
Explicit	$T_t \propto N_x N_t N_{\text{rhs}} = (L/\Delta L)^{D+1+r}$	$T_t \propto N_x^2 N_t N_{\text{rhs}} = (L/\Delta L)^{2D-1+r}$; $0 \le r \le 1$
Implicit	$T_t \propto N_x^{2(D-1)/D+1} = (L/\Delta L)^{3D-2}$, $D = 2, 3$; $\propto N_x N_t N_{\text{rhs}} = (L/\Delta L)^{2+r}$, $D = 1$; $0 \le r \le 1$	$T_t \propto N_x^3 = (L/\Delta L)^{3(D-1)}$

Note that D is the number of *spatial* dimensions in the problem and is not necessarily the *sampling* dimensionality d. The distinction is important because when an appropriate Green's function is available, the source integrals are usually one dimension less than the problem dimension, i.e., $d = D - 1$. An exception is an inhomogeneous, penetrable body where $d = D$ when using an IE. We also assume for simplicity that matrix solution is achieved via factorization rather than iteration but that banded matrices are exploited for the DE approach where feasible. The solution-time dependencies given can thus be regarded as upper-bound estimates. See Table 45.10 for further discussion of linear-system solutions.

1. The differential operator is a local rather than global one in contrast to the Green's function upon which the integral operator is based. This means that the spatial variation of the fields must be developed from sampling in as many dimensions as possessed by the problem, rather than one less as the IE model permits if an appropriate Green's function is available.
2. The integral operator includes an explicit radiation condition, whereas the DE does not.
3. The differential operator includes a capability to treat medium inhomogeneities, non-linearities, and time variations in a more straightforward manner than does the integral operator, for which an appropriate Green's function may not be available.

These and other differences between development of IE and DE models are summarized in Table 45.5, with their modeling applicability compared in Table 45.6.

Modal-Expansion Model

Modal expansions are useful for propagating electromagnetic fields because the source-field relationship can be expressed in terms of well-known analytical functions as an alternate way of writing a Green's function for special distributions of point sources. In two dimensions, for example, the propagator can be written in terms of circular harmonics and cylindrical Hankel functions. Corresponding expressions in three dimensions might involve spherical harmonics, spherical Hankel functions, and Legendre polynomials. Expansion in terms of analytical solutions to the wave equation in other coordinate systems can also be used but requires computation

TABLE 45.6 Relative Applicability of IE- and DE-Based Computer Models

Time Domain			Frequency Domain	
DE	IE	Issue	DE	IE
		Medium		
√	√	Linear	√	√
~	x	Dispersive	√	√
√	x	Lossy	√	√
√	~	Anisotropic	√	√
√	x	Inhomogeneous	√	x
√	x	Nonlinear	x	x
√	x	Time-varying	x	x
		Object		
~	√	Wire	~	√
√	√	Closed surface	√	√
√	√	Penetrable volume	√	√
~	√	Open surface	~	√
		Boundary Conditions		
√	√	Interior problem	√	√
~	√	Exterior problem	~	√
√	√	Linear	√	√
√	√	Nonlinear	x	x
√	√	Time-varying	x	x
~	x	Halfspace	~	√
		Other Aspects		
~	~	Symmetry exploitation	√	√
~	√	Far-field evaluation	~	√
x	~	Number of unknowns	~	√
√	~	Length of code	~	x
		Suitability for Hybridizing with Other:		
~	√	Numerical procedures	√	√
x	~	Analytical procedures	~	√
x	~	GTD	x	√

√ signifies highly suited or most advantageous.
~ signifies moderately suited or neutral.
x signifies unsuited or least advantageous.

of special functions that are generally less easily evaluated, such as Mathieu functions for the two-dimensional solution in elliptical coordinates and spheroidal functions for the three-dimensional solution in oblate or prolate spheroidal coordinates.

One implementation of modal propagators for numerical modeling is that due to Waterman [in Mittra, 1973], whose approach uses the extended boundary condition (EBC) whereby the required field behavior is satisfied away from the boundary surface on which the sources are located. This procedure, widely known as the *T-matrix* approach, has evidently been more widely used in optics and acoustics than in electromagnetics. In what amounts to a reciprocal application of EBC, the sources can be removed from the boundary surface on which the field-boundary conditions are applied. These modal techniques seem to offer some computational advantages for certain kinds of problems and might be regarded as using entire-domain basis and testing functions but nevertheless lead to linear systems of equations whose numerical solution is required. Fourier transform solution techniques might also be included in this category since they do involve modal expansions, but that is a specialized area that we do not pursue further here.

Modal expansions are receiving increasing attention under the general name "fast multipole method," which is motivated by the goal of systematically exploiting the reduced complexity of the source-field interactions as their separation increases. The objective is to reduce the significant interactions of a Green's-function based matrix from being proportional to $(N_x)^2$ to of order $N_x \log (N_x)$, thus offering the possibility of decreasing the operation count of iterative solutions.

Geometrical-Optics Model

Geometrical optics and the geometrical theory of diffraction (GTD) are high-frequency asymptotic techniques wherein the fields are propagated using such optical concepts as shadowing, ray tubes, and refraction and diffraction. Although conceptually straightforward, optical techniques are limited analytically by the unavailability of diffraction coefficients for various geometries and material bodies and numerically by the need to trace rays over complex surfaces. There is a vast literature on geometrical optics and GTD, as may be ascertained by examining the yearly and cumulative indexes of such publications as the *Transactions of the IEEE Antennas and Propagation Society.*

45.4 Numerical Issues in Developing a Computer Model

Sampling Functions

At the core of numerical analysis is the idea of polynomial approximation, an observation made by Arden and Astill [1970] in facetiously using the subtitle "Numerical Analysis or 1001 Applications of Taylor's Series." The basic idea is to approximate quantities of interest in terms of sampling functions, often polynomials, that are then substituted for these quantities in various analytical operations. Thus, integral operators are replaced by finite sums, and differential operators are similarly replaced by generalized finite differences. For example, use of a first-order difference to approximate a derivative of the function $F(x)$ in terms of samples $F(x_+)$ and $F(x_-)$ leads to

$$\frac{dF(x)}{dx} \approx \frac{F(x_+) - F(x_-)}{h}; \quad x_- \le x \le x_+ \tag{45.2a}$$

and implies a linear variation for $F(x)$ between x_+ and x_- as does use of the trapezoidal rule

$$\int_{x_-}^{x_+} F(x)dx \approx \frac{h}{2}\,[F(x_+) + F(x_-)] \tag{45.2b}$$

to approximate the integral of $F(x)$, where $h = x_+ - x_-$. The central-difference approximation for the second derivative,

$$\frac{d^2F(x)}{dx^2} \approx \frac{[F(x_+) - 2F(x_0) + F(x_-)}{h^2} \tag{45.2c}$$

similarly implies a quadratic variation for $F(x)$ around $x_0 = x_+ - h/2 = x_- + h/2$, as does use of Simpson's rule

$$\int_{x_-}^{x_+} F(x)dx \approx \frac{h}{6}\,[F(x_+) + 4F(x_0) + F(x_-)] \tag{45.2d}$$

to approximate the integral. Other kinds of polynomials and function sampling can be used, as discussed in a large volume of literature, some examples of which are Abramowitz and Stegun [1964], Acton [1970], and Press et al. [1986]. It is interesting to see that numerical differentiation and integration can be accomplished using the same set of function samples and spacings, differing only in the signs and values of some of the associated weights. Note also that the added degrees of freedom that arise when the function samples can be unevenly spaced, as in Gaussian quadrature, produce a generally more accurate result (for well-behaved functions) for

TABLE 45.7 Sampling Operations Involved in MoM Modeling

Equation	DE Model $L(s')f(s') = g(s')$	IE Model $L(s,s')f(s') = g(s)$
Sampling of:		
Unknown via basis-functions $b_j(s')$ using $f(s') \approx \sum a_j b_j(s')$	Subdomain bases usually of low order are used. Known as FD procedure when pulse basis is used, and as FE approach when bases are linear.	Can use either subdomain or entire-domain bases. Use of latter is generally confined to bodies of rotation. Former is usually of low order, with piecewise linear or sinusoidal being the maximum variation used.
Equation via weight functions $w_i(s)$ $<w_i(s),L(s,s')\sum a_j b_j(s')> = <w_i(s),g(s)>$ to get $Z_{ij}a_j = g_i$	Pointwise matching is commonly employed, using a delta function. Pulse and linear matching are also used.	Pointwise matching is commonly employed, using a delta function. For wires, pulse, linear, and sinusoidal testing is also used. Linear and sinusoidal testing is also used for surfaces.
Operator	Operator sampling for DE models is entwined with sampling the unknown in terms of the difference operators used.	Sampling needed depends on the nature of the integral operator $L(s,s')$. An important consideration whenever the field integrals cannot be evaluated in closed form.
Solution of:		
$Z_{ij}a_j = g_i$ for the a_j	Interaction matrix is sparse. Time-domain approach may be explicit or implicit. In frequency domain, banded-matrix technique usually used.	Interaction matrix is full. Solution via factorization or iteration.

a given number of samples. This suggests the benefits that might be derived from using unequal sample sizes in MoM modeling should a systematic way of determining the best nonuniform sampling scheme be developed.

The Method of Moments

Numerical implementation of the moment method is a relatively straightforward, and an intuitively logical, extension of these basic elements of numerical analysis, as described in the book by Harrington [1968] and discussed and used extensively in CEM [see, for example, Mittra, 1973, 1975; Strait, 1980; Poggio and Miller, 1988]. Whether it is an integral equation, a differential equation, or another approach that is being used for the numerical model, three essential sampling operations are involved in reducing the analytical formulation via the moment method to a computer algorithm as outlined in Table 45.7. We note that operator sampling can ultimately determine the sampling density needed to achieve a desired accuracy in the source–field relationships involving integral operators, especially at and near the "self term," where the observation and source points become coincident or nearly so and the integral becomes nearly singular. Whatever the method used for these sampling operations, they lead to a linear system of equations or matrix approximation of the original integral or differential operators. Because the operations and choices involved in developing this matrix description are common to all moment-method models, we shall discuss them in somewhat more detail.

When using IE techniques, the coefficient matrix in the linear system of equations that results is most often referred to as an impedance matrix because in the case of the E-field form, its multiplication of the vector of unknown currents equals a vector of electric fields or voltages. The inverse matrix similarly is often called an admittance matrix because its multiplication of the electric-field or voltage vector yields the unknown-current vector. In this discussion we instead use the terms *direct matrix* and *solution matrix* because they are more generic descriptions whatever the forms of the originating integral or differential equations. As illustrated in the following, development of the direct matrix and solution matrix dominates both the computer time and storage requirements of numerical modeling.

In the particular case of an IE model, the coefficients of the direct or original matrix are the mutual impedances of the multiport representation which approximates the problem being modeled, and the coefficients of its solution matrix (or equivalent thereof) are the mutual admittances. Depending on whether a subdomain or entire-domain basis has been used (see Basic Function Selection), these impedances and admittances represent either spatial or modal interactions among the N ports of the numerical model. In either case, these

TABLE 45.8 Examples of Generic Basis/Weight-Function Combinations

Method	jth Term of Basis	ith Term of Weight
Galerkin	$a_j b_j(\mathbf{r}')$	$w_i(\mathbf{r}) = b_j(\mathbf{r})$
Least square	$a_j b_j(\mathbf{r}')$	$Q(\mathbf{r})\partial\varepsilon(\mathbf{r})/\partial a_i$
Point matching	$a_j\delta(\mathbf{r} - \mathbf{r}_j)$	$\delta(\mathbf{r} - \mathbf{r}_i)$
General collocation	$a_j b_j(\mathbf{r}')$	$\delta(\mathbf{r} - \mathbf{r}_i)$
Subsectional collocation	$U(\mathbf{r}_j)\sum a_{jk} b_k(\mathbf{r}')$	$\delta(\mathbf{r} - \mathbf{r}_i)$
Subsectional Galerkin	$U(\mathbf{r}_j)\sum a_{jk} b_k(\mathbf{r}')$	$U(\mathbf{r}_i)\sum b_i(\mathbf{r})$

$\mathbf{r}'$ and $\mathbf{r}$ denote source and observation points respectively; a_j, a_{jk} are unknown constants associated with the jth basis function (entire domain) or the kth basis function of the jth subsection (subdomain); $U(\mathbf{r}_k)$ is the unit sampling function which equals 1 on the kth subdomain and is 0 elsewhere; $b_j(\mathbf{r}')$ is the jth basis function; $w_i(\mathbf{r})$ is the ith testing function; $\delta(\mathbf{r} - \mathbf{r}_i)$ is the Dirac delta function; $Q(\mathbf{r})$ is a positive-definite function of position; and $\varepsilon(\mathbf{r})$ is the residual or equation error [from Poggio and Miller in Mitra (1973)].

coefficients possess a physical relatability to the problem being modeled and ultimately provide all the information available concerning any electromagnetic observables that are subsequently obtained.

Similar observations might also be made regarding the coefficients of the DE models but whose multiport representations describe local rather than global interactions. Because the DE model almost always leads to a larger, albeit less dense, direct matrix, its inverse (or equivalent) is rarely computed. It is worth noting that there are two widely used approaches for DE modeling, finite-difference (FD) and finite-element (FE) methods. They differ primarily in how the differential operators are approximated and the differential equations are satisfied, i.e., in the order of the basis and weight functions, although the FE method commonly starts from a variational viewpoint, while the FD approach begins from the defining differential equations. The FE method is generally better suited for modeling problems with complicated boundaries to which it provides a piecewise linear or higher order approximation as opposed to the cruder stairstep approximation of FD.

Factors Involved in Choosing Basis and Weight Functions

Basis and weight function selection plays a critical role in determining the accuracy and efficiency of the resulting computer model. One goal of the basis and weight function selection is to minimize computer time while maximizing accuracy for the problem set to which the model is to be applied. Another, possibly conflicting, goal might be that of maximizing the collection of problem sets to which the model is applicable. A third might be to replicate the problem's physical behavior with as few samples as possible. Some of the generic combinations of bases and weights that are used for MoM models are listed in Table 45.8 [Poggio and Miller from Mittra, 1973].

Basis Function Selection. We note that there are two classes of bases used in MoM modeling, subdomain and entire-domain functions. The former involves the use of bases that are applied in a repetitive fashion over subdomains or sections (segments for wires, patches for surfaces, cells for volumes) of the object being modeled. The simplest example of a subdomain basis is the single-term basis given by the pulse or stairstep function, which leads to a single, unknown constant for each subdomain. Multiterm bases involving two or more functions on each subdomain and an equivalent number of unknowns are more often used for subdomain expansions.

The entire-domain basis, on the other hand, uses multiterm expansions extending over the entire object, for example, a circular harmonic expansion in azimuth for a body of revolution. As for subdomain expansions, an unknown is associated with each term in the expansion. Examples of hybrid bases can also be found, where subdomain and entire-domain bases are used on different parts of an object.

Although subdomain bases are probably more flexible in terms of their applicability, they have a disadvantage generally not exhibited by the entire-domain form, which is the discontinuity that occurs at the domain boundaries. This discontinuity arises because an n_s-term subdomain function can provide at most $n_s - 1$ degrees of continuity to an adjacent basis of the unknown it represents, assuming one of the n_s constants is reserved for the unknown itself. For example, the three-term or sinusoidal subdomain basis $a_i + b_i \sin(ks) + c_i \cos(ks)$ used for wire modeling can represent a current continuous at most up to its first derivative. This provides continuous charge density but produces a discontinuous first derivative in charge equivalent to a tripole charge at each junction.

TABLE 45.9 Examples of Specific Basis/Weight-Function Combinations

Application	*j*th Term of Basis	*i*th Term of Weight
1D/wires	Constant—$a_j U_j(s)$	Delta function—$\delta(s - s_i)$
1D/wires	Piecewise linear—$a_{j1}(s - s_j - \delta_j/2) + a_{j2}(s - s_j + \delta_j/2)$	Piecewise linear—$(s - s_i - \delta_i/2) + (s - s_i + \delta_i/2)$
1D/wires	3-term sinusoidal—$a_{j1} + a_{j2}\sin[k(s - s_j)] + a_{j3}\cos[k(s - s_j)]$	Delta function—$\delta(s - s_i)$
1D/wires	Piecewise sinusoidal—$a_{j1}\sin[k(s - s_j - \delta_j/2)] + a_{j2}\sin[k(s - s_j + \delta_j/2)]$	Piecewise sinusoidal—$\sin[k(s - s_i - \delta_i/2)] + \sin[k(s - s_i + \delta_i/2)]$
2D/surfaces	Weighted delta function—$a_j\delta(s - s_j)\Delta_j$	Delta function—$\delta(s - s_i)$
2D/rotational surfaces	Piecewise linear axially, and $\exp(in\phi)$ azimuthally	Same (Galerkin's method)
2D/surfaces	Piecewise linear	Same (Galerkin's method)
2D/surfaces	Piecewise linear subdomain/Fourier series entire domain	Same (Galerkin's method)
3D/volumes	Piecewise linear	Same (Galerkin's method)

δ_k is the length of wire segment k; Δ_k is the area of surface patch k.

As additional terms are used to develop a subdomain basis, higher-order continuity can be achieved in the unknown that the basis represents, assuming still that one unknown is reserved for the amplitude of the multiterm basis function. In the general case of the n_s-term subdomain basis, up to $n_s - 1$ constants can be determined from continuity conditions, with the remainder reserved for the unknown. The kind of basis function employed ultimately determines the degree of fit that the numerical result can provide to the true behavior of the unknown for a given order of matrix. An important factor that should influence basis-function selection, then, is how closely a candidate function might resemble the physical behavior of the unknown it represents. Another consideration is whether a system of equations that is numerically easier to solve might result from a particular choice of basis and weight function, for example, by increasing its diagonal dominance so that an iterative technique will converge more rapidly and/or reduce the number of significant interactions. Various evolving approaches having names such as "impedance-matrix localization." "fast multipole method," "spatial decomposition," and "multilevel matrix-decomposition" are being developed with these goals.

Weight Function Selection. The simplest weight that might be used is a delta function which leads to a point-sampled system of equations, but point sampling of the field operators can be sensitive to any numerical anomalies that might arise as a result of basis function discontinuities. Distributed, multiterm weight functions can also be used on either a subdomain or an entire-domain basis to provide a further smoothing of the final equations to be solved. One example of this is the special case where the same functions are used for both the bases and weights, a procedure known as Galerkin's method. The kind of testing function used ultimately determines the degree to which the equations can be matched for a given basis function and number of unknowns. Some specific examples of basis and weight function combinations used in electromagnetics are summarized in Table 45.9.

Computing the Direct Matrix

We observe that obtaining the coefficients of the direct matrix in IE modeling is generally a two-step process. The first step is that of evaluating the defining integral operator in which the unknown is replaced by the basis functions selected. The second step involves integration of this result multiplied by the weight function selected. When using delta-function weights, this second step is numerically trivial, but when using nondelta weights, such as the case in a Galerkin approach, this second step can be analytically and numerically challenging.

Among the factors affecting the choice of the basis and weight functions, therefore, one of the most important is that of reducing the computational effort needed to obtain the coefficients of the direct matrix. This is one of the reasons, aside from their physical appeal, why sinusoidal bases are often used for wire problems. In this case, where piecewise linear, filamentary current sources are most often used in connection with the thin-wire approximation, field expressions are available in easily evaluated, analytical expressions. This is the case as well where Galerkin's method is used.

TABLE 45.10 Summary of Operation Count for Solution of General Direct Matrix Having N_x Unknowns

Method	To Obtain Solution Matrix	To Obtain Solution	Comments
Cramer's rule	Expand in co-factors leading to →	$\sim N_x!$	Not an advisable procedure but useful to illustrate just how bad the problem could be
Inversion	N_x^3	N_x^2	Provides RHS-independent solution matrix
Factorization	$N_x^3/3$	N_x^2	RHS-independent solution matrix
Iteration			
General	—	$N_x^2 - N_x^3$	Each RHS requires separate solution
With FFT	—	$N_x - N_x^2$	Same, plus applicability to arbitrary problems uncertain
Symmetry			
Reflection	$(1 \text{ to } 2^p) \times (N_x/2^p)^3$	$N_x^2/2^p$	For $p = 1$ to 3 reflection planes
Translation (Toeplitz)	$n_x^3[t(\log_2 t)^2]$	N_x^2	For n_x unknowns per t sections of translation
Rotation (circulant)	$\log_2(N_x)N_x^3/n^2$	N_x	For n rotation sectors and a complete solution
	$m \log_2(N_x)(N_x/n)^3$	m	For $m = 1$ to n modes
Banded			
General	$N_x W^2$	$N_x W$	For a bandwidth of W coefficients
Toeplitz	$N_x \log_2 N_x$	W^2	For a bandwidth of W coefficients

Aside from such special cases, however, numerical evaluation of the direct-matrix coefficients will involve the equivalent of point sampling of whatever order is needed to achieve the desired accuracy as illustrated below. Using a wirelike one-dimensional problem to illustrate this point, we observe that at its most elementary level evaluation of the ijth matrix coefficient then involves evaluating integrals of the form

$$\begin{aligned} Z_{i,j} &= \int_{C(\mathbf{r})} w_i(s) \int_{C(\mathbf{r})} [b_j(s')K(s, s')ds']ds \\ &\approx \sum_{m=1}^{M(i,j)} \sum_{n=1}^{N(i,j)} p_m q_n w_i(s_n) b_j(s'_m) K(s_n, s'_m) \\ &= \sum_{m=1}^{M(i,j)} \sum_{n=1}^{N(i,j)} p_m q_n z(i, j, m, n); \quad i, j = 1, \ldots, N \end{aligned} \tag{45.03}$$

where $K(s,s')$ is the IE kernel function, and s_n and s'_m are the nth and mth locations of the observation and source integration samples. Thus, the final, direct-matrix coefficients can be seen to be constructed from sums of the more elementary coefficients $z(i,j,m,n)$ weighted by the quadrature coefficients p_m and q_n used in the numerical integration, which will be the case whenever analytical expressions are not available for the $Z_{i,j}$. These elementary coefficients, given by $w_i(s_n)b_j(s'_m)K(s_n,s'_m)$, can in turn be seen to be simply products of samples of the IE kernel or operator and sampled basis and testing functions. It should be apparent from this expanded expression for the direct-matrix coefficients that interchanging the basis and weight functions leaves the final problem description unchanged, although the added observation that two different IEs can yield identical matrices when using equivalent numerical treatments is less obvious.

Computing the Solution Matrix

Once the direct matrix has been computed, the solution can be obtained numerically using various approaches, ranging from inversion of the direct matrix to developing a solution via iteration as summarized in Table 45.10. A precautionary comment is in order with respect to the accuracy with which the solution matrix might be obtained. As computer speed and storage have increased, the number of unknowns used in modeling has also increased, from a few tens in earlier years to hundreds of thousands now when using IE models and millions of unknowns when using DE models. The increasing number of operations involved in solving these larger

matrices increases sensitivity of the results to roundoff errors. This is especially the case when the direct matrix is not well conditioned. It is therefore advisable to perform some sensitivity analyses to determine the direct-matrix condition number and to ascertain the possible need for performing some of the computations in double precision.

Obtaining the Solution

When a solution matrix has been developed using inversion or factorization, subsequently obtaining the solution (most often a current) is computationally straightforward, involving multiplication of the right-hand side (RHS) source vector by the solution matrix. When an iterative approach is used, a solution matrix is not computed, but the solution is instead developed from RHS-dependent manipulation of the direct matrix. Motivation for the latter comes from the possibility of reducing the N_x^3 dependency of the direct procedure. As problem size increases, the computation cost will be increasingly dominated by the solution time.

45.5 Some Practical Considerations

Although the overall solution effort has various cost components, perhaps the one most considered is the computer time and storage required to obtain the numerical results desired. With the increasing computer memories becoming available, where even microcomputers and workstations can directly address gigabytes, the memory costs of modeling are becoming generally less important than the time cost, with which we are primarily concerned here. For each model class considered, the computer-time dependence on the number of unknowns is presented in a generic formula followed by the highest-order $(L/\Delta L)$ term in that formula to demonstrate how computer time grows with increasing problem size.

Integral Equation Modeling

Frequency Domain

If we consider an IE model specifically, we can show that, in general, the computer time associated with its application is dependent on the number of unknowns N_x in the frequency domain as

$$T_{\mathrm{IE},\omega} \approx A_{\mathrm{fill}} N_x^2 + A_{\mathrm{solve}} N_x^3 + A_{\mathrm{source}} N_x^2 N_{\mathrm{rhs}} + A_{\mathrm{field}} N_x N_{\mathrm{rhs}} N_{\mathrm{fields}}$$
$$\sim (L/\Delta L)^{3(D-1)} \tag{45.4a}$$

where the A's are computer- and algorithm-dependent coefficients that account for computation of A_{fill}, the direct (impedance) matrix; A_{solve}, the solution (admittance) matrix (assuming inversion or factorization); A_{source}, the source response (currents and charges) for one of N_{rhs} different excitations or right-hand sides (the g term of Table 45.7); and A_{field}, one of N_{field} fields, where $A_{\mathrm{field}} \leq A_{\mathrm{fill}}$, depending on whether a near-field (=) or far-field (<) value is obtained. D is the problem dimensionality (for a wire IE model, $D = 2$ except when used for wire-mesh approximations of surfaces in which case $D = 3$); L is a characteristic length of the object being modeled; and ΔL is the spatial resolution required, being proportional to the wavelength.

Time Domain

A similar relationship holds for a time-domain IE model which uses N_t time steps,

$$T_{\mathrm{IE},t} \approx A_{\mathrm{fill}} N_x^2 + A_{\mathrm{solve}} N_x^3 + A_{\mathrm{source}} N_x^2 N_t N_{\mathrm{rhs}} + A_{\mathrm{field}} N_x N_t N_{\mathrm{rhs}} N_{\mathrm{fields}}$$
$$\sim (L/\Delta L)^{2(D-1)+1+r}, \text{ explicit approach, } 0 \leq r \leq 1$$
$$\sim (L/\Delta L)^{3(D-1)}, \quad \text{implicit approach} \tag{45.4b}$$

with the A's accounting for computation of the time-domain terms equivalent to their frequency-domain counterparts (with different numerical values), and r accounting for the RHS dependency. Although a direct matrix may require solution initially before time-stepping the model, that is normally avoided by using $\delta t \leq$

$\Delta x/c$, which yields an explicit solution, in which case $A_{\text{solve}} = 0$. As can be appreciated from these expressions, the number of unknowns that is required for these computations to be acceptably accurate has a strong influence on the computer time eventually needed.

Differential Equation Modeling

Frequency Domain

DE modeling is less commonly used in the frequency domain primarily because the order of the matrix that results depends on $(L/\Delta L)^D$ rather than the usual $(L/\Delta L)^{D-1}$ dependency of an IE model. On the other hand, the matrix coefficients require less computation whether the DE model is based on a finite-difference or finite-element treatment. Furthermore, the matrix is very sparse because a differential operator is a local rather than a global one, as is the integral operator. Matrix fill time is therefore generally not of concern, and the overall computer time is given approximately by

$$T_{\text{DE},\omega} \approx A_{\text{solve}} N_x W^2 + A_{\text{source}} N_x W N_{\text{rhs}} + A_{\text{field}} N_x^{(D-1)/D} N_{\text{rhs}} N_{\text{fields}}$$
$$\sim (L/\Delta L)^{3D-2} \tag{45.4c}$$

which exhibits a dominance by the matrix-solution term. Note that the banded nature of the DE direct matrix has been taken into account where the bandwidth W varies as N_x^0, $N_x^{1/2}$, and $N_x^{2/3}$ respectively ($N_x^{[(D-1)/D]}$), in one, two, and three dimensions.

Time Domain

Time-domain DE modeling can use either implicit or explicit solution methods for developing the time variation of the solution. An explicit technique is one whereby the update at each time step is given in terms of solved-for past values of the unknowns and the present excitation, with no interaction permitted between unknowns within the same time step, and is the approach used in a technique known as finite-difference time domain (FDTD). An implicit technique, on the other hand, does allow for interaction of unknowns within the same time step but can therefore require the solution of a matrix equation. In spite of this disadvantage, implicit techniques are important because they are not subject to Courant instability when $c\delta t > \Delta\chi$ as is an explicit approach. Books entirely devoted to FDTD and its applications are now becoming available, one of which is by Kunz and Luebbers (1993).

The solution time for the explicit case is approximated by

$$T_{\text{DE},t} \approx A_{\text{source}} N_x N_t N_{\text{rhs}} + A_{\text{field}} N_x^{(D-1)/D} N_{\text{rhs}} N_{\text{fields}}$$
$$\sim (L/\Delta L)^{D+1+r}, \text{ explicit approach; } 0 \le r \le 1 \tag{45.4d}$$
$$\approx A_{\text{solve}} N_x W^2 + A_{\text{source}} N_x W N_t N_{\text{rhs}} + A_{\text{field}} N_x^{(D-1)/D} N_{\text{rhs}} N_{\text{fields}}$$
$$\sim (L/\Delta L)^{3D-2}, \text{ for } D = 2, 3$$

and

$$\sim (L/\Delta L)^{2+r}, \text{ for } D = 1, \text{ implicit approach; } 0 \le r \le 1$$

assuming a banded matrix is used to solve the implicit direct matrix.

Discussion

It should be recognized that the above computer-time estimates assume solutions are obtained via matrix factorization, an N^3 process, and that iterative techniques when applicable should be expected to reduce the

TABLE 45.11 Nominal Sampling Requirements for Various Field Quantities

Quantity	Value
N_x, total number of spatial samples (per scalar unknown)	$\sim(L/\Delta L)^d = (2\pi L/\lambda)^d$
N_t, number of time steps for time-domain model	$\sim (L/\Delta L) = (2\pi L/\lambda)$
N_f, number of frequency steps to characterize spectral response from frequency-domain model	$\sim (L/2\Delta L) = N_t/2$
N_{rhs}, number of excitation sources for monostatic radar cross section in one plane[a]	$\sim (4L/\Delta L) = 8\pi L/\lambda$
N_{fields}, number of far fields needed for bistatic pattern in one observation plane[a]	$\sim N_{\text{rhs}} = (4L/\Delta L)$

λ is the wavelength at the highest frequency of interest; ΔL is the spatial resolution being sought; L is object maximum object dimension or dimension in observation plane; d is the number of *spatial* dimensions being sampled and is not necessarily the *problem* dimensionality D. The distinction is important because when an appropriate Green's function is available, the source integrals are usually one dimension less than the problem dimension, i.e., $d = D - 1$. An exception is an inhomogeneous, penetrable body where $d = D$ when using an integral equation.

[a]Assuming ~6 samples per lobe of the scattering pattern are needed.

maximum order of the $(L/\Delta L)$ dependency but at the cost, however, of requiring the computation to be repeated for each RHS. We also emphasize that these comparisons consider only problems involving homogeneous objects, thereby providing a more favorable situation for IE models because their sampling dimensionality $d = D - 1$ for a problem dimensionality of D but which increases to $d = D$ when an inhomogeneous object is modeled. Because of these and other factors that can lead to many different combinations of formulation and numerical treatment, the foregoing results should be viewed as only generic guidelines, with the computational characteristics of each specific model requiring individual analysis to obtain numerical values for the various A_x coefficients and their $(L/\Delta L)$ dependency. It is relevant to observe that the lowest-order size dependency for three-dimensional problems is exhibited by the DE explicit time-domain model which is on the order of $(L/\Delta L)^4$.

An additional factor that should be considered when choosing among computer models is the information needed for a particular application relative to the information provided by the model. A time-domain model, for example, can intrinsically provide a frequency response over a band of frequencies from a single calculation, whereas a frequency-domain model requires repeated evaluation at each of the frequencies required to define the wideband response. Iterative solution of the direct matrix may be preferable for problems involving only one, or a few, excitations such as is the case for antenna modeling, to avoid computing all N_x^2 admittances of the solution matrix when only a single column of that matrix is needed. A DE-based model necessarily provides the "near" fields throughout the region being modeled, while an IE-based model requires additional computations essentially the same as those done in filling the impedance matrix once the sources have been obtained to evaluate the near fields. For applications that require modest computer time and storage, these considerations may be relatively less important than those that strain available computer resources. Clearly, the overall objective from an applications viewpoint is to obtain the needed information at the required level of accuracy for the minimum overall cost.

Sampling Requirements

We may estimate the number of samples needed to adequately model the spatial, temporal, and angular variation of the various quantities of interest in terms of an object characteristic length L and sampling dimension d. This may be done from knowledge of the typical spatial and temporal densities determined from computer experiments and/or from invocation of Nyquist-like sampling rates for field variations in angle as a function of aperture size. The resulting estimates are summarized in Table 45.11 and apply to both IE and DE models.

These may be regarded as *wavelength-driven* sampling rates, in contrast with the *geometry-driven* sampling rates that can arise because of problem variations that are small in scale compared with λ. Geometry-driven sampling would affect primarily N_x, resulting in larger values than those indicated above.

We note that the computer time is eventually dominated by computation of the solution matrix and can grow as $(L/\Delta L)^3$, $(L/\Delta L)^6$, and $(L/\Delta L)^9$ (or f^3, f^6, and f^9), respectively, for wire, surface, and volume objects modeled using integral equations and matrix factorization or inversion. Thus, in spite of the fact that mainframe computer power has grown by a factor of about 10^6 from the UNIVAC-1 to the CRAY2, a growth that is anticipated to continue during the near future as shown in Fig. 45.2, the growth in problem size is much less

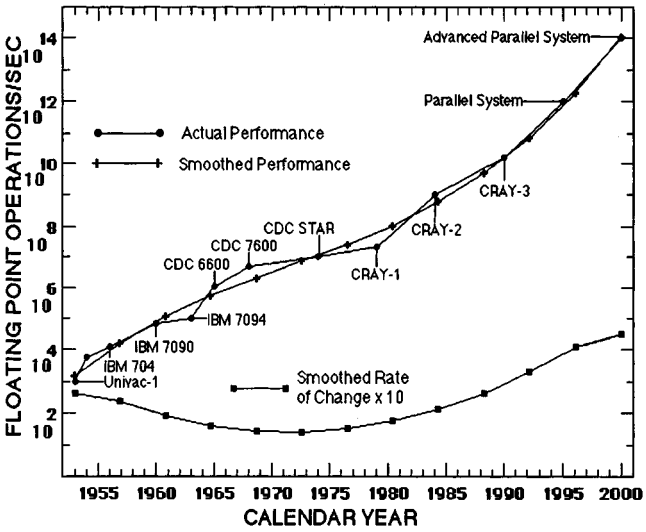


FIGURE 45.2 Raw and smoothed FLOP (floating-point operation) rates of mainframe computers and smoothed rate-of-change in speed, at year of introduction from the UNIVAC-1 to the projected performance of an advanced parallel system at year 2000. Future growth is increasingly dependent on computer architecture, requiring increasing parallelism as improvements due to component performance reach physical speed limits.

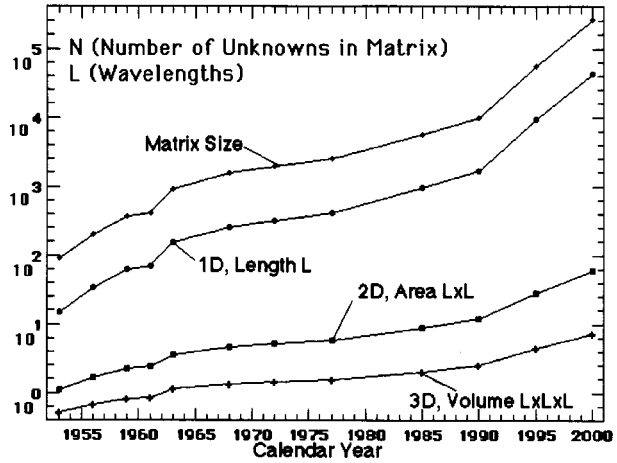


FIGURE 45.3 Time development of IE-based modeling capability for one-dimensional (e.g., a wire), two-dimensional (e.g., a plate), and three-dimensional (e.g., a penetrable, inhomogeneous cube) sampling of a problem of characteristic dimension L in wavelengths and matrix order N solvable in 1 h of computer time using mainframe computers introduced in the years indicated. Linear-systems solution using LU decomposition (an N^3 dependency) is assumed with number of unknowns proportional to L, L^2 and L^3, respectively, without any problem symmetry being exploited. These results should be viewed as upper bounds on solution time and might be substantially reduced by advances in linear-system solution procedures.

impressive, as illustrated by Fig. 45.3. The curves on this graph demonstrate emphatically the need for finding faster ways of performing the model computations, a point further emphasized by the results shown in Fig. 45.4 where the computer time required to solve a reference problem using various standard models is plotted as a function of frequency.

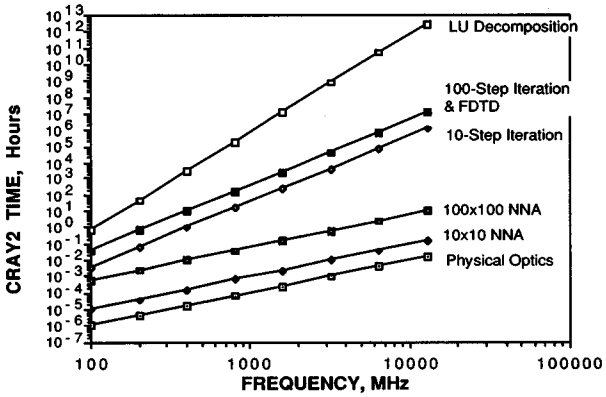


FIGURE 45.4 An illustration of the frequency dependence of the CRAY2 computer time required for some standard computer models applied to the reference problem of a perfectly conducting, space-shuttle-sized object having a surface area of 540 m^2 [Miller, 1988]. At a sampling density of $100/\lambda^2$, a total of 6,000 surface samples is assumed for an IE model at 100 MHz for which LU decomposition of the direct (impedance) matrix requires about 1 h of CRAY2 time. The top, LU, curve has a slope of f^6 as discussed in the text. The next two curves have slopes of f^4, the upper corresponding to use of a TD DE model (FDTD) as well as an iterative solution of the direct IE matrix, assuming acceptable convergence occurs in 100 iteration steps. The third curve is for a 10-step iterative solution of the IE matrix. The bottom three curves have f^2 slopes. The upper two of these are for 100- and 10-step iterative solutions used in connection with a near-neighbor approximation (NNA), wherein only the 100 and 10 largest interaction coefficients are retained in the matrix, respectively. The bottom curve is for the physical-optics approximation, in which the induced current is computed from the incident magnetic field. The effects of these different frequency slopes on the computer time can be seen to be extreme, emphasizing the need for developing more efficient solution procedures.

45.6 Ways of Decreasing Computer Time

The obvious drawback of direct moment-method models as N_x increases with increasing problem size and/or complexity suggests the need for less computationally intensive alternatives. There are various alternatives for decreasing the computer cost associated with solving electromagnetic problems using the method of moments. The basic intent in any case is either to reduce the direct cost of performing a given modeling computation or to reduce the number of modeling computations needed to obtain a desired result. An example for achieving the latter is to employ lower-order models that can accurately enough represent the behavior of observables as a function of space, angle, frequency, or time so that the sampling density of the first-principles model can be reduced. A specific example is used of the rational functions to approximate frequency-domain transfer functions. These might include analytical, computational, and experimental approaches or combinations thereof, about which further discussion and references may be found in Miller [1988].

45.7 Validation, Error Checking, and Error Analysis

Modeling Uncertainties

The process of proceeding from an original physical problem to computed results is one that is subject to numerous uncertainties caused by a variety of factors. Perhaps foremost among these factors is the degree of arbitrariness associated with many of the choices that are made by the code developer and/or modeler in the course of eventually obtaining numerical results. Whereas the numerical evaluation of classical boundary-value problems such as scattering from a sphere is numerically robust in the sense that different workers using different

TABLE 45.12 Error Types that Occur in Computational Electromagnetics

Category	Definition
Physical modeling error, ε_P	Arises because the numerical model used is normally an idealized mathematical representation of the actual physical reality
Numerical modeling error, ε_N	Arises because the numerical results obtained are only approximate solutions to that idealized representation and consists of two components: (1) Solution error—The difference that can exist between the computed results and an exact solution even were the linear system of equations to be solved exactly, due to using a finite number of unknowns (2) Equation error—The equation mismatch that can occur in the numerical solution because of roundoff due to finite-precision computations or when using an iterative technique because of limited solution convergence

computers and different software can obtain results in agreement to essentially as many significant figures as they wish, the same observation cannot be made for moment-method modeling.

Modeling uncertainties can be assigned to two basic error categories, a physical **modeling error** ε_P and a numerical modeling error ε_N as outlined in Table 45.12. The former is due to the fact that for most problems of practical interest varying degrees of approximation are needed in developing a simplified or idealized problem representation that will be compatible with the computer code to be used for the modeling computations. The latter is due to the fact that the numerical results obtained are almost invariably only approximate solutions to that idealized representation. We note that although an analytical expression may in principle represent a formally exact solution, the process of obtaining numerical results in that case is still one that inevitably involves finite-precision evaluation of the formal solution.

By its very nature, the physical modeling error requires some kind of measurement for its determination, except for those few problems whose analytical solution in principle involves no physical idealization or subsequent numerical approximation. One example of such problems is that of determining the scattering or radiating properties of the perfectly conducting or dielectric sphere.

The numerical modeling error is itself composed of two components in general, the determination of which would normally involve one or more kinds of computation. The first and generally more important of these components is the solution error that arises because the computer model used, even if solved exactly, would not provide an exact solution for the idealized problem representation. The solution error arises essentially because the computer model is solved using a finite number of unknowns. The other, generally less important contributor to the numerical modeling error is the equation error that arises because the numerical results obtained from the computer model used may not numerically satisfy the modeling equations. The equation error may be caused both by roundoff due to the computer word size as well as the solution algorithm used, as in the case of iteration, for example. The impact of equation error can be expected to increase with increasing condition number of the direct matrix.

Validation and Error Checking

One of the most time consuming and long lasting of the tasks associated with any model development is that of validation. Long after work on the model has been completed, questions will continue to arise about whether a given result is valid or whether the model can be applied to a given problem. There are essentially two kinds of validation procedures that can be considered to answer such questions: (1) internal validation, a check that can be made concerning solution validity within the model itself; and (2) external validation, a check that utilizes information from other sources which could be analytical, experimental, or numerical.

Existing computer models often do not perform internal checks on the results they produce but instead leave that as an exercise for the user. It would be of extremely great potential value if a variety of such checks could be built into the code and exercised as desired by the modeler. The topic of error checking and validation is an active one in CEM and receives a great deal of ongoing attention, for which the technical literature provides a good point of departure for the reader interested in more detail.

45.8 Concluding Remarks

In the preceding discussion we have presented a selective survey of computational electromagnetics. Attention has been directed to radiation and scattering problems solved using the method of moments, a general procedure applicable to differential- and integral-equation formulations developed by either the frequency domain or the time domain. Beginning from the viewpoint of electromagnetics as a transfer-function process, we concluded that the basic problem is one of developing source-field relationships or field propagators. Of the various ways by which these propagators might be expressed, we briefly discussed the Maxwell curl equations and Green's-function source integrals as providing the analytical basis for moment-method computer models. We then considered at more length some of the numerical issues involved in developing a computer model, including the idea of sampling functions used both to represent the unknowns to be solved for and to approximate the equations that they must satisfy. Some of the factors involved in choosing these sampling functions and their influence on the computational requirements were examined. Next, we discussed some ways of decreasing the needed computer time based on either analytical or numerical approaches. Some closing comments were directed to the important problems of validation, error checking, and error analysis. Throughout our discussion, emphasis has been given to implementation issues involved in developing and using computer models as opposed to exploring analytical details.

Defining Terms

Computer model: Based on a numerical solution of some appropriate analytical formulation that describes a physical phenomenon of interest. The model is realized in an *algorithm* or computer *code* that reduces the formulation to a series of operations suitable for computer solution.

Field propagator: The analytical description of how electromagnetic fields are related to the sources that cause them. Common field propagators in electromagnetics are the defining Maxwell equations that lead to differential equation models, Green's functions that produce integral equation models, optical propagators that lead to optics models, and multipole expansions that lead to modal models.

Integral equation: An analytical relationship in which the quantity whose solution is sought (the unknown) appears under an integral sign. When this is the only place that the unknown appears, the integral equation is commonly called a first-kind equation, while if the unknown also appears outside the integral, it is a second-kind integral equation.

Method of moments: A general technique for reducing integral, differential (including partial), and integrodifferential equations to a linear system of equations or matrix. The moment method involves discretizing, sampling, and approximating the defining equations using *basis* or *expansion* functions to replace the unknown and *testing* or *weighting* functions to satisfy the defining equations. The matrix that results may be full (all coefficients nonzero) or sparse (only a few per row are nonzero), depending on whether the model is an integral or differential equation.

Modeling errors: In essentially all computer modeling, there are two basic kinds of errors. One, the *physical modeling error*, arises from replacing some real-world physical problems with some idealized mathematical representation. The other, the *numerical modeling error*, comes from obtaining only an approximate solution to that idealized representation. Usually, the numerical modeling error can be reduced below the physical modeling error if enough unknowns, i.e., a large enough matrix, are used to model the problem of interest.

Sampling: The process of replacing a continuous physical quantity by some sequence of sampled values. These values are associated with the analytical function used to approximate the behavior of the physical quantity whose solution is sought and are the unknowns of the moment-method matrix. Sampling is also involved in determining how well the defining equations are to be satisfied. A common approach for equation sampling is *point* sampling, where the equations are explicitly satisfied at a series of discrete points in some prescribed region of space. Unknown sampling can involve localized basis functions, an approach called *subdomain* sampling, while if the basis functions reside over the entire region occupied by the unknown, the approach is called *entire-domain* sampling.

Solution domain: Electromagnetic fields can be represented as a function of time, or a *time-domain* description, or as a function of frequency using a (usually) Fourier transform, which produces a *frequency-domain* description.

Related Topics

35.1 Maxwell Equations • 44.3 Numerical Methods • 44.4 Modern Design Environment

References

M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Applied Mathematics Series, vol. 55, Washington, D.C.: National Bureau of Standards, 1964.

F.S. Acton, *Numerical Methods that Work*, New York: Harper and Row, 1970.

B.W. Arden and K.N. Astill, *Numerical Algorithms: Origins and Applications*, Reading, Mass.: Addison-Wesley, 1970.

R.F. Harrington, *Time-Harmonic Electromagnetic Fields*, New York: McGraw-Hill, 1961.

R.F. Harrington, *Field Computation by Moment Methods*, New York: Macmillan, 1968.

K.S. Kunz and R.J. Luebbers, *The Finite Difference Time Domain Method for Electromagnetics,* Boca Raton, Fla.: CRC Press, 1993.

E. K. Miller, "A selective survey of computational electromagnetics," *IEEE Trans. Antennas Propagat.*, vol. AP-36, pp. 1281–1305, 1988.

E.K. Miller, "Solving bigger problems—by decreasing the operation count and increasing the computation bandwidth," invited article in special issue of *IEEE Proc. Electromagnets*, vol. 79, no. 10, pp. 1493–1504, 1991.

E.K. Miller, L. Medgyesi-Mitschang, and E.H. Newman, *Computational Electromagnetics: Frequency-Domain Method of Moments,* New York: IEEE Press, 1991.

R. Mittra, ed., *Computer Techniques for Electromagnetics*, New York: Pergamon Press, 1973.

R. Mittra, ed., *Numerical and Asymptotic Techniques in Electromagnetics,* New York: Springer-Verlag, 1975.

J. Moore and R. Pizer, *Moment Methods in Electromagnetics: Techniques and Applications*, New York: Wiley, 1984.

A.J. Poggio and E.K. Miller, "Low frequency analytical and numerical methods for antennas," in *Antenna Handbook*, Y. T. Lo and S. W. Lee, eds., New York: Van Nostrand Reinhold, 1988.

B.D. Popovic, M.B. Dragovic, and A.R. Djordjevic, *Analysis and Synthesis of Wire Antennas*, Letchworth, Hertfordshire, England: Research Studies Press, 1982.

W.H. Press, B.R. Flannery, S.A. Teukolsky, and W.T. Vettering, *Numerical Recipes*, London: Cambridge University Press, 1986.

B.J. Strait, ed., *Applications of the Method of Moments to Electromagnetic Fields,* St. Cloud, Fla.: SCEEE Press, 1980.

J.A. Stratton, *Electromagnetic Theory,* New York: McGraw-Hill, 1941.

W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, New York: John Wiley, 1981.

P. L. E. Uslenghi, ed., *Electromagnetic Scattering*, New York: Academic Press, 1978.

J.H. Wang, *Generalized Moment Methods in Electromagnetics*, New York: Wiley Interscience, 1991.

Further Information

The *International Journal of Numerical Modeling*, published by Wiley four times per year, includes numerous articles on modeling electronic networks, devices, and fields. Information concerning subscriptions should be addressed to Subscription Department, John Wiley & Sons Ltd., Baffins Lane, Chichester, Sussex PO19 1UD, England.

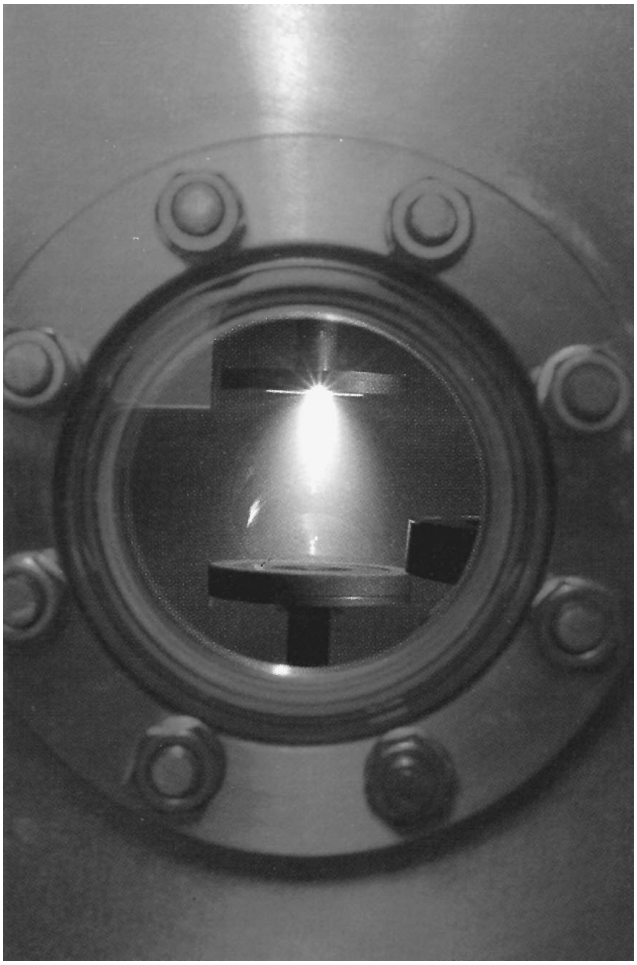
The *Journal of the Acoustical Society of America* is published by the American Institute of Physics on a monthly basis. Most issues contain articles about the numerical solution of acoustics problems which have much in common with problems in electromagnetics. Information about the society and journal can be obtained from Acoustical Society of America, 500 Sunnyside Blvd., Woodbury, NY 11797.

The *Journal of the Applied Computational Electromagnetics Society* is published two or three times a year, accompanied by a newsletter published about four times per year. The focus of the society and journal is the application of computer models, their validation, information about available software, etc. Membership and subscription information can be obtained from Dr. R.W. Adler, Secretary, Applied Computational Electromagnetics Society, Naval Postgraduate School, Code ECAB, Monterey, CA 93943.

The *Journal of Electromagnetic Waves and Applications* is published by VNU Science Press. It contains numerous articles dealing with the numerical solution of electromagnetic problems. Information about the journal can be obtained from its editor-in-chief, Professor J.A. Kong, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

The *Proceedings of the IEEE, Transactions on Microwave Theory and Techniques of the IEEE, Transactions on Antennas and Propagation of the IEEE,* and *Transactions on Electromagnetic Compatibility of the IEEE* all are periodicals published by the Institute of Electrical and Electronics Engineers, about which information can be obtained from IEEE Service Center, 445 Hoes Lane, PO Box 1331, Piscataway, NJ 08855-1331.

Feisel, L.D. "Section V – Electrical Effect s and Devices"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



Ever since the discovery of superconductivity in 1911, researchers have sought to raise the temperature at which superconductivity occurs. With the advent of high temperature superconducting (HTS) materials in 1986, superconductors have begun to emerge from the laboratory and appear in practical applications. A pioneer in this explosively advancing technology is Superconducting Technologies, Inc., Santa Barbara, California. This company uses thallium, the highest temperature material for making high temperature superconductors. Thallium remains conductive at temperatures above 77°K and can be cooled to working temperature by a liquid nitrogen system instead of the more difficult and more expensive helium method. Shown above is a high temperature superconductor being produced by a laser ablation system. (Photo courtesy of National Aeronautics and Space Administration.)

V

Electrical Effects and Devices

46 Electroacoustic Devices *P. H. Rogers*
Transduction Mechanisms • Sensitivity and Source Level • Reciprocity • Canonical Equations and Electroacoustic Coupling • Radiation Impedance • Directivity

47 Surface Acoustic Wave Filters *D. C. Malocha*
SAW Material Properties • Basic Filter Specifications • SAW Transducer Modeling • Distortion and Second-Order Effects • Bidirectional Filter Response • Multiphase Unidirectional Transducers • Single-Phase Unidirectional Transducers • Dispersive Filters • Coded SAW Filters • Resonators

48 Ultrasound *G. W. Farnell*
Propagation in Solids • Piezoelectric Excitation • One-Dimensional Propagation • Transducers

49 Ferroelectric and Piezoelectric Materials *K. F. Etzold*
Mechanical Characteristics • Ferroelectric Materials • Ferroelectric and High Epsilon Thin Films

50 Electrostriction *V. Sundar and R. E. Newnham*
Defining Equations • PMN-PT—A Prototype Electrostrictive Material

51 Piezoresistivity *A. Amin*
Equation of State • Effect of Crystal Point Group on $\prod_{ijkl}$ • Geometric Corrections and Elastoresistance Tensor • Multivalley Semiconductors • Longitudinel Piezoresistivity Π_l and Maximum Sensitivity Direction • Semiconducting (PTCR) Perovskites • Thick Film Resistors • Design Considerations

52 The Hall Effect *A. C. Ehrlich*
Theoretical Background • Relation to the Electronic Structure—(i) $\omega_c\tau << 1$ • Relation to the Electronic Structure—(ii) $\omega_c\tau >> 1$

53 Superconductivity *K. A. Delin, T. P. Orlando*
General Electromagnetic Properties • Superconducting Electronics • Types of Superconductors

54 Pyroelectric Materials and Devices *R. W. Whatmore*
Polar Dielectrics • The Pyroelectric Effect • Pyroelectric Materials and Their Selection

55 Dielectrics and Insulators *R. Bartnikas*
Dielectric Losses • Dielectric Breakdown • Insulation Aging • Dielectric Materials

56 Sensors *R. L. Smith*
Physical Sensors • Chemical Sensors • Biosensors • Microsensors

57 Magnetooptics *D. Young, Y. Pu*
Classification of Magnetooptic Effects • Applications of Magnetooptic Effects

58 Smart Materials *P. S. Neelakanta*
Smart/Intelligent Structures • Objective-Based Classification of Smart/Intelligent Materials • Material Properties Conducive for Smart Material Applications • State-of-the-Art Smart Materials • Smart Sensors • Examples of Smart/Intelligent Systems • High-Tech Application Potentials

Lyle D. Feisel
State University of New York, Binghamton

Every high school student who takes a course in physics or even general science is—or at least should be—familiar with the first-order, linear electrical effects such as resistance, inductance, capacitance, etc. The more esoteric effects, however, are often neglected, even in otherwise comprehensive undergraduate electrical engineering curricula. These effects, though, are not only fascinating in their manifestations but are also potentially—and in some cases, currently—exceedingly useful in application. This section will describe many of these higher-order electrical and magnetic effects and some of the devices that are based upon them. Readers are invited not only to study the current applications but to let their imaginations extrapolate to other uses as yet unproposed.

A number of phenomena are related to the interaction of mechanical energy with electrical energy. The field of *acoustics* deals with those situations where that mechanical energy takes the form of sound waves. Acoustic applications have been particularly fruitful, especially during the last two decades. Surface acoustic wave (SAW) filters are among the more useful applications. These elegant devices are a marriage of sophisticated signal theory and piezoelectricity, consummated on the bed of thin-film technology. Unlike some elegant devices, they have been commercially successful as well.

A special class of acoustoelectric devices deals with acoustic frequencies beyond the range of human hearing. The field of *ultrasonics* and its related devices and systems are finding broad application in the area of nondestructive testing. Of course, one of the testing applications where the nondestructive property is especially important is in investigating the human body. Medical imaging has provided considerable impetus for advances in ultrasonics in the last few years.

Most people know that if a sample of certain types of material (e.g., iron) is subjected to a magnetic field, it will exhibit a retained magnetic behavior. Few, however, realize that some materials exhibit a similar retention effect when an electric field is applied. *Ferroelectricity* is the phenomenon in which certain crystalline or polycrystalline materials retain electric polarization after an external electric field has been applied and removed. Since the direction of the polarization depends upon the direction of the applied field and since the polarization is quite persistent, memory devices can be based on this effect. Other applications have also been suggested.

For decades, the frequencies of radio transmitters have been stabilized with "crystals." In recent years, the effect called *piezoelectricity*—in which a mechanical strain induces an electric field and vice versa—has found many other applications. Like ferroelectrics, piezoelectric materials can be either crystalline or polycrystalline and can be fabricated in a variety of shapes.

If an electric charge is moved with a velocity at some angle to a magnetic field, the charge will experience a force at right angles to both the charge velocity and the magnetic field. If the charge is inside a solid material, a charge inhomogeneity is created and an electric field results. This is the well-known *Hall effect*, which finds practical application in such devices as magnetic field meters and in more basic uses as measuring and understanding the properties of semiconductors.

Probably the second electrical phenomenon observed by humans (lightning was probably the first), *ferromagnetism* deals with the interaction of molecular magnetic dipoles with external and internal magnetic fields. Ferromagnetic materials retain some polarization after an external field is removed—a desirable property if the application is a permanent magnet or a recording device—but one which causes losses in a transformer. These materials have improved as the demands of magnetic recording have increased.

If certain materials get cold enough, their resistivity goes to zero—not to some very small value but, as nearly as we can tell, zero. *Superconductivity* has been known as an interesting phenomenon for many years, but applications have been limited because the phenomenon only occurred at temperatures within a few degrees of absolute zero. Recent advances, however, have produced materials which exhibit superconductive behavior at substantially higher temperatures, and there is renewed interest in developing applications. This is certainly an area to watch in the next few years.

Some very elegant devices have been developed to exploit the interactions between electric fields and photons or optical waves. *Electrooptics* is the key to many of the recent and, indeed, future advances in optical communication. The phenomena are generally higher-order, nonintuitive, and exceedingly interesting, and the devices are generally quite elegant but simple.

We have come a long way since the first Atlantic Cable was fabricated using gutta-percha, tarred hemp, and pitch for insulation. *Dielectrics and insulators* are now better understood and controlled for a wide variety of applications. At one time the only property of real interest was dielectric strength, the insulator's ability to stand up to high voltage. Today, many other properties, as well as ease and economy of fabrication, are at least as important.

The word *application* appears many times in the preceding paragraphs. What are these applications? Many of the practical uses of the phenomena described in this section are in measuring the variables that define the phenomena. Thus, *sensors* constitute a primary application. For instance, the Hall effect can be used to measure magnetic fields, and mechanical strain can be measured using the phenomenon of piezoelectricity.

Just as photons will interact with electric fields, so, too, will they affect and be affected by magnetic fields. *Magnetooptics* is the study and application of these interactions. As with electrooptics, the increased activity in optical communications has provided renewed interest in this field.

The use of *smart materials* may solve a variety of engineering problems. In general, these are materials which change their properties to adapt to their environments, thereby doing their jobs better. This promises to be an area of increased activity in the future.

Again, the reader is admonished not only to understand the applications presented in the following chapters but to understand, at least at the phenomenological level, the phenomena upon which the applications are based. Such understanding is likely to lead to even broader applications in the future.

Nomenclature

Symbol	Quantity	Unit
α	attenuation constant	Np/m
χ_o	magnetic susceptibility of free space	
D	diffraction constant	
E	transducer efficiency	
ε	dielectric constant	
ϵ	complex permittivity	F/m
G_T	thermal conductance	W/K
η	viscosity	Poise
η	emissivity	
k	quantum mechanical wave factor	m^{-1}
k^2	SAW coupling factor	
K	thermal conductivity of pyroelectric	$W/m^2/K$
m	molar mass	kg
R	Hall coefficient	m^3/C
S	strain	
σ	conductivity	S/m
T	stress	N/m^2
τ_T	thermal time constant of element	s
θ_f	Faraday rotation coefficient	
V	phase velocity	m/s
V	Verdet constant	
W	electromagnetic energy density	W/m^2
Z_R	radiation impedance	Ω

Rogers, P.H. “Electroacoustic Devices”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

46

Electroacoustic Devices[1]

46.1 Introduction
46.2 Transduction Mechanisms
Piezoelectricity • Magnetostriction • Electrodynamic • Electrostatic • Magnetic • Hydraulic • Fiber Optic • Parametric Transducers • Carbon Microphones
46.3 Sensitivity and Source Level
46.4 Reciprocity
46.5 Canonical Equations and Electroacoustic Coupling
46.6 Radiation Impedance
46.7 Directivity

Peter H. Rogers
Georgia Institute of Technology

46.1 Introduction

Electroacoustics is concerned with the transduction of acoustical to electrical energy and vice versa. Devices which convert acoustical signals into electrical signals are referred to as "**microphones**" or "hydrophones" depending on whether the acoustic medium is air or water. Devices which convert electrical signals into acoustical waves are referred to as "loudspeakers" (or earphones) in air and "projectors" in water.

46.2 Transduction Mechanisms

Piezoelectricity

Certain crystals produce charge on their surfaces when strained or conversely become strained when placed in an electric field. Important piezoelectric crystals include quartz, ADP, lithium sulphate, rochelle salt, and tourmaline. Lithium sulphate and tourmaline are "volume expanders," that is, their volume changes when subjected to an electric field in the proper direction. Such crystals can detect hydrostatic pressure directly. Crystals which are not volume expanders must have one or more surfaces shielded from the pressure field in order to convert the pressure to a uniaxial strain which can be detected. Tourmaline is relatively insensitive and used primarily in blast gauges, while quartz is used principally in high Q ultrasonic transducers.

Certain ceramics such as lead zirconate titanate (PZT), barium titanate, and lead metaniobate become piezoelectric when polarized. They exhibit relatively high electromechanical coupling, are capable of producing very large forces, and are used extensively as sources and receivers for underwater sound. PZT and barium titanate have only a small volume sensitivity; hence they must have one or more surfaces shielded in order to detect sound efficiently. Piezoelectric ceramics have extraordinarily high dielectric coefficients and hence high capacitance, and they are thus capable of driving long cables without preamplifiers.

[1]This chapter is adapted from R. M. Besançon, *Encyclopedia of Physics*, 3rd ed., New York: Chapman & Hall, 1985, pp. 337–341. With permission.

Recently, it has been discovered that certain polymers, notably polyvinylidene fluoride, are piezoelectric when stretched. Such piezoelectric polymers are finding use in directional microphones and ultrasonic hydrophones.

Magnetostriction

Some ferromagnetic materials become strained when subjected to a magnetic field. The effect is quadratic in the field, so a bias field or dc current is required for linear operation. Important magnetostrictive metals and alloys include nickel and permendur. At one time, magnetostrictive transducers were used extensively in active sonars but have now been largely replaced by ceramic transducers. Magnetostrictive transducers are rugged and reliable but inefficient and configurationally awkward. Recently, it has been discovered that certain rare earth iron alloys such as terbium-dysprosium-iron possess extremely large magnetostrictions (as much as 100 times that of nickel). They have relatively low eddy current losses but require large bias fields, are fragile, and have yet to find significant applications. Metallic glasses have also recently been considered for magnetostrictive transducers.

Electrodynamic

Electrodynamic transducers exploit the forces produced on a current-carrying conductor in a magnetic field and, conversely, the currents produced by a conductor moving in a magnetic field. Direct radiation moving coil transducers dominate the loudspeaker field. Prototypes of high-power underwater projectors have been constructed using superconducting magnets. Electrodynamic microphones, particularly the directional ribbon microphones, are also common.

Electrostatic

Electrostatic sources utilize the force of attraction between charged capacitor plates. The force is independent of the sign of the voltage, so a bias voltage is necessary for linear operation. Because the forces are relatively weak, a large area is needed to obtain significant acoustic output. The effect is reciprocal, with the change in the separation of the plates (i.e., the capacitance) produced by an incident acoustic pressure generating a voltage. The impedance of a condenser microphone, however, is high, so a preamplifier located close to the sensor is required. Condenser microphones are very flat and extremely sensitive. The change in capacitance induced by an acoustic field can also be detected by making the capacitor a part of a bridge circuit or, alternatively, a part of an oscillator circuit. The acoustic signal will then appear as either an amplitude or frequency modulation of some ac carrier. The charge storage properties of electrets have been exploited to produce electrostatic microphones which do not require a bias voltage.

Magnetic

Magnetic transducers utilize the force of attraction between magnetic poles and, reciprocally, the voltages produced when the reluctance of a magnetic circuit is changed. Magnetic speakers are used extensively in telephone receivers.

Hydraulic

Nonreversible, low-frequency, high-power underwater projectors can be constructed utilizing hydraulic forces acting to move large pistons. Electroacoustic transduction is achieved by modulating the hydraulic pressure with a spool valve actuated by an electrostrictive (PZT) stack.

Fiber Optic

An acoustic field acting on an optical fiber will change the optical path length by changing the length and index of refraction of the fiber. Extremely sensitive hydrophones and microphones can be made by using a fiber exposed to an acoustic field as one leg of an optical interferometer. Path length changes of the order of 10^{-6} optical wavelengths can be detected. The principal advantages of such sensors are their configurational flexibility,

their sensitivity, and their suitability for use with fiber optic cables. Fiber optic sensors which utilize amplitude modulation of the light (microbend transducers) are also being developed.

Parametric Transducers

The nonlinear interaction of sound waves can be used to produce highly directional sound sources with no side lobes and small physical apertures. In spite of their inherent inefficiency, substantial source levels can be achieved and such "parametric sonars" have found a number of underwater applications. Parametric receivers have also been investigated but practical applications have yet to be found.

Carbon Microphones

Carbon microphones utilize a change in electrical resistance with pressure and are used extensively in telephones.

46.3 Sensitivity and Source Level

A microphone or hydrophone is characterized by its free-field voltage sensitivity, M, which is defined as the ratio of the output voltage, E, to the free-field amplitude of an incident plane acoustic wave. That is, for an incident wave which *in the absence of the transducer* is given by

$$P = P_0 \cos(\mathbf{k} \cdot \mathbf{R} - \omega t) \tag{46.1}$$

M is defined by

$$M = E/P_0 \tag{46.2}$$

In general, M will be a function of frequency and the orientation of the transducer with respect to the wave vector $\mathbf{k}$ (i.e., the direction of incidence of the wave). Thus, for a given frequency, M is proportional to the directivity of the transducer. It is usually desirable for a microphone or hydrophone to have a flat (i.e., frequency independent) free-field voltage sensitivity over the broadest possible range of frequencies to assure fidelity of the output electrical signal.

A loudspeaker or projector is characterized in a similar manner by its transmitting current response, S, which is defined as the ratio of the acoustic source level to the driving current, I. In the farfield of a transducer the acoustic pressure is a spherical wave which can be expressed as

$$P(R) = P_s(\theta, \phi)(R_0/R) \cos(kR - \omega t) \tag{46.3}$$

where θ and ϕ are elevation and azimuth angles and R_0 an arbitrary reference distance (usually 1 meter). $P_s(\theta, \phi)$ is defined as the source level. Thus S is given by

$$S = P_s(\theta, \phi)/I \tag{46.4}$$

which is a function of θ and ϕ and the frequency ω. For high-fidelity sound reproduction S should be as flat as possible over the broadest possible bandwidth. For some purposes, however, such as ultrasonic cleaning or long-range underwater acoustic propagation, fidelity is unnecessary and high Q resonant transducers are employed to produce high-intensity sound over a narrow bandwidth.

46.4 Reciprocity

Most conventional transducers are *reversible*, that is, they can be used as either sources or receivers of sound (a carbon microphone and a fiber optic hydrophone are examples of transducers which are *not* reversible). A transducer is said to be *linear* if the input and output variables are linearly proportional (hot-wire microphones

and unbiased magnetostrictive transducers are examples of *nonlinear* transducers). A transducer is said to be *passive* if the only source of energy is the input electrical or acoustical signal (a microphone with a built-in preamplifier and a parametric projector are examples of *nonpassive* transducers). Most transducers which are linear, passive, and reversible exhibit a remarkable property called *reciprocity*. For a *reciprocal* transducer of any kind (moving coil, piezoelectric, magnetostrictive, electrostatic, magnetic, etc.) the ratio of the free-field voltage sensitivity to the transmitting current response is equal to the reciprocity factor J which is independent of the geometry and construction of the transducer. That is:

$$\frac{M(\omega, \theta, \phi)}{S(\omega, \theta, \phi)} = J(\omega) = \frac{4\pi R_0}{\rho_0 \omega} \tag{46.5}$$

where ρ_0 is the density of the medium and R_0 is the reference distance used in defining the source level. Equation (46.5) has a number of useful consequences: (1) the receiving and transmitting beam patterns of a reciprocal transducer are identical, (2) a transducer cannot be simultaneously flat as a receiver and transmitter since S has an additional factor of ω, and (3) Eq. (46.5) provides the basis for the three-transducer reciprocity calibration technique whereby an absolute calibration of a hydrophone or microphone can be obtained from purely electrical measurements.

46.5 Canonical Equations and Electroacoustic Coupling

Simple acoustic transducers can be characterized by the following canonical equations:

$$E = Z_e I + T_{em} V \tag{46.6}$$

$$F = T_{me} I + Z_m V \tag{46.7}$$

where V is the velocity of the radiating or receiving surface, F is the total force acting on the surface (including acoustic reaction forces), Z_e is the blocked ($V = 0$) electrical impedance, Z_m is the open-circuit mechanical impedance, and T_{em} and T_{me} are the electromechanical coupling coefficients. For reciprocal transducers $T_{em} = \pm T_{me}$. For example, for a moving coil transducer where the "motor" is coil in a radial magnetic field, B,

$$T_{em} = -T_{me} = BL \tag{46.8}$$

where L is the length of the wire in the coil and the electrical impedance Z_e is largely inductive. For a piston transducer with a piezoelectric "motor"

$$T_{me} = T_{em} = -id_{33}/(\epsilon^{\mathrm{T}} s \omega) \tag{46.9}$$

where d_{33} is the piezoelectric strain coefficient, s is the compliance, ϵ^{T} is the permittivity at constant stress, and the electrical impedance Z_e is largely capacitive.

If a piston transducer is placed in an acoustic field such that the average pressure over the surface of the piston is P_B, then $F = P_B A$, where A is the area of the piston, and for a receiver $I = 0$, so

$$E = (T_{em} A / Z_m) P_B \tag{46.10}$$

If the transducer is small compared with an acoustic wavelength $P_E \approx P_0$ (in general $P_B = DP_0$ where D is the diffraction constant) and the free-field voltage sensitivity is given by

$$mM = T_{em} A / Z_m \tag{46.11}$$

From Eq. (46.5) the transmitting current response is

$$S = \frac{\rho_0 \omega T_{em} A}{4\pi R_0 Z_m} \tag{46.12}$$

From these simple considerations a number of principles of practical transducer design can be deduced. The mechanical impedance Z_m is in general given by

$$Z_m = \frac{K_m}{i\omega} + i\omega M + R_m \tag{46.13}$$

where K_m is an effective spring constant, M the mass, and R_m the mechanical resistance. For a piezoelectric transducer [Eq. (46.9)] T_{em} is inversely proportional to frequency; hence from Eqs. (46.10) and (46.11) we see that a piezoelectric transducer will have a flat receiving sensitivity below resonance (i.e., where its behavior is controlled by stiffness). On the other hand, a moving coil microphone must have a resistive mechanical impedance to have a flat response. From Eq. (46.12) we derive the fundamental tenet of loudspeaker design, that a moving coil loudspeaker will have a flat transmitting current response above resonance (i.e., where it is mass controlled). Accordingly, moving coil loudspeakers are designed to have the lowest possible resonant frequency (by means of a high compliance since the output is inversely proportional to the mass) and piezoelectric hydrophones are designed to have the highest possible resonant frequency.

An interesting and important consequence of electromechanical coupling is the effect of the motion of the transducer on the electrical impedance. In the absence of external forces (including radiation reactance) from Eqs. (46.6) and (46.7)

$$E = \left(Z_e - \frac{T_{em} T_{me}}{Z_m} \right) I \tag{46.14}$$

That is, the electrical impedance has a "motional" component given by $T_{em}T_{me}/Z_m$. The motional component can be quite significant near resonance where Z_m is small. This effect is the basis of crystal-controlled oscillators.

46.6 Radiation Impedance

An oscillating surface produces a reaction force F_R on its surface given by

$$F_R = -Z_R V \tag{46.15}$$

where Z_R is the radiation impedance. We can thus rewrite Eq. (46.7) as

$$F_{ext} = T_{em} I + (Z_R + Z_m) V \tag{46.16}$$

where F_{ext} now includes only external forces. For an acoustically small baffled circular piston of radius a,

$$Z_R = \pi a^4 \rho_0 \omega^2 / 2c - i(8/3)\omega \rho_0 a^3 \tag{46.17}$$

The radiation impedance thus has a mass-like reactance with an equivalent "radiation mass" of $(8/3)\rho_0 a^3$ and a small resistive component proportional to ω^2 responsible for the radiated power. A transducer will thus have a lower resonant frequency when operated underwater than when operated in air or vacuum. The total radiated power of the piston transducer is given by

$$\pi = \mathrm{Re}Z_r|V|^2 = (\pi a^4 \rho_0 \omega^2/2c)V^2 \tag{46.18}$$

Most transducers are displacement limited, so for a direct-radiating transducer V in Eq. (46.18) is limited. To obtain the most output power the piston should have the largest possible surface area consistent with keeping the transducer omnidirectional (the transducer will become directional when $a \geq \lambda$). This is easy to do in air but difficult in water since it is hard to make pistons which are both lightweight and stiff enough to hold their shape in water. Alternatively, the driver can be placed at the apex of a horn. For a conical horn, the fluid velocity at the end of the horn (where the radius is a_e) will be reduced to $V(a/a_e)$ but the radiating piston will now have an effective radius of a_e so the radiated power will increase by a factor of $(a_e/a)^2$. For high-power operation at a single frequency, the driver can be placed at the end of a quarter wave resonator.

46.7 Directivity

It is often desirable for transducers to be directional. Directional sound sources are needed in diagnostic and therapeutic medical ultrasonics, for acoustic depth sounders; and to reduce the power requirements and reverberation in active sonars, etc. Directional microphones are useful to reduce unwanted noise (e.g., to pick up the voice of a speaker and not the audience); directional hydrophones or hydrophone arrays increase signal-to-noise and aid in target localization. One way to achieve directionality is to make the radiating surface large. A baffled circular piston has a directivity given by

$$D_e = 2J_1(ka \sin\theta)/ka \sin\theta \tag{46.19}$$

D_e equals unity for $\theta = 0$ and 1/2 when $ka \sin\theta = 2.2$. For small values of ka, D_e is near unity for all angles.

Some transducers respond to the gradient of the acoustic pressure rather than pressure, for example, the ribbon microphone which works by detecting the motion of a thin conducting strip orthogonal to a magnetic field. Such transducers have a directivity which is dipole in nature, i.e.,

$$D_e = \cos\theta \tag{46.20}$$

Note that since the force in this case is proportional not to P_0 but to kP_0, a ribbon microphone (which like a moving coil microphone is electrodynamic) will have flat receiving sensitivity when its impedance is mass controlled. By combining a dipole receiver with a monopole receiver one obtains a unidirectional cardioid receiver with

$$D_e = (1 + \cos\theta) \tag{46.21}$$

Defining Terms

Electroacoustics: Concerned with the transduction of acoustical to electrical energy and vice versa.
Microphones: Devices which convert acoustical signals into electrical signals.

Related Topic

49.1 Introduction

References

J.A. Bucaro, H.D. Dardy, and E.F. Carome, "Fiber optic hydrophone," *J. Acoust. Soc. Am.,* vol. 62, p. 1302, 1977.
R.J. Bobber, "New types of transducer," in *Underwater Acoustics and Signal Processing,* L. Bjorno (Ed.), Dordrecht, Holland: D. Riedel, 1981.
R.J. Bobber, Underwater Electroacoustic Measurements, Washington, D.C.: Government Printing Office, 1969.

J.V. Bouyoucos, "Hydroacoustic transduction," *J. Acoust. Soc. Am.,* vol. 57, p. 1341, 1975.
F.V. Hunt, *Electroacoustics*, Cambridge: Harvard University Press, and New York: Wiley, 1954.
S.W. Meeks and R.W. Timme, "Rare earth iron magnetostrictive underwater sound transducer," *J. Acoust. Soc. Am.,* vol. 62, p. 1158, 1977.
M.B. Moffett and R.M. Mellon, "Model for parametric acoustic sources," *J. Acoust. Soc. Am.,* vol. 61, p. 325, 1977.
D. Ricketts, "Electroacoustic sensitivity of piezoelectric polymer cylinders," *J. Acoust. Soc. Am.,* vol. 68, p. 1025, 1980.
G.M. Sessler and J.E. West, "Applications," in *Electrets,* G.M. Sessler (Ed.), New York: Springer-Verlag, 1980.

Further Information

IEEE Transactions on Acoustics, Speech, and Signal Processing.

Malocha, D.C. “Surface Acoustic Wave Filters”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

47

Surface Acoustic Wave Filters

47.1 Introduction
47.2 SAW Material Properties
47.3 Basic Filter Specifications
47.4 SAW Transducer Modeling
The SAW Superposition Impulse Response Transducer Model • Apodized SAW Transducers
47.5 Distortion and Second-Order Effects
47.6 Bidirectional Filter Response
47.7 Multiphase Unidirectional Transducers
47.8 Single-Phase Unidirectional Transducers
47.9 Dispersive Filters
47.10 Coded SAW Filters
47.11 Resonators

Donald C. Malocha
University of Central Florida

47.1 Introduction

A **surface acoustic wave (SAW),** also called a Rayleigh wave, is composed of a coupled compressional and shear wave in which the SAW energy is confined near the surface. There is also an associated electrostatic wave for a SAW on a piezoelectric substrate which allows electroacoustic coupling via a transducer. SAW technology's two key advantages are its ability to electroacoustically access and tap the wave at the crystal surface and that the wave velocity is approximately 100,000 times slower than an electromagnetic wave. Assuming an electromagnetic wave velocity of 3×10^8 m/s and an acoustic wave velocity of 3×10^3 m/s, Table 47.1 compares relative dimensions versus frequency and delay. The SAW wavelength is on the same order of magnitude as line dimensions which can be photolithographically produced and the lengths for both small and long delays are achievable on reasonable size substrates. The corresponding E&M transmission lines or waveguides would be impractical at these frequencies.

Because of SAWs' relatively high operating frequency, linear delay, and tap weight (or sampling) control, they are able to provide a broad range of signal processing capabilities. Some of these include linear and dispersive filtering, coding, frequency selection, convolution, delay line, time impulse response shaping, and others. There are a very broad range of commercial and military system applications which include components for radars, front-end and IF filters, CATV and VCR components, cellular radio and pagers, synthesizers and analyzers, navigation, computer clocks, tags, and many, many others [Campbell, 1989; Matthews, 1977].

There are four principal SAW properties: transduction, reflection, regeneration and nonlinearities. Nonlinear elastic properties are principally used for convolvers and will not be discussed. The other three properties are present, to some degree, in all SAW devices, and these properties must be understood and controlled to meet device specifications.

TABLE 47.1 Comparison of SAW and E&M Dimensions versus Frequency and Delay, Where Assumed Velocities are v_{SAW} = 3000 m/s and $v_{EM} = 3 \times 10^8$ m/s

Parameter	SAW	E&M
F_0 = 10 MHz	λ_{SAW} = 300 μm	λ_{EM} = 30 m
F_0 = 2 GHz	λ_{SAW} = 1.5 μm	λ_{EM} = 0.15 m
Delay = 1 ns	L_{SAW} = 3 μm	L_{EM} = 0.3 m
Delay = 10 μs	L_{SAW} = 30 mm	L_{EM} = 3000 m

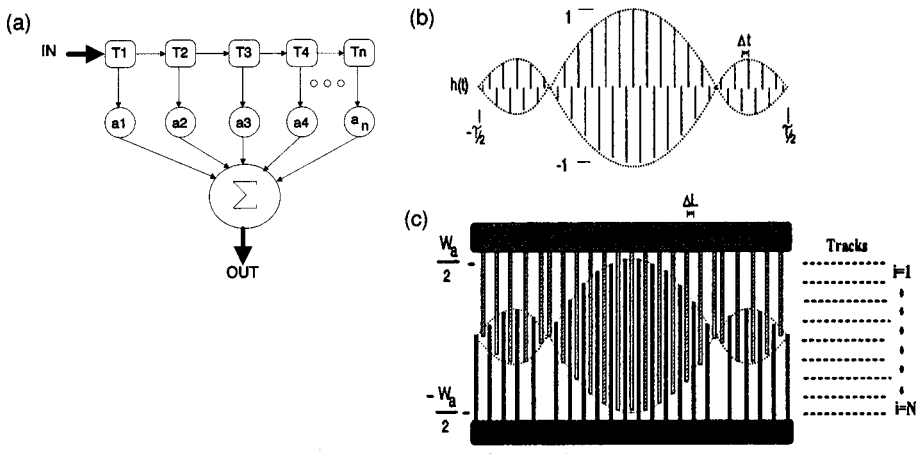


FIGURE 47.1 (a) Schematic of a finite-impulse response (FIR) filter. (b) An example of a sampled time function; the envelope is shown in the dotted lines. (c) A SAW transducer implementation of the time function $h(t)$.

A finite-impulse response (FIR) or transversal filter is composed of a series of cascaded time delay elements which are sampled or "tapped" along the delay line path. The sampled and delayed signal is summed at a junction which yields the output signal. The output time signal is finite in length and has no feedback. A schematic of an FIR filter is shown in Fig. 47.1.

A SAW transducer is able to implement an FIR filter. The electrodes or fingers provide the ability to sample or "tap" the SAW and the distance between electrodes provides the relative delay. For a uniformly sampled SAW transducer, the delay between samples, Δt, is given by $\Delta t = \Delta L / v_a$, where ΔL is the electrode period and v_a is the acoustic velocity. The typical means for providing attenuation or weighting is to vary the overlap between adjacent electrodes which provides a spatially weighted sampling of a uniform wave. Figure 47.1 shows a typical FIR time response and its equivalent SAW transducer implementation. A SAW filter is composed of a minimum of two transducers and possibly other SAW components. A schematic of a simple SAW bidirectional filter is shown in Fig. 47.2. A **bidirectional transducer** radiates energy equally from each side of the transducer (or port). Energy not being received is absorbed to eliminate spurious reflections.

47.2 SAW Material Properties

There are a large number of materials which are currently being used for SAW devices. The most popular single-crystal piezoelectric materials are quartz, lithium niobate ($LiNbO_3$), and lithium tantalate ($LiTa_2O_5$). The materials are anisotropic, which will yield different material properties versus the cut of the material and the direction of propagation. There are many parameters which must be considered when choosing a given material for a given device application. Table 47.2 shows some important material parameters for consideration for four of the most popular SAW materials [Datta, 1986; Morgan, 1985].

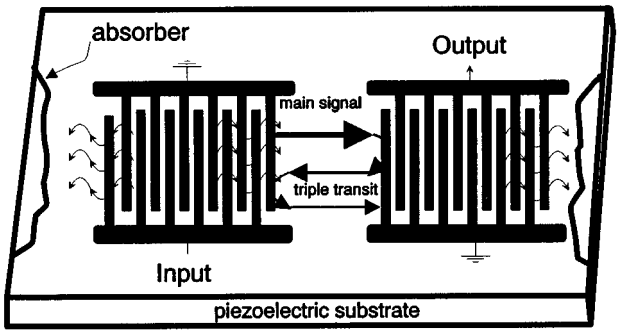


FIGURE 47.2 Schematic diagram of a typical SAW bidirectional filter consisting of two interdigital transducers. The transducers need not be identical. The input transducer launches waves in either direction and the output transducer converts the acoustic energy back to an electrical signal. The device exhibits a minimum 6-dB insertion loss. Acoustic absorber damps unwanted SAW energy to eliminate spurious reflections which could cause distortions.

TABLE 47.2 Common SAW Material Properties

Parameter/Material	ST-Quartz	YZ $LiNbO_3$	128° YX $LiNbO_3$	YZ $LiTa_2O_3$
k^2 (%)	0.16	4.8	5.6	0.72
C_s (pf/cm-pair)	0.05	4.6	5.4	4.5
v_0 (m/s)	3,159	3,488	3,992	3,230
Temp. coeff. of delay (ppm/°C)	0	94	76	35

The coupling coefficient, k^2, determines the electroacoustic coupling efficiency. This determines the fractional bandwidth versus minimum insertion loss for a given material and filter. The static capacitance is a function of the transducer electrode structure and the dielectric properties of the substrate. The values given in the table correspond to the capacitance per pair of electrodes having quarter wavelength width and one-half wavelength period. The free surface velocity, v_0, is a function of the material, cut angle, and propagation direction. The temperature coefficient of delay (TCD) is an indication of the frequency shift expected for a transducer due to a change of temperature and is also a function of cut angle and propagation direction.

The substrate is chosen based on the device design specifications and includes consideration of operating temperature, fractional bandwidth, and insertion loss. Second-order effects such as diffraction and beam steering are considered important on high-performance devices [Morgan, 1985]. Cost and manufacturing tolerances may also influence the choice of the substrate material.

47.3 Basic Filter Specifications

Figure 47.3 shows a typical time domain and frequency domain device performance specification. The basic frequency domain specification describes frequency bands and their desired level with respect to a given reference. Time domain specifications normally define the desired impulse response shape and any spurious time responses. The overall desired specification may be defined by combinations of both time and frequency domain specifications. Since time, $h(t)$, and frequency, $H(\omega)$, domain responses form unique Fourier transform pairs, given by

$$h(t) = 1/2\pi \int_{-\infty}^{\infty} H(\omega) e^{j\omega t} d\omega \tag{47.1}$$

$$H(\omega) = \int_{-\infty}^{\infty} h(t) e^{-j\omega t} dt \tag{47.2}$$

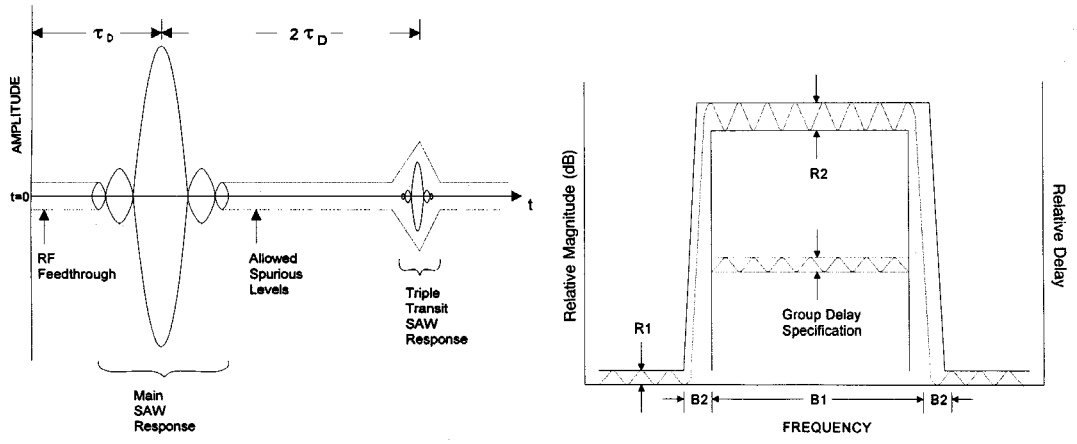


FIGURE 47.3 Typical time and frequency domain specification for a SAW filter.The filter bandwidth is B_1, the transition bandwidth is B_2, the inband ripple is R_2 and the out-of-band sidelobe level is R_1.

it is important that combinations of time and frequency domain specifications be self-consistent.

The electrodes of a SAW transducer act as sampling points for both transduction and reception. Given the desired modulated time response, it is necessary to sample the time waveform. For symmetrical frequency responses, sampling at twice the center frequency, $f_s = 2f_0$, is sufficient, while nonsymmetric frequency responses require sampling at twice the highest frequency of interest. A very popular approach is to sample at $f_s = 4f_0$. The SAW frequency response obtained is the convolution of the desired frequency response with a series of impulses, separated by f_s, in the frequency domain. The net effect of sampling is to produce a continuous set of harmonics in the frequency domain in addition to the desired response at f_0. This periodic, time-sampled function can be written as

$$g(t_n) = \sum_{-N/2}^{N/2} a_n \cdot \delta(t - t_n) \tag{47.3}$$

where a_n represents the sample values, $t_n = n\Delta t$, $n = n$th sample, and Δt = time sample separation. The corresponding frequency response is given by

$$G(f) = \sum_{-N/2}^{N/2} g(t_n) e^{-j2\pi f t_n} = \sum_{-N/2}^{N/2} g(t_n) e^{-j2\pi n f / f_s} \tag{47.4}$$

where $f_s = 1/\Delta t$. The effect of sampling in the time domain can be seen by letting $f = f + mf_s$, where m is an integer, which yields $G(f + mf_s) = G(f)$ which verifies the periodic harmonic frequency response.

Before leaving filter design, it is worth noting that a SAW filter is composed of two transducers which may have different center frequencies, bandwidth, and other filter specifications. This provides a great deal of flexibility in designing a filter by allowing the product of two frequency responses to achieve the total filter specification.

47.4 SAW Transducer Modeling

The four most popular and widely used models include the transmission line model, the coupling of modes model, the impulse response model, and the superposition model. The superposition model is an extension of the impulse response model and is the principal model used for the majority of SAW bidirectional and

multiphase filter synthesis which do not have inband, interelectrode reflections. As is the case for most technologies, many models may be used in conjunction with each other for predicting device performance based on ease of synthesis, confidence in predicted parameters, and correlation with experimental device data.

The SAW Superposition Impulse Response Transducer Model

The impulse response model was first presented by Hartmann et al. [1973] to describe SAW filter design and synthesis. For a linear causal system, the Fourier transform of the device's frequency response is the device impulse time response. Hartmann showed that the time response of a SAW transducer is given by

$$h(t) = 4k\sqrt{C_s}\, f_i^{3/2}(t)\sin[\theta(t)] \quad \text{where } \theta(t) = 2\pi\int_0^t f_i(\tau)d\tau \tag{47.5}$$

and where the following definitions are k^2 = SAW coupling coefficient, C_s = electrode pair capacitance per unit length (pf/cm-pair), and $f_i(t)$ = instantaneous frequency at a time, t. This is the general form for a uniform beam transducer with arbitrary electrode spacing. For a uniform beam transducer with periodic electrode spacing, $f_i(t) = f_0$ and $\sin\theta(t) = \sin\omega t$. This expression relates a time response to the physical device parameters of the material coupling coefficient and the electrode capacitance.

Given the form of the time response, energy arguments are used to determine the device equivalent circuit parameters. Assume a delta function voltage input, $v_{in}(t) = \delta(t)$, then $V_{in}(\omega) = 1$. Given $h(t)$, $H(\omega)$ is known and the energy launched as a function of frequency is given by $E(\omega) = 2\cdot|H(\omega)|^2$. Then

$$E(\omega) = V_{in}^2(\omega)\cdot G_a(\omega) = 1\cdot G_a(\omega) \tag{47.6}$$

or

$$G_a(\omega) = 2\cdot|H(\omega)|^2 \tag{47.7}$$

There is a direct relationship between the transducer frequency transfer function and the transducer conductance. Consider an **interdigital transducer (IDT)** with uniform overlap electrodes having N_p interaction pairs. Each gap between alternating polarity electrodes is considered a localized SAW source. The SAW impulse response at the fundamental frequency will be continuous and of duration τ, where $\tau = N\cdot\Delta t$, and $h(t)$ is given by

$$h(t) = \kappa\cdot\cos(\omega_0 t)\cdot rect(t/\tau) \tag{47.8}$$

where $\kappa = 4k\sqrt{C_s}\, f_0^{3/2}$ and f_0 is the carrier frequency. The corresponding frequency response is given by

$$H(\omega) = \frac{\kappa\tau}{2}\left\{\frac{\sin(x_1)}{x_1} + \frac{\sin(x_2)}{x_2}\right\} \tag{47.9}$$

where $x_1 = (\omega - \omega_0)\cdot\tau/2$ and $x_2 = (\omega + \omega_0)\cdot\tau/2$.

This represents the ideal SAW continuous response in both time and frequency. This can be related to the sampled response by a few substitutions of variables. Let

$$\Delta t = \frac{1}{2\cdot f_0}, \qquad t_n = n\cdot\Delta t, \qquad N\cdot\Delta t = \tau, \qquad N_p\cdot\Delta t = \tau/2 \tag{47.10}$$

Assuming a frequency bandlimited response, the negative frequency component centered around $-f_0$ can be ignored. Then the frequency response, using Eq. (47.9), is given by

$$H(\omega) = \kappa\left\{\frac{\pi N_p}{\omega_0}\right\} \cdot \frac{\sin(x_n)}{x_n} \tag{47.11}$$

where

$$x_n = \frac{(\omega - \omega_0)}{\omega_0}\pi N_p = \frac{(f - f_0)}{f_0}\pi N_p$$

The conductance, given using Eqs. (47.6) and (47.10), is

$$G_a(f) = 2\kappa^2\left\{\frac{\pi N_p}{2\pi f_0}\right\}^2 \frac{\sin^2(x_n)}{x_n^2} = 8k^2 f_0 C_s N_p^2 \cdot \frac{\sin^2(x_n)}{x_n^2} \tag{47.12}$$

This yields the frequency-dependent conductance per unit width of the transducer. Given a uniform transducer of width, W_a, the total transducer conductance is obtained by multiplying Eq. (47.12) by W_a. Defining the center frequency conductance as

$$G_a(f_0) = G_0 = 8k^2 f_0 C_s W_a N_p^2 \tag{47.13}$$

the transducer conductance is

$$G_a(f_0) = G_0 \cdot \frac{\sin^2(x_n)}{x_n^2} \tag{47.14}$$

The transducer electrode capacitance is given as

$$C_e = C_s W_a N_p \tag{47.15}$$

Finally, the last term of the SAW transducer's equivalent circuit is the frequency-dependent susceptance. Given any system where the frequency-dependent real part is known, there is an associated imaginary part which must exist for the system to be real and causal. This is given by the Hilbert transform susceptance, defined as B_a, where [Datta, 1986]

$$B_a(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{G_a(u)}{(u - \omega)}\,du = G_a(\omega) * 1/\omega \tag{47.16}$$

where "*" indicates convolution.

These three elements compose a SAW transducer equivalent circuit. The equivalent circuit, shown in Fig. 47.4, is composed of one lumped element and two frequency-dependent terms which are related to the substrate material parameters, transducer electrode number, and the transducer configuration. Figure 47.5 shows the

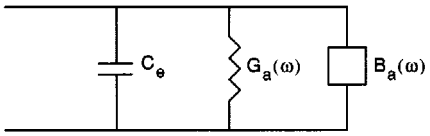


FIGURE 47.4 Electrical equivalent circuit model.

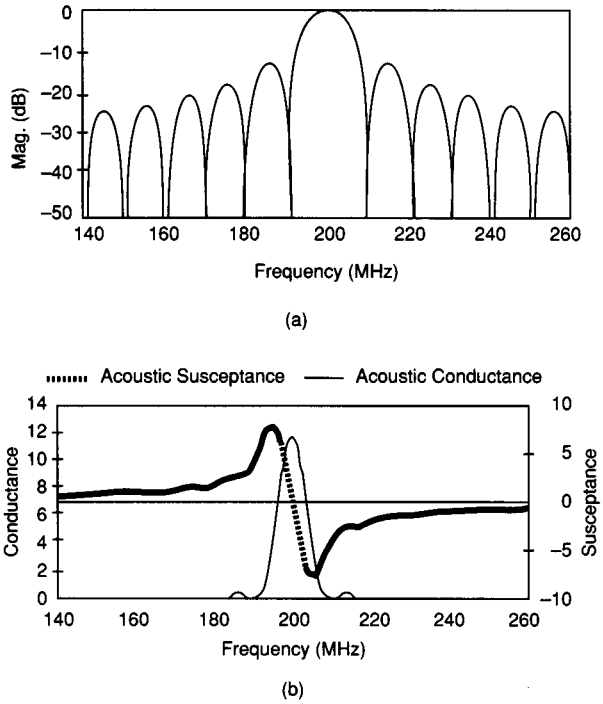


FIGURE 47.5 (a) Theoretical frequency response of a $rect(t/\tau)$ time function having a time length of 0.1 μs and a 200-MHz carrier frequency. (b) Theoretical conductance and susceptance for a SAW transducer implementing the frequency response. The conductance and susceptance are relative and are given in millisiemens.

time and frequency response for a uniform transducer and the associated frequency-dependent conductance and Hilbert transform susceptance. The simple impulse model treats each electrode as an ideal impulse; however, the electrodes have a finite width which distorts the ideal impulse response. The actual SAW potential has been shown to be closely related to the electrostatic charge induced on the transducer by the input voltage. The problem is solved assuming a quasi-static and electrostatic charge distribution, assuming a semi-infinite array of electrodes, solving for a single element, and then using superposition and convolution. The charge distribution solution for a single electrode with all others grounded is defined as the basic charge distribution function (BCDF). The result of a series of arbitrary voltages placed on a series of electrodes is the summation of scaled, time-shifted BCDFs. The identical result is obtained if an array factor, $a(x)$, defined as the ideal impulses localized at the center of the electrode or gap, is convolved with the BCDF, often called the element factor. This is very similar to the analysis of antenna arrays. Therefore, the ideal frequency transfer function and conductance given by the impulse response model need only be modified by multiplying the frequency-dependent element factor. The analytic solution to the BCDF is given in Datta [1986] and Morgan [1985], and is shown to place a small perturbation in the form of a slope or dip over the normal bandwidths of interest. The BCDF also predicts the expected harmonic frequency responses.

Apodized SAW Transducers

Apodization is the most widely used method for weighting a SAW transducer. The desired time-sampled impulse response is implemented by assigning the overlap of opposite polarity electrodes at a given position to a normalized sample weight at a given time. A tap having a weight of unity has an overlap across the entire beamwidth while a small tap will have a small overlap of adjacent electrodes. The time impulse response can be broken into tracks which have uniform height but whose time length and impulse response may vary. Each of these time tracks is implemented spatially across the transducer's beamwidth by overlapped electrode sections at the proper positions. This is shown in Fig. 47.1. The smaller the width of the tracks, the more exact the approximation of uniform time samples. There are many different ways to implement the time-to-spatial transformation; Fig. 47.1 shows just one such implementation.

The impulse response can be represented, to any required accuracy, as the summation of uniform samples located at the proper positions in time in a given track. Mathematically this is given by

$$h(t) = \sum_{i=1}^{I} h_i(t) \tag{47.17}$$

and

$$H(\omega) = \sum_{i=1}^{I} H_i(\omega) = \sum_{i=1}^{I} \left\{ \int_{-\tau/2}^{\tau/2} h_i(t) e^{-j\omega t} dt \right\} \tag{47.18}$$

The frequency response is the summation of the individual frequency responses in each track, which may be widely varying depending on the required impulse response. This spatial weighting complicates the calculations of the equivalent circuit for the transducer. Each track must be evaluated separately for its acoustic conductance, acoustic capacitance, and acoustic susceptance. The transducer elements are then obtained by summing the individual track values yielding the final transducer equivalent circuit parameters. These parameters can be solved analytically for simple impulse response shapes (such as the rect, triangle, cosine, etc.) but are usually solved numerically on a computer [Richie et al., 1988].

There is also a secondary effect of apodization when attempting to extract energy. Not all of the power of a nonuniform SAW beam can be extracted by an a uniform transducer, and reciprocally, not all of the energy of a uniform SAW beam can be extracted by an apodized transducer. The transducer efficiency is calculated at center frequency as

$$E = \frac{\left| \sum_{i=1}^{I} H(\omega_0) \right|^2}{I \cdot \sum_{i=1}^{I} H^2(\omega_0)} \tag{47.19}$$

The apodization loss is defined as

$$\text{apodization loss} = 10 \cdot \log(E) \tag{47.20}$$

Typical apodization loss for common SAW transducers is 1 dB or less.

Finally, because an apodized transducer radiates a nonuniform beam profile, the response of two cascaded apodized transducers is not the product of each transducer's individual frequency responses, but rather is given by

$$H_{12}(\omega) = \sum_{i=1}^{I} H_{1i}(\omega) \cdot H_{2i}(\omega) \neq \sum_{i=1}^{I} H_{1i}(\omega) \cdot \sum_{i=1}^{I} H_{2i}(\omega) \tag{47.21}$$

In general, filters are normally designed with one apodized and one uniform transducer or with two apodized transducers coupled with a spatial-to-amplitude acoustic conversion component, such as a multistrip coupler [Datta, 1986].

47.5 Distortion and Second-Order Effects

In SAW devices there are a number of effects which can distort the desired response from the ideal response. The most significant distortion in SAW transducers is called the **triple transit echo (TTE)** which causes a delayed signal in time and an inband ripple in the amplitude and delay of the filter. The TTE is primarily due to an electrically regenerated SAW at the output transducer which travels back to the input transducer, where it induces a voltage across the electrodes which in turn regenerates another SAW which arrives back at the output transducer. This is illustrated schematically in Fig. 47.2. Properly designed and matched **unidirectional transducers** have acceptably low levels of TTE due to their design. Bidirectional transducers, however, must be mismatched in order to achieve acceptable TTE levels. To first order, the TTE for a bidirectional two-transducer filter is given as

$$\text{TTE} \approx 2 \cdot IL + 6 \text{ dB} \tag{47.22}$$

where IL = filter insertion loss, in dB [Matthews, 1977]. As examples, the result of TTE is to cause a ghost in a video response and intersymbol interference in data transmission.

Another distortion effect is electromagnetic feedthrough which is due to direct coupling between the input and output ports of the device, bypassing any acoustic response. This effect is minimized by proper device design, mounting, bonding, and packaging.

In addition to generating a SAW, other spurious acoustic modes may be generated. Bulk acoustic waves (BAW) may be both generated and received, which causes passband distortion and loss of out-of-band rejection. BAW generation is minimized by proper choice of material, roughening of the crystal backside to scatter BAWs, and use of a SAW track changer, such as a multistrip coupler.

Any plane wave which is generated from a finite aperture will begin to diffract. This is exactly analogous to light diffracting through a slit. Diffraction's principal effect is to cause effective shifts in the filter's tap weights and phase which results in increased sidelobe levels in the measured frequency response. Diffraction is minimized by proper choice of substrate and filter design.

Transducer electrodes are fabricated from thin film metal, usually aluminum, and are finite in width. This metal can cause discontinuities to the surface wave which cause velocity shifts and frequency-dependent reflections. In addition, the films have a given sheet resistance which gives rise to a parasitic electrode resistance loss. The electrodes are designed to minimize these distortions in the device.

47.6 Bidirectional Filter Response

A SAW filter is composed of two cascaded transducers. In addition, the overall filter function is the product of two acoustic transfer functions, two electrical transfer functions, and a delay line function, as illustrated in Fig. 47.6. The acoustic filter functions are as designed by each SAW transducer. The delay line function is dependent on several parameters, the most important being frequency and transducer separation. The propagation path transfer function, $D(\omega)$, is normally assumed unity, although this may not be true for high frequencies ($f >$ 500 MHz) or if there are films in the propagation path. The electrical networks may cause distortion of the acoustic response and are typically compensated in the initial SAW transducer's design.

The SAW electrical network is analyzed using the SAW equivalent circuit model plus the addition of packaging parasitics and any tuning or matching networks. Figure 47.7 shows a typical electrical network which is computer analyzed to yield the overall transfer function for one port of the two-port SAW filter [Morgan, 1985]. The second port is analyzed in a similar manner and the overall transfer function is obtained as the product of the electrical, acoustic, and propagation delay line effects.

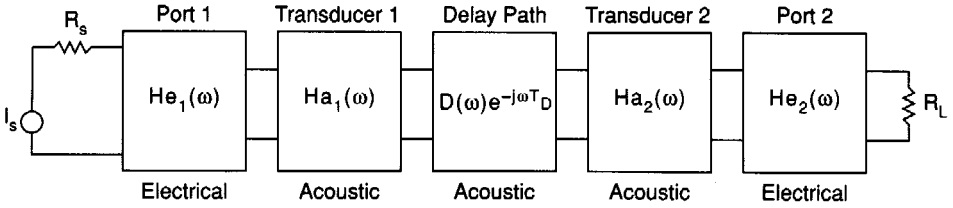


FIGURE 47.6 Complete transfer function of a SAW filter including the acoustic, electrical, and delay line transfer functions. The current generator is I_s, and R_s and R_L are the source and generator resistances, respectively.

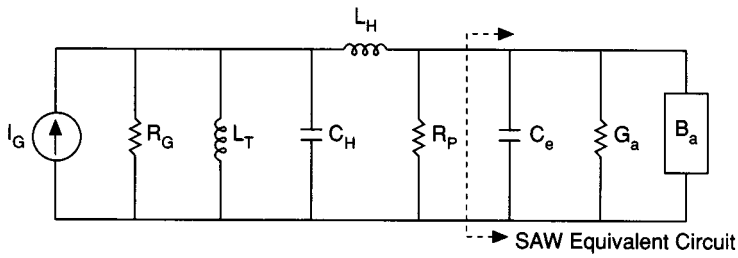


FIGURE 47.7 Electrical network analysis for a SAW transducer. I_G and R_G represent the generator source and impedance, L_T is a tuning inductor, C_H and L_H are due to the package capacitance and bond wire, respectively, and R_P represents a parasitic resistance due to the electrode transducer resistance. The entire network, including the frequency-dependent SAW network, is solved to yield the single-port transfer function.

47.7 Multiphase Unidirectional Transducers

The simplest SAW transducers are single-phase bidirectional transducers. Because of their symmetrical nature, SAW energy is launched equally in both directions from the transducer. In a two- transducer configuration, half the energy (3 dB) is lost at the transmitter, and reciprocally, only half the energy can be received at the receiver. This yields a net 6-dB loss in a filter. However, by adding nonsymmetry into the transducer, either by electrical multiphases or nonsymmetry in reflection and regeneration, energy can be unidirectionally directed yielding a theoretical minimum 0-dB loss.

The most common SAW UDTs are called the three-phase UDT (3PUDT) and the group type UDT (GUDT). The 3PUDT has the broadest bandwidth and requires multilevel metal structures with crossovers. The GUDT uses a single-level metal but has a narrower unidirectional bandwidth due to its structure. In addition, there are other UDT or equivalent embodiments which can be implemented but will not be discussed [Morgan, 1985]. The basic structure of a 3PUDT is shown in Fig. 47.8. A unit cell consists of three electrodes, each connected to a separate bus bar, where the electrode period is $\lambda_0/3$. One bus bar is grounded and the other two bus bars will be driven by an electrical network where $V_1 = V_2 \angle 60°$. The transducer analysis can be accomplished similar to a simple IDT by considering the 3PUDT as three collinear IDTs with a spatial phase shift, as shown in Fig. 47.8. The electrical phasing network, typically consisting of one or two reactive elements, in conjunction with the spatial offset results in energy being launched in only one direction from the SAW transducer. The transducer can then be matched to the required load impedance with one or two additional reactive elements. The effective unidirectional bandwidth of the 3PUDT is typically 20% or less, beyond which the transducer behaves as a normal bidirectional transducer. Figure 47.9 shows a 3PUDT filter schematic consisting of two transducers and their associated matching and phasing networks. The overall filter must be analyzed with all external electrical components in place for accurate prediction of performance. The external components can be miniaturized and may be fabricated using only printed circuit board material and area. This type of device has demonstrated as low as 2 dB insertion loss.

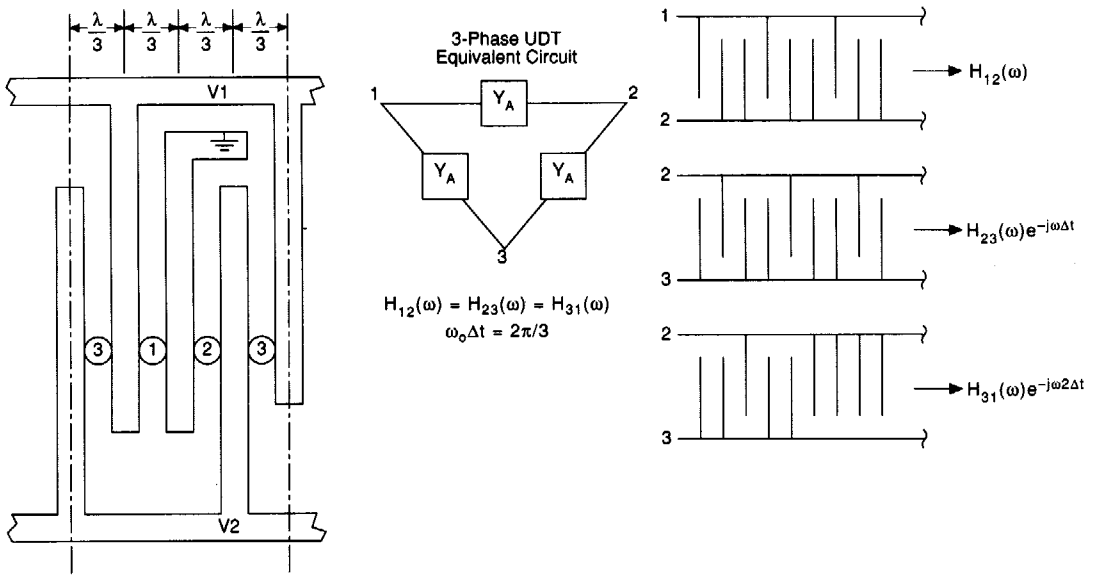


FIGURE 47.8 Schematic of a unit cell of a 3PUDT and the basic equivalent circuit. The 3PUDT can be analyzed as three collinear transducers with a spatial offset.

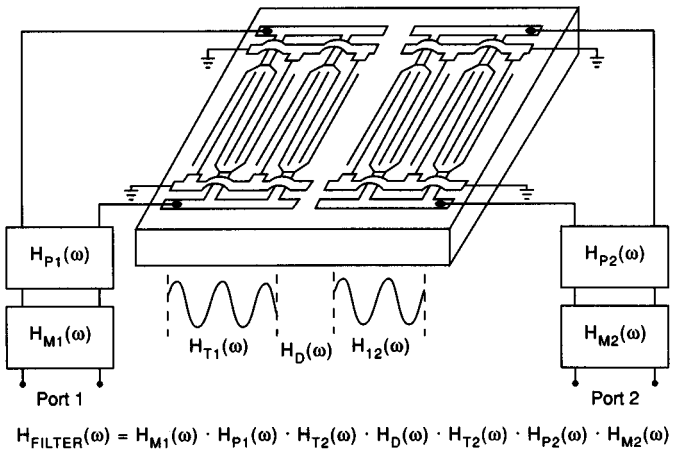


FIGURE 47.9 Schematic diagram of a 3PUDT which requires the analysis of both the acoustic transducer responses as well as electrical phasing and matching networks.

47.8 Single-Phase Unidirectional Transducers

Single-phase unidirectional transducers (SPUDT) use spatial offsets between mechanical electrode reflections and electrical regeneration to launch a SAW in one direction. A reflecting structure may be made of metal electrodes, dielectric strips, or grooved reflectors which are properly placed within a transduction structure. Under proper design and electrical matching conditions, the mechanical reflections can exactly cancel the electrical regeneration in one direction of the wave over a moderate band of frequencies. This is schematically illustrated in Fig. 47.10 which shows a reflector structure and a transduction structure merged to form a SPUDT. The transducer needs to be properly matched to the load for optimum operation. The mechanical reflections can be controlled by modifying the width, position, or height of the individual reflector. The regenerated SAW is primarily controlled by the electrical matching to the load of the transduction structure. SPUDT filters have

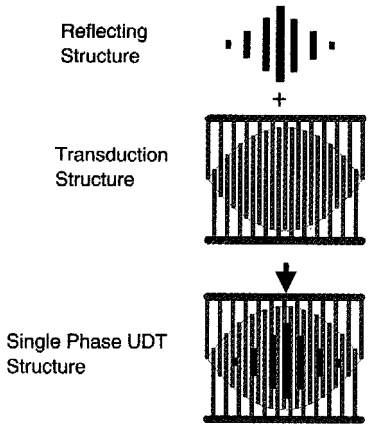


FIGURE 47.10 Schematic representation of a SPUDT which is a combination of transduction and reflecting structures to launch a SAW in one direction over moderate bandwidths.

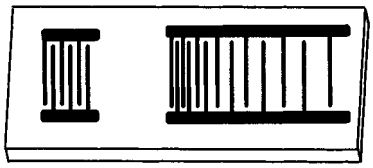


FIGURE 47.11 A SAW dispersive filter consisting of a uniform transducer and a "down chirp" dispersive transducer. The high frequencies have a shorter delay than the low frequencies in this example.

exhibited as low as 3 dB loss over fractional bandwidths of 5% or less and have the advantage of not needing phasing networks when compared to the multiphase UDTs.

47.9 Dispersive Filters

SAW filters can also be designed and fabricated using nonuniformly spaced electrodes in the transducer. The distance between adjacent electrodes determines the "local" generated frequency. As the spacing between the electrodes changes, the frequency is slowly changed either up (decreasing electrode spacing) or down (increasing electrode spacing) as the position progresses along the transducer. This slow frequency change with time is often called a "chirp." Figure 47.11 shows a typical dispersive filter consisting of a chirped transducer in cascade with a uniform transducer. Filters can be designed with either one or two chirped transducers and the rate of the chirp is variable within the design. These devices have found wide application in radar systems due to their small size, reproducibility, and large time bandwidth product.

47.10 Coded SAW Filters

Because of the ability to control the amplitude and phase of the individual electrodes or taps, it is easy to implement coding in a SAW filter. Figure 47.12 shows an example of a coded SAW filter implementation. By changing the phase of the taps, it is possible to generate an arbitrary code sequence. These types of filters are used in secure communication systems, spread spectrum communications, and tagging, to name a few [Matthews, 1977].

SAW devices can also be used to produce time impulse response shapes for use in modulators, equalizers, and other applications. An example of a SAW modulator used for generating a cosine envelope for a minimum shift keyed (MSK) modulator is shown in Fig. 47.13 [Morgan, 1985].

47.11 Resonators

Another very important class of devices is SAW resonators. Resonators can be used as frequency control elements in oscillators, as notch filters, and as narrowband filters, to name a few. Resonators are typically fabricated on piezoelectric quartz substrates due to its low TCD which yields temperature-stable devices. A resonator uses one or two transducers for coupling energy in/out of the device and one or more distributed reflector arrays to store energy in the device. This is analogous to an optical cavity with the distributed reflector arrays acting

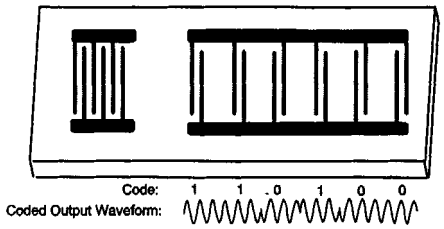


FIGURE 47.12 Example of a coded SAW tapped delay line.

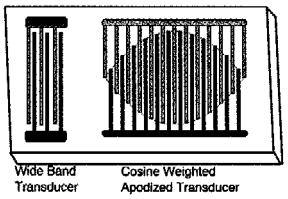


FIGURE 47.13 A SAW filter for implementing an MSK waveform using a wideband input transducer and a cosine envelope apodized transducer.

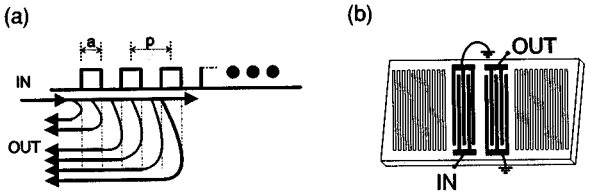


FIGURE 47.14 (a) SAW reflector array illustrating synchronous distributed reflections at center frequency. Individual electrode width (a) is 1/4 wavelength and the array period is 1/2 wavelength at center frequency. (b) A schematic of a simple single-pole, single-cavity two-port SAW resonator.

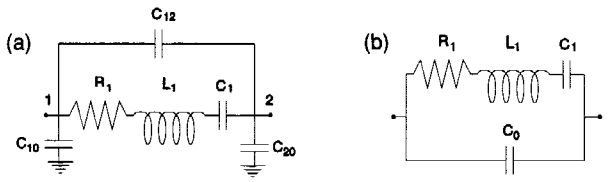


FIGURE 47.15 (a) Two-port resonator equivalent circuit and (b) one-port resonator equivalent circuit.

as the mirrors. A localized acoustic mirror, such as a cleaved edge, is not practical for SAW because of spurious mode coupling at edge discontinuities which causes significant losses.

A distributive reflective array is typically composed of a series of shorted metal electrodes, etched grooves in the substrate, or dielectric strips. There is a physical discontinuity on the substrate surface due to the individual reflectors. Each reflector is one-quarter wavelength wide and the periodicity of the array is one-half wavelength. This is shown schematically in Fig. 47.14. The net reflections from all the individual array elements add synchronously at center frequency, resulting in a very efficient reflector. The reflection from each array element is small and very little spurious mode coupling results.

Figure 47.14 shows a typical single-pole, single-cavity, two-port SAW resonator. Resonators can be made multipole by addition of multiple cavities, which can be accomplished by inline acoustic coupling, transverse acoustic coupling, and by electrical coupling. The equivalent circuit for SAW two-port and one-port resonators is shown in Fig. 47.15. SAW resonators have low insertion loss and high electrical Q's of several thousand [Campbell, 1989; Datta, 1986; Morgan, 1985].

Defining Terms

Bidirectional transducer: A SAW transducer which launches energy from both acoustic ports which are located at either end of the transducer structure.

Interdigital transducer: A series of collinear electrodes placed on a piezoelectric substrate for the purpose of launching a surface acoustic wave.

Surface acoustic wave (SAW): A surface acoustic wave (also known as a Rayleigh wave) is composed of a coupled compressional and shear wave. On a piezoelectric substrate there is also an electrostatic wave which allows electroacoustic coupling. The wave is confined at or near the surface and decays away rapidly from the surface.

Triple transit echo (TTE): A multiple transit echo received at three times the main SAW signal delay time. This echo is caused due to the bidirectional nature of SAW transducers and the electrical and/or acoustic mismatch at the respective ports. This is a primary delayed signal distortion which can cause filter distortion, especially in bidirectional transducers and filters.

Unidirectional transducer (UDT): A transducer which is capable of launching energy from primarily one acoustic port over a desired bandwidth of interest.

Related Topics

2.1 Step, Impulse, Ramp, Sinusoidal, Exponential, and DC Signals • 5.3 Distortion • 10.2 Ideal Filters • 49.2 Mechanical Characteristics

References

D.S. Ballintine, *Acoustic Wave Sensors,* San Diego, Calif.: Academic Press, 1995.

C. Campbell, *Surface Acoustic Wave Devices and their Signal Processing Applications,* San Diego, Calif.: Academic Press, 1989.

S. Datta, *Surface Acoustic Wave Devices*, Englewood Cliffs, N.J.: Prentice-Hall, 1986.

C.S. Hartmann, D.T. Bell, and R.C. Rosenfeld, "Impulse model design of acoustic surface wave filters," *IEEE Transactions on Microwave Theory and Techniques,* vol. 21, pp. 162–175, 1973.

H. Matthews, *Surface Wave Filters,* New York: Wiley Interscience, 1977.

D.P. Morgan, *Surface Wave Devices for Signal Processing*, New York: Elsevier, 1985.

S.M. Richie, B.P. Abbott, and D.C. Malocha, "Description and development of a SAW filter CAD system," *IEEE Transactions on Microwave Theory and Techniques,* vol. 36, no. 2, 1988.

Further Information

The *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* provides excellent information and detailed articles on SAW technology.

The *IEEE Ultrasonics Symposium Proceeding* provides information on ultrasonic devices, systems, and applications for that year. Articles present the latest research and developments and include invited articles from eminent engineers and scientists.

The *IEEE Frequency Control Symposium Proceedings* provides information on frequency control devices, systems, and applications (including SAW) for that year. Articles present the latest research and developments and include invited articles from eminent engineers and scientists.

For additional information, see the following references:

IEEE Transaction on Microwave Theory and Techniques, vol. 21, no. 4, 1973, special issue on SAW technology.

IEEE Proceedings, vol. 64, no. 5, special issue on SAW devices and applications.

Joint Special Issue of *IEEE Transaction on Microwave Theory and Techniques* and *IEEE Transactions on Sonics and Ultrasonics,* MTT-vol. 29, no. 5, 1981, on SAW device systems.

M. Feldmann and J. Henaff, *Surface Acoustic Waves for Signal Processing*, Norwood, Mass.: Artech House, 1989.

B.A. Auld, *Acoustic Fields and Waves in Solids,* New York: Wiley, 1973.

V.M. Ristic, *Principles of Acoustic Devices*, New York: Wiley, 1983.

A. Oliner, *Surface Acoustic Waves,* New York: Springer-Verlag, 1978.

Farnell, G.W. “Ultrasound”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

48

Ultrasound

48.1 Introduction
48.2 Propagation in Solids
48.3 Piezoelectric Excitation
48.4 One-Dimensional Propagation
48.5 Transducers

Gerald W. Farnell
McGill University

48.1 Introduction

In electrical engineering, the term *ultrasonics* usually refers to the study and use of waves of mechanical vibrations propagating in solids or liquids with frequencies in the megahertz or low gigahertz ranges. Such waves in these frequency ranges have wavelengths on the order of micrometers and thus can be electrically generated, directed, and detected with transducers of reasonable size. These ultrasonic devices are used for signal processing directly in such applications as filtering and pulse compression and indirectly in acousto-optic processors; for flaw detection in optically opaque materials; for resonant circuits in frequency control applications; and for medical imaging of human organs, tissue, and blood flow.

48.2 Propagation in Solids

If the solid under consideration is elastic (linear), homogeneous, and nonpiezoelectric, the components, u_i, of the displacement of an infinitesimal region of the material measured along a set of Cartesian axes, x_i, are interrelated by an equation of motion:

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \sum_j \sum_k \sum_l c_{ijkl} \frac{\partial^2 u_j}{\partial x_k \partial x_l}, \qquad \text{Form: } \rho \frac{\partial^2 u}{\partial t^2} = c \frac{\partial^2 u}{\partial x^2} \tag{48.1}$$

where ρ is the mass density of the material and c_{ijkl} $(i, j, k, l = 1, 2, 3)$ is called the stiffness tensor. It is the set of proportionality constants between the components of the stress tensor T and the strain tensor S in a three-dimensional Hooke's law (form: $T = cS$ with $S = \partial u/\partial x$). In Eq. (48.1) and in the subsequent equations the form of the equation is shown without the clutter of the many subscripts. The form is useful for discussion purposes; moreover, it gives the complete equation for cases in which the propagation can be treated as one dimensional, i.e., with variations in only one direction, one component of displacement, and one relevant c.

In an infinite medium, the simplest solutions of Eq. (48.1) are plane waves given by the real part of

$$u_i = U_i e^{-jk\left(\sum_j L_j x_j - Vt\right)} \qquad \text{Form: } u = U e^{j(\omega t - kx)} \tag{48.2}$$

where the polarization vector has components U_i along the axes. The **phase velocity** of the wave V is measured along the propagation vector **k** whose direction cosines with respect to these axes are given by L_i. Substituting

the assumed solutions of Eq. (48.2) into Eq. (48.1) gives the third-order eigenvalue equations, usually known as the Christoffel equations:

$$\sum_j \sum_k \sum_l L_k L_l c_{ijkl} U_j = \rho V^2 U_i, \qquad \text{Form: } (c - \rho V^2)U = 0 \tag{48.3}$$

The three eigenvalues in Eq. (48.3) give three values of ρV^2 and hence the phase velocities of three waves propagating in the direction of positive k and three propagating in the negative k direction. The eigenvectors of the three forward solutions give the polarization vector for each, and they form a mutually perpendicular triad. The polarization vector of one of the plane waves will be parallel, or almost parallel, to the k vector, and it is called the longitudinal wave, or quasi-longitudinal if the displacement is not exactly parallel to k. The other two waves will have mutually perpendicular polarization vectors, which will each be perpendicular, or almost perpendicular, to the k vector. If the polarization is perpendicular, the wave is called a transverse or shear wave; if almost perpendicular, it is called quasi-shear. The three waves propagate independently through the solid, and their respective amplitudes depend on the exciting source.

In an isotropic medium where there are only two independent values of c_{ijkl} in Eq. (48.1), there are one longitudinal wave and two degenerate shear waves. The phase velocities of these waves are independent of the direction of propagation and are given by

$$V_l = \sqrt{\frac{c_{1111}}{\rho}} \quad \text{and} \quad V_s = \sqrt{\frac{c_{1212}}{\rho}} \tag{48.4}$$

The phase velocities in isotropic solids are often expressed in terms of the so-called Lame constants defined by $\mu = c_{1212}$ and $\lambda = c_{1111} - 2c_{1212}$. The longitudinal velocity is larger than the shear velocity. Exact velocity values depend on fabrication procedures and purity, but Table 48.1 gives typical values for some materials important in ultrasonics.

In signal processing applications of ultrasonics, the propagating medium is often a single crystal, and thus a larger number of independent stiffness constants is required to describe the mechanical properties of the medium, e.g., three in a cubic crystal, five in a hexagonal, and six in a trigonal. Note that while the number of independent constants is relatively small, a large number of the c_{ijkl} are nonzero but are related to each other by the symmetry characteristics of the crystal. The phase velocities of each of the three independent plane waves in an anisotropic medium depend on the direction of propagation. Rather than plotting V as a function of angle of propagation, it is more common to use a **slowness surface** giving the reciprocal of V (or $\mathbf{k} = \omega/V$ for a given ω) as a function of the direction of k. Usually planar cuts of such slowness surfaces are plotted as shown in Figs. 48.1(a) and (b).

In anisotropic materials the direction of energy flow (the ultrasonic equivalent of the electromagnetic Poynting vector) in a plane wave is not parallel to **k.** Thus the direction of **k** is set by the transducer but the energy flow or beam direction is normal to the tangent to the slowness surface at the point corresponding to **k.** The direction of propagation (of **k**) in Fig. 48.1 lies in the basal plane of a cubic crystal, here silicon. At each angle there are three waves—one is pure shear polarized perpendicular to this plane, one is quasilongitudinal for most angles, while the third is quasi-shear. For the latter two, the tangent to the slowness curves at an arbitrary angle is not normal to the radius vector, and thus there is an appreciable angle between the direction of energy flow and the direction of **k**. This angle is shown on the diagram by the typical **k** and **P** vectors, the latter being the direction of energy flow in an acoustic beam with this **k.** Along the cubic axes in a cubic crystal, the two shear waves are degenerate, and for all three waves the energy flow is parallel to **k.** When the particle displacement of a mode is either parallel to the propagation vector or perpendicular to it and the energy flow is parallel to **k,** the mode is called a **pure mode.** The propagation vector in Fig. 48.1(b) lies in the basal plane of a trigonal crystal, quartz.

When ultrasonic waves propagate in a solid, there are various losses that attenuate the wave. Usually the attenuation per wavelength is small enough that one can neglect the losses in the initial calculation of the

TABLE 48.1 Typical Acoustic Properties

Material	Velocity (km/s)		Impedance ($kg/m^2 s \times 10^6$)		Density ($kg/m^3 \times 10^3$)	Comments
	Longitudinal	Shear	Longitudinal	Shear		
Alcohol, methanol	1.103		0.872		0.791	Liq. 25°C
Aluminum, rolled	6.42	3.04	17.33	8.21	02.70	Isot.
Brass, 70% Cu, 30% Zn	4.70	2.10	40.6	18.14	8.64	Isot.
Cadmium sulphide	4.46	1.76	21.5	8.5	4.82	Piez crys *Z*-dir
Castor oil	1.507		1.42		0.942	Liq. 20°C
Chromium	6.65	4.03	46.6	28.21	7.0	Isot.
Copper, rolled	5.01	2.27	44.6	20.2	8.93	Isot.
Ethylene glycol	1.658		1.845		1.113	Liq. 25°C
Fused quartz	5.96	3.76	13.1	8.26	2.20	Isot.
Glass, crown	5.1	2.8	11.4	6.26	2.24	Isot.
Gold, hard drawn	3.24	1.20	63.8	23.6	19.7	Isot.
Iron, cast	5.9	3.2	46.4	24.6	7.69	Isot.
Lead	2.2	0.7	24.6	7.83	11.2	Isot.
Lithium niobate, $LiNbO_3$	6.57	4.08	30.9	19.17	4.70	Piez crys *X*-dir
		4.79		22.53		
Nickel	5.6	3.0	49.5	26.5	8.84	Isot.
Polystyrene, styron	2.40	1.15	2.52	1.21	1.05	Isot.
PZT-5H	4.60	1.75	34.5	13.1	7.50	Piez ceram *Z*
Quartz	5.74	3.3	15.2	8.7	2.65	Piez crys *X*-dir
		5.1		13.5		
Sapphire Al_2O_3	11.1	6.04	44.3	25.2	3.99	Cryst. *Z*-axis
Silver	3.6	1.6	38.0	16.9	10.6	Isot.
Steel, mild	5.9	3.2	46.0	24.9	7.80	Isot.
Tin	3.3	1.7	24.2	12.5	7.3	Isot.
Titanium	6.1	3.1	27.3	13.9	4.48	Isot.
Water	1.48		1.48		1.00	Liq. 20°C
YAG $Y_3Al_{15}O_{12}$	8.57	5.03	39.0	22.9	4.55	Cryst. *Z*-axis
Zinc	4.2	2.4	29.6	16.9	7.0	Isot.
Zinc oxide	6.37	2.73	36.1	15.47	5.67	Piez crys *Z*-dir

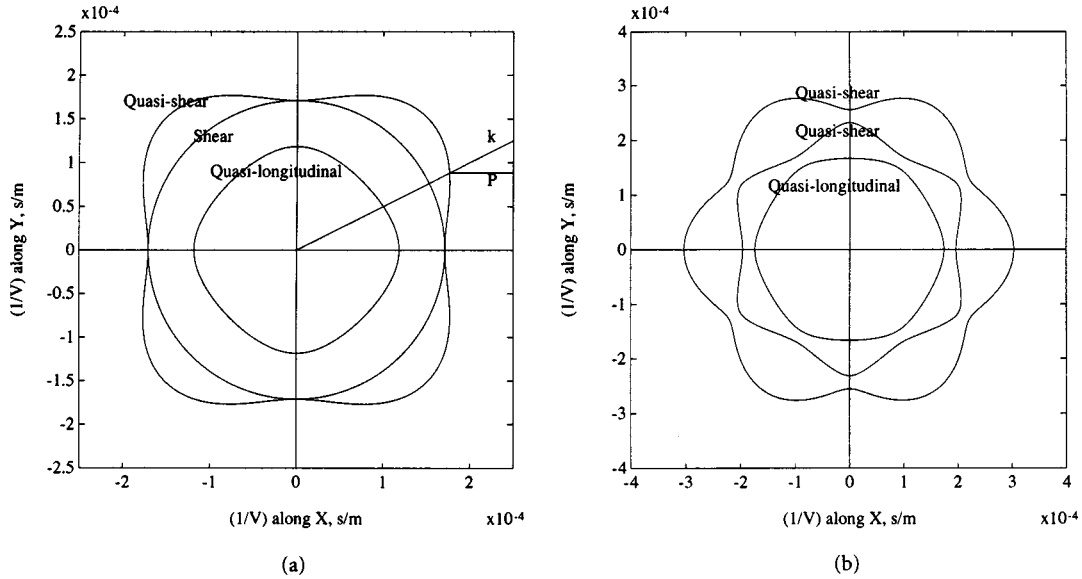


FIGURE 48.1 (a) Slowness curves, basal plane, cubic crystal, silicon. (b) Slowness curves, basal plane, trigonal crystal, quartz.

propagation characteristics of the material and the excitation, and then multiply the resulting propagating wave by a factor of the form $\exp[-\alpha x]$ where x is in the direction of k and α is called the *attenuation constant*. One loss mechanism is the viscosity of the material and due to it the attenuation constant is

$$\alpha = \eta \frac{\omega^2}{2V^3\rho} \tag{48.5}$$

in which η is the coefficient of viscosity. It should be noted that the attenuation constant for viscous loss increases as the square of the frequency. In polycrystalline solids there is also loss due to scattering from dislocation and grain structure; thus, for the same material the loss at high frequencies is much higher in a polycrystalline form than in a crystalline one. As a result, in high-frequency applications of ultrasound, such as for signal processing, the propagation material is usually in single-crystal form.

48.3 Piezoelectric Excitation

When a piezoelectric material is stressed, an electric field is generated in the stressed region; similarly, if an electric field is applied, there will be an induced stress on the material in the region of the field. Thus, there is a coupling between mechanical motion and time-varying electric fields. Analysis of wave propagation in piezoelectric solids should thus include the coupling of the mechanical equations such as Eq. (48.1) with Maxwell's equations. In most ultrasonic problems, however, the velocity of the mechanical wave solutions is slow enough that the electric fields can be described by a scalar potential ϕ. This is called the *quasi-static approximation*. Within this approximation, the equations of motion in a piezoelectric solid become

$$\rho\frac{\partial^2 u_i}{\partial t^2} - \sum_j\sum_k\sum_l c_{ijkl}\frac{\partial^2 u_j}{\partial x_k \partial x_l} = \sum_j\sum_k e_{ijk}\frac{\partial^2 \phi}{\partial x_j \partial x_k} \qquad \text{Form: } \rho\frac{\partial^2 u}{\partial t^2} - c\frac{\partial^2 u}{\partial x^2} = e\frac{\partial^2 \phi}{\partial x^2}$$
$$\sum_i\sum_j \epsilon_{ij}\frac{\partial^2 \phi}{\partial x_i \partial x_j} = \sum_i\sum_j\sum_k e_{ijk}\frac{\partial^2 u_j}{\partial x_i \partial x_k} \qquad \epsilon\nabla^2\phi = e\frac{\partial^2 u}{\partial x^2} \tag{48.6}$$

The piezoelectric coupling constants e_{ijk} form a third-rank tensor property of the solid and are the proportionality constants between the components of the electric field and the components of the stress. Similarly ϵ_{ij} is the second-rank permittivity tensor, giving the proportionality constants between the components of the electric field E and of the electric displacement D. If the material is nonpiezoelectric $e_{ijk} = 0$, then the first three equations of Eq. (48.6) reduce to the corresponding three of Eq. (48.1), whereas the fourth equation becomes the anisotropic Laplace equation. In a piezoelectric, these mechanical and electrical components are coupled.

The plane wave solution of Eq. (48.6) then has the three mechanical components of Eq. (48.2) and in addition has a potential given by

$$\phi = \Phi e^{-jk\left(\sum_j L_j x_j - Vt\right)} \qquad \text{Form: } \phi = \Phi e^{j(\omega t - kx)} \tag{48.7}$$

Thus, for the quasi-static approximation there is a wave of potential that propagates with an acoustic phase velocity V in synchronism with the mechanical variations. As will be seen in Section 48.5, it is possible to use the corresponding electric field, $-\nabla\phi$, to couple to electrode configurations and thus excite or detect the ultrasonic wave from external electric circuits.

Rather than substituting Eq. (48.7) and Eq. (48.2) into Eq. (48.6) to obtain a set of four equations similar to Eq. (48.3), it is frequently more convenient to substitute Eq. (48.7) into the fourth equation in the set of Eq. (48.6). Because there are no time derivatives involved, this substitution gives the potential as a linear combination of the components of the mechanical displacement:

$$\Phi = \frac{\sum_i \sum_j \sum_k e_{ijk} L_i L_k U_j}{\sum_i \sum_j \epsilon_{ij} L_i L_j} \qquad \text{Form:} \quad \Phi = \frac{e}{\epsilon} U \tag{48.8}$$

When this combination is substituted into the first three equations of Eq. (48.6) and terms gathered, they become identical to Eq. (48.1) but with each c_{ijkl} replaced by

$$\bar{c}_{ijkl} = c_{ijkl} + \frac{\sum_m \sum_n e_{mij} e_{nkl} L_m L_n}{\sum_m \sum_n \epsilon_{mn} L_m L_n} \qquad \text{Form:} \quad \bar{c} = c\,(1 + K^2) \text{ with } K^2 = \frac{e^2}{c\epsilon} \tag{48.9}$$

Using these so-called stiffened elastic constants, we obtain the same third-order eigenvalue equation, Eq. (48.3), and hence the velocities of each of the three modes and the corresponding mechanical displacement components. The potential is obtained from Eq. (48.8). The velocities obtained for the piezoelectric material are usually at most a few percent higher than would be obtained with the piezoelectricity ignored. The parameter K in Eq. (48.9) is called the electromechanical coupling constant.

48.4 One-Dimensional Propagation

If an acoustic plane wave as in Eq. (48.2) propagating within one medium strikes an interface with another medium, there will be reflection and transmission, much as in the corresponding case in optics. To satisfy the boundary conditions at the interface, it will be necessary in general to generate three transmitted modes and three reflected modes. Thus, the concepts of reflection and transmission coefficients for planar interfaces between anisotropic media are complicated. In many propagation and excitation geometries, however, one can consider only one independent pure mode with energy flow parallel to **k** and particle displacement polarized along **k** or perpendicular to it. This mode (plane wave) then propagates along the axis or its negative in Eq. (48.2). Discussion of the generation, propagation, and reflection of this wave is greatly assisted by considering analogies to the one-dimensional electrical transmission line.

With the transmission line model operating in the sinusoidal steady state, the particle displacement u_i of Eq. (48.2) is represented by a phasor, u. The time derivative of the particle displacement is the particle velocity and is represented by a phasor, $v = j\omega u$, which is taken as analogous to the current on the one-dimensional electrical transmission line. The negative of the stress, or the force per unit area, caused by the particle displacement is represented by a phasor, $(-T) = jkcu$, which is taken as analogous to the voltage on the transmission line. Here c is the appropriate stiffened elastic constant for the mode in question in Eq. (48.3). With these definitions, the general impedance, the characteristic impedance, the phase velocity, and the wave vector, respectively, of the equivalent line are given by

$$Z = \frac{(-T)}{v} \qquad Z_0 = \sqrt{\rho c} \qquad V = \sqrt{\frac{c}{\rho}} \qquad k = \frac{\omega}{V} \tag{48.10}$$

Some typical values of the characteristic impedance of acoustic media are given in Table 48.1. The characteristic impedance corresponding to a mode is given by the product of the density and the phase velocity, ρV, even in the anisotropic case where the effective stiffness c in Eq. (48.10) is difficult to determine.

As an example of the use of the transmission line model, consider a pure longitudinal wave propagating in an isotropic solid and incident normally on the interface with a second isotropic solid. There would be one reflected wave and one transmitted wave, both longitudinally polarized. The relative amplitudes of the stresses in these waves would be given, with direct use transmission line concepts, by the voltage reflection and transmission coefficients

$$\Gamma_R = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}} \quad \text{and} \quad \Gamma_T = \frac{2Z_{02}}{Z_{02} + Z_{01}} \tag{48.11}$$

When an acoustic wave meets a discontinuity or a mismatch, part of the wave is reflected. For an incident mode, an interface represents a lumped impedance. If the medium on the other side of the interface is infinitely deep, that lumped impedance is the characteristic impedance of the second medium. However, if the second medium is of finite depth h in the direction of propagation and it in turn is terminated by a lumped impedance Z_{L2} the impedance seen by the incident wave at the interface is given, as in transmission line theory, by

$$Z_{in} = Z_{02} \frac{Z_{L2} \cos k_2 h + jZ_{02} \sin k_2 h}{Z_{02} \cos k_2 h + jZ_{L2} \sin k_2 h} \tag{48.12}$$

Thus, as with transmission lines, an intervening layer can be used to match from one transmitting medium to another. For example, if the medium following the layer is infinite and of characteristic impedance Z_{03}, i.e., $Z_{L2} = Z_{03}$, the interface will look like Z_{01} to the incident wave if $kh = \pi/2$, quarter-wave thickness, and the layer characteristic impedance is $Z^2_{02} = Z_{01}Z_{03}$. This matching, which provides complete power transfer from medium 1 to medium 3, is valid only at the frequency for which $kh = \pi/2$. For matching over a band of frequencies, multiple matching layers are required.

48.5 Transducers

Electrical energy is converted to acoustic waves in ultrasonic applications by means of electro-acoustic transducers. Most transducers are reciprocal in that they will also convert the mechanical energy in acoustic waves into electrical energy. The form of the transducer is very application dependent. Categories of applications include imaging, wherein one transducer is used to create an acoustic beam, discontinuities in the propagating medium scatter this beam, and the scattered energy is captured by the same or another transducer [see Fig. 48.4(b)]. From the changes of the scattered energy as the beam is moved, characteristics of the scatterer are determined. This is the process in the use of ultrasonics for nondestructive evaluation (NDE), flaw detection, for example, and in ultrasonic images for medical diagnosis. These are radar-like applications and are practical at reasonable frequencies because most solids and liquids support acoustic waves with tolerable losses and the wavelength is short enough that the resolution is adequate for practical targets. By recording both the amplitude and phase of the scattered signal as the transmitter-receiver combination is rotated about a target, one can generate tomographic-type images of the target.

A second category of transducer provides large acoustic standing waves at a particular frequency and, as a result, has a resonant electrical input impedance at this frequency and can be used as a narrowband filter in electrical circuits. In a third category of transducer, the object is to provide an acoustic beam that distorts the medium, as it passes through, in a manner periodic in space and time, and thus provides a dynamic diffraction grating that will deflect or modulate an optical beam that is passed through it [see Fig. 48.4(c)]. Such acousto-optic devices are used in broadband signal processing.

Another category of transducer uses variation of the shape of the electrodes and the geometry of the electroacoustic coupling region so that the transfer function between a transmitting and a receiving transducer is made to have a prescribed frequency response. Such geometries find wide application in filtering and pulse compression applications in the frequency range up to a few gigahertz. Because of the ease of fabrication of complicated electrode geometries, special forms of the solution of the wave equation, Eq. (48.1), called surface acoustic waves (SAW) are dominant in such applications. Because surface acoustic waves are discussed in another section of this handbook, here we will confine the discussion to transducers that generate or detect acoustic waves that are almost plane and usually single mode, the so-called bulk modes.

The prototype geometry for a bulk-mode transducer is shown in Fig. 48.2. The active region is the portion of the piezoelectric slab between the thin metal electrodes, which can be assumed to be circular or rectangular

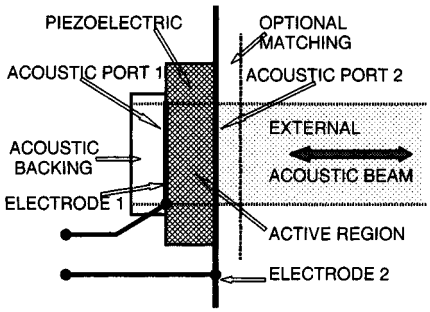


FIGURE 48.2 Prototype transducer geometry.

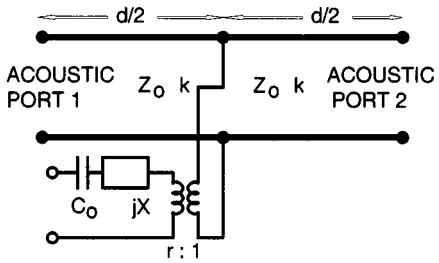


FIGURE 48.3 Model of active region.

in shape. Connections to these electrodes form the electrical port for the transducer and the voltage between them creates a spatially uniform electric field in the active region, and this time-varying electric field couples to the acoustic waves propagating between the electrodes. If the planar electrodes are many wavelengths in transverse dimensions and the active region is much thinner, and if the axial direction is a pure mode direction for the piezoelectric, the waves in the active region can be considered as plane waves. We then have the one-dimensional geometry considered earlier. The transducer may be in contact with another elastic medium on either side, as indicated in Fig. 48.2, so that the plane waves propagate in and out of the active regions in the cross-sectional region shown. Thus, the transducer has in general two acoustic ports for coupling to the outside world as well as the electrical port.

In the absence of piezoelectric coupling, the active region could be represented by a one-dimensional transmission line as discussed in the previous section and as indicated by the heavy lines in Fig. 48.3. With piezoelectricity there will be the stiffening of the appropriate stiffness constants as discussed in Eq. (48.9) with the concomitant perturbation of the characteristic impedance Z_{0p} and the phase velocity V_p, but more important there will also be coupling to the electrical port. One model including the latter coupling is shown in Fig. 48.3 in which the parameters are defined by

$$C_0 = \frac{\epsilon A}{d}; \quad jX = \frac{j}{\omega C_0} K^2 \frac{\sin(\pi\omega/\omega_0)}{\pi\omega/\omega_0} \quad r = \frac{2e/\epsilon}{\omega A Z_0} \sin(\pi\omega/2\omega_0) \tag{48.13}$$

Here C_0 is the capacity that would be measured between the electrodes if there were no mechanical strain on the piezoelectric, A is the cross-sectional area of the active region, and X is an effective reactance. The quantity r is the transformer ratio (with dimensions) of an ideal transformer coupling the electrical port to the center of the acoustic transmission line. K is the electromechanical coupling constant for the material as defined in Eq. (48.9). The so-called resonant frequency ω_0 is that angular frequency at which the length d of the active region is one-half of the stiffened wavelength, $\omega_0 = \pi V/d$. In the physical configuration of Fig. 48.4(a), the transducer has zero stress on the surfaces of the active region and hence both acoustic ports of Fig. 48.3 are terminated in short circuits and the line is mechanically resonant at the angular frequency ω_0. At this frequency the secondary of the transformer of Fig. 48.3 is open circuited if there are no losses, and thus the electrical input impedance is infinite at this frequency and behaves like a parallel resonant circuit for neighboring frequencies. This configuration can be used as a high-Q resonant circuit if the mechanical losses can be kept low, as they are in single crystals of such piezoelectric materials as quartz. It should be noted, however, that the behavior is not as simple as that of a simple L-C parallel resonant circuit, primarily because of the frequency dependence of the effective reactance X and of the transformer ratio in the equivalent circuit. The electrical input impedance is given by

$$Z_{in} = \frac{1}{j\omega C_0}\left(1 - K^2 \frac{\tan kd/2}{kd/2}\right) \tag{48.14}$$

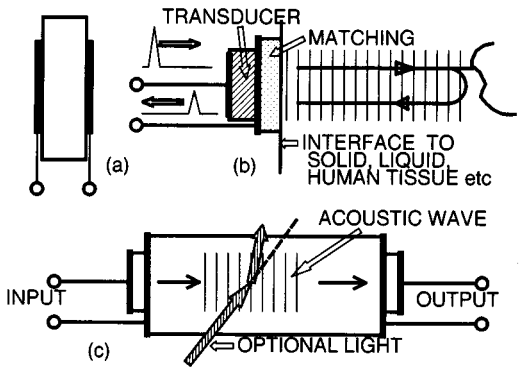


FIGURE 48.4 (a) Resonator structure; (b) acoustic probe; (c) acoustic delay line or optical modulator.

Thus, while the input impedance is infinite as in a parallel resonant circuit at ω_o, it is zero as in a series resonant circuit at a slightly lower frequency where the bracketed term in Eq. (48.14) is zero. When losses are present or there is radiation out of an acoustic port, a resistive term is included in the reactive expression of Eq. (48.14).

Behavior analogous to that of coupled tuned electrical circuits for multipole filters can be achieved by subdividing the electrodes of Fig. 48.4(a) into different areas, each of which will act separately as a tuned circuit, but if they are close enough together there will be acoustic coupling between the different radiators. By controlling this coupling, narrowband filters of very high Q and of somewhat tailored frequency response can be built in the megahertz and low gigahertz range.

The basic geometry of Fig. 48.4(c) gives an electric-to-electric delay line whose delay is given by the length of the medium between the transducers divided by the phase velocity of the acoustic wave and would be on the order of 2 μs/cm. Since the solid has little dispersion, the bandwidth of the delay line is determined by that of the transducers. Here it is necessary to choose the characteristic impedances and thicknesses of the backing and matching layers in Fig. 48.2 in such a manner that the conversion of the electrical energy incident on the electrical port to the acoustic energy out of acoustic port 2 of Fig. 48.3 is independent of frequency over a large range about the resonant frequency of the **piezoelectric transducer** itself. Varying the matching and backing layers is equivalent to varying the terminating impedances on the acoustic line of Fig. 48.3. The matching is often assisted by lumped elements in the external electrical circuit.

The geometry of Fig. 48.4(c) is also the prototype form for acousto-optic interactions. Here the second transducer is not relevant and can be replaced by an acoustic absorber so that there is no reflected wave present in the active region. An optical wave coming into the crystal as shown in Fig. 48.4(c) sees a propagating periodic perturbation of the medium, and if the photoelastic coefficients of the solid are large, the wave sees appreciable variations in the refractive index and hence a moving diffraction grating. The angle of deflection of the output optical beam and its frequency as produced by the grating depend on the amplitude of the various frequency components in the acoustic beam when the optical beam traversed it. Thus, for example, the intensity versus angular position of the emerging optical beam is a measure of the frequency spectrum of any information modulated on the acoustic beam.

As noted previously, ultrasonic waves are often used as probes when the wavelength and attenuation are appropriate. For these radar-like applications, the acoustic beam is generated by a transducer and propagates in the medium containing the scatterer to be investigated as shown in Fig. 48.4(b). The acoustic wave is scattered by any discontinuity in the medium, and energy is returned to the same or to another transducer. If the outgoing signal is pulsed, then the delay for the received pulse is a measure of the distance to the scatterer. If the transducer is displaced or rotated, the change in delay of the echo gives a measure of the shape of the scatterer. Any movement of the scatterer, for example, flowing blood in an artery, causes a Doppler shift of the echo, and this shift, along with the known direction of the returned beam, gives a map of the flow pattern. Phasing techniques with multiple transducers or multiple areas of one transducer can be used to produce focused beams or beams electrically swept in space by differential variation of the phases of the excitation of the component areas of the transducer.

Defining Terms

Characteristic impedance: Ratio of the negative of the stress to the particle velocity in an ultrasonic plane wave.

Form: Term used to indicate the structure and dimensions of a multiterm equation without details within component terms.

Phase velocity: Velocity of propagation of planes of constant phase.

Piezoelectric transducers: Devices that convert electric signals to ultrasonic waves, and vice versa, by means of the piezoelectric effect in solids.

Pure longitudinal and shear waves (modes): Ultrasonic plane waves in which the particle motion is parallel or perpendicular, respectively, to the wave vector and for which energy flow is parallel to the wave vector.

Slowness surface: A plot of the reciprocal of the phase velocity as a function of direction in an anisotropic crystal.

Related Topics

15.2 Speech Enhancement and Noise Reduction • 49.2 Mechanical Characteristics

References

B.A. Auld, *Acoustic Fields and Waves in Solids,* 2nd ed., Melbourne, Fla.: Robert E. Krieger, 1990.

E.A. Gerber and A. Ballato, *Precision Frequency Control*, vol.1, *Acoustic Resonators and Filters,* Orlando, Fla.: Academic Press, 1985.

G.S. Kino, *Acoustic Waves: Devices Imaging and Analog Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

Landolt-Bornstein, *Numerical Data and Functional Relationships in Science and Technology: Gp III Crystal and Solid State Physics*, vol. 11, *Elastic, Piezoelectric, Pyroelectric and Piezooptic Constants of Crystals,* Berlin: Springer-Verlag, 1979.

W.P. Mason and R.N. Thurston (Eds.), *Physical Acoustics, Principles and Methods,* multivolume series, New York: Academic Press.

H.B. Meire, *Basic Ultrasound,* New York: Wiley, 1995.

J.F. Rosenbaum, *Bulk Acoustic Wave Theory and Devices,* Boston: Artech House, 1988.

Further Information

The main conferences in the ultrasonics area are the annual Ultrasonics Symposium sponsored by the IEEE Ultrasonics, Ferroelectrics and Frequency Control Society and the biannual Ultrasonics International Conference organized by the journal *Ultrasonics,* both of which publish proceedings. The periodicals include the *Transactions of the IEEE Ultrasonics, Ferroelectrics and Frequency Control Society,* the journal *Ultrasonics* published by Butterworth & Co., and the *Journal of the Acoustical Society of America.* The books by Kino and by Rosenbaum in the References provide general overviews of the field.

Etzold, K.F. “Ferroelectric and Piezoelectric Materials”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

49

Ferroelectric and Piezoelectric Materials

49.1 Introduction
49.2 Mechanical Characteristics
Applications • Structure of Ferroelectric and Piezoelectric Materials
49.3 Ferroelectric Materials
Electrical Characteristics
49.4 Ferroelectric and High Epsilon Thin Films

K. F. Etzold
IBM T. J. Watson Research Center

49.1 Introduction

Piezoelectric materials have been used extensively in actuator and ultrasonic receiver applications, while **ferroelectric** materials have recently received much attention for their potential use in nonvolatile (NV) memory applications. We will discuss the basic concepts in the use of these materials, highlight their applications, and describe the constraints limiting their uses. This chapter emphasizes properties which need to be understood for the effective use of these materials but are often very difficult to research. Among the properties which are discussed are **hysteresis** and **domains**.

Ferroelectric and piezoelectric materials derive their properties from a combination of structural and electrical properties. As the name implies, both types of materials have electric attributes. A large number of materials which are ferroelectric are also piezoelectric. However, the converse is not true. Pyroelectricity is closely related to ferroelectric and piezoelectric properties via the symmetry properties of the crystals.

Examples of the classes of materials that are technologically important are given in Table 49.1. It is apparent that many materials exhibit electric phenomena which can be attributed to ferroelectric, piezoelectric, and **electret** materials. It is also clear that vastly different materials (organic and inorganic) can exhibit ferroelectricity or piezoelectricity, and many have actually been commercially exploited for these properties.

As shown in Table 49.1, there are two dominant classes of ferroelectric materials, ceramics and organics. Both classes have important applications of their piezoelectric properties. To exploit the ferroelectric property, recently a large effort has been devoted to producing thin films of **PZT** (lead [Pb] zirconate titanate) on various substrates for silicon-based memory chips for nonvolatile storage. In these devices, data is retained in the absence of external power as positive and negative **polarization**. Organic materials have not been used for their ferroelectric properties. Liquid crystals in display applications are used for their ability to rotate the plane of polarization of light and not their ferroelectric attribute.

It should be noted that the prefix *ferro* refers to the permanent nature of the electric polarization in analogy with the magnetization in the magnetic case. It does not imply the presence of iron, even though the root of the word means iron. The root of the word piezo means pressure; hence the original meaning of the word piezoelectric implied "pressure electricity"—the generation of electric field from applied pressure. This definition ignores the fact that these materials are reversible, allowing the generation of mechanical motion by applying a field.

TABLE 49.1 Ferroelectric, Piezoelectric, and Electrostrictive Materials

Type	Material Class	Example	Applications
Electret	Organic	Waxes	No recent
Electret	Organic	Fluorine based	Microphones
Ferroelectric	Organic	PVF2	No known
Ferroelectric	Organic	Liquid crystals	Displays
Ferroelectric	Ceramic	PZT thin film	NV-memory
Piezoelectric	Organic	PVF2	Transducer
Piezoelectric	Ceramic	PZT	Transducer
Piezoelectric	Ceramic	PLZT	Optical
Piezoelectric	Single crystal	Quartz	Freq. control
Piezoelectric	Single crystal	$LiNbO_3$	SAW devices
Electrostrictive	Ceramic	PMN	Actuators

49.2 Mechanical Characteristics

Materials are acted on by forces (stresses) and the resulting deformations are called strains. An example of a strain due to a force to the material is the change of dimension parallel and perpendicular to the applied force. It is useful to introduce the coordinate system and the numbering conventions which are used when discussing these materials. Subscripts 1, 2, and 3 refer to the *x*, *y*, and *z* directions, respectively. Displacements have single indices associated with their direction. If the material has a preferred axis, such as the poling direction in PZT, the axis is designated the *z* or 3 axis. Stresses and strains require double indices such as *xx* or *xy*. To make the notation less cluttered and confusing, contracted notation has been defined. The following mnemonic rule is used to reduce the double index to a single index:

1	6	5
xx	*xy*	*xz*
	2	4
	yy	*yz*
		3
		zz

This rule can be thought of as a matrix with the diagonal elements having repeated indices in the expected order, then continuing the count in a counterclockwise direction. Note that *xy* = *yx*, etc. so that subscript 6 applies equally to *xy* and *yx*.

Any mechanical object is governed by the well-known relationship between stress and strain,

$$\mathbf{S} = \mathbf{sT} \tag{49.1}$$

where **S** is the strain (relative elongation), **T** is the stress (force per unit area), and **s** contains the coefficients connecting the two. All quantities are tensors; **S** and **T** are second rank, and **s** is fourth rank. Note, however, that usually contracted notation is used so that the full complement of subscripts is not visible. PZT converts electrical fields into mechanical displacements and vice versa. The connection between the two is via the *d* and *g* coefficients. The *d* coefficients give the displacement when a field is applied (transmitter), while the *g* coefficients give the field across the device when a stress is applied (receiver). The electrical effects are added to the basic Eq. (49.1) such that

$$\mathbf{S} = \mathbf{sT} + \mathbf{dE} \tag{49.2}$$

where **E** is the electric field and **d** is the tensor which contains the coupling coefficients. The latter parameters are reported in Table 49.2 for representative materials. One can write the matrix equation [Eq. (49.2)],

TABLE 49.2 Properties of Well-Known PZT Formulations (Based on the Original Navy Designations and Now Used by Commercial Vendor Vernitron)

	Units	PZT4	PZT5A	PZT5H	PZT8
ε_{33}	—	1300	1700	3400	1000
d_{33}	10^{-2} Å/V	289	374	593	225
d_{13}	10^{-2} Å/V	–123	–171	–274	–97
d_{15}	10^{-2} Å/V	496	584	741	330
g_{33}	10^{-3} Vm/N	26.1	24.8	19.7	25.4
k_{33}	—	70	0.705	0.752	0.64
T_Θ	°C	328	365	193	300
Q	—	500	75	65	1000
ρ	g/cm^3	7.5	7.75	7.5	7.6
Application	—	High signal	Medium signal	Receiver	Highest signal

$$\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & & & \\ s_{12} & s_{11} & s_{13} & & 0 & \\ s_{13} & s_{13} & s_{33} & & & \\ & & & s_{44} & & \\ & 0 & & & s_{44} & \\ & & & & & 2(s_{11} - s_{12}) \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{bmatrix} + \begin{bmatrix} 0 & 0 & d_{13} \\ 0 & 0 & d_{13} \\ 0 & 0 & d_{33} \\ 0 & d_{15} & 0 \\ d_{15} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} \tag{49.3}$$

Note that **T** and **E** are shown as column vectors for typographical reasons; they are in fact row vectors. This equation shows explicitly the stress-strain relation and the effect of the electromechanical conversion.

A similar equation applies when the material is used as a receiver:

$$\mathbf{E} = -\mathbf{gT} + (\varepsilon^{\mathrm{T}})^{-1}\mathbf{D} \tag{49.4}$$

where T is the transpose and **D** the electric displacement. For all materials the matrices are not fully populated. Whether a coefficient is nonzero depends on the symmetry. For PZT, a ceramic which is given a preferred direction by the poling operation (the z-axis), only d_{33}, d_{13}, and d_{15} are nonzero. Also, again by symmetry, $d_{13} = d_{23}$ and $d_{15} = d_{25}$.

Applications

Historically the material which was used earliest for its piezoelectric properties was single-crystal quartz. Crude sonar devices were built by Langevin using quartz transducers, but the most important application was, and still is, frequency control. Crystal oscillators are today at the heart of every clock that does not derive its frequency reference from the ac power line. They are also used in every color television set and personal computer. In these applications at least one (or more) "quartz crystal" controls frequency or time. This explains the label "quartz" which appears on many clocks and watches. The use of quartz resonators for frequency control relies on another unique property. Not only is the material piezoelectric (which allows one to excite mechanical vibrations), but the material has also a very high mechanical "Q" or quality factor (Q >100,000). The actual value depends on the mounting details, whether the crystal is in a vacuum, and other details. Compare this value to a Q for PZT between 75 and 1000. The Q factor is a measure of the rate of decay and thus the mechanical losses of an excitation with no external drive. A high Q leads to a very sharp resonance and thus tight frequency control. For frequency control it has been possible to find orientations of cuts of quartz which reduce the influence of temperature on the vibration frequency.

Ceramic materials of the PZT family have also found increasingly important applications. The piezoelectric but not the ferroelectric property of these materials is made use of in transducer applications. PZT has a very high efficiency (electric energy to mechanical energy coupling factor k) and can generate high-amplitude ultrasonic waves in water or solids. The coupling factor is defined by

$$k^2 = \frac{\text{energy stored mechanically}}{\text{total energy stored electrically}} \tag{49.5}$$

Typical values of k_{33} are 0.7 for PZT 4 and 0.09 for quartz, showing that PZT is a much more efficient transducer material than quartz. Note that the energy is a scalar; the subscripts are assigned by finding the energy conversion coefficient for a specific vibrational mode and field direction and selecting the subscripts accordingly. Thus k_{33} refers to the coupling factor for a longitudinal mode driven by a longitudinal field.

Probably the most important applications of PZT today are based on ultrasonic echo ranging. Sonar uses the conversion of electrical signals to mechanical displacement as well as the reverse transducer property, which is to generate electrical signals in response to a stress wave. Medical diagnostic ultrasound and nondestructive testing systems devices rely on the same properties. Actuators have also been built but a major obstacle is the small displacement which can conveniently be generated. Even then, the required voltages are typically hundreds of volts and the displacements are only a few hundred angstroms. For PZT the strain in the z-direction due to an applied field in the z-direction is (no stress, $\mathbf{T} = 0$)

$$s_3 = d_{33}E_3 \tag{49.6}$$

or

$$s_3 = \frac{\Delta d}{d} = d_{33}\frac{V}{d} \tag{49.7}$$

where s is the strain, E the electric field, and V the potential; d_{33} is the coupling coefficient which connects the two. Thus

$$\Delta d = d_{33}V \tag{49.8}$$

Note that this expression is independent of the thickness d of the material but this is true only when the applied field is parallel to the displacement. Let the applied voltage be 100 V and let us use PZT8 for which d_{33} is 225 (from Table 49.2). Hence $\Delta d = 225$ Å or 2.25 Å/V, a small displacement indeed. We also note that Eq. (49.6) is a special case of Eq. (49.2) with the stress equal to zero. This is the situation when an actuator is used in a force-free environment, for example, as a mirror driver. This arrangement results in the maximum displacement. Any forces which tend to oppose the free motion of the PZT will subtract from the available displacement with the reduction given by the normal stress-strain relation, Eq. (49.1).

It is possible to obtain larger displacements with mechanisms which exhibit mechanical gain, such as laminated strips (similar to bimetallic strips). The motion then is typically up to about 1 millimeter but at a cost of a reduced available force. An example of such an application is the video head translating device to provide tracking in VCRs.

There is another class of ceramic materials which recently has become important. **PMN** (lead [Pb], magnesium niobate), typically doped with ≈10% lead titanate) is an **electrostrictive** material which has seen applications where the absence of hysteresis is important. For example, deformable mirrors require repositioning of the reflecting surface to a defined location regardless of whether the old position was above or below the original position.

Electrostrictive materials exhibit a strain which is quadratic as a function of the applied field. Producing a displacement requires an internal polarization. Because the latter polarization is induced by the applied field

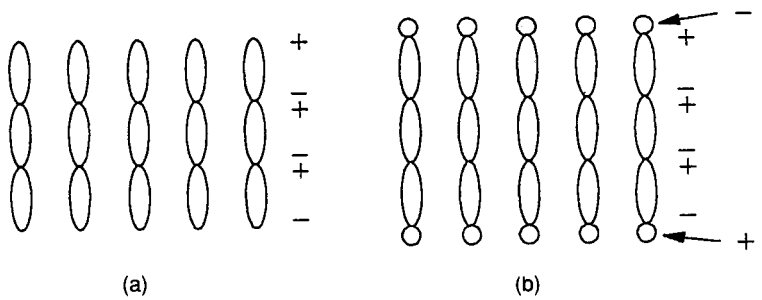


FIGURE 49.1 Charge configurations in ferroelectric model materials: (a) uncompensated and (b) compensated dipole arrays.

and is not permanent, as it is in the ferroelectric materials, electrostrictive materials have essentially no hysteresis. Unlike PZT, electrostrictive materials are not reversible; PZT will change shape on application of a field and generate a field when a strain is induced. Electrostrictive materials only change shape on application of a field and, therefore, cannot be used as receivers. PZT has inherently large hysteresis because of the domain nature of the polarization.

Organic electrets have important applications in self-polarized condenser (or capacitor) microphones where the required electric bias field in the gap is generated by the diaphragm material rather than by an external power supply.

Structure of Ferroelectric and Piezoelectric Materials

Ferroelectric materials have, as their basic building block, atomic groups which have an associated electric field, either as a result of their structure or as result of distortion of the charge clouds which make up the groups. In the first case, the field arises from an asymmetric placement of the individual ions in the group (these groupings are called unit cells). In the second case, the electronic cloud is moved with respect to the ionic core. If the group is distorted permanently, then a permanent electric field can be associated with each group. We can think of these distorted groups as represented by electric dipoles, defined as two equal but opposite charges which are separated by a small distance. Electric dipoles are similar to magnetic dipoles which have the familiar north and south poles. The external manifestation of a magnetic dipole is a magnetic field and that of an electric dipole an electric field.

Figure 49.1(a) represents a hypothetical slab of material in which the dipoles are perfectly arranged. In actual materials the atoms are not as uniformly arranged, but, nevertheless, from this model there would be a very strong field emanating from the surface of the crystal. The common observation, however, is that the fields are either absent or weak. This effective charge neutrality arises from the fact that there are free, mobile charges available which can be attracted to the surfaces. The polarity of the mobile charges is opposite to the charge of the free dipole end. The added charges on the two surfaces generate their own field, equal and opposite to the field due to the internal dipoles. Thus the effect of the internal field is canceled and the external field is zero, as if no charges were present at all [Fig. 49.1(b)].

In ferroelectric materials a crystalline asymmetry exists which allows electric dipoles to form. In their absence the dipoles are absent and the internal field disappears. Consider an imaginary horizontal line drawn through the middle of a dipole. We can see readily that the dipole is not symmetric about that line. The asymmetry thus requires that there be no center of inversion when the material is in the ferroelectric state.

All ferroelectric and piezoelectric materials have phase transitions at which the material changes crystalline symmetry. For example, in PZT there is a change from tetragonal or rhombohedral symmetry to cubic as the temperature is increased. The temperature at which the material changes **crystalline phases** is called the **Curie temperature**, T_Θ. For typical PZT compositions the Curie temperature is between 250 and 450°C.

A consequence of a phase transition is that a rearrangement of the lattice takes place when the material is cooled through the transition. Intuitively we would expect that the entire crystal assumes the same orientation throughout as we pass through the transition. By orientation we mean the direction of the preferred axis (say

the tetragonal axis). Experimentally it is found, however, that the material breaks up into smaller regions in which the preferred direction and thus the polarization is uniform. Note that cubic materials have no preferred direction. In tetragonal crystals the polarization points along the *c*-axis (the longer axis) whereas in rhombohedral lattices the polarization is along the body diagonal. The volume in which the preferred axis is pointing in the same direction is called a domain and the border between the regions is called a domain wall. The energy of the multidomain state is slightly lower than the single-domain state and is thus the preferred configuration. The direction of the polarization changes by either 90° or 180° as we pass from one uniform region to another. Thus the domains are called 90° and 180° domains. Whether an individual crystallite or grain consists of a single domain depends on the size of the crystallite and external parameters such as strain gradients, impurities, etc. It is also possible that the domain extend beyond the grain boundary and encompasses two or more grains of the crystal.

Real materials consist of large numbers of unit cells, and the manifestation of the individual charged groups is an internal and an external electric field when the material is stressed. Internal and external refer to inside and outside of the material. The interaction of an external electric field with a charged group causes a displacement of certain atoms in the group. The macroscopic manifestation of this is a displacement of the surfaces of the material. This motion is called the piezoelectric effect, the conversion of an applied field into a corresponding displacement.

49.3 Ferroelectric Materials

PZT ($PbZr_xTi_{(1-x)}O_3$) is an example of a ceramic material which is ferroelectric. We will use PZT as a prototype system for many of the ferroelectric attributes to be discussed. The concepts, of course, have general validity. The structure of this material is ABO_3 where A is lead and B is one or the other atoms, Ti or Zr. This material consists of many randomly oriented crystallites which vary in size between approximately 10 nm and several microns. The crystalline symmetry of the material is determined by the magnitude of the parameter x. The material changes from rhombohedral to tetragonal symmetry when $x > 0.48$. This transition is almost independent of temperature. The line which divides the two phases is called a **morphotropic phase boundary** (change of symmetry as a function of composition only). Commercial materials are made with $x \approx 0.48$, where the d and g sensitivity of the material is maximum. It is clear from Table 49.2 that there are other parameters which can be influenced as well. Doping the material with donors or acceptors often changes the properties dramatically. Thus niobium is important to obtain higher sensitivity and resistivity and to lower the Curie temperature. PZT typically is a *p*-type conductor and niobium will significantly decrease the conductivity because of the electron which Nb^{5+} contributes to the lattice. The Nb ion substitutes for the **B-site** ion Ti^{4+} or Zr^{4+}. The resistance to depolarization (the hardness of the material) is affected by iron doping. Hardness is a definition giving the relative resistance to depolarization. It should not be confused with mechanical hardness. Many other dopants and admixtures have been used, often in very exotic combinations to affect aging, sensitivity, etc.

The designations used in Table 49.2 reflect very few of the many combinations which have been developed. The PZT designation types were originally designed by the U.S. Navy to reflect certain property combinations. These can be obtained with different combinations of compositions and dopants. The examples given in the table are representative of typical PZT materials, but today essentially all applications have their own custom formulation. The name PZT has become generic for the lead zirconate titanates and does not reflect Navy or proprietary designations.

When PZT ceramic material is prepared, the crystallites and domains are randomly oriented, and therefore the material does not exhibit any piezoelectric behavior [Fig. 49.2(a)]. The random nature of the displacements for the individual crystallites causes the net displacement to average to zero when an external field is applied. The tetragonal axis has three equivalent directions 90° apart and the material can be poled by reorienting the polarization of the domains into a direction nearest the applied field. When a sufficiently high field is applied, some but not all of the domains will be rotated toward the electric field through the allowed angle 90° or 180°. If the field is raised further, eventually all domains will be oriented as close as possible to the direction of the field. Note however, that the polarization will not point exactly in the direction of the field [Fig. 49.2(b)]. At this point, no further domain motion is possible and the material is saturated. As the field is reduced, the majority of domains retain the orientation they had with the field on leaving the material in an oriented state which now has a net polarization. Poling is accomplished for commercial PZT by raising the temperature to

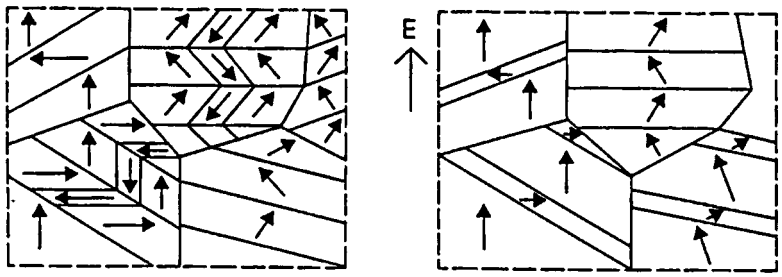


FIGURE 49.2 Domains in PZT, as prepared (a) and poled (b).

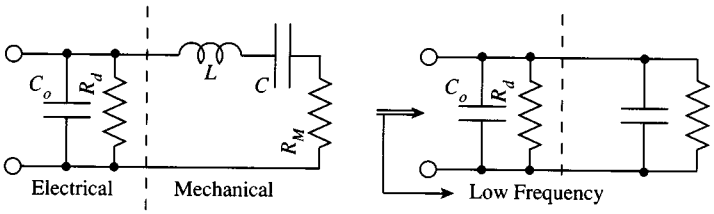


FIGURE 49.3 Equivalent circuit for a piezoelectric resonator. The reduction of the equivalent circuit at low frequencies is shown on the right.

about 150°C (to lower the **coercive field**, E_c) and applying a field of about 30–60 kV/cm for several minutes. The temperature is then lowered but it is not necessary to keep the field on during cooling because the domains will not spontaneously rerandomize.

Electrical Characteristics

Before considering the dielectric properties, we will consider the equivalent circuit for a slab of ferroelectric material. In Fig. 49.3 the circuit shows a mechanical (acoustic) component and the static or clamped capacity C_o (and the dielectric loss R_d) which are connected in parallel. The acoustic components are due to their motional or mechanical equivalents, the compliance (capacity, C) and the mass (inductance, L). There will be mechanical losses, which are indicated in the mechanical branch by R. The electrical branch has the clamped capacity C_o and a dielectric loss (R_d), distinct from the mechanical losses. This configuration will have a resonance which is usually assumed to correspond to the mechanical thickness mode but can represent other modes as well. This simple model does not show the many other modes a slab (or rod) of material will have. Thus transverse, plate, and flexural modes are present. Each can be represented by its own combination of L, C, and R. The presence of a large number of modes often causes difficulties in characterizing the material since some parameters must be measured either away from the resonances or from clean, nonoverlapping resonances. For instance, the clamped capacity (or clamped dielectric constant) of a material is measured at high frequencies where there are usually a large number of modes present. For an accurate measurement these must be avoided and often a low-frequency measurement is made in which the material is physically clamped to prevent motion. This yields the static, nonmechanical capacity, C_o. The circuit can be approximated at low frequencies by ignoring the inductor and redefining R and C. Thus, the coupling constant can be extracted from the value of C and C_o. From the previous definition of k we find

$$k^2 = \frac{\text{energy stored mechanically}}{\text{total energy stored electrically}} = \frac{CV^2/2}{(C + C_o)V^2/2} = \frac{1}{\dfrac{C_o}{C}} + 1 \qquad (49.9)$$

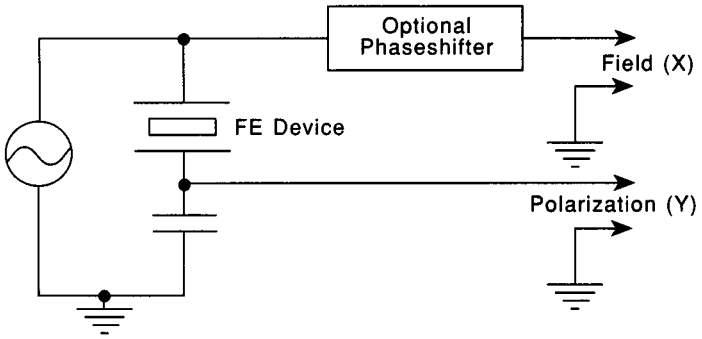


FIGURE 49.4 Sawyer Tower circuit.

It requires charge to rotate or flip a domain. Thus, there is charge flow associated with the rearrangement of the polarization in the ferroelectric material. If a bipolar, repetitive signal is applied to a ferroelectric material, its hysteresis loop is traced out and the charge in the circuit can be measured using the Sawyer Tower circuit (Fig. 49.4). In some cases the drive signal to the material is not repetitive and only a single cycle is used. In that case the starting point and the end point do not have the same polarization value and the hysteresis curve will not close on itself.

The charge flow through the sample is due to the rearrangement of the polarization vectors in the domains (the polarization) and contributions from the static capacity and losses (C_o and R_d in Fig. 49.3). The charge is integrated by the measuring capacitor which is in series with the sample. The measuring capacitor is sufficiently large to avoid a significant voltage loss. The polarization is plotted on a X-Y scope or plotter against the applied voltage and therefore the applied field.

Ferroelectric and piezoelectric materials are lossy. This will distort the shape of the hysteresis loop and can even lead to incorrect identification of materials as ferroelectric when they merely have nonlinear conduction characteristics. A resistive component (from R_d in Fig. 49.3) will introduce a phase shift in the polarization signal. Thus the display has an elliptical component, which looks like the beginnings of the opening of a hysteresis loop. However, if the horizontal signal has the same phase shift, the influence of this lossy component is eliminated, because it is in effect subtracted. Obtaining the exact match is the function of the optional phase shifter, and in the original circuits a bridge was constructed which had a second measuring capacitor in the comparison arm (identical to the one in series with the sample). The phase was then matched with adjustable high-voltage components which match C_o and R_d.

This design is inconvenient to implement and modern Sawyer Tower circuits have the capability to shift the reference phase either electronically or digitally to compensate for the loss and static components. A contemporary version, which has compensation and no voltage loss across the integrating capacitor, is shown in Fig. 49.5. The op-amp integrator provides a virtual ground at the input, reducing the voltage loss to negligible values. The output from this circuit is the sum of the polarization and the capacitive and loss components. These contributions can be canceled using a purely real (resistive) and a purely imaginary (capacitive, 90° phaseshift) compensation component proportional to the drive across the sample. Both need to be scaled (magnitude adjustments) to match them to the device being measured and then have to be subtracted (adding negatively) from the output of the op amp. The remainder is the polarization. The hysteresis for typical ferroelectrics is frequency dependent and traditionally the reported values of the polarization are measured at 50 or 60 Hz.

The improved version of the Sawyer Tower (Fig. 49.6) circuit allows us to cancel C_o and R_d and the losses, thus determining the active component. This is important in the development of materials for ferroelectric memory applications. It is far easier to judge the squareness of the loop when the inactive components are canceled. Also, by calibrating the "magnitude controls" the value of the inactive components can be read off directly. In typical measurements the resonance is far above the frequencies used, so ignoring the inductance in the equivalent circuit is justified.

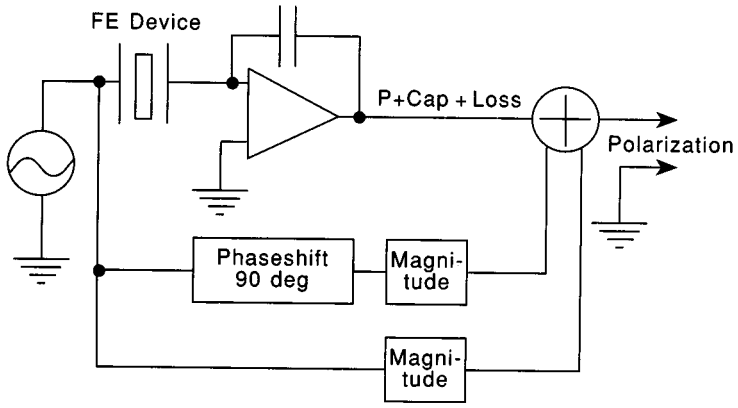


FIGURE 49.5 Modern hysteresis circuit. An op amp is used to integrate the charge; loss and static capacitance compensation are included.

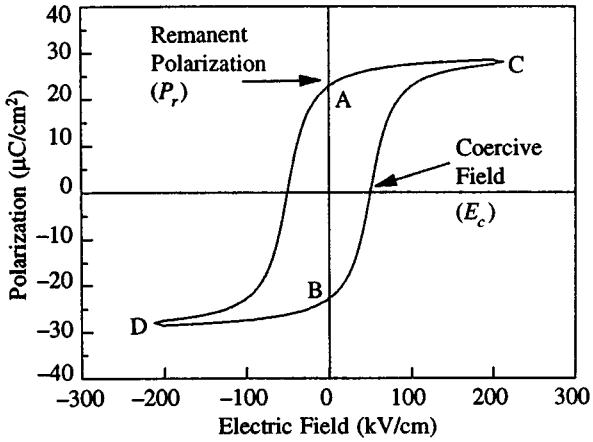


FIGURE 49.6 Idealized hysteresis curve for typical PZT materials. Many PZT materials display offsets from the origin and have asymmetries with respect to the origin. The curve shows how the remanent polarization ($\bar{P}_r$) and the coercive field ($\bar{E}_c$) are defined. While the loop is idealized, the values given for the polarization and field are realistic for typical PZT materials.

The measurement of the dielectric constant and the losses is usually very straightforward. A slab with a circular or other well-defined cross section is prepared, electrodes are applied, and the capacity and loss are measured (usually as a function of frequency). The dielectric constant is found from

$$C = \varepsilon_o \varepsilon \frac{A}{t} \tag{49.10}$$

where A is the area of the device and t the thickness. In this definition (also used in Table 49.2) ε is the relative dielectric constant and ε_o is the permittivity of vacuum. Until recently, the dielectric constant, like the polarization, was measured at 50 or 60 Hz (typical powerline frequencies). Today the dielectric parameters are typically specified at 1 kHz, which is possible because impedance analyzers with high-frequency capability are readily available. To avoid low-frequency anomalies, even higher frequencies such as 1 MHz are often selected. This is especially the case when evaluating PZT thin films. Low frequency anomalies are not included in the equivalent circuit (Fig. 49.3) and are due to interface layers. These layers will cause both the resistive and reactive components to rise at low frequencies producing readings which are not representative of the dielectric properties.

TABLE 49.3 Material Properties and Applications Areas

	Ferroelectric	Epsilon	Polarization	Coercive Field	Leakage	Aging	Electro-Optical	Electro-Mechanical
NV RAM	X		X	X	X	X		
DRAM		X			X	X		
Actuator				X				X
Display	X				X	X	X	
Optical Modulator	X				X	X	X	

A piezoelectric component often has a very simple geometric shape, especially when it is prepared for measurement purposes. There will be mechanical resonances associated with the major dimensions of a sample piece. The resonance spectrum will be more or less complicated, depending on the shape of a sample piece. If the object has a simple shape, then some of the resonances will be well separated from each other and can be associated with specific vibrations and dimensions (modes). Each of these resonances has an electrical equivalent, and inspection of the equivalent circuit shows that there will be a resonance (minimum impedance) and an antiresonance (maximum impedance). Thus an impedance plot can be used to determine the frequencies and also the coupling constants and mechanical parameters for the various modes.

49.4 Ferroelectric and High Epsilon Thin Films

While PZT and other ferroelectric (FE) bulk materials have had major commercial importance, thin films prepared from these materials have only recently been the focus of significant research efforts. In this section the material properties and process issues will be discussed. Because of the potentially large payoff, major efforts have been directed at developing the technologies for depositing thin films of ferroelectric and non-ferroelectric but high epsilon (high dielectric constant) thin films.

A recent trend has been the ever increasing density of dynamic random access memory (DRAM). The storage capacitor in these devices is becoming a major limiting factor because the dielectric has to be very thin in order to achieve the desired capacitance values to yield, in turn, a sufficient signal for the storage cell. It is often also desirable to have nonvolatile operation (no data loss on power loss). These two desires have, probably more than anything else, driven the development of high epsilon and FE thin films. Of course, these are not the only applications of FE films. Table 49.3 lists the applications of FE (nonvolatile, NV) and high epsilon films (volatile) and highlights which of the properties are important for their use. It is seen that the memory application is very demanding. Satisfying all these requirements simultaneously has produced significant challenges in the manufacture of these films.

Perhaps the least understood and to some extent still unsolved problem is that of fatigue. In nonvolatile memory applications the polarization represents the memory state of the material (up ≡ bit 1; down ≡ bit 0). In use the device can be switched at the clock rate, say 100 MHz. Thus for a lifetime of 5 years the material must withstand $\simeq 10^{16}$ polarization reversals or large field reversals. Typical materials for ferroelectric applications are PZTs with the ratio of zirconium to titanium adjusted to yield the maximum dielectric constant and polarization. This maximum will be near the morphotropic phase boundary for PZT. Small quantities of other materials can be added, such as lanthanum or niobium to modify optical or switching characteristics. The Sol-Gel method discussed below is particularly suitable for incorporating these additives. Devices made from materials at the current state of the art loose a significant portion of their polarization after 10^{10} to 10^{12} cycles, rendering them useless for their intended memory use because of the associated signal loss. This is a topic of intensive investigation and only one proprietary material has emerged which might be suitable for memory use (Symetric Corporation). High epsilon nonferroelectric materials are of great interest for DRAM applications. As an example, major efforts are extant to produce thin films of mixtures of barium and strontium titanate (BST). Dielectric constants of 600 and above have been achieved (compared to 4–6 for silicon oxides and nitrides).

In applications for FE films, significant opportunities also exist for electro-optical modulators for fiber-optic devices and light valves for displays. Another large scale application is actuators and sensors. For the latter the

TABLE 49.4 Deposition Methods for PZT and Perovskites

	Process Type	Rate nm/min	Substrate Temperature	Anneal Temperature	Target/Source
Wet	Sol-Gel	100 nm/coat	RT	450–750	Metal organic
Wet	MOD	300 nm/coat	RT	500–750	Metal organic
Dry	RF sputter	.5–5	RT–700	500–700	Metals and oxides
Dry	Magnetron sputter	5–30	RT–700	500–700	Metals and oxides
Dry	Ion beam sputter	2–10	RT–700	500–700	Metals and oxides
Dry	Laser sputter	5–100	RT–700	500–700	Oxide
Dry	MOCVD	5–100	400–800	500–700	MO vapor and carrier gas

electro-mechanical conversion property is used and large values of d_{33} (the conversion coefficient) are desirable. However, economically the importance of all other applications are, and probably will be in the foreseeable future, less significant than that of memory devices.

Integration of ferroelectric or nonferroelectric materials with silicon devices and substrates has proved to be very challenging. Contacts and control of the crystallinity and crystal size and the stack structure of the capacitor device are the principal issues. In both volatile and nonvolatile memory cells the dielectric material tends to interact with the silicon substrate. Thus an appropriate barrier layer must be incorporated while at the same time obtaining a suitable substrate on which to grow the dielectric films. A typical device structure starts with an oxide layer (SiO_x) on the silicon substrate followed by a thin titanium layer which prevents diffusion of the final substrate layer, platinum (the actual growth substrate).

Significant differences have been observed in the quality of the films depending on the nature of the substrate. The quality can be described by intrinsic parameters such as the crystallinity (i.e., the degree to which non-crystalline phases are present). The uniformity of the orientation of the crystallites also seems to play a role in determining the electrical properties of the films. In the extreme case of perfect alignment of the crystallites of the film with the substrate and the formation of large single crystal areas, an epitaxial film is obtained. These films tend to have the best electrical properties. In addition to amorphous material, other crystalline but nonferroelectric phases can be present. An example is the pyrochlore phase in PZT. These phases often form incidentally to the growth process of the desired film and usually degrade one or more of the desired properties of the film (for instance the dielectric constant). The pyrochlore and other oxide materials can accumulate between the Pt electrode and the desired PZT or BST layer. The interface layer is then electrically in series with the desired dielectric layer and degrades its properties. The apparent reduction of the dielectric constant which is often observed in these films as the thickness is reduced can be attributed to the presence of these low dielectric constant layers.

There are many growth methods for these films. Table 49.4 lists the most important techniques along with some of the critical parameters. Wet methods use metal organic compounds in liquid form. In the Sol-Gel process the liquid is spun onto the substrate. The wafer is then heated, typically to a lower, intermediate temperature (around 300°C). This spin-on and heat process is repeated until the desired thickness is reached. At this temperature only an amorphous film forms. The wafer is then heated to between 500 and 700°C usually in oxygen and the actual crystal growth takes place. Instead of simple long term heating (order of hours), rapid thermal annealing (RTA) is often used. In this process the sample is only briefly exposed to the elevated temperature, usually by a scanning infrared beam. It is in the transition between the low decomposition temperature and the firing temperature that the pyrochlore tends to form. At the higher temperatures the more volatile components have a tendency to evaporate, thus producing a chemically unbalanced compound which also has a great propensity to form one or more of the pyrochlore phases. In the case of PZT, 5 to 10% excess lead is usually incorporated which helps to form the desired perovskite material and compensates for the loss. In preparing Sol-Gel films it is generally easy to prepare the compatible liquid compounds of the major constituents and the dopants. The composition is then readily adjusted by appropriately changing the ratio of the constituents. Very fine quality films have been prepared by this method, including epitaxial films.

The current semiconductor technology is tending toward dry processing. Thus, in spite of the advantages of the Sol-Gel method, other methods using physical vapor deposition (PVD) are being investigated. These methods use energetic beams or plasma to move the constituent materials from the target to the heated substrate.

The compound then forms *in situ* on the heated wafer (≃500°C). Even then, however, a subsequent anneal is often required. With PVD methods it is much more difficult to change the composition since now the oxide or metal ratios of the target have to be changed or dopants have to be added. This involves the fabrication of a new target for each composition ratio. MOCVD is an exception here; the ratio is adjusted by regulating the carrier gas flow. However, the equipment is very expensive and the substrate temperatures tend to be high (up to 800°, uncomfortably high for semiconductor device processing). The laser sputtering method is very attractive and it has produced very fine films. The disadvantage is that the films are defined by the plume which forms when the laser beam is directed at the source. This produces only small areas of good films and scanning methods need to be developed to cover full size silicon wafers. Debris is also a significant issue in laser deposition. However, it is a convenient method to produce films quickly and with a small investment. In the long run MOCVD or Sol-Gel will probably evolve as the method of choice for realistic DRAM devices with state of the art densities.

Defining Terms

A-site: Many ferroelectric materials are oxides with a chemical formula ABO_3. The A-site is the crystalline location of the A atom.

B-site: Analogous to the definition of the A-site.

Coercive field: When a ferroelectric material is cycled through the hysteresis loop the coercive field is the electric field value at which the polarization is zero. A material has a negative and a positive coercive field and these are usually, but not always, equal in magnitude to each other.

Crystalline phase: In crystalline materials the constituent atoms are arranged in regular geometric ways; for instance in the cubic phase the atoms occupy the corners of a cube (edge dimensions ≈2–15 Å for typical oxides).

Curie temperature: The temperature at which a material spontaneously changes its crystalline phase or symmetry. Ferroelectric materials are often cubic above the Curie temperature and tetragonal or rhombohedral below.

Domain: Domains are portions of a material in which the polarization is uniform in magnitude and direction. A domain can be smaller, larger, or equal in size to a crystalline grain.

Electret: A material which is similar to ferroelectrics but charges are macroscopically separated and thus are not structural. In some cases the net charge in the electrets is not zero, for instance when an implantation process was used to embed the charge.

Electrostriction: The change in size of a nonpolarized, dielectric material when it is placed in an electric field.

Ferroelectric: A material with permanent charge dipoles which arise from asymmetries in the crystal structure. The electric field due to these dipoles can be observed external to the material when certain conditions are satisfied (ordered material and no charge on the surfaces).

Hysteresis: When the electric field is raised across a ferroelectric material the polarization lags behind. When the field is cycled across the material the hysteresis loop is traced out by the polarization.

Morphotropic phase boundary (MPB): Materials which have a MPB assume a different crystalline phase depending on the composition of the material. The MPB is sharp (a few percent in composition) and separates the phases of a material. It is approximately independent of temperature in PZT.

Piezoelectric: A material which exhibits an external electric field when a stress is applied to the material and a charge flow proportional to the strain is observed when a closed circuit is attached to electrodes on the surface of the material.

PLZT: A PZT material with a lanthanum doping or admixture (up to approximately 15% concentration). The lanthanum occupies the A-site.

PMN: Generic name for electrostrictive materials of the lead (Pb) magnesium niobate family.

Polarization: The polarization is the amount of charge associated with the dipolar or free charge in a ferroelectric or an electret, respectively. For dipoles the direction of the polarization is the direction of the dipole. The polarization is equal to the external charge which must be supplied to the material to produce a polarized state from a random state (twice that amount is necessary to reverse the polarization). The statement is rigorously true if all movable charges in the material are reoriented (i.e., saturation can be achieved).

PVF2: An organic polymer which can be ferroelectric. The name is an abbreviation for polyvinyledene difluoride.

PZT: Generic name for piezoelectric materials of the lead (Pb) zirconate titanate family.

Remanent polarization: The residual or remanent polarization of a material after an applied field is reduced to zero. If the material was saturated, the remanent value is usually referred to as the polarization, although even at smaller fields a (smaller) polarization remains.

Related Topics

47.2 SAW Material Properties • 48.3 Piezoelectric Excitation • 58.4 Material Properties Conducive for Smart Material Applications

References

J. C. Burfoot and G. W. Taylor, *Polar Dielectrics and Their Applications*, Berkeley: University of California Press, 1979.

H. Diamant, K. Drenck, and R. Pepinsky, *Rev. Sci. Instrum.*, vol. 28, p. 30, 1957.

T. Hueter and R. Bolt, *Sonics*, New York: John Wiley and Sons, 1954.

B. Jaffe, W. Cook, and H. Jaffe, *Piezoelectric Ceramics*, London: Academic Press, 1971.

M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectric Materials*, Oxford: Clarendon Press, 1977.

R. A. Roy and K. F. Etzold, "Ferroelectric film synthesis, past and present: a select review," *Mater. Res. Soc. Symp. Proc.,* vol. 200, p. 141, 1990.

C. B. Sawyer and C. H. Tower, *Phys. Rev.*, vol. 35, p. 269, 1930.

Z. Surowiak, J. Brodacki, and H. Zajosz, *Rev. Sci. Instrum.*, vol. 49, p. 1351, 1978.

Further Information

IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control (UFFC).

IEEE Proceedings of International Symposium on the Application of Ferroelectrics (ISAF) (these symposia are held at irregular intervals).

Materials Research Society, Symposium Proceedings, vols. 191, 200, and 243 (this society holds symposia on ferroelectric materials at irregular intervals).

K.-H. Hellwege, Ed., *Landolt-Bornstein: Numerical Data and Functional Relationships in Science and Technology,* New Series, Gruppe III, vols. 11 and 18, Berlin: Springer-Verlag, 1979 and 1984 (these volumes have elastic and other data on piezoelectric materials).

American Institute of Physics Handbook, 3rd ed., New York: McGraw-Hill, 1972.

Sundar, V., Newnham, R.E. “Electrostriction”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

50

Electrostriction

V. Sundar and
R.E. Newnham
Intercollege Materials Research Laboratory, The Pennsylvania State University

50.1 Introduction
50.2 Defining Equations
Piezoelectricity and Electrostriction • Electrostriction and Compliance Matrices • Magnitudes and Signs of Electrostrictive Coefficients
50.3 PMN–PT — A Prototype Electrostrictive Material
50.4 Applications
50.5 Summary

50.1 Introduction

Electrostriction is the basic electromechanical coupling mechanism in centric crystals and amorphous solids. It has been recognized as the primary electromechanical coupling in centric materials since early in the 20th century [Cady, 1929]. Electrostriction is the quadratic coupling between the strain developed in a material and the electric field applied, and it exists in all insulating materials. **Piezoelectricity** is a better-known linear coupling mechanism that exists only in materials without a center of symmetry.

Electrostriction is a second-order property that is tunable and nonlinear. Electrostrictive materials exhibit a reproducible, nonhysteretic, and tunable strain response to electric fields, which gives them an advantage over piezoelectrics in micropositioning applications. While most electrostrictive actuator materials are **perovskite** ceramics, there has been much interest in large electrostriction effects in such polymer materials as polyvinylidene fluoride (PVDF) copolymers recently.

This chapter discusses the three electrostrictive effects and their applications. A discussion of the sizes of these effects and typical electrostrictive coefficients is followed by an examination of lead magnesium niobate (PMN) as a prototype electrostrictive material. The electromechanical properties of some common electrostrictive materials are also compared. A few common criteria used to select relaxor ferroelectrics for electrostrictive applications are also outlined.

50.2 Defining Equations

Electrostriction is defined as the quadratic coupling between strain (**x**) and electric field (**E**), or between strain and polarization (**P**). It is a fourth-rank tensor defined by the following relationship:

$$x_{ij} = M_{ijmn} E_m E_n \tag{50.1}$$

where x_{ij} is the strain tensor, E_m and E_n components of the electric field vector, and M_{ijmn} the fourth-rank field-related electrostriction tensor. The M coefficients are defined in units of m^2/V^2.

Ferroelectrics and related materials often exhibit nonlinear dielectric properties with changing electric fields. To better express the quadratic nature of electrostriction, it is useful to define a polarization-related electrostriction coefficient Q_{ijmn}, as

$$x_{ij} = Q_{ijmn} P_m P_n \tag{50.2}$$

Q coefficients are defined in units of m^4/C^2. The M and Q coefficients are equivalent. Conversions between the two coefficients are carried out using the field-polarization relationships:

$$P_m = \eta_{mn} E_n, \quad \text{and} \quad E_n = \chi_{mn} P_m \tag{50.3}$$

where η_{mn} is the dielectric susceptibility tensor and χ_{mn} is the inverse dielectric susceptibility tensor.

Electrostriction is not a simple phenomenon but manifests itself as three thermodynamically related effects [Sundar and Newnham, 1992]. The first is the well-known variation of strain with polarization, called the direct effect ($d^2x_{ij}/dE_k\, dE_l = M_{ijkl}$). The second is the stress (X_{kl}) dependence of the dielectric stiffness χ_{mn}, or the reciprocal dielectric susceptibility, called the first converse effect ($d\chi_{mn}/dX_{kl} = M_{mnkl}$). The third effect is the polarization dependence of the piezoelectric voltage coefficient g_{jkl}, called the second converse effect ($dg_{jkl}/dP_i = \chi_{mk}\chi_{nl}M_{ijmn}$).

Piezoelectricity and Electrostriction

Piezoelectricity is a third-rank tensor property found only in acentric materials and is absent in most materials. The noncentrosymmetric point groups generally exhibit piezoelectric effects that are larger than the electrostrictive effects and obscure them. The electrostriction coefficients M_{ijkl} or Q_{ijkl} constitute fourth-rank tensors which, like the elastic constants, are found in all insulating materials, regardless of symmetry.

Electrostriction is the origin of piezoelectricity in ferroelectric materials, in both conventional ceramic ferroelectrics such as $BaTiO_3$ as well as in organic polymer ferroelectrics such as PVDF copolymers [Furukawa and Seo, 1990]. In a ferroelectric material, that exhibits both spontaneous and induced polarizations, P_i^s and P_i', the strains arising from spontaneous polarizations, piezoelectricity, and electrostriction may be formulated as

$$x_{ij} = Q_{ijkl} P_k^s P_l^s + 2Q_{ijkl} P_k^s P_l' + Q_{ijkl} P_k' P_l' \tag{50.4}$$

In the paraelectric state, we may express the strain as $x_{ij} = Q_{ijkl}P_kP_l$, so that $dx_{ij}/dP_k = g_{ijk} = 2Q_{ijkl}P_l$. Converting to the commonly used d_{ijk} coefficients,

$$d_{ijk} = \chi_{mk} g_{ijm} = 2\chi_{mk} Q_{ijmn} P_n \tag{50.5}$$

This origin of piezoelectricity in electrostriction provides us an avenue into nonlinearity. In this case, it is the ability to tune the piezoelectric coefficient and the dielectric behavior of a transducer. The piezoelectric coefficient varies with the polarization induced in the material, and may be controlled by an applied electric field. The electrostrictive element may be tuned from an inactive to a highly active state. The electrical impedance of the element may be tuned by exploiting the dependence of permittivity on the biasing field for these materials, and the saturation of polarization under high fields [Newnham, 1990].

Electrostriction and Compliance Matrices

The fourth-rank electrostriction tensor is similar to the **elastic compliance** tensor, but is not identical. Compliance is a more symmetric fourth-rank tensor than is electrostriction. For compliance, in the most general case,

$$s_{ijkl} = s_{jikl} = s_{ijlk} = s_{jilk} = s_{klij} = s_{lkij} = s_{klji} = s_{lkij} \tag{50.6}$$

but for electrostriction:

$$M_{ijkl} = M_{jikl} = M_{ijlk} = M_{jilk} \neq M_{klij} = M_{lkij} = M_{klji} = M_{lkij} \tag{50.7}$$

This means that for most point groups the number of independent electrostriction coefficients exceeds those for elasticity. M and Q coefficients may also be defined in a matrix (Voigt) notation. The electrostriction and elastic compliance matrices for point groups $6/mmm$ and ∞/mm are compared below.

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} & 0 & 0 & 0 \\ S_{12} & S_{11} & S_{13} & 0 & 0 & 0 \\ S_{13} & S_{13} & S_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & S_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & S_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\left(S_{44} - S_{12}\right) \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} & M_{13} & 0 & 0 & 0 \\ M_{12} & M_{11} & M_{13} & 0 & 0 & 0 \\ M_{31} & M_{31} & M_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & M_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & M_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \left(M_{11} - M_{12}\right) \end{bmatrix}$$

Compliance coefficients s_{13} and s_{31} are equal, but M_{13} and M_{31} are not. The difference arises from an energy argument which requires the elastic constant matrix to be symmetric.

It is possible to define sixth-rank and higher-order electrostriction coupling coefficients. The electrostriction tensor can also be treated as a complex quantity, similar to the dielectric and the piezoelectric tensors. The imaginary part of the electrostriction is also a fourth-rank tensor. Our discussion is confined to the real part of the quadratic electrostriction tensor.

Magnitudes and Signs of Electrostrictive Coefficients

The values of M coefficients range from about 10^{-24} m^2/V^2 in low-permittivity materials to 10^{-16} m^2/V^2 in high-permittivity actuator materials made from **relaxor ferroelectrics** such as PMN–lead titanate (PMN–PT) compositions. Large strains of the order of strains in ferroelectric piezoelectric materials such as lead zirconate titanate (PZT) may be induced in these materials. Q values vary in an opposite way to M values. Q ranges from 10^{-3} m^4/C^2 in relaxor ferroelectrics to greater than 1 m^4/C^2 in low-permittivity materials. Since the strain is directly proportional to the square of the induced polarization, it is also proportional to the square of the dielectric permittivity. This implies that materials with large dielectric permittivities, like relaxor ferroelectrics, can produce large strains despite having small Q coefficients.

As a consequence of the quadratic nature of the electrostriction effect, the sign of the strain produced in the material is independent of the polarity of the field. This is in contrast with linear piezoelectricity where reversing the direction of the field causes a change in the sign of the strain. The sign of the electrostrictive strain depends only on the sign of the electrostriction coefficient. In most oxide ceramics, the longitudinal electrostriction coefficients are positive. The transverse coefficients are negative as expected from Poisson ratio effects. Another consequence is that electrostrictive strain occurs at twice the frequency of an applied ac field. In acentric materials, where both piezoelectric and electrostrictive strains may be observed, this fact is very useful in separating the strains arising from piezoelectricity and from electrostriction.

50.3 PMN–PT — A Prototype Electrostrictive Material

Most commercial applications of electrostriction involve high-permittivity materials such as relaxor ferroelectrics. PMN ($Pb(Mg_{1/3}Nb_{2/3})O_3$) relaxor ferroelectric compounds were first synthesized more than 30 years ago. Since then, the PMN system has been well characterized in both single-crystal and ceramic forms, and may be considered the prototype ferroelectric electrostrictor [Jang et al., 1980]. Lead titanate ($PbTiO_3$, PT) and other materials are commonly added to PMN to shift T_{max} or increase the maximum dielectric constant. The addition of PT to PMN gives rise to a range of compositions, the PMN–PT system, that have a higher Curie range and superior electromechanical coupling coefficients. The addition of other oxide compounds, mostly other ferroelectrics, is a widely used method to tailor the electromechanical properties of electrostrictors [Voss et al., 1983]. Some properties of the PMN–PT system are listed here.

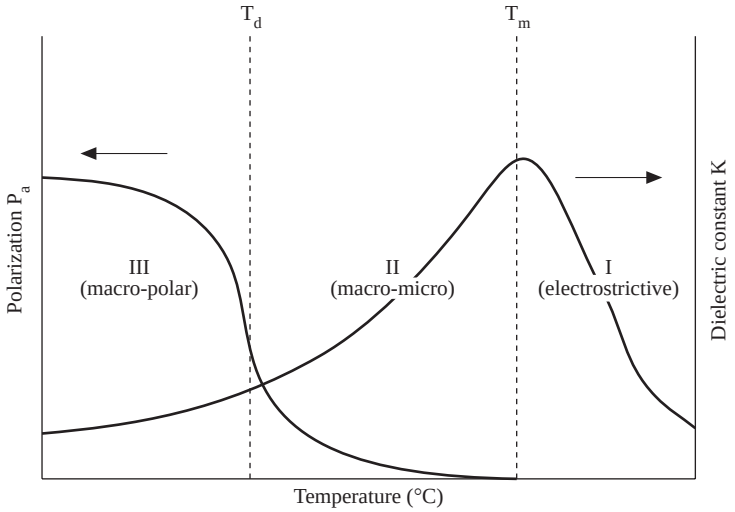


FIGURE 50.1 Polarization and dielectric behavior of a relaxor ferroelectric as a function of temperature, showing the three temperature regimes.

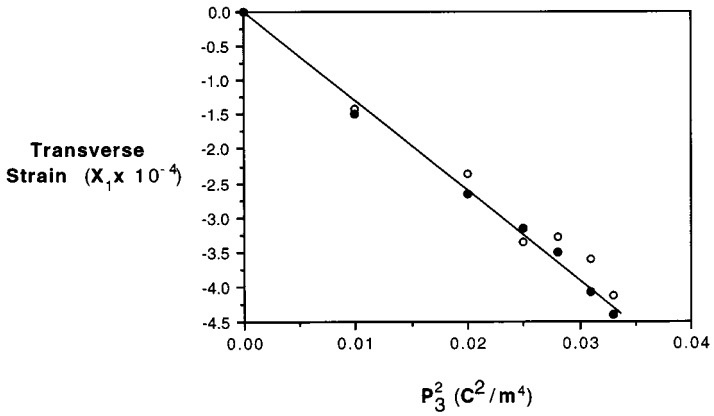


FIGURE 50.2 Transverse strain as a function of the square of the polarization in ceramic 0.9PMN–0.1PT, at RT. The quadratic ($x = QP^2$) nature of electrostriction is illustrated. Shaded circles indicate strain measured while increasing polarization and unshaded circles indicate decreasing polarization.

Based on dielectric constant vs. temperature plots, the electromechanical behavior of a relaxor ferroelectric may divided into three regimes (Fig. 50.1). At temperatures less than T_d, the depolarization temperature, the relaxor material is macropolar, exhibits a stable remanent polarization, and behaves as a piezoelectric. T_{max} is the temperature at which the maximum dielectric constant is observed. Between T_d and T_{max}, the material possesses nanometer-scale microdomains that strongly influence the electromechanical behavior. Large dielectric permittivities and large electrostrictive strains arising from micro–macrodomain reorientation are observed. Above T_{max}, the material is a "true electrostrictor" in that it is paraelectric and exhibits nonhysteretic, quadratic strain-field behavior. Since macroscale domains are absent, no remanent strain is observed. Improved reproducibility in strain and low-loss behavior are achieved.

Figure 50.2 illustrates the quadratic dependence of the transverse strain on the induced polarization for ceramic 0.9PMN–0.1PT. Figure 50.3a and b show the longitudinal strain as a function of the applied electric field for the same composition. The strain-field plots are not quadratic, and illustrate essentially anhysteretic nature of electrostrictive strain. The transverse strain is negative, as expected.

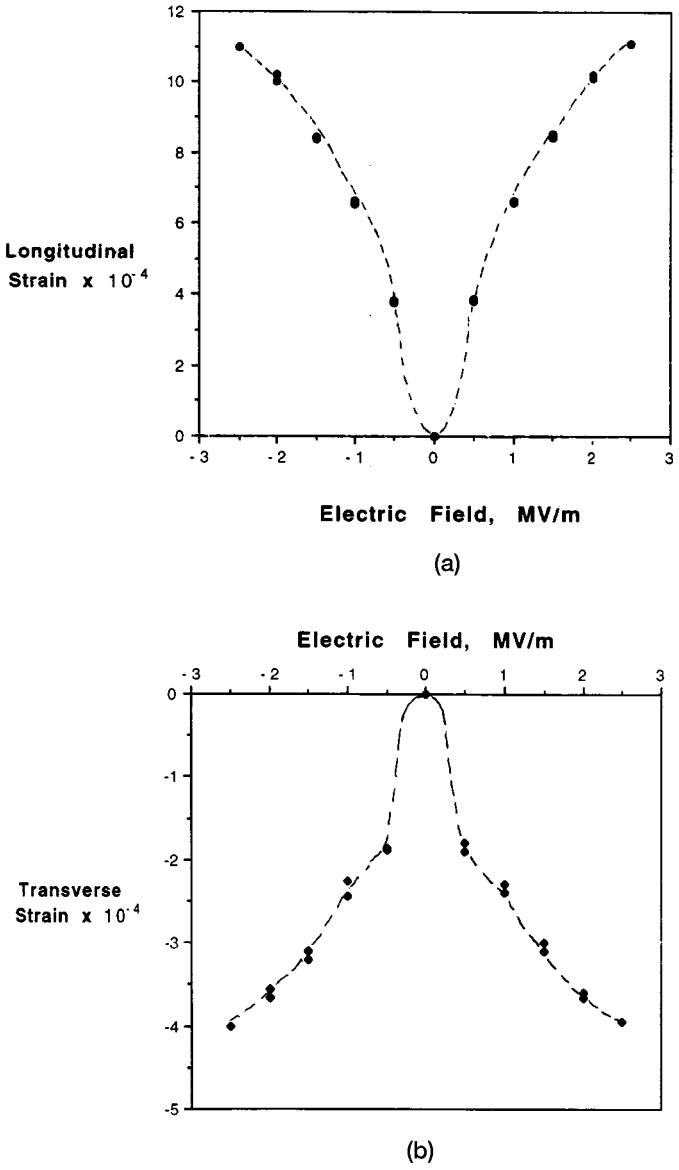


FIGURE 50.3 Longitudinal (a) and transverse (b) strains as a function of applied electric field in 0.9PMN–0.1PT, at RT. x is not quadratic with E except at low fields.

The averaged longitudinal and transverse electrostriction coefficients have been measured for poled ceramic PMN to be $Q_{33} \sim 2.3 \times 10^{-2}$ m^4/C^2, $Q_{13} \sim -0.64 \times 10^{-2}$ m^4/C^2. The corresponding field-related coefficients are $M_{33} \sim 1.50 \times 10^{-16}$ m^2/V^2 and $M_{13} \sim -4.19 \times 10^{-17}$ m^2/V^2. Induced strains of the order of 10^{-4} may be achieved with moderate electric fields of ~40 kV/cm. These strains are much larger than thermal expansion strains, and are in fact equivalent to thermal expansion strains induced by a temperature change of ~1000°C. M_{33} values for some other common ferroelectrics and a PVDF copolymer are listed in Table 50.1.

The mechanical quality factor Q_M for PMN is 8100 (at a field of ~200 kV/m) compared with 300 for poled barium titanate or 75 for poled PZT 5-A [Nomura and Uchino, 1981]. The induced piezoelectric coefficients d_{33} and d_{31} can vary with field (Fig. 50.4). The maxima in the induced piezoelectric coefficients for PMN as a function of biasing electric field are at E ~ 1.2 MV/m, with $d_{33} = 240$ pC/N and $-d_{31} = 72$ pC/N. $Pb(Mg_{0.3}Nb_{0.6}Ti_{0.1})O_3$ is a very active composition, with a maximum $d_{33} = 1300$ pC/N at a biasing field of 3.7 kV/cm.

TABLE 50.1 Electrostrictive and Dielectric Data for Some Common Actuator Materials[a]

Composition	$M_{33} \times 10^{-17}$ m²/V²	Dielectric Constant K	Ref.
$Pb(Mg_{1/3}Nb_{2/3})O_3$ (PMN)	15.04	9140	Nomura and Uchino, 1983
$(Pb_{1-x}La_{2x/3})(Zr_{1-y}Ti_y)O_3$ (PLZT 11/65/35)	1.52	5250	Landolt-Bornstein
$BaTiO_3$ (poled)	1.41	1900	Nomura and Uchino, 1983
$PbTiO_3$	1.65	1960	Landolt-Bornstein
$SrTiO_3$	5.61×10^{-2}	247	Landolt-Bornstein
PVDF/TrFE copolymer	43	12	Elhami et al., 1995

[a] At room temperature, low frequency (<100 Hz) and low magnitude electric fields (<0.1 MV/m).

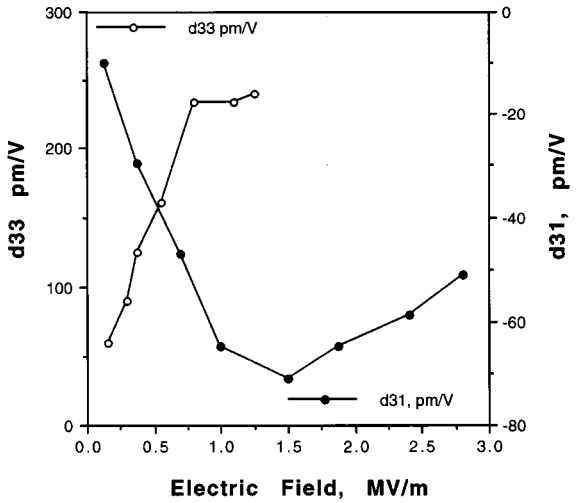


FIGURE 50.4 Induced piezoelectric coefficients d_{33} and $-d_{31}$ as a function of applied biasing field for ceramic PMN, 18°C.

50.4 Applications

The advantages that electrostrictors have over other actuator materials include low hysteresis of the strain-field response, no remanent strain (walk off), reduced aging and creep effects, a high response speed (<10 ms), and strain values (>0.03%) achievable at realizable electric fields. Displacement ranges of several tens of microns may be achieved with ±0.01μ reproducibility. Most actuator applications of electrostrictors as servotransducers and micropositioning devices take advantage of these characteristics.

Mechanical applications range from stacked actuators through inchworms, microangle adjusting devices, and oil pressure servovalves. Multilayer actuators produce large displacements and high forces at low drive voltages. The linear change in capacitance with applied stress of an electrostrictor can be used as a capacitive stress gauge [Sundar and Newnham, 1992]. Electrostrictors may also be used as used in field-tunable piezoelectric transducers. Recently, electrostrictive materials have been integrated into ultrasonic motors and novel flextensional transducers.

Electrostrictors have also been integrated into **"smart"** optical systems such as bistable optical devices, interferometric dilatometers, and deformable mirrors. Electrostrictive correction of optical aberrations is a significant tool in active optics. Electrostrictors also find applications in **"very smart" systems** such as sensor–actuator active vibration-suppression elements. A shape memory effect arising from inverse hysteretic behavior and electrostriction in PZT family antiferroelectrics is also of interest. A recent survey [Uchino, 1993] predicts that the market share of piezoelectric and electrostrictive transducers is expected to increase to more than $10 billion by 1998.

TABLE 50.2 Selection Critera for Relaxor Ferroelectrics for Electrostrictive Devices

Desirable Properties	Material Behavior
• Large strain, induced polarization, and induced piezoelectricity	• Large dielectric constants
• Large operating temperature range	• $T_{max} - T_d$ is large • Broad dielectric transition
• Low-loss, low-joule heating, minimal hysteresis, no remanent polarization	• Operation in paraelectric regime ($T > T_{max}$)

In selecting electrostrictive relaxor ferroelectrics for actuator and sensor applications, the following criteria are commonly used. A large dielectric constant and field stability in the K vs. E relations are useful in achieving large electrostrictive strains. These criteria also lead to large induced polarizations and large induced piezoelectric coefficients through the second converse effect.

Broad dielectric transitions allow for a large operating temperature range. In the case of relaxors, this implies a large difference between T_{max} and T_d. Minimal E–P hysteresis and no remanent polarization are useful in achieving a low-loss material that is not susceptible to joule heating effects. These factors are listed in Table 50.2.

50.5 Summary

Electrostriction is a fundamental electromechanical coupling effect. In ceramics with large dielectric constants and in some polymers, large electrostrictive strains may be induced that are comparable in magnitude with piezoelectric strains in actuator materials such as PZT. The converse electrostrictive effect, which is the change in dielectric susceptibility with applied stress, facilitates the use of the electrostrictor as a stress gauge. The second converse effect may be used to tune the piezoelectric coefficients of the material as a function of the applied field. Electrostrictive materials offer tunable nonlinear properties that are suitable for application in very smart systems.

Defining Terms

Elastic compliance: A fourth-rank tensor (s_{ijkl}) relating the stress (X) applied on a material and the strain (x) developed in it, $x_{ij} = s_{ijkl}X_{kl}$. Its inverse is the elastic stiffness tensor (c_{ijkl}).

Electrostriction: The quadratic coupling between strain and applied field or induced polarization. Conversely, it is the linear coupling between dielectric susceptibility and applied stress. It is present in all insulating materials.

Ferroelectricity: The phenomenon by which a material exhibits a permanent spontaneous polarization that can be reoriented (switched) between two or more equilibrium positions by the application of a realistic electric field (i.e., less than the breakdown field of the material).

Perovskite: A crystal structure with the formula ABO_3, with A atoms at the corners of a cubic unit cell, B atoms at the body-center position, and O atoms at the centers of the faces. Many oxide perovskites are used as transducers, capacitors, and thermistors.

Piezoelectricity: The linear coupling between applied electric field and induced strain in acentric materials. The converse effect is the induction of polarization when stress is applied.

Relaxor ferroelectric: Relaxor ferroelectric materials exhibit a diffuse phase transition between paraelectric and ferroelectric phases, and a frequency dependence of the dielectric properties.

Smart and very smart systems: A system that can sense a change in its environment, and tune its response suitably to the stimulus. A system that is only smart can sense a change in its environment and react to it.

Related Topic

58.5 State-of-the-Art Smart Materials

References

W. G. Cady, *International Critical Tables*, vol. 6, p. 207, 1929.
K. Elhami, B. Gauthier-Manuel, J. F. Manceau, and F. Bastien, *J. Appl. Phys.,* vol. 77, p. 3987, 1995.
T. Furukawa and N. Seo, *Jpn. J. Appl. Phys.*, vol. 29, p. 675, 1990.
S. J. Jang, K. Uchino, S. Nomura, and L. E. Cross, *Ferroelectrics*, vol. 27, p. 31, 1980.
Landolt-Bornstein, *Numerical Data and Functional Relationships in Science and Technology*, New Series, Gruppe III, vols. 11 and 18, Berlin: Springer-Verlag, 1979, 1984.
R. E. Newnham, *Chemistry of Electronic Ceramic Materials*, in *Proc. Intl. Conf.,* Jackson, Wyo., 1990; NIST Special Publication 804, 39, 1991.
S. Nomura and K. Uchino, *Ferroelectrics*, vol. 50, p. 197, 1983.
V. Sundar and R. E. Newnham, *Ferroelectrics*, vol. 135, p. 431, 1992.
K. Uchino, *MRS Bull.*, vol. 18, pp. 42, 1993.
D. J. Voss, S. L. Swartz, and T. R. Shrout, *Ferroelectrics*, vol. 50, p. 1245, 1983.

Further Information

IEEE Proceedings of the International Symposium on the Applications of Ferroelectrics (ISAF)
IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control (UFFC)
American Institute of Physics Handbook, 3rd ed., New York: McGraw-Hill, 1972
M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectric Materials,* Oxford: Clarendon Press, 1977

Amin, A. “Piezoresistivity”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

51

Piezoresistivity

51.1 Introduction
51.2 Equation of State
51.3 Effect of Crystal Point Group on $\prod_{ijkl}$
51.4 Geometric Corrections and Elastoresistance Tensor
51.5 Multivalley Semiconductors
51.6 Longitudinal Piezoresistivity $\prod_l$ and Maximum Sensitivity Directions
51.7 Semiconducting (PTCR) Perovskites
51.8 Thick Film Resistors
51.9 Design Considerations

Ahmed Amin
Texas Instruments, Inc.

51.1 Introduction

Piezoresistivity is a linear coupling between mechanical stress X_{kl} and electrical resistivity ρ_{ij}. Hence, it is represented by a fourth rank polar tensor $\prod_{ijkl}$. The piezoresistance properties of semiconducting silicon and germanium were discovered by Smith [1953] when he was verifying the form of their energy surfaces. Piezoresistance measurements can provide valuable insights concerning the conduction mechanisms in solids such as strain-induced carrier repopulation and intervalley scattering in multivalley semiconductors [Herring and Vogt, 1956], barrier tunneling in thick film resistors [Canali et al., 1980] and barrier raising in semiconducting positive temperature coefficient of resistivity (PTCR) perovskites [Amin, 1989]. Piezoresistivity has also been investigated in compound semiconductors, thin metal films [Rajanna et al., 1990], polycrystalline silicon and germanium thin films [Onuma and Kamimura, 1988], heterogeneous solids [Carmona et al., 1987], and high T_c superconductors [Kennedy et al., 1989]. Several sensors that utilize this phenomenon are commercially available.

51.2 Equation of State

The equation of state of a crystal subjected to a stress X_{kl} and an electric field E_i is conveniently formulated in the isothermal representation. The difference between isothermal and adiabatic changes, however, is negligible [Keyes, 1960]. Considering only infinitesimal deformations, where the linear theory of elasticity is valid, the electric field E_i is expressed in terms of the current density I_j and applied stress X_{kl} as [Mason and Thurston, 1957].

$$E_i = E_i\,(I_j,\ X_{kl}) \qquad i,j,k,l = 1,2,3 \tag{51.1}$$

In what follows the summation convention over repeated indices in the same term is implied, and the letter subscripts assume the values 1, 2, and 3 unless stated otherwise. Expanding in a McLaurin's series about the origin (state of zero current and stress)

$$dE_i = (\partial E_i/\partial I_j)\ dI_j + (\partial E_i/\partial X_{kl})\ dX_{kl}$$
$$+ (1/2!)\ [(\partial^2 E_i/\partial I_j \partial I_m)\ dI_j\ dI_m$$
$$+ (\partial^2 E_i/\partial X_{kl}\partial X_{no})\ dX_{kl}\ dX_{no}$$
$$+ 2\ (\partial^2 E_i/\partial X_{kl}\partial I_j)\ dX_{kl}\ dI_j] + \ldots \text{H.O.T} \tag{51.2}$$

The partial derivatives in the expansion Eq. (51.2) have the following meanings: $(\partial E_i/\partial I_j) = \rho_{ij}$(electric resistivity tensor); $(\partial E_i/\partial X_{kl}) = d_{ikl}$(converse piezoelectric tensor); $(\partial^2 E_i/\partial X_{kl}\partial I_j) = (\partial/\partial X_{kl})\ (\partial E_i/\partial I_j) = \Pi_{ijkl}$(piezoresistivity tensor); $(\partial^2 E_i/\partial I_j \partial I_m) = \rho_{ijm}$(nonlinear resistivity tensor); $(\partial^2 E_i/\partial X_{kl}\partial X_{no}) = \delta_{iklno}$(nonlinear piezoelectric tensor).

Replacing the differentials in Eq. (51.2) by the components themselves, we get

$$E_i = \rho_{ij} I_j + d_{ikl}\ X_{kl} + 1/2\ \rho_{ijm}\ I_j I_m + 1/2\ \delta_{iklno}\ X_{kl}\ X_{no} + \prod_{ijkl}\ X_{kl}\ I_j \tag{51.3}$$

Most of the technologically important piezoresistive materials, e.g., silicon, germanium, and polycrystalline films, are centrosymmetric. The effect of center of symmetry (i.e., the inversion operator) on Eq. (51.3) is to force all odd rank tensor coefficients to zero; hence, the only contribution to the resistivity change under stress will result from the piezoresistive term. Therefore, Eq. (51.3) takes the form

$$E_i = \Sigma_j \rho_{ij}\ I_j + \Sigma_j \Sigma_k \Sigma_l \Pi_{ijkl}\ X_{kl}\ I_j \tag{51.4}$$

taking the partial derivatives of Eq. (51.4) with respect to the current density I_j and rearranging

$$\partial E_i/\partial I_j = \rho_{ij}(X)\ 2\ \rho_{ij}(0) = \Sigma_k \Sigma_l \Pi_{ijkl}\ X_{kl}$$

Thus, the specific change in resistivity with stress is given by

$$(\delta\rho_{ij}/\rho_0) = \Pi_{ijkl}\ X_{kl} \tag{51.5}$$

the piezoresistivity tensor Π_{ijkl} in Eq. (51.5) has the dimensions of reciprocal stress (square meters per newton in the MKS system of units). The effects of the intrinsic symmetry of the piezoresistivity tensor and the crystal point group are discussed next.

51.3 Effect of Crystal Point Group on Π_{ijkl}

The transformation law of Π_{ijkl} (a fourth rank polar tensor) is as follows:

$$\Pi'_{ijkl} = (\partial x'_i/\partial x_m)(\partial x'_j/\partial x_n)(\partial x'_k/\partial x_o)(\partial x'_l/\partial x_p)\Pi_{mnop} \tag{51.6}$$

where the primed and unprimed components refer to the new and old coordinate systems, respectively, and the determinants of the form $\|\partial x'_i/\partial x_m\|$, . . . etc. are the Jacobian of the transformation. A general fourth rank tensor has 81 independent components. The piezoresistivity tensor Π_{ijkl} has the following internal symmetry:

$$\Pi_{ijkl} = \Pi_{ijlk} = \Pi_{jilk} = \Pi_{jikl} \tag{51.7}$$

which reduces the number of independent tensor components from 81 to 36 for the most general triclinic point group $C_1(1)$. It is convenient to use the reduced (two subscript) matrix notation

$$\Pi_{ijkl} = \Pi_{mn} \tag{51.8}$$

where $m,n = 1, 2, 3, \ldots, 6$. The relation between the subscripts in both notations is

Tensor:	11	22	33	23, 32	13, 31	12, 21
Matrix:	1	2	3	4	5	6

and

$$\Pi_{mn} = 2\Pi_{ijkl}, \text{ for } m \text{ and/or } n = 4, 5, 6$$

Thus, for example, $\Pi_{1111} = \Pi_{11}$, $\Pi_{1122} = \Pi_{12}$, $2\Pi_{2323} = \Pi_{44}$, $2\Pi_{1212} = \Pi_{66}$, and $2\Pi_{1112} = \Pi_{16}$. Hence, Eq. (51.5) takes the form

$$(\delta\rho_i/\rho_0) = \Pi_{ij} X_j, \qquad (i, j = 1, 2, \ldots, 6) \tag{51.9}$$

Further reduction of the remaining 36 piezoresistivity tensor components is obtained by applying the generating elements of the point group to the piezoresistivity tensor transformation law Eq. (51.6) and demanding invariance. The following are two commonly encountered piezoresistivity matrices:

1. Cubic $O_h(m3m)$: single crystal silicon and germanium

$$\begin{matrix} \Pi_{11} & \Pi_{12} & \Pi_{12} & 0 & 0 & 0 \\ & \Pi_{11} & \Pi_{12} & 0 & 0 & 0 \\ & & \Pi_{11} & 0 & 0 & 0 \\ & & & \Pi_{44} & 0 & 0 \\ & & & & \Pi_{44} & 0 \\ & & & & & \Pi'_{44} \end{matrix}$$

2. Spherical ($\infty \infty mmm$): polycrystalline silicon and germanium and films

$$\begin{matrix} \Pi_{11} & \Pi_{12} & \Pi_{12} & 0 & 0 & 0 \\ & \Pi_{11} & \Pi_{12} & 0 & 0 & 0 \\ & & \Pi_{11} & 0 & 0 & 0 \\ & & & \Pi_{44} & 0 & 0 \\ & & & & \Pi_{44} & 0 \\ & & & & & \Pi_{44} \end{matrix}$$

where $\Pi_{44} = 2(\Pi_{11} - \Pi_{12})$. Thus, three coefficients Π_{11}, Π_{12}, and Π_{44} are required to completely specify the piezoresistivity tensor for silicon and germanium single crystals, and only two, Π_{11} and Π_{12}, for polycrystalline films. Under hydrostatic pressure conditions, the piezoresistivity coefficient Π_h for the preceding two symmetry groups is a linear combination of the longitudinal Π_{11} and transverse Π_{12} components, $\Pi_h = \Pi_{11} + 2\Pi_{12}$. Unlike the elastic stiffness c_{ij} (a fourth rank polar tensor), the piezoresistivity tensor Π_{mn} is not symmetric, i.e., $\Pi_{mn} \neq \Pi_{nm}$, except for the following point groups, $C_{\infty v}(\infty \infty mmm)$, $O_h(m3m)$, $T_d(\overline{4}3m)$, and $O(432)$.

51.4 Geometric Corrections and Elastoresistance Tensor

The experimentally derived quantity is the piezoresistance coefficient $1/R_0(\partial R/\partial X)$. This must be corrected for the dimensional changes to obtain the piezoresistivity coefficient $1/\rho_0(\partial\rho/\partial X)$ as follows:

1. Uniaxial tensile stress parallel to current flow

$$1/R_0(\partial R/\partial X) - (s_{11} - 2s_{12}) = 1/\rho_0(\partial\rho/\partial x) = \Pi_{11} \tag{51.10}$$

TABLE 51.1 Numerical Values of Π_{ij} and M_{ij} for Selected Materials

Material	Resistivity (Unstrained and RT)	(10^{-11} m²/N) Π_{11}	Π_{12}	Π_{44}	Dimensionless M_{11}	M_{12}	M_{44}
Silicon							
n-type	11.7 (Ω-cm)	–102.2	53.4	–13.6	–72.6	86.4	–10.8
p-type	7.8 (Ω-cm)	6.6	–1.1	138.1	10.5	2.7	110
$Ba_{.648}Sr_{.35}$							
$La_{.002}TiO_3$	≈100 (Ω-cm)	250	250				
Thin films							
Si					15		
Ge					30		
Mn	160 (μΩ-cm)				3		
Thick film resistors							
DP 1351, main constituent $Bi_2Ru_2O_7$							
	100 ($K\Omega/\square$)				13.5		
ESL 2900	100 (KΩ/□)				13.8	11.6	

2. Uniaxial tensile stress perpendicular to current flow

$$1/R_0(\partial R/\partial X) + s_{11} = 1/\rho_0(\partial\rho/\partial X) = \Pi_{12} \tag{51.11}$$

3. Hydrostatic pressure

$$1/R_0(\partial R/\partial p) - (s_{11} + 2s_{12}) = 1/\rho_0(\partial\rho/\partial p) = \Pi_h \tag{51.12}$$

where s_{11} and s_{12} are the elastic compliances that appear in the linear elasticity equation $x_{ij} = s_{ijkl}\, X_{kl}$, with x_{ij} the infinitesimal strain components. Details on the different geometries and methods of measuring the piezoresistance effect can be found in the References. Equation (51.9) could be written in terms of the strain conjugate x_o as follows

$$(\delta\rho/\rho_0) = M_{io}\, x_o \tag{51.13}$$

the dimensionless quantity M_{io} is the elastoresistance tensor (known as the gage factor in the sensors literature). It is related to the piezoresistivity Π_{ik} and the elastic stiffness c_{ko} tensors by

$$M_{io} = \Pi_{ik}\, c_{ko} \tag{51.14}$$

thus, the 3 independent elastoresistance components (gauge factors) can be expressed as follows

$$M_{11} = \Pi_{11}\, c_{11} + 2\,\Pi_{12}\, c_{12}$$
$$M_{12} = \Pi_{11}\, c_{12} + \Pi_{12}\,(c_{11} + c_{12})$$
$$M_{44} = \Pi_{44}\, c_{44}$$

51.5 Multivalley Semiconductors

For a multivalley semiconductor, e.g., *n*-type silicon, the energy minima (ellipsoids of revolutions) of the unstrained state in momentum space are along the six <100> cubic symmetry directions; they possess the symmetry group $O_h(m3m)$. A tensile stress in the *x*-direction, for example, will strain the lattice in the *xy*-plane and destroy the three-fold symmetry, thereby lifting the degeneracy of the energy minima. However, the four-fold symmetry along the *x*-direction will be preserved. Thus, the two valleys along the direction of stress will be shifted relative to the four valleys in the perpendicular directions.

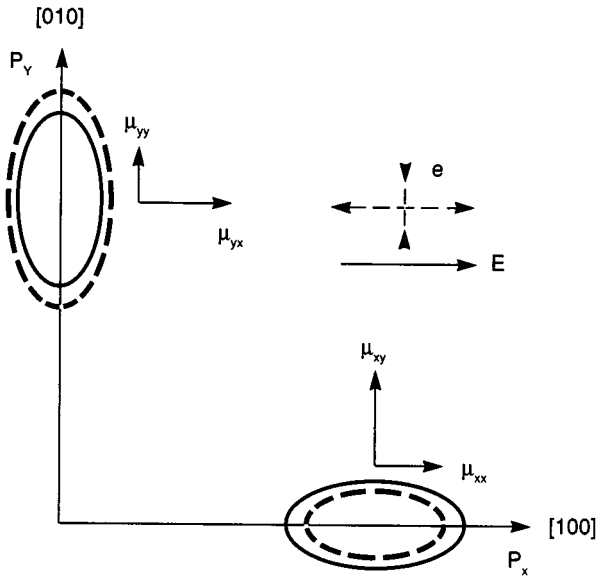


FIGURE 51.1 Two-dimensional representation of the constant energy surfaces in momentum space of a multivalley semiconductor (e.g., *n*-type silicon) showing only one quadrant, point group symmetry $C_{4v}(4mm)$. (*Source:* C.S. Smith, Piezoresistance effect in germanium and silicon, *Phys. Rev.,* vol. 94, p. 42, 1953. With permission.)

According to the deformation potential theory, the strain will shift the energy of all the states in a given band extremum by the same amount, i.e., the valley moves along the energy scale as a whole by an amount (the deformation potential constant) which is linearly proportional to the strain. Let's assume that the energy of those on the *y*- and *z*-axes are lowered with respect to those on the *x*-axis. This effect is represented by dashed lines in Fig. 51.1. As a result, there will be electron transfer from the high to low energy valleys. The components of the mobility tensor μ_{xy} ($= e\,\tau/m_{xy}$, where e is the electron charge, τ is the relaxation time, and m_{xy} is the effective mass) are illustrated by arrows in Fig. 51.1. The mobility anisotropy is due to the curvature of the conduction band near the bottom. The effective mass is inversely proportional to this curvature ($1/m_{xy} = (h/2\pi)^{-2}(\partial^2 E/\partial k_x \partial k_y)$), which is larger for a direction perpendicular to the valley. For an applied E field parallel to the stress, the conductivity will increase (i.e., the resistivity decreases) relative to the unstressed state because of the increase in the number of electrons in the four valleys (*yz*-plane) for which the mobility is large in the field direction. If the field is perpendicular to the stress, the conductivity will decrease (i.e., the resistivity increases) with stress. Therefore, the piezoresistivity components Π_{11} and Π_{12} have opposite signs.

A shear stress about the crystallographic axes will not lift the degeneracy; hence, $\Pi_{44} = 0$. Similarly, a tensile stress along the <111> does not destroy the three-fold symmetry, and the degeneracy will not be lifted; thus, no piezoresistance should be there. Calculations based on the deformation potential model show that $\Pi_{11} = 22\,\Pi_{12}$, and $\Pi_{44} = 0$.

Information concerning the symmetry properties of the valleys can be derived from the representation surface of the longitudinal Π_{11} piezoresistance component. This surface can be constructed by measuring the dependence of Π_{11} on the crystallographic direction. Smith showed that Π_{11} is maximum in the <001> directions of *n*-type silicon and not quite zero in the <111> directions. Reasons for the deviation from the deformation potential model of piezoresistivity in multivalley semiconductors are discussed in Keyes [1960]. For *n*-type germanium Π_{11} is maximum in the <111> directions. This is consistent with the loci of the valleys in these two materials. Qualitatively, a piezoresistance effect is produced whenever the stress destroys the symmetry elements that are responsible for the degeneracy of the valleys.

Intervalley scattering contribution to the piezoresistance of multivalley semiconductors may be comparable to that of the strain-induced electron repopulation. In this scattering process, the initial and final electron states are in different valleys. The effect of intervalley scattering can be deduced from the T^{-1} dependence of the elastoresistance tensor.

The influence of hydrostatic pressure on the electrical resistivity can provide additional insights on the transport properties. Some of the noted features include (1) high pressures (in the GPa range, versus MPa for tensile stresses) can be applied without destroying the crystal; (2) it does not destroy the symmetry, provided no phase transition is involved; hence, the symmetry degeneracies in the band structure are not lifted; (3) band edges which are not degenerate for symmetry reasons will be shifted; and (4) nonlinear effects could be discerned.

51.6 Longitudinal Piezoresistivity Π_l and Maximum Sensitivity Directions

Consider a long thin bar "strain gauge" cut from a piezoresistive crystal with the bar length parallel to an arbitrary direction in the crystal. Let Π_l, θ, and φ be the spherical coordinates of the longitudinal piezoresistivity tensor measured along the length of the bar. For the cubic symmetry group O_h ($m3m$) of Si and Ge, Π_l is given by (Mason et al. 1957)

$$\Pi_l = \Pi_{11} + 2(\Pi_{44} + \Pi_{12} - \Pi_{11})\ [\sin^2\theta \cos^2\theta + \cos^4\theta \cos^2\varphi \sin^2\varphi].$$

$$= \Pi_{11} + 2(\Pi_{44} + \Pi_{12} - \Pi_{11})\ F[\theta,\varphi]$$

The variation of Π_l with direction may be considered as a property surface. The distance from the center to any point in the surface is equal to the magnitude of Π_l. The function $F[\theta,\varphi]$ has a maximum for $\theta = 54° 40`$ and $\varphi = 45°$ which is the <111> family of directions, for which Π_l takes the following form,

$$\Pi_l = \Pi_{11} + 2/3\ (\Pi_{44} + \Pi_{12} - \Pi_{11})$$

If $(\Pi_{44} + \Pi_{12} - \Pi_{11})$ and Π_{11} have the same sign or $2/3 \left| (\Pi_{44} + \Pi_{12} - \Pi_{11}) \right| > \Pi_{11}$ then the maximum sensitivity direction occurs along <111>. If $(\Pi_{44} + \Pi_{12} - \Pi_{11}) = 0$ the longitudinal effect is isotropic and equal to Π_{11} in all directions, otherwise it occurs along a crystal axis. The maximum sensitivity directions are shown in Fig. 51.2 for Si and Ge.

51.7 Semiconducting (PTCR) Perovskites

Large hydrostatic piezoresistance Π_h coefficients (two orders of magnitude larger than those of silicon and germanium) have been observed in this class of polycrystalline semiconductors [Sauer et al., 1959]. PTCR compositions are synthesized by donor doping ferroelectric barium titanate $BaTiO_3$, $(Ba,Sr)TiO_3$, or $(Ba,Pb)TiO_3$ with a trivalent element (e.g., yttrium) or a pentavalent element (e.g., niobium). Below the ferroelectric transition temperature T_c, Schottky barriers between the conductive ceramic grains are neutralized by the spontaneous polarization P_s associated with the ferroelectric phase transition. Above T_c the barrier height increases rapidly with temperature (hence the electrical resistivity) because of the disappearance of P_s and the decrease of the paraelectric state dielectric constant. Analytic expressions that permit the computation of barrier heights under different elastic and thermal boundary conditions have been developed [Amin, 1989].

51.8 Thick Film Resistors

Thick film resistors consist of a conductive phase, e.g., rutile (RuO_2), perovskite ($BaRuO_3$), or pyrochlore ($Pb_2Ru_2O_{7-x}$), and an insulating phase (e.g., lead borosilicate) dispersed in an organic vehicle. They are formed by screen printing on a substrate, usually alumina, followed by sintering at ≈850°C for 10 min.

The increase of the piezoresistance properties of a commercial thick film resistor (ESL 2900 series) with sheet resistivity is illustrated in Fig. 51.3. The experimentally observed properties such as the resistance increase and decrease with tensile and compressive strains, respectively, and the increase of the elastoresistance tensor with sheet resistivity seem to support a barrier tunneling model [Canali et al., 1980].

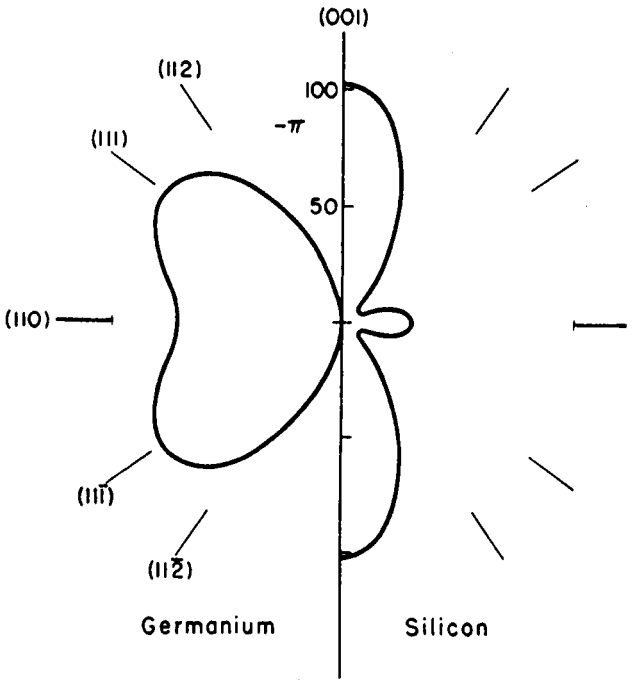


FIGURE 51.2 Section of the longitudinal piezoresistivity surface, the maximum sensitivity directions in Si and Ge are shown [Keys, 1960].

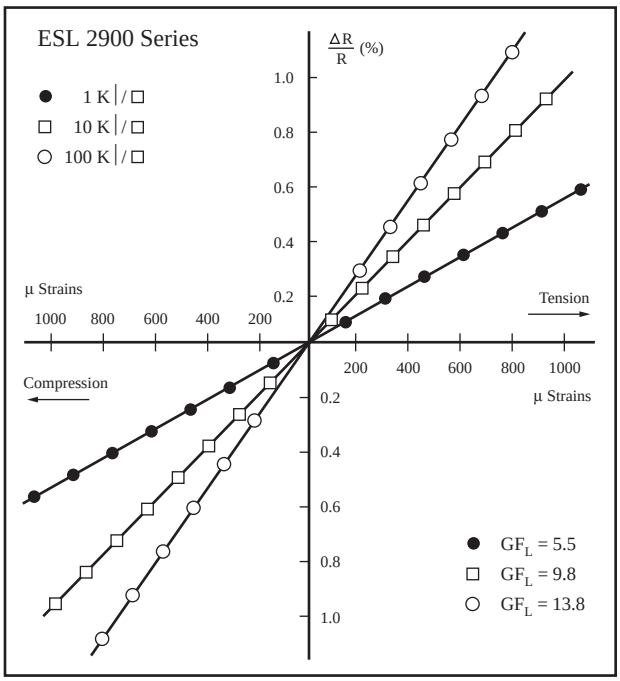


FIGURE 51.3 Relative changes of resistance for compressive and tensile strain applied parallel to the current direction. Note the increase of gage factor with sheet resistivity.

51.9 Design Considerations

Many commercially available sensors (pressure, acceleration, vibration,… etc.) are fabricated from piezoresistive materials (see for example, Chapter 56 in this handbook.) The most commonly used geometry for pressure sensors is the edge clamped diaphragm. Four resistors are usually deposited on the diaphragm and connected to form a Wheatstone bridge. The deposition technique varies depending upon the piezoresistive material: standard IC technology and micro-machining for Si type diaphragms; sputtering for thin film metal strain gauges; bonding for wire strain gauges, and screen printing for thick film resistors. Different types of diaphragms (sapphire, metallic, ceramic,… etc.) have been reported in the literature for hybrid sensors.

To design a highly accurate and sensitive sensor, it is necessary to analyze the stress–strain response of the diaphragm using plate theory and finite element techniques to take into account: (1) elastic anisotropy of the diaphragm, (2) large deflections of plate (elastic non linearities), and (3) maximum sensitivity directions of the piezoresistivity coefficient. Signal conditioning must be provided to compensate for temperature drifts of the gauge offset and sensitivity.

Defining Terms

$\boldsymbol{\rho}_{ij}$: Electric resistivity tensor
$\boldsymbol{d}_{ikl}$: Converse piezoelectric tensor
$\boldsymbol{\Pi}_{ijkl}$: Piezoresistivity tensor
$\boldsymbol{\rho}_{ijm}$: Nonlinear resistivity tensor
$\boldsymbol{\delta}_{iklno}$: Nonlinear piezoelectric tensor

Related Topic

1.1 Resistors

References

A. Amin, "Numerical computation of the piezoresistivity matrix elements for semiconducting perovskite ferroelectrics," *Phys. Rev. B*, vol. 40, 11603, 1989.

C. Canali, D. Malavasi, B. Morten, M. Prudenziati, and A.Taroni, "Piezoresistive effect in thick-film resistors," *J. Appl. Phys.,* vol. 51, 3282, 1980.

F. Carmona, R. Canet, and P. Delhaes, "Piezoresistivity in heterogeneous solids," *J. Appl. Phys.,* vol. 61, 2550, 1987.

C. Herring and E. Vogt, "Transport and deformation-potential theory for many valley semiconductors with anisotropic scattering," *Phys. Rev.,* vol. 101, 944, 1956.

R. J. Kennedy, W. G. Jenks, and L. R. Testardi, "Piezoresistance measurements of $YBa_2Cu_3O_{7-x}$ showing large magnitude temporal anomalies between 100 and 300 K," *Phys. Rev. B*, vol. 40, 11313, 1989.

R. W. Keyes, "The effects of elastic deformation on the electrical conductivity of semiconductors," *Solid State Phys.,* vol. 11, 149, 1960.

W. P. Mason and R. N. Thurston, "Use of piezoresistive materials in the measurement of displacement, force, and torque," *J. Acoust. Soc. Am.,* vol. 10, 1096, 1957.

Y. Onuma and K. K. Kamimura, "Piezoresistive elements of polycrystalline semiconductor thin films," *Sensors and Actuators,* vol. 13, 71, 1988.

K. Rajanna, S. Mohan, M. M. Nayak, and N. Gunasekaran, "Thin film pressure transducer with manganese film as the strain gauge," *Sensor and Actuators,* vol. A 24, 35, 1990.

H. A. Sauer, S. S. Flaschen, and D. C. Hoestery, "Piezoresistance and piezocapacitance effect in barium strontium titanate ceramics," *J. Am. Ceram. Soc.,* vol. 42, 363, 1959.

C. S. Smith, "Piezoresistance effect in germanium and silicon," *Phys. Rev.,* vol. 94, 42, 1953.

Further Information

M. Neuberger and S. J. Welles, *Silicon,* Electronic Properties Information Center, Hughes Aircraft Co., Culver City, Calif., 1969. This reference contains a useful compilation of the piezoresistance properties of silicon.

Electronic databases such as *Chemical Abstracts* will provide an update on the current research on piezoresistance materials and properties.

Ehrlich A.C. “The Hall Effect”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

52

The Hall Effect

Alexander C. Ehrlich
U.S. Naval Research Laboratory

52.1 Introduction
52.2 Theoretical Background
52.3 Relation to the Electronic Structure—(i) $\omega_c\tau << 1$
52.4 Relation to the Electronic Structure—(ii) $\omega_c\tau >> 1$

52.1 Introduction

The Hall effect is a phenomenon that arises when an electric current and magnetic field are simultaneously imposed on a conducting material. Specifically, in a flat plate conductor, if a current density, J_x, is applied in the x direction and (a component of) a magnetic field, B_z, in the z direction, then the resulting electric field, E_y, transverse to J_x and B_z is known as the Hall electric field E_H (see Fig. 52.1) and is given by

$$E_y = RJ_xB_z \tag{52.1}$$

where R is known as the Hall coefficient. The Hall coefficient can be related to the electronic structure and properties of the **conduction bands** in metals and semiconductors and historically has probably been the most important single parameter in the characterization of the latter. Some authors choose to discuss the Hall effect in terms of the Hall angle, ϕ, shown in Fig. 52.1, which is the angle between the net electric field and the imposed current. Thus,

$$\tan\phi = E_H/E_x \tag{52.2}$$

For the vast majority of Hall effect studies that have been carried out, the origin of E_H is the Lorentz force, F_L, that is exerted on a charged particle as it moves in a magnetic field. For an electron of charge e with velocity v, F_L is proportional to the vector product of **v** and **B**; that is,

$$F_L = e\mathbf{vxB} \tag{52.3}$$

In these circumstances a semiclassical description of the phenomenon is usually adequate. This description combines the classical Boltzmann transport equation with the Fermi–Dirac distribution function for the charge carriers (electrons or holes) [Ziman, 1960], and this is the point of view that will be taken in this chapter. Examples of Hall effect that cannot be treated semiclassically are the spontaneous (or extraordinary) Hall effect that occurs in ferromagnetic conductors [Berger and Bergmann, 1980], the quantum Hall effect [Prange and Girvin, 1990], and the Hall effect that arises in conjuction with hopping conductivity [Emin, 1977].

In addition to its use as an important tool in the study of the nature of electrically conducting materials, the Hall effect has a number of direct practical applications. For example, the sensor in some commercial devices for measuring the magnitude and orientation of magnetic fields is a Hall sensor. The spontaneous Hall effect has been used as a nondestructive method for exploring the presence of defects in steel structures. The quantum Hall effect has been used to refine our knowledge of the magnitudes of certain fundamental constants such as the ratio of e^2/h where h is Planck's constant.

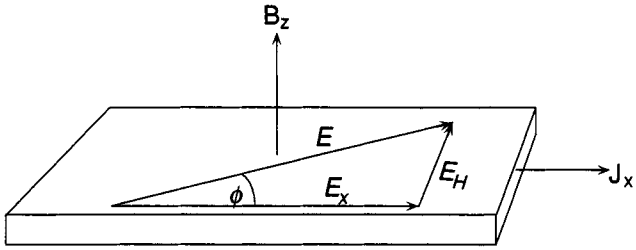


FIGURE 52.1 Typical Hall effect experimental arrangement in a flat plate conductor with current J_x and magnetic field B_z. The Hall electric field $E_H = E_y$ in this geometry arises because of the Lorentz force on the conducting charges and is of just such a magnitude that in combination with the Lorentz force there is no net current in the y direction. The angle ϕ between the current and net electric field is called the Hall angle.

52.2 Theoretical Background

The Boltzmann equation for an electron gas in a homogeneous, isothermal material that is subject to constant electric and magnetic fields is [Ziman, 1960]

$$e[\mathbf{E}+\mathbf{vXB}]\left(\frac{1}{\hbar}\right)\nabla_{\mathbf{k}} f(\mathbf{k})-\left(\frac{\partial f}{\partial t}\right)_s = 0 \tag{52.4}$$

Here $\mathbf{k}$ is the quantum mechanical wave vector, $\hbar$ is Planck's constant divided by 2π, t is the time, f is the electron distribution function, and "s" is meant to indicate that the time derivative of f is a consequence of scattering of the electrons. In static equilibrium (E = 0, B = 0) f is equal to f_0 and f_0 is the Fermi–Dirac distribution function

$$f_0 = \frac{1}{e^{\left(\mathscr{E}(\mathbf{k})-\zeta\right)/KT}+1} \tag{52.5}$$

where $\mathscr{E}(\mathbf{k})$ is the energy, ζ is the chemical potential, K is Boltzmann's constant, and T is the temperature. Each term in Eq. (52.4) represents a time rate of change of f and in dynamic equilibrium their sum has to be zero. The last term represents the effect of collisions of the electrons with any obstructions to their free movement such as lattice vibrations, crystallographic imperfections, and impurities. These collisions are usually assumed to be representable by a relaxation time, $\tau(\mathbf{k})$, that is

$$\left(\frac{\partial f}{\partial t}\right)_c = \frac{-\left(f - f_0\right)}{\tau(\mathbf{k})} = \frac{(\partial f_0 / \partial \mathscr{E})\, g(\mathbf{k})}{\tau(\mathbf{k})} \tag{52.6}$$

where $f - f_0$ is written as $(\partial f_0/\partial\varepsilon)g(\mathbf{k})$, which is essentially the first term in an expansion of the deviation of f from its equilibrium value, f_0. Eqs. (52.6) and (52.4) can be combined to give

$$e[\mathbf{E} + \mathbf{vXB}]\,\frac{1}{\hbar}\,\nabla_{\mathbf{k}} f(\mathbf{k}) = \frac{(\partial f_0/\partial \mathscr{E})\, g(\mathbf{k})}{\tau(\mathbf{k})} \tag{52.7}$$

If Eq. (52.7) can be solved for $g(\mathbf{k})$, then expressions can be obtained for both the E_H and the magnetoresistance (the electrical resistance in the presence of a magnetic field). Solutions can in fact be developed that are linear

in the applied electric field (the regime where Ohm's law holds) for two physical situations: (i) when $\omega_c\tau << 1$ [Hurd, 1972, p. 69] and (ii) when $\omega_c\tau >> 1$ [Hurd, 1972; Lifshitz et al., 1956] where $\omega_c = Be/m$ is the cyclotron frequency. Situation (ii) means the electron is able to complete many cyclotron orbits under the influence of **B** in the time between scatterings and is called the high (magnetic) field limit. Conversely, situation (i) is obtained when the electron is scattered in a short time compared to the time necessary to complete one cyclotron orbit and is known as the low field limit. In effect, the solution to Eq. (52.7) is obtained by expanding $g(\boldsymbol{k})$ in a power series in $\omega_c\tau$ or $1/\omega_c\tau$ for (i) and (ii), respectively. Given $g(\mathbf{k})$ the current vector, J_l ($l = x,y,z$) can be calculated from [Blatt, 1957]

$$J_l = \left(\frac{e}{4\pi^3}\right)\int v_l(\mathbf{k})\,g(\mathbf{k})\,(\partial f_0/\partial \mathscr{E})\,d^3k \tag{52.8}$$

where $v_l(\mathbf{k})$ is the velocity of the electron with wave vector **k.** Every term in the series defining J_l is linear in the applied electric field, **E,** so that the conductivity tensor σ_{lm} is readily obtained from $J_l = \sigma_{lm}E_m$ [Hurd, 1972, p. 9] This matrix equation can be inverted to give $E_l = \rho_{lm}J_m$. For the same geometry used in defining Eq. (52.1)

$$E_y = E_H = \rho_{yx}J_x \tag{52.9}$$

where ρ_{21} is a component of the resistivity tensor sometimes called the Hall resistivity. Comparing Eqs. (52.1) and (52.9) it is clear that the B dependence of E_H is contained in ρ_{12}. However, nothing in the derivation of ρ_{12} excludes the possibility of terms to the second or higher powers in B. Although these are usually small, this is one of the reasons that experimentally one usually obtains R from the measured transverse voltage by reversing magnetic fields and averaging the measured E_H by calculating $(1/2)[E_H(\mathbf{B}) - E_H(-\mathbf{B})]$. This eliminates the second-order term in B and in fact all even power terms contributing to the E_H. Using the Onsager relation [Smith and Jensen, 1989, p. 60] $\rho_{12}(\mathbf{B}) = \rho_{21}(-\mathbf{B})$, it is also easy to show that in terms of the Hall resistivity

$$R = \frac{1}{2}\frac{1}{B}[\rho_{12}(\mathbf{B}) + \rho_{21}(\mathbf{B})] \tag{52.10}$$

Strictly speaking, in a single crystal the electric field resulting from an applied electric current and magnetic field, both of arbitrary direction relative to crystal axes and each other, cannot be fully described in terms of a second-order resistivity tensor. [Hurd, 1972, p. 71] On the other hand, Eqs. (52.1), (52.9), and (52.10) do define the Hall coefficient in terms of a second-order resistivity tensor for a polycrystalline (assumed isotropic) sample or for a cubic single crystal or for a lower symmetry crystal when the applied fields are oriented along major symmetry directions. In real world applications the Hall effect is always treated in this manner.

52.3 Relation to the Electronic Structure — (i) $\omega_c\tau << 1$

General expressions for R in terms of the parameters that describe the electronic structure can be obtained using Eqs. (52.7)–(52.10) and have been given by Blatt [Blatt, 1957] for the case of crystals having cubic symmetry. An even more general treatment has been given by McClure [McClure, 1956]. Here the discussion of specific results will be restricted to the free electron model wherein the material is assumed to have one or more conducting bands, each of which has a quadratic dispersion relationship connecting $\mathscr{E}$ and **k;** that is

$$\mathscr{E}_i = \frac{\hbar^2 k_i^2}{2m_i} \tag{52.11}$$

where the subscript specifies the band number and m_i, the **effective mass** for each band. These masses need not be equal nor the same as the free electron mass. In effect, some of the features lost in the free electron

approximation are recovered by allowing the masses to vary. The **relaxation times**, τ_i, will also be taken to be isotropic (not **k** dependent) within each band but can be different from band to band. Although extreme, these approximations are often qualitatively correct, particularly in polycrystalline materials, which are macroscopically isotropic. Further, in semiconductors these results will be strictly applicable only if τ_i is energy independent as well as isotropic.

For a single spherical band, R_H is a direct measure of the number of current carriers and turns out to be given by [Blatt, 1957]

$$R_H = \frac{1}{ne} \tag{52.12}$$

where n is the number of conduction carriers/volume. R_H depends on the sign of the charge of the current carriers being negative for electrons and positive for **holes**. This identification of the carrier sign is itself a matter of great importance, particularly in semiconductor physics. If more than one band is involved in electrical conduction, then by imposing the boundary condition required for the geometry of Fig. 52.1 that the total current in the y direction from all bands must vanish, $J_y = 0$, it is easy to show that [Wilson, 1958]

$$R_H = (1/\sigma)^2 \Sigma[\sigma_i^2 R_i] \tag{52.13}$$

where R_i and σ_i are the Hall coefficient and electrical conductivity, respectively, for the ith band ($\sigma_i = n_i e^2 \tau_i / m_i$), $\sigma = \Sigma\sigma_i$ is the total conductivity of the material, and the summation is taken over all bands. Using Eq. (52.12), Eq. (52.13) can also be written

$$R_H = \frac{1}{en_{\text{eff}}} = \frac{1}{e} \sum \left[\frac{1}{n_i} \left(\frac{\sigma_i}{\sigma} \right)^2 \right] \tag{52.14}$$

where n_{eff} is the effective or apparent number of electrons determined by a Hall effect experiment. (Note that some workers prefer representing Eqs. (52.13) and (52.14) in terms of the current carrier mobility for each band, μ_i, defined by $\sigma_i = n_i e \mu_i$.)

The most commonly used version of Eq. (52.14) is the so-called two-band model, which assumes that there are two spherical bands with one composed of electrons and the other of holes. Eq. (52.14) then takes the form

$$R_H = \frac{1}{e} \left[\frac{1}{n_e} \left(\frac{\sigma_e}{\sigma} \right)^2 - \frac{1}{n_h} \left(\frac{\sigma_h}{\sigma} \right)^2 \right] \tag{52.15}$$

From Eq. (52.14) or (52.15) it is clear that the Hall effect is dominated by the most highly conducting band. Although for fundamental reasons it is often the case that $n_e = n_h$ (a so-called compensated material), R_H would rarely vanish since the conductivities of the two bands would rarely be identical. It is also clear from any of Eqs. (52.12), (52.14), or (52.15) that, in general, the Hall effect in semiconductors will be orders of magnitude larger than that in metals.

52.4 Relation to the Electronic Structure — (ii) $\omega_c\tau >> 1$

The high field limit can be achieved in metals only in pure, crystalographically well-ordered materials and at low temperatures, which circumstances limit the electron scattering rate from impurities, crystallographic

imperfections, and lattice vibrations, respectively. In semiconductors, the much longer relaxation time and smaller effective mass of the electrons makes it much easier to achieve the high field limit. In this limit the result analogous to Eq. (52.15) is [Blatt, 1968, p. 290]

$$R_H = \frac{1}{e}\frac{1}{n_e - n_h} \tag{52.16}$$

Note that the individual band conductivities do not enter in Eq. (52.16). Eq. (52.16) is valid provided the cyclotron orbits of the electrons are closed for the particular direction of **B** used. It is not necessary that the bands be spherical or the τ's isotropic. Also, for more than two bands R_H depends only on the net difference between the number of electrons and the number of holes. For the case where $n_e = n_h$, in general, the lowest order dependence of the Hall electric field on B is B^2 and there is no simple relationship of R_H to the number of current carriers. For the special case of the two-band model, however, R_H is a constant and is of the same form as Eq. (52.15) [Fawcett, 1964].

Metals can have geometrically complicated Fermi surfaces wherein the Fermi surface contacts the Brillouin zone boundary as well as encloses the center of the zone. This leads to the possibility of open electron orbits in place of the closed cyclotron orbits for certain orientations of **B.** In these circumstances R can have a variety of dependencies on the magnitude of B and in single crystals will generally be dependent on the exact orientation of **B** relative to the crystalline axes [Hurd, 1972, p. 51; Fawcett, 1964]. R will not, however, have any simple relationship to the number of current carriers in the material.

Semiconductors have too few electrons to have open orbits but can manifest complicated behavior of their Hall coefficient as a function of the magnitude of B. This occurs because of the relative ease with which one can pass from the low field limit to the high field limit and even on to the so-called quantum limit with currently attainable magnetic fields. (The latter has not been discussed here.) In general, these different regimes of B will not occur at the same magnitude of B for all the bands in a given semiconductor, further complicating the dependence of R on B.

Defining Terms

Conducting band: The band in which the electrons primarily responsible for the electric current are found.

Effective mass: An electron in a lattice responds differently to applied fields than would a free electron or a classical particle. One can, however, often describe a particular response using classical equations by defining an effective mass whose value differs from the actual mass. For the same material the effective mass may be different for different phenomena; e.g., electrical conductivity and cyclotron resonance.

Electron band: A range or band of energies in which there is a continuum (rather than a discrete set as in, for example, the hydrogen atom) of allowed quantum mechanical states partially or fully occupied by electrons. It is the continuous nature of these states that permits them to respond almost classically to an applied electric field.

Hole or hole state: When a conducting band, which can hold two electrons/unit cell, is more than half full, the remaining unfilled states are called *holes*. Such a band responds to electric and magnetic fields as if it contained positively charged carriers equal in number to the number of holes in the band.

Relaxation time: The time for a distribution of particles, out of equilibrium by a measure Φ, to return exponentially toward equilibrium to a measure Φ/e out of equilibrium when the disequilibrating fields are removed (e is the natural logarithm base).

Related Topic

22.1 Physical Properties

References

L. Berger and G. Bergmann, in *The Hall Effect and Its Applications,* C. L. Chien and C. R. Westlake, Eds., New York: Plenum Press, 1980, p. 55.

F. L. Blatt in *Solid State Physics,* vol. 4, F. Seitz and D. Turnbull, Eds., New York: Academic Press, 1957, p. 199.

F. L. Blatt, *Physics of Electronic Conduction in Solids,* New York: McGraw-Hill, 1968, p. 290. See also N. W. Ashcroft and N. D. Mermin in *Solid State Physics,* New York: Holt, Rinehart and Winston, 1976, p. 236.

D. Emin, *Phil. Mag.,* vol. 35, p. 1189, 1977.

E. Fawcett, *Adv. Phys.* vol. 13, p. 139, 1964.

C. M. Hurd, *The Hall Effect in Metals and Alloys,* New York: Plenum Press, 1972, p. 69.

I. M. Lifshitz, M. I. Azbel, and M. I. Kaganov, *Zh. Eksp. Teor. Fiz.,* vol. 31, p. 63, 1956 [*Soviet Phys. JETP* (Engl. Trans.), vol. 4, p. 41, 1956].

J. W. McClure, *Phys. Rev.,* vol. 101, p. 1642, 1956.

R. E. Prange and S. M. Girvin, Eds., *The Quantum Hall Effect,* New York: Springer-Verlag, 1990.

H. Smith, and H. H. Jensen, *Transport Phenomena,* Oxford: Oxford University Press, 1989, p. 60.

A. H. Wilson, *The Theory of Metals,* London: Cambridge University Press, 1958, p. 212.

J. M. Ziman, *Electrons and Phonons,* London: Oxford University Press, 1960. See also N. W. Ashcroft and N. D. Mermin in *Solid State Physics,* New York: Holt, Rinehart and Winston, 1976, chapters 12 and 16.

Further Information

In addition to the texts and review article cited in the references, an older but still valid article by J. P. Jan, in *Solid State Physics* (edited by F. Seitz and D. Turnbull, New York: Academic Press, 1957, p. 1) can provide a background in the various thermomagnetic and galvanomagnetic properties in metals. A parallel background for semiconductors can be found in the monograph by E. H. Putley, *The Hall Effect and Related Phenomena* (Boston: Butterworths, 1960).

Examples of applications of the Hall effect can be found in the book *Hall Generators and Magnetoresistors,* by H. H. Wieder, edited by H. J. Goldsmid (London: Pion Limited, 1971).

An index to the most recent work on or using any aspect of the Hall effect reported in the major technical journals can be found in *Physics Abstracts* (Science Abstracts Series A).

Delin, K.A., Orlando, T.P. “Superconductivity”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

53

Superconductivity

Kevin A. Delin
Jet Propulsion Laboratory

Terry P. Orlando
Massachusetts Institute of Technology

53.1 Introduction
53.2 General Electromagnetic Properties
53.3 Superconducting Electronics
53.4 Types of Superconductors

53.1 Introduction

The fundamental idea behind all of a superconductor's unique properties is that **superconductivity** is a quantum mechanical phenomenon on a macroscopic scale created when the motions of individual electrons are correlated. According to the theory developed by John Bardeen, Leon Cooper, and Robert Schrieffer (BCS theory), this correlation takes place when two electrons couple to form a Cooper pair. For our purposes, we may therefore consider the electrical charge carriers in a superconductor to be Cooper pairs (or more colloquially, superelectrons) with a mass m^* and charge q^* twice those of normal electrons. The average distance between the two electrons in a Cooper pair is known as the coherence length, ξ. Both the coherence length and the binding energy of two electrons in a Cooper pair, 2Δ, depend upon the particular superconducting material. Typically, the coherence length is many times larger than the interatomic spacing of a solid, and so we should not think of Cooper pairs as tightly bound electron molecules. Instead, there are many other electrons between those of a specific Cooper pair allowing for the paired electrons to change partners on a time scale of $h/(2\Delta)$ where h is Planck's constant.

If we prevent the Cooper pairs from forming by ensuring that all the electrons are at an energy greater than the binding energy, we can destroy the superconducting phenomenon. This can be accomplished, for example, with thermal energy. In fact, according to the BCS theory, the critical temperature, T_c, associated with this energy is

$$\frac{2\Delta}{k_B T_c} \approx 3.5 \tag{53.1}$$

where k_B is Boltzmann's constant. For low critical temperature (conventional) superconductors, 2Δ is typically on the order of 1 meV, and we see that these materials must be kept below temperatures of about 10 K to exhibit their unique behavior. High critical temperature superconductors, in contrast, will superconduct up to temperatures of about 100 K, which is attractive from a practical view because the materials can be cooled cheaply using liquid nitrogen. Other types of depairing energy are kinetic, resulting in a critical current density J_c, and magnetic, resulting in a critical field H_c. To summarize, a superconductor must be maintained under the appropriate temperature, electrical current density, and magnetic field conditions to exhibit its special properties. An example of this phase space is shown in Fig. 53.1.

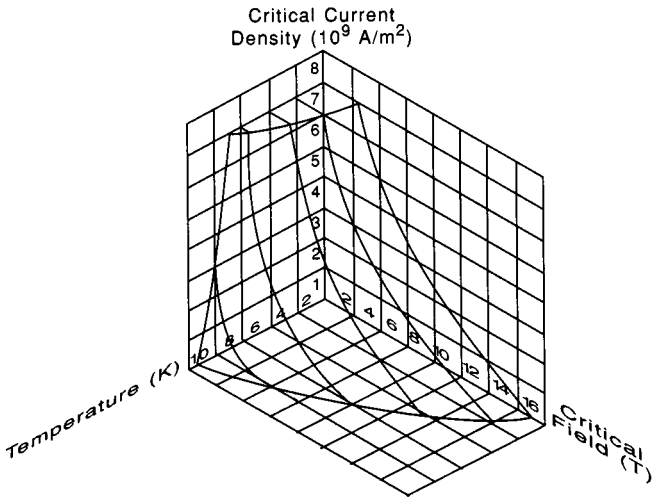


FIGURE 53.1 The phase space for the superconducting alloy niobium–titanium. The material is superconducting inside the volume of phase space indicated.

53.2 General Electromagnetic Properties

The hallmark electromagnetic properties of a superconductor are its ability to carry a static current without any resistance and its ability to exclude a static magnetic flux from its interior. It is this second property, known as the Meissner effect, that distinguishes a superconductor from merely being a perfect conductor (which conserves the magnetic flux in its interior). Although superconductivity is a manifestly quantum mechanical phenomenon, a useful classical model can be constructed around these two properties. In this section, we will outline the rationale for this classical model, which is useful in engineering applications such as waveguides and high-field magnets.

The zero dc resistance criterion implies that the superelectrons move unimpeded. The electromagnetic energy density, w, stored in a superconductor is therefore

$$w = \frac{1}{2}\varepsilon \mathrm{E}^2 + \frac{1}{2}\mu_o \mathrm{H}^2 + \frac{n^\star}{2} m^\star v_s^2 \tag{53.2}$$

where the first two terms are the familiar electric and magnetic energy densities, respectively. (Our electromagnetic notation is standard: ε is the permittivity, μ_o is the permeability, E is the electric field, and the magnetic flux density, B, is related to the magnetic field, H, via the constitutive law $\mathrm{B} = \mu_o \mathrm{H}$.) The last term represents the kinetic energy associated with the undamped superelectrons' motion ($n^\star$ and v_s are the superelectrons' density and velocity, respectively). Because the supercurrent density, J_s, is related to the superelectron velocity by $\mathrm{J}_s = n^\star q^\star \mathrm{v}_s$, the kinetic energy term can be rewritten

$$n^\star \left(\frac{1}{2} m^\star \mathrm{v}_s^2 \right) = \frac{1}{2} \Lambda \mathrm{J}_s^2 \tag{53.3}$$

where Λ is defined as

$$\Lambda = \frac{m^\star}{n^\star \left(q^\star \right)^2} \tag{53.4}$$

Assuming that all the charge carriers are superelectrons, there is no power dissipation inside the superconductor, and so Poynting's theorem over a volume V may be written

$$-\int_V \nabla \cdot \left(\mathbf{E} \times \mathbf{H}\right) dv = \int_V \frac{\partial w}{\partial t}\, dv \tag{53.5}$$

where the left side of the expression is the power flowing into the region. By taking the time derivative of the energy density and appealing to Faraday's and Ampère's laws to find the time derivatives of the field quantities, we find that the only way for Poynting's theorem to be satisfied is if

$$\mathbf{E} = \frac{\partial}{\partial t}\left(\Lambda \mathbf{J}_s\right) \tag{53.6}$$

This relation, known as the first London equation (after the London brothers, Heinz and Fritz), is thus necessary if the superelectrons have no resistance to their motion.

Equation (53.6) reveals that the superelectrons' inertia creates a lag between their motion and that of an applied electric field. As a result, a superconductor will support a time-varying voltage drop. The impedance associated with the supercurrent is therefore inductive and it will be useful to think of Λ as a kinetic inductance created by the correlated motion of the Cooper pairs.

If the first London equation is substituted into Faraday's law, $\nabla \times \mathrm{E} = -(\partial \mathrm{B}/\partial t)$, and integrated with respect to time, the second London equation results:

$$\nabla \times \left(\Lambda \mathbf{J}_s\right) = -\mathbf{B} \tag{53.7}$$

where the constant of integration has been defined to be zero. This choice is made so that the second London equation is consistent with the Meissner effect as we now demonstrate. Taking the curl of the quasi-static form of Ampère's law, $\nabla \times \mathrm{H} = \mathrm{J}_s$, results in the expression $\nabla^2 \mathrm{B} = -\mu_o \nabla \times \mathrm{J}_s$, where a vector identity, $\nabla \times \nabla \times \mathrm{C} = \nabla(\nabla \cdot \mathrm{C}) - \nabla^2 \mathrm{C}$; the constitutive relation, $\mathrm{B} = \mu_o \mathrm{H}$; and Gauss's law, $\nabla \cdot \mathrm{B} = 0$, have been used. By now appealing to the second London equation, we obtain the vector Helmholtz equation

$$\nabla^2 \mathbf{B} - \frac{1}{\lambda^2}\mathbf{B} = 0 \tag{53.8}$$

where the penetration depth is defined

$$\lambda \equiv \sqrt{\frac{\Lambda}{\mu_o}} = \sqrt{\frac{m^\star}{n^\star \left(q^\star\right)^2 \mu_o}} \tag{53.9}$$

From Eq. (53.8), we find that a flux density applied parallel to the surface of a semi-infinite superconductor will decay away exponentially from the surface on a spatial length scale of order λ. In other words, a bulk superconductor will exclude an applied flux as predicted by the Meissner effect.

The London equations reveal that there is a characteristic length λ over which electromagnetic fields can change inside a superconductor. This penetration depth is different from the more familiar skin depth of electromagnetic theory, the latter being a frequency-dependent quantity. Indeed, the penetration depth at zero temperature is a distinct material property of a particular superconductor.

Notice that λ is sensitive to the number of correlated electrons (the superelectrons) in the material. As previously discussed, this number is a function of temperature and so only at $T = 0$ do *all* the electrons that usually conduct ohmically participate in the Cooper pairing. For intermediate temperatures, $0 < T < T_c$, there

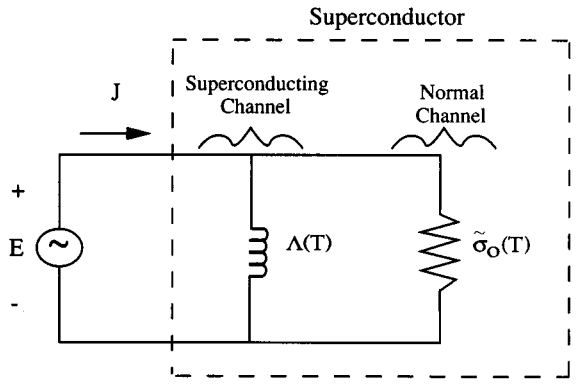


FIGURE 53.2 A lumped element model of a superconductor.

are actually two sets of interpenetrating electron fluids: the uncorrelated electrons providing ohmic conduction and the correlated ones creating supercurrents. This two-fluid model is a useful way to build temperature effects into the London relations.

Under the two-fluid model, the electrical current density, **J**, is carried by both the uncorrelated (normal) electrons and the superelectrons: $\mathbf{J} = \mathbf{J}_n + \mathbf{J}_s$ where $\mathbf{J}_n$ is the normal current density. The two channels are modeled in a circuit as shown in Fig. 53.2 by a parallel combination of a resistor (representing the ohmic channel) and an inductor (representing the superconducting channel). To a good approximation, the respective temperature dependences of the conductor and inductor are

$$\tilde{\sigma}_o(T) = \sigma_o(T_c)\left(\frac{T}{T_c}\right)^4 \quad \text{for } T \leq T_c \tag{53.10}$$

and

$$\Lambda(T) = \Lambda(0)\left(\frac{1}{1-(T/T_c)^4}\right) \quad \text{for } T \leq T_c \tag{53.11}$$

where σ_o is the dc conductance of the normal channel. (Strictly speaking, the normal channel should also contain an inductance representing the inertia of the normal electrons, but typically such an inductor contributes negligibly to the overall electrical response.) Since the temperature-dependent penetration depth is defined as $\lambda(T) = \sqrt{\Lambda(T)/\mu_o}$, the effective conductance of a superconductor in the sinusoidal steady state is

$$\sigma = \tilde{\sigma}_o + \frac{1}{j\omega\mu_o\lambda^2} \tag{53.12}$$

where the explicit temperature dependence notation has been suppressed.

It should be noted that the temperature dependencies given in Equations (53.10) and (53.11) are not precisely correct for the high-T_c materials. It has been suggested that this is because the angular momentum of the electrons forming a Cooper pair in high-T_c materials is different from that in low-T_c ones. Nevertheless, the two-fluid picture of transport and its associated constitutive law, Eq. (53.12), are still valid for high-T_c superconductors.

Most of the important physics associated with the classical model is embedded in Eq. (53.12). As is clear from the lumped element model, the relative importance of the normal and superconducting channels is a

function not only of temperature but also of frequency. The familiar *L/R* time constant, here equal to $\Lambda\tilde{\sigma}_o$, delineates the frequency regimes where most of the total current is carried by J_n (if $\omega\Lambda\tilde{\sigma}_o >> 1$) or J_s (if $\omega\Lambda\tilde{\sigma}_o << 1$). This same result can also be obtained by comparing the skin depth associated with the normal channel, $\delta = \sqrt{2/(\omega\mu_o\tilde{\sigma}_o)}$, to the penetration depth to see which channel provides more field screening. In addition, it is straightforward to use Eq. (53.12) to rederive Poynting's theorem for systems that involve superconducting materials:

$$-\int_V \nabla \cdot \left(\mathbf{E} \times \mathbf{H}\right) dv = \frac{d}{dt}\int_V \left(\frac{1}{2}\varepsilon \mathbf{E}^2 + \frac{1}{2}\mu_o \mathbf{H}^2 + \frac{1}{2}\Lambda(T)\mathbf{J}_s^2\right) dv + \int_V \frac{1}{\tilde{\sigma}_o(T)}\mathbf{J}_n^2\, dv \tag{53.13}$$

Using this expression, it is possible to apply the usual electromagnetic analysis to find the inductance (L_o), capacitance (C_o), and resistance (R_o) per unit length along a parallel plate transmission line. The results of such analysis for typical cases are summarized in Table 53.1.

53.3 Superconducting Electronics

The macroscopic quantum nature of superconductivity can be usefully exploited to create a new type of electronic device. Because all the superelectrons exhibit correlated motion, the usual wave–particle duality normally associated with a single quantum particle can now be applied to the entire ensemble of superelectrons. Thus, there is a spatiotemporal phase associated with the ensemble that characterizes the supercurrent flowing in the material.

If the overall electron correlation is broken, this phase is lost and the material is no longer a superconductor. There is a broad class of structures, however, known as weak links, where the correlation is merely perturbed locally in space rather than outright destroyed. Coloquially, we say that the phase "slips" across the weak link to acknowledge the perturbation.

The unusual properties of this phase slippage were first investigated by Brian Josephson and constitute the central principles behind superconducting electronics. Josephson found that the phase slippage could be defined as the difference between the macroscopic phases on either side of the weak link. This phase difference, denoted as ϕ, determined the supercurrent, i_s, through and voltage, v, across the weak link according to the Josephson equations,

$$i_s = I_c \sin\phi \tag{53.14}$$

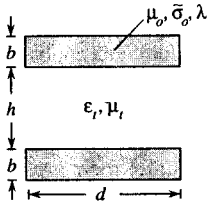
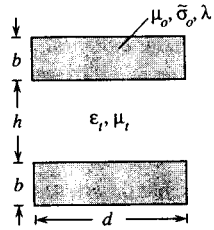
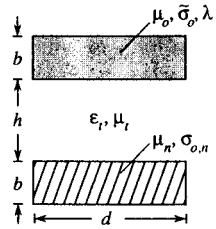
and

$$v = \frac{\Phi_o}{2\pi}\frac{d\phi}{dt} \tag{53.15}$$

where I_c is the critical (maximum) current of the junction and Φ_o is the quantum unit of flux. (The flux quantum has a precise definition in terms of Planck's constant, h, and the electron charge, e: $\Phi_o \equiv h/(2e) \approx 2.068 \times 10^{-15}$ Wb). As in the previous section, the correlated motion of the electrons, here represented by the superelectron phase, manifests itself through an inductance. This is straightforwardly demonstrated by taking the time derivative of Eq. (53.14) and combining this expression with Eq. (53.15). Although the resulting inductance is nonlinear (it depends on $\cos\phi$), its relative scale is determined by

$$L_j = \frac{\Phi_o}{2\pi I_c} \tag{53.16}$$

TABLE 53.1 Lumped Circuit Element Parameters Per Unit Length for Typical Transverse Electromagnetic Parallel Plate Waveguides*

Transmission Line Geometry	*Lo*	*Co*	*Ro*
 Two identical, thin ($\lambda \gg b$) superconducting plates	$\frac{\mu_t h}{d} + \frac{2\mu_o \lambda^2}{db}$	$\frac{\varepsilon_t d}{h}$	$\frac{8}{db\tilde{\sigma}_o}\left(\frac{\lambda}{\delta}\right)^4$
 Two identical, thick ($\lambda \ll b$) superconducting plates	$\frac{\mu_t h}{d} + \frac{2\mu_o \lambda}{d}$	$\frac{\varepsilon_t d}{h}$	$\frac{4}{d\delta\tilde{\sigma}_o}\left(\frac{\lambda}{\delta}\right)^3$
 One thick ($\lambda \ll b$) superconducting plate and one thick ($\lambda \ll b$) ohmic plate	$\frac{\mu_t h}{d} + \frac{\mu_o \lambda}{d} + \frac{\mu_n \delta_n}{2d}$	$\frac{\varepsilon_t d}{h}$	$\frac{1}{d\delta_n \sigma_{o,n}}$

*The subscript *n* refers to parameters associated with a normal (ohmic) plate. Using these expressions, line input impedance, attenuation, and wave velocity can be calculated.

a useful quantity for making engineering estimates. For example, the energy scale associated with Josephson coupling is $L_j I_c^2 = (I_c \Phi_o)/2\pi$.

A common weak link, known as the Josephson tunnel junction, is made by separating two superconducting films with a very thin (typically 20 Å) insulating layer. Such a structure is conveniently analyzed using the resistively and capacitively shunted junction (RCSJ) model shown in Fig. 53.3. Under the RCSJ model an ideal lumped junction [described by Eqs. (53.14) and (53.15)] and a resistor R_j represent how the weak link structure influences the respective phases of the super and normal electrons. A capacitor C_j represents the physical capacitance of the sandwich structure. Parasitic capacitance created by the fields around a device interacting with a dielectric substrate is also included in this lumped element. If the ideal lumped junction portion of the circuit is treated as an inductor-like element, many Josephson tunnel junction properties can be calculated with the familiar circuit time constants associated with the model. For example, the quality factor Q of the RCSJ circuit can be expressed as

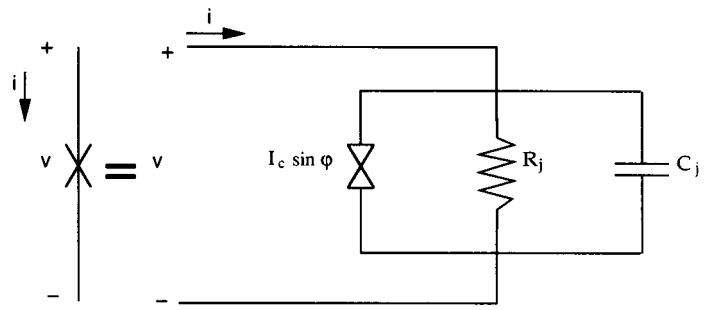


FIGURE 53.3 A real Josephson tunnel junction can be modeled using ideal lumped circuit elements.

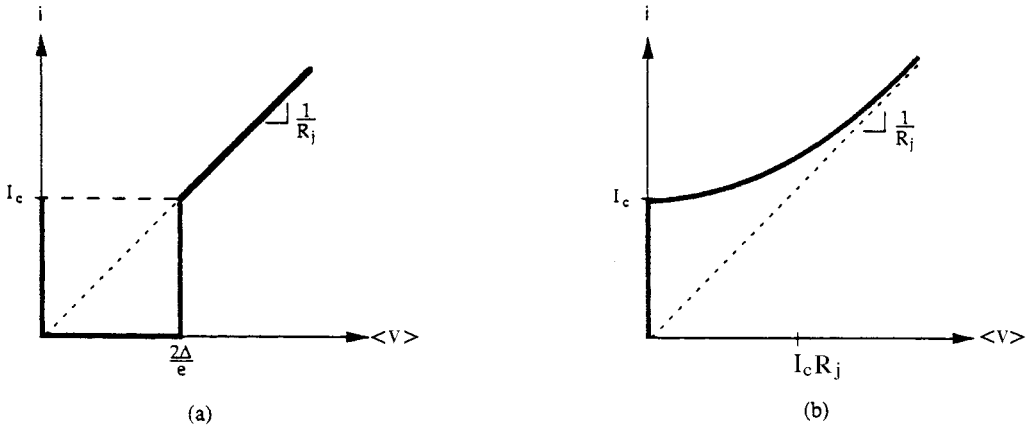


FIGURE 53.4 The *i-v* curves for a Josephson junction: (a) $\beta >> 1$, and (b) $\beta << 1$.

$$Q^2 = \frac{R_j C_j}{L_j / R_j} = \frac{2\pi I_c R_j^2 C_j}{\Phi_o} \equiv \beta \tag{53.17}$$

where β is known as the Stewart-McCumber parameter. Clearly, if $\beta >> 1$, the capacitive time constant R_jC_j dominates the dynamics of the circuit. Thus, as the bias current is raised from zero, no time-average voltage is created until the critical current I_c is exceeded. At this point, the junciton switches to a voltage consistent with the breaking of the Cooper pairs, $2\Delta/e$, with a time constant $\sqrt{L_jC_j}$. Once the junction has latched in the voltage state, however, the capacitor has charged up and the only way for it to discharge is to lower the bias current to zero again. As a result, a device with $\beta >>1$ will have a hysteretic current-voltage curve as shown in Fig. 53.4a. Conversely, $\beta << 1$ implies that the capacitance of the device is unimportant and so the current-voltage curve is not hysteretic (see Fig. 53.4b). In fact, the time-averaged voltage $\langle v \rangle$ for such an RSJ device is

$$\langle v \rangle = iR_j \sqrt{1 - \left(\frac{I_c}{i}\right)^2} \qquad \text{for } |i| > I_c. \tag{53.18}$$

In other words, once the supercurrent channel carries its maximum amount of current, the rest of the current is carried through the normal channel.

Just as the correlated motion of the superelectrons creates the frequency-independent Meissner effect in a bulk superconductor through Faraday's law, so too the macroscopic quantum nature of superconductivity

allows the possibility of a device whose output voltage is a function of a static magnetic field. If two weak links are connected in parallel, the lumped version of Faraday's law gives the voltage across the second weak link as $v_2 = v_1 + (d\Phi/dt)$, where Φ is the total flux threading the loop between the links. Substituting Eq. (53.15), integrating with respect to time, and again setting the integration constant to zero yields

$$\phi_2 - \phi_1 = \left(2\pi\Phi\right)/\Phi_o \tag{53.19}$$

showing that the spatial change in the phase of the macroscopic wavefunction is proportional to the local magnetic flux. The structure described is known as a *superconducting quantum interference device (SQUID)* and can be used as a highly sensitive magnetometer by biasing it with current and measuring the resulting voltage as a function of magnetic flux. From this discussion, it is apparent that a duality exists in how fields interact with the macroscopic phase: electric fields are coupled to its rate of change in time and magnetic fields are coupled to its rate of change in space.

53.4 Types of Superconductors

The macroscopic quantum nature of superconductivity also affects the general electromagnetic properties previously discussed. This is most clearly illustrated by the interplay of the characteristic lengths ξ, representing the scale of quantum correlations, and λ, representing the scale of electromagnetic screening. Consider the scenario where a magnetic field, H, is applied parallel to the surface of a semi-infinite superconductor. The correlations of the electrons in the superconductor must lower the overall energy of the system or else the material would not be superconducting in the first place. Because the critical magnetic field H_c destroys all the correlations, it is convenient to define the energy density gained by the system in the superconducting state as $(1/2)\mu_o H_c^2$. The electrons in a Cooper pair are separated on a length scale of ξ, however, and so the correlations cannot be fully achieved until a distance roughly ξ from the boundary of the superconductor. There is thus an energy per unit area, $(1/2)\mu_o H_c^2 \xi$, that is lost because of the presence of the boundary. Now consider the effects of the applied magnetic field on this system. It costs the superconductor energy to maintain the Meissner effect, $B = 0$, in its bulk; in fact the energy density required is $(1/2)\mu_o H^2$. However, since the field can penetrate the superconductor a distance roughly λ, the system need not expend an energy per unit area of $(1/2)\mu_o H^2 \lambda$ to screen over this volume. To summarize, more than a distance ξ from the boundary, the energy of the material is lowered (because it is superconducting), and more than a distance λ from the boundary the energy of the material is raised (to shield the applied field).

Now, if $\lambda < \xi$, the region of superconducting material greater than λ from the boundary but less than ξ will be higher in energy than that in the bulk of the material. Thus, the surface energy of the boundary is positive and so costs the total system some energy. This class of superconductors is known as type I. Most elemental superconductors, such as aluminum, tin, and lead, are type I. In addition to having $\lambda < \xi$, type I superconductors are generally characterized by low critical temperatures ($\sim$5 K) and critical fields ($\sim$0.05 T). Typical type I superconductors and their properties are listed in Table 53.2.

TABLE 53.2 Material Parameters for Type I Superconductors*

Material	T_c (K)	λ_o (nm)	ξ_o (nm)	Δ_o (meV)	$\mu_0 H_{co}$ (mT)
Al	1.18	50	1600	0.18	110.5
In	3.41	65	360	0.54	123.0
Sn	3.72	50	230	0.59	130.5
Pb	7.20	40	90	1.35	180.0
Nb	9.25	85	40	1.50	198.0

*The penetration depth λ_o is given at zero temperature, as are the coherence length ξ_o, the thermodynamic critical field H_{co}, and the energy gap Δ_o.

Source: R.J. Donnelly, "Cryogenics," in *Physics Vade Mecum,* H.L. Anderson, Ed., New York: American Institute of Physics, 1981. With permission.

Conversely, if $\lambda > \xi$, the surface energy associated with the boundary is negative and lowers the total system energy. It is therefore thermodynamically favorable for a normal–superconducting interface to form inside these type II materials. Consequently, this class of superconductors does not exhibit the simple Meissner effect as do type I materials. Instead, there are now two critical fields: for applied fields below the lower critical field, H_{c1}, a type II superconductor is in the Meissner state, and for applied fields greater than the upper critical field, H_{c2}, superconductivity is destroyed. The three critical field are related to each other by $H_c \approx \sqrt{H_{c1}H_{c2}}$.

In the range $H_{c1} < H < H_{c2}$, a type II superconductor is said to be in the vortex state because now the applied field can enter the bulk superconductor. Because flux exists in the material, however, the superconductivity is destroyed locally, creating normal regions. Recall that for type II materials the boundary between the normal and superconducting regions lowers the overall energy of the system. Therefore, the flux in the superconductor creates as many normal–superconducting interfaces as possible without violating quantum criteria. The net result is that flux enters a type II superconductor in quantized bundles of magnitude Φ_o known as vortices or fluxons (the former name derives from the fact that current flows around each quantized bundle in the same manner as a fluid vortex circulates around a drain). The central portion of a vortex, known as the core, is a normal region with an approximate radius of ξ. If a defect-free superconductor is placed in a magnetic field, the individual vortices, whose cores essentially follow the local average field lines, form an ordered triangular array, or flux lattice. As the applied field is raised beyond H_{c1} (where the first vortex enters the superconductor), the distance between adjacent vortex cores decreases to maintain the appropriate flux density in the material. Finally, the upper critical field is reached when the normal cores overlap and the material is no longer superconducting. Indeed, a precise calculation of H_{c2} using the phenomenological theory developed by Vitaly Ginzburg and Lev Landau yields

$$H_{c2} = \frac{\Phi_o}{2\pi\mu_o\xi^2} \tag{53.20}$$

which verifies our simple picture. The values of typical type II material parameters are listed in Tables 53.3 and 53.4.

Type II superconductors are of great technical importance because typical H_{c2} values are at least an order of magnitude greater than the typical H_c values of type I materials. It is therefore possible to use type II materials to make high-field magnet wire. Unfortunately, when current is applied to the wire, there is a Lorentz-like force on the vortices, causing them to move. Because the moving vortices carry flux, their motion creates a static voltage drop along the superconducting wire by Faraday's law. As a result, the wire no longer has a zero dc

TABLE 53.3 Material Parameters for Conventional Type II Superconductors*

Material	T_c (K)	$\lambda_{GL}(0)$ (nm)	$\xi_{GL}(0)$ (nm)	Δ_o (meV)	$\mu_0 H_{c2,o}$ (T)
Pb-In	7.0	150	30	1.2	0.2
Pb-Bi	8.3	200	20	1.7	0.5
Nb-Ti	9.5	300	4	1.5	13.0
Nb-N	16.0	200	5	2.4	15.0
$PbMo_6S_8$	15.0	200	2	2.4	60.0
V_3Ga	15.0	90	2–3	2.3	23.0
V_3Si	16.0	60	3	2.3	20.0
Nb_3Sn	18.0	65	3	3.4	23.0
Nb_3Ge	23.0	90	3	3.7	38.0

*The values are only representative because the parameters for alloys and compounds depend on how the material is fabricated. The penetration depth $\lambda_{GL}(0)$ is given as the coefficient of the Ginzburg–Landau temperature dependence as $\lambda_{GL}(T) = \lambda_{GL}(0)(1 - T/T_c)^{-1/2}$; likewise for the coherence length where $\xi_{GL}(T) = \xi_{GL}(0)(1 - T/T_c)^{-1/2}$. The upper critical field $H_{c2,o}$ is given at zero temperature as well as the energy gap Δ_o.

Source: R.J. Donnelly, "Cryogenics," in *Physics Vade Mecum,* H.L. Anderson, Ed., New York: American Institute of Physics, 1981. With permission.

TABLE 53.4 Type II (High-Temperature Superconductors)

Material	T_c (K)	$\lambda_{a,b}$ (nm)	λ_c (nm)	$\xi_{a,b}$ (nm)	ξ_c (nm)
$LuNi_2B_2C$	17	71		6	
Rb_3C_{60}	33	300		3	
$YBa_2Cu_3O_7$	95	150	1350	3	0.2
$Bi_2Sr_2CaCu_2O_8$	85	25	500	4.5	0.2
$Bi_2Sr_2Ca_2Cu_3O_{10}$	110				
$Tl_2Ba_2Ca_2Cu_3O_{10}$	125				
$HgBaCaCu_2O_6$	115	150		2.5	
$HgBa_2Ca_2Cu_3O_8$	135				

resistance, even though the material is still superconducting. To fix this problem, type II superconductors are usually fabricated with intentional defects, such as impurities or grain boundaries, in their crystalline structure to pin the vortices and prevent vortex motion. The pinning is created because the defect locally weakens the superconductivity in the material, and it is thus energetically favorable for the normal core of the vortex to overlap the nonsuperconducting region in the material. Critical current densities usually quoted for practical type II materials, therefore, really represent the depinning critical current density where the Lorentz-like force can overcome the pinning force. (The depinning critical current density should not be confused with the depairing critical current density, which represents the current when the Cooper pairs have enough kinetic energy to overcome their correlation. The depinning critical current density is typically an order of magnitude less than the depairing critical current density, the latter of which represents the theoretical maximum for J_c.)

By careful manufacturing, it is possible to make superconducting wire with tremendous amounts of current-carrying capacity. For example, standard copper wire used in homes will carry about 10^7 A/m^2, whereas a practical type II superconductor like niobium–titanium can carry current densities of 10^{10} A/m^2 or higher even in fields of several teslas. This property, more than a zero dc resistance, is what makes superconducting wire so desirable.

Defining Terms

Superconductivity: A state of matter whereby the correlation of conduction electrons allows a static current to pass without resistance and a static magnetic flux to be excluded from the bulk of the material.

Related Topic

35.1 Maxwell Equations

References

A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect,* New York: Wiley, 1982.

R. J. Donnelly, "Cryogenics," in *Physics Vade Mecum,* H.L. Anderson, Ed., New York: American Institute of Physics, 1981.

S. Foner and B. B. Schwartz, *Superconducting Machines and Devices,* New York: Plenum Press, 1974.

S. Foner and B. B. Schwartz, *Superconducting Materials Science,* New York: Plenum Press, 1981.

J. Knuutila, M. Kajola, H. Seppä, R. Mutikainen, and J. Salmi, Design, optimization, and construction of a DC SQUID with complete flux transformer circuits, *J. Low. Temp. Phys.,* 71, 369–392, 1988.

K. K. Likharev, *Dynamics of Josephson Junctions and Circuits,* Philadelphia, Pa.: Gordon and Breach Science Publishers, 1986.

T. P. Orlando and K. A. Delin, *Foundations of Applied Superconductivity,* Reading, Mass.: Addison-Wesley, 1991.

S. T. Ruggiero and D. A. Rudman, *Superconducting Devices,* Boston: Academic Press, 1990.

B. B. Schwartz and S. Foner, *Superconducting Applications: SQUIDs and Machines,* New York: Plenum Press, 1977.

T. Van Duzer and C. W. Turner, *Principles of Superconductive Devices and Circuits,* New York: Elsevier North Holland, 1981.

H. Weinstock and R. W. Ralston, *The New Superconducting Electronics,* Boston, Mass: Kluwer Academic Publishers, 1993.
J. P. Wikswo, SQUID magnetometers for biomagnetism and non-destructive testing: important questions and initial answers, *IEEE Trans. Appl. Supercond.,* 5, 74–120, 1995.
M. N. Wilson, *Superconducting Magnets,* Oxford: Oxford University Press, 1983.

Further Information

Every two years an Applied Superconductivity Conference is held devoted to practical technological issues. The proceedings of these conferences have been published every other year from 1977 to 1991 in the *IEEE Transactions on Magnetics.*

In 1991, the *IEEE Transactions on Applied Superconductivity* began publication. This quarterly journal focuses on both the science and the technology of superconductors and their applications, including materials issues, analog and digital circuits, and power systems. The proceedings of the Applied Superconductivity Conference now appear in this journal.

Whatmore, R.W. “Pyroelectric Materials and Devices”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

54

Pyroelectric Materials and Devices

54.1 Introduction
54.2 Polar Dielectrics
54.3 The Pyroelectric Effect
54.4 Pyroelectric Materials and Their Selection

Roger W. Whatmore
Cranfield University

54.1 Introduction

It was known over 2000 years ago that certain minerals such as tourmaline would attract small objects when heated. It was understood over 200 years ago that this attraction was a manifestation of the appearance of electrical charges on the surface as a consequence of the temperature change. This is called the **pyroelectric** effect and over the last 15 years has become the basis of a major worldwide industry manufacturing detectors of infrared radiation. These are exploited in such devices as "people detectors" for intruder alarms and energy conservation systems, fire and flame detectors, spectroscopic gas analyzers—especially looking for pollutants from car exhausts—and, more recently, devices for thermal imaging. Such thermal imagers can be used for night vision and, by exploiting the smoke-penetrating properties of long-wavelength infrared radiation, in devices to assist firefighters in smoke-filled spaces. The major advantages of the devices in comparison with the competing infrared detectors that exploit narrow bandgap semiconductors are that no cooling is necessary and that they are cheap and consume little power.

The pyroelectric effect appears in any material which possesses a polar symmetry axis. This chapter describes the basic effect, gives a brief account of how it can be used in radiation detection, and discusses the criteria by which materials can be selected for use in this application, concluding with a comparison of the properties of several of the most commonly used materials.

54.2 Polar Dielectrics

A polar material is one whose crystal structure contains a unique axis, along which an electric dipole moment will exist. There are 10 polar crystal classes:

• Triclinic	1	• Monoclinic	2, m
• Tetragonal	4, 4mm	• Orthorhombic	mm2
• Hexagonal	6, 6mm	• Trigonal	3, 3m

All crystals whose structures possess one of these symmetry groups will exhibit both pyroelectric and **piezoelectric** characteristics. In **ferroelectrics,** which are a subset of the set of pyroelectrics, the orientation of the polar axis can be changed by application of an electric field of sufficient magnitude. The original and final states of the crystal are symmetrically related. It is important to note that

TABLE 54.1 Spontaneous Polarizations and Curie Temperatures for a Range of Ferroelectrics

Material	T_c(k)	P_s(cm-2)	T(k)
KH_2PO_4 (KDP)	123	0.053	96
Triglycine sulphate	322	0.028	293
Polyvinylidene fluoride (PVDF)	>453	0.060	293
DOBAMBC (liquid crystal)	359	$\sim 3 \times 10^{-5}$	354
$PbTiO_3$	763	0.760	293
$BaTiO_3$	393	0.260	296

1. Not all polar materials are ferroelectric.
2. There is a set of point groups which lack a center of symmetry, without possessing a polar axis. The crystals belonging to these groups (222, $\bar{4}$, 422, $\bar{4}2$m, 32, $\bar{6}$, $\bar{6}$m2, 23, and $\bar{4}3$m) are piezoelectric without being pyroelectric. (432 is a noncentrosymmetric, nonpiezoelectric class.)

Typical values of spontaneous polarizations (P_s) and Curie temperatures (T_c) for a range of ferroelectrics are given in Table 54.1.

A very wide range of materials exhibit ferroelectric, and thus pyroelectric, behavior. These range from crystals, such as potassium dihydrogen phosphate and triglycine sulphate, to polymers, such as polyvinylidene fluoride, and liquid crystals such as DOBAMBC and ceramics, such as barium titanate and lead zirconate titanate.

Most ferroelectrics exhibit a Curie temperature (T_c) at which the spontaneous plarization goes to zero. (A few ferroelectrics, such as the polymer polyvinylidene fluoride [PVDF] melt before this temperature is reached.)

The fact that the orientation of the polar axis in ferroelectrics can be changed by the application of a field has a very important consequence for ceramic materials. If a polycrystalline body is made of a polar material, then the crystal axes will, in general, be randomly oriented. It cannot therefore show pyroelectricity. However, if an electric field greater than the **coercive field** (E_c) is applied to a ferroelectric ceramic, then the polar axes within the grains will tend to be reoriented so that they each give a component along the direction of the applied field. This process is called "poling." The resulting ceramic is polar (with a point symmetry ∞m) and will show both piezoelectricity and pyroelectricity.

54.3 The Pyroelectric Effect

The pyroelectric effect is described by:

$$P_i = p_i \Delta T \tag{54.1}$$

where P_i is the change in the coefficient of the polarization vector due to a change in temperature ΔT and p_i is the pyroelectric coefficient, which is a vector. The effect and its applications have been extensively reviewed in Whatmore [1986]. The effect of a temperature change on a pyroelectric material is to cause a current, i_p, to flow in an external circuit, such that

$$i_p = ApdT/dt \tag{54.2}$$

where A is the electroded area of the material, p the component of the pyroelectric coefficient normal to the electrodes, and dT/dt the rate of change of temperature with time.

Pyroelectric devices detect changes in temperature in the sensitive material and as such are detectors of supplied energy. It can be seen that the pyroelectric current is proportional to the rate of change of the material with time and that in order to obtain a measurable signal, it is necessary to modulate the source of energy. As energy detectors, they are most frequently applied to the detection of incident electromagnetic energy, particularly in the infrared wavebands.

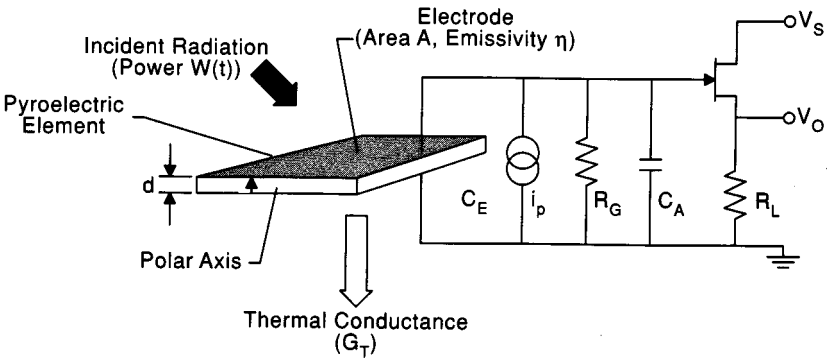


FIGURE 54.1 Pyroelectric detector with FET amplifier.

Typically, a pyroelectric detector element will consist of a thin chip of the pyroelectric material cut perpendicular to the polar axis of the material, electroded with a conducting material such as an evaporated metal and connected to a low-noise, high-input impedance amplifier—for example, a junction field-effect transistor (JFET) or metal-oxide gate transistor (MOSFET)—as shown in Fig. 54.1. In some devices, the radiation is absorbed directly in the element. In this case the front electrode will be a thin metal layer matched to the permittivity of free space with an electrical surface resistivity of 367 Ω/square. However, in most high-performance devices, the element is coated with a layer designed to absorb the radiation of interest. The element itself must be thin to minimize the thermal mass and, in most cases, well isolated thermally from its environment. These measures are designed to increase the temperature change for a given amount of energy absorbed and thus the electrical signal generated. The necessary modulation of the radiation flux can be achieved either by deliberately only "looking" for moving objects or other radiation sources (e.g., flickering flames for a flame detector) or by interposing a mechanical radiation "chopper" such as a rotating blade.

The voltage responsivity of a device such as this is defined as $R_v = V_o/W$, where V_o is the output voltage and W is the input radiation power. For radiation sinusoidally modulated at a frequency ω, R_v is given by

$$R_v = \frac{R_G \eta p A \omega}{G_T (1 + \omega^2 \tau_T^2)^{1/2} (1 + \omega^2 \tau_E^2)^{1/2}} \tag{54.3}$$

where G_T is the thermal conductance from the element to the environment, τ_T is the thermal time constant of the element, τ_E is the electrical time constant of the element, R_G is the electrical resistance across the element, η is the emissivity of the element for the radiation being detected, and A is the sensitive area of the element.

It is easy to show that the response of a pyroelectric device maximizes at a frequency equal to the inverse of the geometric mean of the two time constants and that above and below the two frequencies given by τ_T^{-1} and τ_E^{-1}, R_v falls as ω^{-1}. The consequence of this is that pyroelectric detectors have their sensitivities maximized by having fairly long electrical time constants (0.1 to 10 s) and that such detectors thus work best at low frequencies (0.1 to 100 Hz). However, if high sensitivity is not required, extremely large bandwidths with little sensitivity variation can be obtained by shortening these time constants (making R_G and C_E low and G_T high). In this way, detectors have been made which give picosecond time responses for tracking fast laser pulses.

There are several noise sources in a pyroelectric device. These are discussed in detail in Whatmore [1986]. In many cases of interest, the dominant noise source is the Johnson noise generated by the ac conductance in the capacitance of the detector element. This noise is given by ΔV_j, where

$$\Delta V_j = \left\{ 4kT \frac{\tan\delta}{C_E} \right\}^{1/2} \omega^{-1/2} \qquad \text{for } C_E >> C_A \tag{54.4}$$

where k is Boltzmann's constant, T is the absolute temperature, tanδ is the dielectric loss tangent of the detector material, C_E is the electrical capacitance of the element, and C_A is the input capacitance of the detector element amplifier.

The input radiation power required to give an output equal to the noise at a given frequency in unity bandwidth is known as the noise equivalent power (NEP). This is given by

$$NEP = V_n / R_v \tag{54.5}$$

where V_n is the total RMS voltage noise from all sources.

A performance figure of merit frequently used when discussing infrared detectors is the detectivity, usually designated as D^*. This is given by

$$D^* = A^{1/2}/NEP \tag{54.6}$$

Thus, the detectivity of a pyroelectric detector can be derived from Eqs. (54.3) to (54.6) and is given by

$$D^* = \frac{\eta d}{(4kT)^{1/2}} \cdot \frac{p}{c'(\epsilon\epsilon_o \tan\delta)^{1/2}} \cdot \frac{1}{\omega^{1/2}} \tag{54.7}$$

where c' is the volume specific heat, ϵ is the dielectric constant of the pyroelectric, and d is the thickness of the pyroelectric element. The roll-off in D^* at high frequencies is thus $1/\omega^{1/2}$.

Pyroelectric single-element IR detectors come in many different varieties. A typical commercial device will have the sensitive element made from a material of the type discussed in the next section, such as a piece of lithium tantalate crystal or a ferroelectric ceramic. The element size will be a few millimeters square. Typical performance figures at about 10 Hz would be a responsivity of a few hundred volts per watt of input radiation, a noise equivalent power of about 8×10^{-9} W/Hz$^{1/2}$, and a detectivity of about 2×10^8 cm Hz$^{1/2}$ W^{-1} for unity bandwidth. The detector can be fitted with a wide variety of windows, depending upon the wavelength of the radiation to be detected.

As noted above, pyroelectric devices have also been used for thermal imaging. In this application, their main advantage when compared with photon detector materials such as mercury cadmium telluride (CMT) (which are more sensitive) is that they can be used at room temperature. All the photon detectors require cooling, typically to 77 K. A very successful device for pyroelectric thermal imaging is the pyroelectric vidicon which uses a thin plate of pyroelectric material contained in a vacuum tube. The thermal image is focused onto the surface of the material using a germanium lens. This causes the formation of a pattern of pyroelectric charges, which are "read" by means of an electron beam. Typical sensitivities for such devices are between 0.50 and 1 K temperature differences in the scene for an f/1 lens. This compares with <0.10 K for a cooled CMT detector-based imager. Recently, a solid-state approach to pyroelectric thermal imaging has been developed. In this, an array of many thousands of very small identical detectors, each between 50 and 100 μm square, depending on the array design, are linked to a silicon amplifier/multiplexer circuit which allows the signals from all the elements to be read onto a single output line. These devices have been primarily developed for thermal imaging applications and excellent sensitivities (close to those achieved by many cooled systems) have been demonstrated.

54.4 Pyroelectric Materials and Their Selection

There are many different types of pyroelectrics and the selection of a material depends strongly upon the application. It is possible to formulate from the given equations a number of figures of merit which describe the contribution of the physical properties of a material to the performance of a device. For example, the current responsivity is proportional to F_i:

$$F_i = p/c' \tag{54.8}$$

TABLE 54.2 Pyroelectric Properties of Selected Materials

Material (Temperature)	Pyroelectric Coefficient P 10^{-4} cm^{-2} K^{-1}	Dielectric Properties (1 kHz) ϵ	tanδ	Volume-Specific Heat c' 10^{6} Jm^{-3} K^{-1}	Thermal Conductivity K 10^{-7} m^{2} s^{-1}	F_v m^{2} C^{-1}	F_D 10^{-5} Pa$^{-1/2}$	F_{vid} 10^{6} sC^{-1}
TGS (35°C)	5.5	55	0.025	2.6	3.3	0.43	6.1	1.3
DTGS (40°C)	5.5	43	0.020	2.4	3.3	0.60	8.3	1.8
PVDF polymer	0.27	12	0.015	2.43	0.62	0.10	0.88	1.6
$LiTaO_3$ crystal	2.3	47	0.005	3.2	13.0	0.17	4.9	0.13
Modified PZ ceramic	3.8	290	0.003	2.5		0.06	5.8	
Modified PT ceramic	3.8	220	0.011	2.5		0.08	3.3	

PZ = $PbZrO_3$, PT = $PbTiO_3$.

The voltage response for a pyroelectric element feeding into a high-input impedance, unity gain amplifier (such as a source follower FET) as shown in Fig. 54.1 is proportional to F_v:

$$F_v = p/c'\epsilon\epsilon_o \tag{54.9}$$

The detectivity is proportional to F_D:

$$F_D = p/\{c'(\epsilon\epsilon_o \tan\delta)^{1/2}\} \tag{54.10}$$

For the pyroelectric vidicon, thermal spreading of the pattern on the target is important and the relevant figure of merit is F_{vid}:

$$F_{vid} = F_v/K \tag{54.11}$$

where K is the thermal conductivity of the pyroelectric. It should be noted that the use of these merit figures must be tempered with a knowledge of the type of detector the material is to be used in. It is necessary, if possible, to match the capacitance of the detector to the input capacitance of the amplifier. Hence, low-permittivity materials are better suited to large-area detectors, and conversely arrays of small-area detectors are better served by materials with high permittivities.

Table 54.2 lists the pyroelectric properties of several different materials, single crystals, ceramics, and polymers. It can be seen that triglycine sulphate (TGS) and its deuterated isomorph (DTGS) exhibit the highest value of F_v and are frequently used for high-performance single-element detectors. These are the preferred materials for pyroelectric vidicon targets. However, they are water soluble, difficult to handle, and show poor long-term stability, both chemically and electrically, because of their low **Curie temperatures.** Furthermore, their dielectric loss is rather high, so that the F_D figures are not so favorable. Lithium tantalate, on the other hand, is an oxide single-crystal material which possesses a relatively low value of F_v but a very low loss so that F_D is favorable. The material is very stable and is now widely used for single-element detectors. Its thermal conductivity is quite high so that it is not a good material for the pyroelectric vidicon. The ferroelectric polymers possess relatively low pyroelectric coefficients and low dielectric constants with high losses, so that their figures of merit are also quite low. Their low thermal conductivities make them quite favorable for use in the pyroelectric vidicon and the fact that they are commercially available in thin sections (down to 6 μm) at low cost, removing any requirement for expensive lapping and polishing, makes them attractive for some low-cost detectors. Their low permittivities make them particularly well suited to large-area detectors.

The ceramic materials modified lead zirconate and modified lead titanate are interesting in that they possess high pyroelectric coefficients with relatively high permittivities and low losses. The modified lead zirconate is a solid solution of lead zirconate with lead iron niobate and lead titanate, with small additions of uranium as a stabilizing dopant. The use of uranium in this material minimizes the dielectric constant and loss (thus maximizing F_D) while also permitting control over the electrical resistivity, allowing the gate bias resistor in

Fig. 54.1 to be designed into the sensor element. The modified lead titanate is doped with calcium titanate and lead cobalt tungstate. The use of hot pressing in ceramic manufacture permits the fabrication of very low porosity material, which can be lapped and polished to very thin sections (as low as 20 μm) while being mechanically strong enough to be placed on a mount which provides support only over a small area, permitting the fabrication of detectors with maximum sensitivity. While the F_v values are relatively small in these materials, the F_D values are as good as most of the single-crystal materials. They are very well suited to small-area detectors because their high dielectric constants enable the element capacitance to be matched to that of the amplifier. Pyroelectric ceramics are now finding use in a wide range of the infrared detector market, from low-cost intruder alarms to high-value imaging arrays.

Recently, a new class of pyroelectric materials which use the effect in the region of T_c have been developed [Whatmore, 1991]. In these materials a bias field must be applied to stabilize the effect, but F_D values as high as 10 to 15 $\times$ 10^{-5} $Pa^{-1/2}$ have been recorded in such materials as barium strontium titanate or lead scandium tantalate, both perovskite ceramics. This mode of operation is usually called "dielectric bolometer."

Defining Terms

Coercive field (E_c): The field required to invert a sufficient proportion of the polarization of a body of a poled ferroelectric such that the net measurable external dipole moment is zero.

Curie temperature (T_c): The temperature at which the spontaneous polarization of a ferroelectric goes to zero.

Ferroelectric: A polar dielectric in which the crystallographic orientation of the internal dipole moment can be changed by the application of an electric field.

Paraelectric: The nonpolar phase into which the ferroelectric transforms above T_c, frequently called the paraelectric phase.

Piezoelectric: A material which possesses a noncentrosymmetric crystal structure which will generate charge on the application of a mechanical stress. As in the case of a pyroelectric, this can be detected as either a potential difference or as a charge flowing in an external circuit.

Pyroelectric: A polar dielectric material in which the internal dipole moment is temperature dependent. This leads to a change in the charge balance at the surface of the material which can be detected as either a potential difference or as a charge flowing in an external circuit.

Remanent polarization: The value to which the externally measured polarization of a ferroelectric body relaxes after it has been subjected to an electric field much greater than the coercive field, which is then removed.

Saturation polarization: The value to which the externally measured electrical dipole moment of a ferroelectric body tends when subjected to an external electrical field greater than the coercive field.

Spontaneous polarization: The value of the electrical dipole moment of a ferroelectric crystal due to the separation of positive and negative charges within the unit cell.

Related Topic

55.1 Introduction

References

R. W. Whatmore, "Pyroelectric devices and materials," *Rep. Prog. Phys.,* vol. 49, pp. 1335–1386, 1986.

R. W. Whatmore, "Pyroelectic ceramics and devices for thermal infrared detection and imaging," *Ferroelectrics,* vol. 118, pp. 241–259, 1991.

B. M. Kulwicki, A. Amin, H. R. Beratan, and C. M. Hanson, "Pyroelectric imaging", Proc. 8th IEEE International Symposium on Applications of Ferroelectrics, p. 1–10, (IEEE Cat. No. 92CH3080-9), 1992.

A. Hadni, "Applications of the pyroelectric effect," *J. Phys. E: Sci. Instrum.,* 14, 1233–1240, 1981.

W.-S. Zhu, J. R. Izatt, and B. K. Deka, "Pyroelectric detection of submicrosecond laser pulses between 230 and 530 μm," *Appl. Opt.,* 28, 3647–3651, 1989.

Bartnikas, R. “Dielectrics and Insulators”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

55

Dielectrics and Insulators

55.1 Introduction
55.2 Dielectric Losses
55.3 Dielectric Breakdown
55.4 Insulation Aging
55.5 Dielectric Materials
Gases • Insulating Liquids • Solid Insulating Materials • Solid-Liquid Insulating Systems

R. Bartnikas
Institut de Recherche d'Hydro-Québec

55.1 Introduction

Dielectrics are materials that are used primarily to isolate components electrically from each other or ground or to act as capacitive elements in devices, circuits, and systems. Their insulating properties are directly attributable to their large energy gap between the highest filled valence band and the conduction band. The number of electrons in the conduction band is extremely low, because the energy gap of a dielectric (5 to 7 eV) is sufficiently large to maintain most of the electrons trapped in the lower band. As a consequence, a dielectric, subjected to an electric field, will evince only an extremely small conduction or loss current; this current will be caused by the finite number of free electrons available in addition to other free charge carriers (ions) associated usually with contamination by electrolytic impurities as well as dipole orientation losses arising with polar molecules under ac conditions. Often the two latter effects will tend to obscure the miniscule contribution of the relatively few free electrons available. Unlike solids and liquids, vacuum and gases (in their nonionized state) approach the conditions of a perfect insulator—i.e., they exhibit virtually no detectable loss or leakage current.

Two fundamental parameters that characterize a dielectric material are its **conductivity** σ and the value of the real permittivity or dielectric constant ε'. By definition, σ is equal to the ratio of the leakage current density J_l to the applied electric field E,

$$\sigma = \frac{J_l}{E} \tag{55.1}$$

Since J_l is in A cm^{-2} and E in V cm^{-1}, the corresponding units of σ are in S cm^{-1} or Ω^{-1} cm^{-1}. Alternatively, when only mobile charge carriers of charge e and mobility μ, in cm^2 V^{-1} s^{-1}, with a concentration of n per cm^3 are involved, the conductivity may be expressed as

$$\sigma = e\mu n \tag{55.2}$$

The conductivity is usually determined in terms of the measured insulation resistance R in Ω; it is then given by $\sigma = d/RA$, where d is the insulation thickness in cm and A the surface area in cm^2. Most practical insulating materials have conductivities ranging from 10^{-6} to 10^{-20} S cm^{-1}. Often dielectrics may be classified in terms of their resistivity value ρ, which by definition is equal to the reciprocal of σ.

The real value of the permittivity or dielectric constant ε' is determined from the ratio

$$\varepsilon' = \frac{C}{C_o} \tag{55.3}$$

where C represents the measured capacitance in F and C_o is the equivalent capacitance *in vacuo,* which is calculated for the same specimen geometry from $C_o = \varepsilon_o A/d$; here ε_o denotes the permittivity *in vacuo* and is equal to 8.854×10^{-14} F cm^{-1} (8.854×10^{-12} F m^{-1} in SI units) or more conveniently to unity in the Gaussian CGS system. In practice, the value of ε_o in free space is essentially the same as that for a gas (e.g., for air, ε_o = 1.000536). The majority of liquid and solid dielectric materials, presently in use, have dielectric constants extending from approximately 2 to 10.

55.2 Dielectric Losses

Under ac conditions **dielectric losses** arise mainly from the movement of free charge carriers (electrons and ions), space charge polarization, and dipole orientation [Bartnikas and Eichhorn, 1983]. Ionic, space charge, and dipole losses are temperature- and frequency-dependent, a dependency which is reflected in the measured values of σ and ε'. This necessitates the introduction of a complex permittivity ε defined by

$$\varepsilon = \varepsilon' - j\varepsilon'' \tag{55.4}$$

where ε'' is the imaginary value of the permittivity, which is equal to σ/ω. Note that the conductivity σ determined under ac conditions may include the contributions of the dipole orientation, space charge, and ionic polarization losses in addition to that of the drift of free charge carriers (ions and electrons) which determine its dc value.

The complex permittivity, ε, is equal to the ratio of the dielectric displacement vector $\overline{D}$ to the electric field vector $\overline{E}$, i.e., $\varepsilon = \overline{D}/\overline{E}$. Since under ac conditions the appearance of a loss or leakage current is manifest as a phase angle difference δ between the $\overline{D}$ and $\overline{E}$ vectors, then in complex notation $\overline{D}$ and $\overline{E}$ may be expressed as $D_o \exp\,[j(\omega t - \delta)]$ and $E_o \exp[j\omega t]$, respectively, where ω is the radial frequency term, t the time, and D_o and E_o the respective magnitudes of the two vectors. From the relationship between $\overline{D}$ and $\overline{E}$, it follows that

$$\varepsilon' = \frac{D_o}{E_o} \cos\delta \tag{55.5}$$

and

$$\varepsilon'' = \frac{D_o}{E_o} \sin\delta \tag{55.6}$$

It is customary under ac conditions to assess the magnitude of loss of a given material in terms of the value of its **dissipation factor,** tanδ; it is apparent from Eqs. (55.5) and (55.6), that

$$\tan\delta = \frac{\varepsilon''}{\varepsilon'} = \frac{\sigma}{\omega\varepsilon'} \tag{55.7}$$

Examination of Eq. (55.7) suggests that the behavior of a dielectric material may also be described by means of an equivalent electrical circuit. It is most commonplace and expedient to use a parallel circuit representation, consisting of a capacitance C in parallel with a large resistance R as delineated in Fig. 55.1. Here C represents

the capacitance and R the resistance of the dielectric. For an applied voltage V across the dielectric, the leakage current is $I_l = V/R$ and the displacement current is $I_C = j\omega CV$; since $\tan\delta = I_l/I_C$, then

$$\tan\delta = \frac{1}{\omega RC} \tag{55.8}$$

It is to be emphasized that in Eq. (55.8), the quantities R and C are functions of temperature, frequency, and voltage. The equivalence between Eqs. (55.7) and (55.8) becomes more palpable if I_l and I_C are expressed as $\omega\varepsilon'' C_o V$ and $j\omega\varepsilon' C_o V$, respectively.

Every loss mechanism will exhibit its own characteristic $\tan\delta$ loss peak, centered at a particular absorption frequency, ω_o for a given test temperature. The loss behavior will be contingent upon the molecular structure of the material, its thickness, and homogeneity, and the temperature, frequency, and electric field range over which the measurements are performed [Bartnikas and Eichhorn, 1983]. For example, dipole orientation losses will be manifested only if the material contains permanent molecular or side-link dipoles; a considerable overlap may occur between the permanent dipole and ionic relaxation regions. Ionic relaxation losses occur in dielectric structures where ions are able to execute short-range jumps between two or more equilibrium positions. Interfacial or space charge polarization will arise with insulations of multilayered structures where the conductivity and permittivity is different for the individual strata or where one dielectric phase is interspersed in the matrix of another dielectric. Space charge traps also occur at crystalline-amorphous interfaces, crystal defects, and oxidation and localized C-H dipole sites in polymers. Alternatively, space charge losses will occur with mobile charge carriers whose movement becomes limited at the electrodes. This type of mechanism takes place often in thin-film dielectrics and exhibits a pronounced thickness effect. If the various losses are considered schematically on a logarithmic frequency scale at a given temperature, then the $\tan\delta$ and ε' values will appear as functions of frequency as delineated schematically in Fig. 55.2. For many materials the dipole and ionic relaxation losses tend to predominate over the frequency range extending from about 0.5 to 300 MHz, depending upon the molecular structure of the dielectric and temperature. For example, the absorption peak of an oil may occur at 1 MHz, while that of a much lower viscosity fluid such as water may appear at approximately 100 MHz. There is considerable overlap between the dipole and ionic relaxation loses, because the ionic jump distances are ordinarily of the same order of magnitude as the radii of the permanent dipoles. Space charge polarization losses manifest themselves normally over the low-frequency region extending from 10^{-6} Hz to 1 MHz and are characterized by very broad and intense peaks; this behavior is apparent from Eq. (55.7), which indicates that even small conductivities may lead to very large $\tan\delta$ values at very low frequencies. The nonrelaxation-type electronic conduction losses are readily perceptible over the low-frequency spectrum and decrease monotonically with frequency.

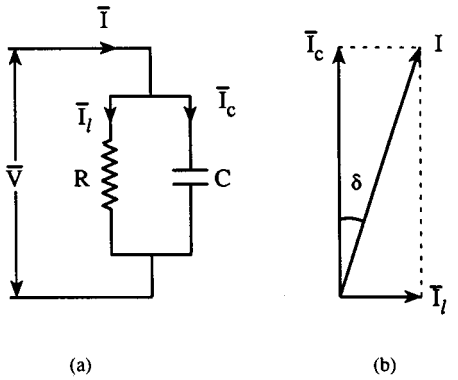


FIGURE 55.1 (a) Parallel equivalent RC circuit and (b) corresponding phasor diagram.

The dielectric loss behavior may be phenomenologically described by the Pellat-Debye equations, relating the imaginary and real values of the permittivity to the relaxation time, τ, of the loss process (i.e., the frequency at which the ε'' peak appears: $f_o = 1/2\ \pi\tau$), the low-frequency or static value of the real permittivity, ε_s, and the high- or optical-frequency value of the real permittivity, ε_∞. Thus, for a loss process characterized by a single relaxation time

$$\varepsilon' = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + \omega^2\tau^2} \tag{55.9}$$

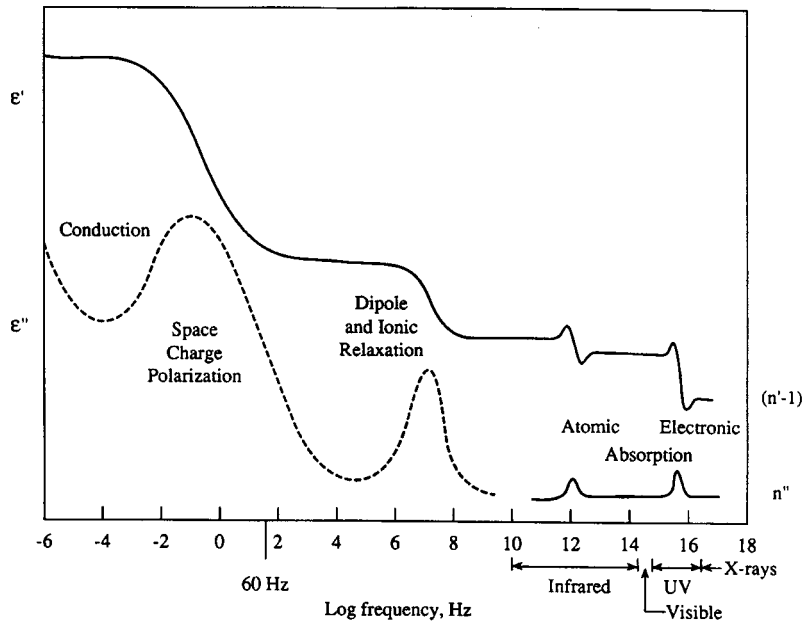


FIGURE 55.2 Schematic representation of different absorption regions [Bartnikas, 1987].

and

$$\varepsilon'' = \frac{(\varepsilon_s - \varepsilon_\infty)\omega\tau}{1 + \omega^2\tau^2} \tag{55.10}$$

In practice Eqs. (55.9) and (55.10) are modified due to a distribution in the relaxation times either because several different loss processes are involved or as a result of interaction or cooperative movement between the discrete dipoles or the trapped and detrapped charge carriers in their own particular environment. Since the relaxation processes are thermally activated, an increase in temperature will cause a displacement of the loss peak to higher frequencies. In the case of ionic and dipole relaxation, the relaxation time may be described by the relation

$$\tau = \frac{h}{kT}\exp\left[\frac{\Delta H}{RT}\right]\exp\left[-\frac{\Delta S}{R}\right] \tag{55.11}$$

where h is the Planck constant (6.624×10^{-34} J s^{-1}), k the Boltzmann constant (1.38×10^{-23} J K^{-1}), ΔH the activation energy of the relaxation process, R the universal gas constant (8.314×10^3 J K^{-1} kmol^{-1}), and ΔS the entropy of activation. For the ionic relaxation process, τ may alternatively be taken as equal to $1/2\Gamma$, where Γ denotes the ion jump probability between two equilibrium positions. Also for dipole orientation in liquids, τ may be approximately equated to the Debye term $\eta/4\pi r^3$T, where η represents the macroscopic viscosity of the liquid and r is the dipole radius [Bartnikas, 1994]. With interfacial or space charge polarization, which may arise due to a pile-up of charges at the interface of two contiguous dielectrics of different conductivity and permittivity, Eq. (55.10) must be rewritten as [von Hippel, 1956]

$$\varepsilon'' = \varepsilon_\infty\left(\frac{\tau}{\omega\tau_1\tau_2} + \frac{K\omega\tau}{1 + \omega^2\tau^2}\right) \tag{55.12}$$

where the Wagner absorption factor K is given by

$$K = \frac{(\tau_1 + \tau_2 - \tau)\tau - \tau_1\tau_2}{\tau_1\tau_2} \tag{55.13}$$

where τ_1 and τ_2 are the relaxation times of the two contiguous layers or strata of respective thicknesses d_1 and d_2; τ is the overall relaxation time of the two-layer combination and is defined by $\tau = (\varepsilon'_1\ d_2 + \varepsilon'_2\ d_1)/(\sigma_1 d_2 + \sigma_2 d_1)$, where ε'_1, ε'_2, σ_1, and σ_2 are the respective real permittivity and conductivity parameters of the two discrete layers. Note that since ε'_1 and ε'_2 are temperature- and frequency-dependent and σ_1 and σ_2 are, in addition, also voltage-dependent, the values of τ and ε'' will in turn also be influenced by these three variables. Space charge processes involving electrons are more effectively analyzed, using dc measurement techniques. If retrapping of electrons in polymers is neglected, then the decay current as a function of time t, arising from detrapped electrons, assumes the form [Watson, 1995]

$$i(t) = \frac{kT}{\nu t} n(E) \tag{55.14}$$

where n(E) is the trap density and ν is the attempt jump frequency of the electrons. The electron current displays the usual t^{-1} dependence and the plot of i(t)t versus kTln(νt) yields the distribution of trap depths. Eq. (55.14) represents an approximation, which underestimates the current associated with the shallow traps and overcompensates for the current due to the deep traps. The mobility of the free charge carriers is determined by the depth of the traps, the field resulting from the trapped charges, and the temperature. As elevated temperatures and low space charge fields, the mobility is proportional to exp[–ΔH/kT] and at low temperatures to $(T)^{1/4}$ [LeGressus and Blaise, 1992]. A high trapped charge density will create intense fields, which will in turn exert a controlling influence on the mobility and the charge distribution profile. In polymers, shallow traps are of the order of 0.5 to 0.9 eV and deep traps are ca. 1.0 to 1.5 eV, while the activation energies of dipole orientation and ionic conduction in solid and liquid dielectrics fall within the same range. It has been known that most charge trapping in the volume occurs in the vicinity of the electrodes; this can now be confirmed by measurement, using thermal and electrically stimulated acoustical pulse methods [Bernstein, 1992]. In the latter method this involves the application of a rapid voltage pulse across a dielectric specimen. The resulting stress wave propagates at the velocity of sound and is detected by a piezoelectric transducer. This wave is assumed not to disturb the trapped charge; the received electrical signal is then correlated with the acoustical wave to determine the profile of the trapped charge. Errors in the measurement would appear to be principally caused by the electrode surface charge effects and the inability to distinguish between the polarization of polar dipoles and that of the trapped charges [Wintle, 1990].

Temperature influences the real value of the permittivity or dielectric constant ε' insofar as it affects the density of the dielectric material. As the density diminishes with temperature, ε' falls with temperature in accordance with the Clausius-Mossotti equation

$$[P] = \frac{(\varepsilon' - 1)}{(\varepsilon' + 2)} \frac{M}{d_o} \tag{55.15}$$

where $[P]$ represents the polarization per mole, M the molar mass, d_o the density at a given temperature, and $\varepsilon' = \varepsilon_s$. Equation (55.14) is equally valid, if the substitution $\varepsilon' = (n')^2$ is made; here n' is the real value of the index of refraction. In fact, the latter provides a direct connection with the dielectric behavior at optical frequencies. In analogy with the complex permittivity, the index of refraction is also a complex quantity, and its imaginary value n'' exhibits a loss peak at the absorption frequencies; in contrast with the ε' value which can only fall with frequency, the real index of refraction n' exhibits an inflection-like behavior at the absorption frequency. This is illustrated schematically in Fig. 55.2, which depicts the kn' or n'' and $n' - 1$ values as a function of frequency over the optical frequency regime. The absorption in the infrared results from atomic

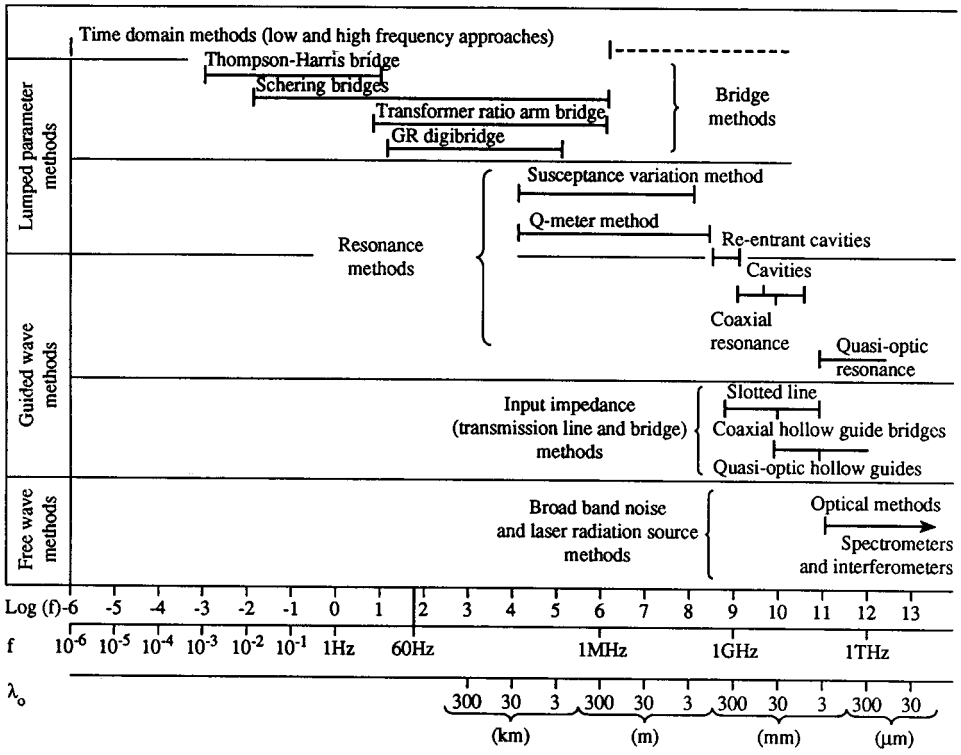


FIGURE 55.3 Frequency rangse of various dielectric test methods [Bartnikas, 1987].

resonance that arises from a displacement and vibration of atoms relative to each other, while an electronic resonance absorption effect occurs over the ultraviolet frequencies as a consequence of the electrons being forced to execute vibrations at the frequency of the external field.

The characterization of dielectric materials must be carried out in order to determine their properties for various applications over different parts of the electromagnetic frequency spectrum. There are many techniques and methods available for this purpose that are too numerous and detailed to attempt to present here even in a cursory manner. However, Fig. 55.3 portrays schematically the different test methods that are commonly used to carry out the characterization over the different frequencies up to and including the optical regime. A direct relationship exists between the time and frequency domain test methods via the Laplace transforms.

The frequency response of dielectrics at the more elevated frequencies is primarily of interest in the electrical communications field. In contradistinction for electrical power generation, transmission, and distribution, it is the low-frequency spectrum that constitutes the area of application. Also, the use of higher voltages in the electrical power area necessarily requires detailed knowledge of how the electrical losses vary as a function of the electrical field. Since most electrical power apparatus operates at a fixed frequency of 50 or 60 Hz, the main variable apart from the temperature is the applied or operating voltage. At power frequencies the dipole losses are generally very small and invariant with voltage up to the saturation fields which exceed substantially the operating fields, being in the order of 10^7 kV cm^{-1} or more. However, both the space charge polarization and ionic losses are highly field-dependent. As the electrical field is increased, ions of opposite sign are increasingly segregated; this hinders their recombination and, in effect, enhances the ion charge carrier concentration. As the dissociation rate of the ionic impurities is further augmented by temperature increases, combined rises in temperature and field may lead to appreciable dielectric loss. Thus, for example, for a thin liquid film bounded by two solids, tanδ increases with voltage until at some upper voltage value the physical boundaries begin to finally limit the amplitude of the ion excursions, at which point tanδ commences a downward trend with voltage (Böning–Garton effect). The interfacial or space charge polarization losses may evince a rather intricate field dependence, depending upon the manner in which the discrete conductivities of the contiguous media change

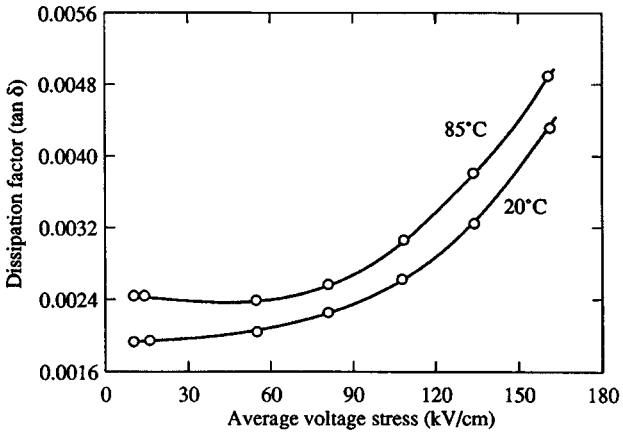


FIGURE 55.4 Loss characteristics of mineral oil-impregnated paper.

with applied voltage and temperature [as is apparent from the nature of Eqs. (55.12) and (55.13)]. The exact frequency value at which the space charge loss exhibits its maximum is contingent upon the value of the relaxation time τ. Figure 55.4 depicts typical tanδ versus applied voltage characteristics for an oil-impregnated paper-insulated cable model at two different temperatures, in which the loss behavior is primarily governed by ionic conduction and space charge effects. The monotonically rising dissipation factor with increasing applied voltage at room temperature is indicative of the predominating ionic loss mechanism, while at 85°C, an incipient decrease in tanδ is suggestive of space charge effects.

55.3 Dielectric Breakdown

As the voltage is increased across a dielectric material, a point is ultimately reached beyond which the insulation will no longer be capable of sustaining any further rise in voltage and breakdown will ensue, causing a short to develop between the electrodes. If the dielectric consists of a gas or liquid medium, the breakdown will be self-healing in the sense that the gas or liquid will support anew a reapplication of voltage until another breakdown recurs. In a solid dielectric, however, the initial breakdown will result in a formation of a permanent conductive channel, which cannot support a reapplication of voltage. The dielectric breakdown processes are distinctly different for the three states of matter.

In the case of solid dielectrics the breakdown is dependent not only upon the molecular structure and morphology of the solid but also upon extraneous variables such as the geometry of the material, the temperature, and the ambient environment. Since breakdown often occurs along some fault of the material, the breakdown voltage displays a readily perceptible decrease with area and thickness of the specimen due to increased incidence of faults over larger volumes. This is indeed part of the reason why thin-film inorganic dielectrics, which are normally evaluated using small-diameter dot counter electrodes, exhibit exceptionally high **dielectric strengths.** With large organic dielectric specimens, recessed electrodes are used to minimize electrode edge effects, leading to greatly elevated breakdown strengths in the order of 10^6 to 10^7 kV cm^{-1}, a range of values which is considered to represent the ultimate breakdown strength of the material or its intrinsic breakdown strength; as the intrinsic breakdown occurs in approximately 10^{-8} to 10^{-6} s, an electronic mechanism is implicated.

The breakdown strength under dc and impulse conditions tends to exceed that at ac fields, thereby suggesting the ac breakdown process may be partially of a thermal nature. An additional factor, which may lower the ac breakdown strength, is that associated with the occurrence of partial discharges either in void inclusions or at the electrode edges; this leads to breakdown values very much less than the intrinsic value. In practice, the breakdowns are generally of an extrinsic nature, and the intrinsic values are useful conceptually insofar as they provide an idea of an upper value that can be attained only under ideal conditions. The intrinsic breakdown theories were essentially developed for crystalline dielectrics, for which it was assumed that a very small number

of thermally activated electrons can be thermally excited to move from the valence to the conduction band and that under the influence of an external field they will be impelled to move in the direction of the field, colliding with the lattice of the crystalline dielectric and dissipating their energy by phonon interactions [Bartnikas and Eichhorn, 1983]. Accordingly, breakdown is said to occur when the average rate of energy gain by the electrons, $A(E, T, T_e, \xi)$, exceeds that lost in collisions with the lattice, $B(T, T_e, \xi)$. Hence, the breakdown criterion can be stated as

$$A(E, T, T_e, \xi) = B(T, T_e, \xi) \tag{55.16}$$

where E is the applied field, T the lattice temperature, T_e the electron temperature, and ξ an energy distribution constant. Thus in qualitative terms as the temperature is increased gradually, the breakdown voltage rises because the interaction between the electrons and the lattice is enhanced as a result of the increased thermal vibrations of the lattice. Ultimately, a critical temperature is attained where the electron–electron interactions surpass in importance those between the electrons and the lattice, and the breakdown strength commences a monotonic decline with temperature; this behavior is borne out in NaCl crystals, as is apparent from Fig. 55.5 [von Hippel and Lee, 1941]. However, with amorphous or partially crystalline polymers, as for example with polyethylene, the maximum in breakdown strength is seen to be absent and only a decrease is observed [Oakes, 1949]; as the crystalline content is increased in amorphous-crystalline solids, the breakdown strength is reduced.

The electron avalanche concept has also been applied to explain breakdown in solids, in particular to account for the observed decrease in breakdown strength with insulation thickness. Since breakdown due to electron avalanches involves the formation of space charge, space charges will tend to modify the conditions for breakdown. Any destabilization of the trapping and detrapping process, such as may be caused by a perturbation of the electrical field, will initiate the breakdown event [LeGressus and Blaise, 1992]. The detrapping of mobile charge carriers will be accompanied by photon emission and formation of the plasma breakdown channel, resulting in the dissipation of polarization energy. If dipole interaction is neglected, the polarization energy due to a trapped charge is of the order of 5χ eV, where χ is the dielectric susceptibility. The release of the polarization energy will be accompanied by electrical tree growth in and melting of the polymer.

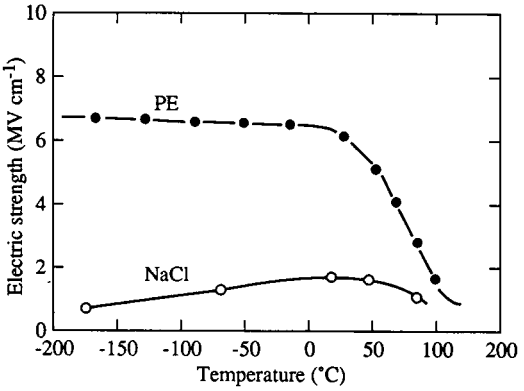


FIGURE 55.5 Dielectric breakdown characteristics of sodium chloride [von Hippel and Lee, 1941] and polyethylene [Oakes, 1949].

The breakdown process in gases is relatively well understood and is explained in terms of the avalanche theory. A free electron, occurring in a gas due to cosmic radiation, will be accelerated in a field and upon collision with neutral molecules in its trajectory will eject, if its energy is sufficient, other electrons that will in turn undergo additional collisions resulting in a production of more free electrons. If the electric field is sufficiently high, the number of free electrons will increase exponentially along the collision route until ultimately an electron avalanche will form. As the fast-moving electrons in the gap disappear into the anode, they leave behind the slower-moving ions, which gradually drift to the cathode where they liberate further electrons with a probability γ. When the height of the positive ion avalanche becomes sufficiently large to lead to a regeneration of a starting electron, the discharge mechanism becomes self-sustaining and a spark bridges the two electrodes. The condition for the Townsend breakdown in a short gap is given by

$$\gamma[\exp(\alpha d) - 1] = 1 \tag{55.17}$$

where d is the distance between the electrodes and α represents the number of ionizing impacts per electron per unit distance. The value of γ is also enhanced by photoemission at the cathode and photon radiation in

the gas volume by the metastable and excited gas atoms or molecules. In fact, in large gaps the breakdown is governed by steamer formation in which photon emission from the avalanches plays a dominant role. Breakdown characteristics of gases are represented graphically in terms of the Paschen curves, which are plots of the breakdown voltage as a function of the product of gas pressure p and the electrode separation d. Each gas is characterized by a well-defined minimum breakdown voltage at one particular value of the pd product.

The breakdown process in liquids is perhaps the least understood due to a lack of a satisfactory theory on the liquid state. The avalanche theory has been applied with limited success to explain the breakdown in liquids, by assuming that electrons injected from an electrode surface exchange energy with the atoms or molecules of the liquid, ultimately causing the atoms and molecules to ionize and thus precipitating breakdown. Recent investigations, utilizing electro-optical techniques, have demonstrated that breakdown involves steamers with tree- or bushlike structures that propagate from the electrodes [Bartnikas, 1994]. The negative streamers emerging from the cathode form due to electron emission, while positive steamers originating at the anode are due to free electrons in the liquid itself. The breakdown of liquids is noticeably affected by electrolytic impurities as well as water and oxygen content; also, macroscopic particles may form bridges between the electrodes along which electrons may hop with relative ease, resulting in a lower breakdown. As in solids, there is a volume effect and breakdown strength decreases with thickness; a slight increase in breakdown voltage is also observed with viscosity.

In both solid and liquid dielectrics, the breakdown strength under dc and impulse fields is markedly greater than that obtained under ac fields, thus suggesting that under ac conditions the breakdown may be partially thermal in nature. Thermal breakdown occurs at localized hot spots where the rate of heat generated exceeds that dissipated by the surrounding medium. The temperature at such hot spots continues to rise until it becomes sufficiently high to induce fusion and vaporization, causing eventually the development of a channel along which breakdown ensues between the opposite electrodes. Since a finite amount of time is required for the heat buildup to occur to lead to the thermal instability, thermally induced breakdown is contingent upon the time of the alternating voltage application and is thus implicated as the leading cause of breakdown in many dielectrics under long-term operating conditions. However, under some circumstances thermal instability may develop over a very short time; for example, some materials have been found to undergo thermal breakdown when subjected to very short repetitive voltage pulses. In low-loss dielectrics, such as polyethylene, the occurrence of thermal breakdown is highly improbable under low operating temperatures, while glasses with significant ionic content are more likely to fail thermally, particularly at higher frequencies.

The condition for thermal breakdown may be stated as

$$KA\,\Delta T/l = \omega\varepsilon' E^2 \tan\delta \tag{55.18}$$

where the left-hand side represents the heat transfer in J s^{-1} along a length l (cm) of sectional area A (cm^2) of the dielectric surface in the direction of the temperature gradient due to the temperature difference ΔT, in °C, such that the units of the thermal conductivity constant K are in J $°C^{-1}$ cm^{-1} s^{-1}. The right-hand side of Eq. (55.18) is equal to the dielectric loss dissipated in the dielectric in J s^{-1}, where E is the external field, ε' the real value of the permittivity, and $\tan\delta$ the dissipation factor at the radial frequency ω.

Other causes of extrinsic breakdown are associated with particular defects in the dielectric or with the environmental conditions under which the dielectric material is employed. For example, some dielectrics may contain gas-filled cavities that are inherent with the porous structure of the dielectric or that may be inadvertently introduced either during the manufacturing process or created under load cycling. If the operating electrical field is sufficiently elevated to cause the gas within the cavities to undergo discharge, the dielectric will be subjected to both physical and chemical degradation by the partial discharges; should the discharge process be sustained over a sufficiently long period, breakdown will eventually ensue.

With overhead line insulators or bushings of electrical equipment, breakdown may occur along the surface rather than in the bulk of the material. Insulator surfaces consisting of porcelain, glass, or polymeric materials (usually elastomers), may become contaminated by either industrial pollutants or salt spray near coastal areas, leading to surface tracking and breakdown below the normal flashover voltage. Surface tracking is enhanced in the presence of moisture, which increases the surface conductivity, particularly in the presence of ionic contaminants [Bartnikas, 1987]. The latter is measured in S or Ω^{-1} and must be distinguished from the volume

conductivity whose units are stated in S cm^{-1} or Ω^{-1} cm^{-1}. Surface tracking may be prevented by cleaning the surface and by applying silicone greases. When insulators are employed under a vacuum environment, charge accumulation will occur in the surface because the charged surface will no longer be able to discharge due to the finite breakdown strength of an ambient gas. Space charge will thus play a dominant role in the surface breakdown mechanism [Miller, 1993].

55.4 Insulation Aging

All insulating materials will undergo varying degrees of aging or deteriorating under normal operating conditions. The rate of aging will be contingent upon the magnitude of the electrical, thermal, and mechanical stresses to which the material is subjected; it will also be influenced by the composition and molecular structure of the material itself as well as the chemical, physical, and radiation environment under which the material must perform. The useful life of an insulating system will thus be determined by a given set and subset of aging variables. For example, the subset of variables in the voltage stress variable are the average and maximum values of the applied voltage, its frequency, and the recurrence rate of superposed impulse or transient voltage surges. For the thermal stress, the upper and lower ambient temperatures, the temperature gradient in the insulation, and the maximum permissible operating temperature constitute the subvariable set. Also, the character of the mechanical stress will differ, depending upon whether torsion, compression, or tension and bending are involved. Furthermore, the aging rate will be differently affected if all stresses (electrical, thermal, and mechanical) act simultaneously, separately, or in some predetermined sequence. The influence exerted on the aging rate by the environment will depend on whether the insulation system will be subjected to corrosive chemicals, petroleum fluids, water or high humidity, air or oxygen, ultraviolet radiation from the sun, and nuclear radiation. Organic insulations, in particular, may experience chemical degradation in the presence of oxygen. For example, polyethylene under temperature cycle will undergo both physical and chemical changes. These effects will be particularly acute at the emergency operating temperatures (90–130°C); at these temperatures partial or complete melting of the polymer will occur and the increased diffusion rate will permit the oxygen to migrate to a greater depth into the polymer. Ultimately the antioxidant will be consumed, resulting in an embrittlement of polymer and in extreme cases in the formation of macroscopic cracks. Subjection of the polymer to many repeated overload cycles will be accompanied by repeated melting and recrystallization of the polymer—a process that will inevitably cause the formation of cavities, which, when subjected to sufficiently high voltages, will undergo discharge, leading eventually to electrical breakdown.

There is a general consensus that electrically induced aging involves the mechanisms of treeing, partial discharge, and dielectric heating. Dielectric heating failures are more characteristic of lossy insulations or when highly conductive contaminants are involved. In the treeing mechanism, a distinction must be made between electrical and water trees [Bartnikas and Eichhorn, 1983]. The former refers to growth, resembling a tree, that occurs under dry conditions in the presence of an electric field; its branches or channels are hollow. In contrast, water trees require the presence of moisture, and their branches consist of fine filamentary channels, joining small cavities, all of which contain water; when placed in a dry environment, they eventually disappear. In a translucent dielectric, water trees are invisible and are rendered visible only when stained with a die, whereas electrical trees once formed remain readily discernible. Water trees are intrinsic to solid polymeric insulation, while electrical trees may occur in both solid and impregnated insulating systems. The actual failure path, when the breakdown current is limited to prevent destruction of evidence, invariably, even in liquids, consists of an electrical treelike structure.

Both electrical and water trees tend to propagate from electrical stress enhancement points, with water trees requiring appreciably lower stresses for their inception (ca. 2 kV/mm). Whereas the electrical tree mechanism usually involves partial discharges, the occurrence of the latter is conspicuously absent in water trees. The fact that water trees may bridge the opposite electrodes without precipitating failure infers a nonconductive nature of the filamentary channels or branches. Yet water trees are implicated in the overall degradation and aging process of the dielectric because of their presence in areas of failure as well as because of the often-observed phenomenon of an electrical tree emerging from some point along a water tree and causing abrupt failure.

The presence of voids or cavities within a solid or solid-liquid insulating system will almost invariably lead to eventual failure, with the proviso that the electrical field is sufficiently elevated to induce them to undergo

either continuous or regularly recurring, though at times intermittent, discharge. Electrical trees may readily ensue from charge injection sites adjacent to discharging cavity inclusions. The time required for electrical tree initiation may be defined quasi-empirically as [Fothergill et al., 1994],

$$t = \frac{N_c}{2bfn_o\left[\exp\left(\alpha l - 1\right)\right]} \tag{55.19}$$

where N_c denotes the critical number of impact ionizations necessary for tree channel formation, l is the avalanche or newly formed tree length, n_o is the number of initiating electrons and b is a constant dependent on the frequency, f. The deterioration rate due to partial discharges is proportional to the power dissipated P by the discharges, which may be expressed as [Bartnikas and McMahon, 1979]

$$P = \sum_{i=1}^{l} \sum_{j=1}^{m} n_{ij}\, \Delta Q_{ij}\, V_{ij} \tag{55.20}$$

where n_{ij} is the recurrence rate of the jth discharge in the ith cavity and ΔQ_{ij} is its corresponding charge release at an applied voltage V_{ij}. Under ac conditions the discharges will tend to recur regularly in each cycle due to the capacitative voltage division across the void. Under dc conditions the discharge rate will be controlled by the time constant required to recharge the cavity following a discharge. The physical damage arising from discharges consists of surface erosion and pitting and is caused by the ion and electron bombardment incident on the void's walls at a given discharge site. Chemical degradation results from molecular chain scission due to particle bombardment of the surface and the reactions between the ambient ionized gases and the gases released due to the molecular chain scission processes. The final chemical composition of the reaction product is generally varied, depending primarily on the molecular structure of the dielectric materials involved and the composition of the ambient ionized gases; one discharge degradation product common to many polymers exposed to discharges in air is that of oxalic acid. Oxalic acid, as a result of its elevated conductivity, when deposited upon the cavity's walls may change the nature of the discharge (e.g., from a spark to a glow type) or may even ultimately extinguish the discharge (i.e., replace the discharge loss by an I^2R-type loss along the cavity's walls).

A number of aging models have been propounded to predict insulation aging under different types of stress. However, there are essentially only two models whose usefulness has been substantiated in practice and which have, therefore, gained wide acceptance. One of these is Dakin's classical thermal degradation model, which is based upon the approach used in chemical reaction rate theory [Bartnikas, 1987],

$$t = A \exp[\Delta H/kT] \tag{55.21}$$

where t represents the time to breakdown, A is a constant, ΔH is the activation energy of the aging process, T is the absolute temperature, and k is the Boltzmann constant (1.38×10^{-23} J K^{-1}). If the log t versus $1/T$ plot represents a straight line that is obtained when the insulation aging is accelerated at various temperatures above the operating temperature, extrapolation of the line to the operating temperature may yield a rudimentary estimate of the aging time or the service life that can be anticipated from the insulation system when operated under normal temperature and load conditions. Deviations from straight-line behavior are indicative of more than one thermal aging mechanism; for example, a polymeric insulation will exhibit such behavior when thermally stressed beyond its melting or phase transition temperature.

Another extremely useful model, applicable to accelerated aging studies under electrical stress, is the so-called inverse power law relationship given by [Bartnikas, 1987]

$$t = BE^{-n} \tag{55.22}$$

where t is the time to breakdown under an electric stress E, B is a constant, and n is an exponent parameter. The relationship is essentially empirical in nature, and its proof of validity rests primarily on experiment and observation. It is found that a single type of electrically induced aging process will generally result in a straight-line relationship between log t and log E. Consequently, aging data obtained at higher electrical stresses with correspondingly shorter times to breakdown when extrapolated to longer times at stresses in the vicinity of the operating stress should yield the value of the effective service life under the operating stress. The slope of the line determines the exponent n, which constitutes an approximate indicator of the type of aging involved. For example, with polymer insulation water treeing ordinarily results in n values less than 4, while under conditions that may involve discharges and electrical trees, the n values may approach 10 or greater.

55.5 Dielectric Materials

Dielectric materials comprise a variety of solids, liquids, and gases. The breakdown strength generally increases with the density of the material so that dielectric solids tend to have higher dielectric strengths than gases. The same tendency is observed also with the dielectric loss and permittivity; for example, the dielectric losses in gases are virtually too small to be measurable and their dielectric constant in most practical applications can be considered as unity. In this section we will describe a number of the more common insulating materials in use.

Gases

The 60-Hz breakdown strength of a 1-cm gap of air at 25°C at atmospheric pressure is 31.7 kV cm^{-1}. Although this is a relatively low value, air is a most useful insulating medium for large electrode separations as is the case for overhead transmission lines. The only dielectric losses in the overhead lines are those due to corona discharges at the line conductor surfaces and leakage losses over the insulator surfaces. In addition, the highly reduced capacitance between the conductors of the lines ensures a small capacitance per unit length, thus rendering overhead lines an efficient means for transmitting large amounts of power over long distances. Nitrogen, which has a slightly higher breakdown strength (33.4 kV cm^{-1}) than air, has been used when compressed in low-loss gas-insulated capacitors. However, as air, which was also used in its compressed form in circuit breakers, it has been replaced by sulfur hexafluoride, SF_6. The breakdown strength of SF_6 is ca. 79.3 kV cm^{-1}. Since the outside metal casing of metal-clad circuit breakers is ordinarily grounded, it has become common practice to have much of the substation interconnected bus and interrupting switching equipment insulated with SF_6. To achieve higher breakdown strengths, the SF_6 gas is compressed at pressures on the order of 6 atm. The use of SF_6 in substation equipment has resulted in the saving of space and the elimination of line insulator pollution problems; however, its extensive usage has also posed some environmental concerns in regard to its flourine content.

The breakdown strength of gases is little affected with increasing frequency until the period of the wave becomes comparable to the transit time of the ions and, finally, electrons across the gap. At this point a substantial reduction in the breakdown strength is observed.

Insulating Liquids

Insulating liquids are rarely used by themselves and are intended for use mainly as impregnants with cellulose or synthetic papers. The 60-Hz breakdown strength of practical insulating liquids exceeds that of gases and for a 1-cm gap separation it is of the order of about 100 kV cm^{-1}. However, since the breakdown strength increases with decreasing gap length and as the oils are normally evaluated using a gap separation of 0.254 cm, the breakdown strengths normally cited range from approximately 138 to 240 kV cm^{-1}. The breakdown values are more influenced by the moisture and particle contents of the fluids than by their molecular structure (cf. Table 55.1).

Mineral oils have been extensively used in high-voltage electrical apparatus since the turn of the century. They constitute a category of hydrocarbon liquids that are obtained by refining petroleum crudes. Their composition consists of paraffinic, naphthenic, and aromatic constituents and is dependent upon the source of the crude as well as the refining procedure followed. The inclusion of the aromatic constituents is desirable

TABLE 55.1 Electrical Properties of a Number of Representative Insulating Liquids [Bartnikas, 1994; Encyclopedia Issue, 1972]

Liquid	Viscosity cSt (37.8°C)	Dielectric Constant (at 60 Hz, 25°C)	Dissipation Factor (at 60 Hz, 100°C)	Breakdown Strength, (kV cm–1)
Capacitor oil	21	2.2	0.001	>118
Pipe cable oil	170	2.15	0.001	>118
Self-contained cable oil	49.7	2.3	0.001	>118
Heavy cable oil	2365	2.23	0.001	>118
Transformer oil	9.75	2.25	0.001	>128
Alkyl benzene	6.0	2.1	0.0004	>138
Polybutene pipe cable oil	110 (SUS)	2.14 (at 1 MHz)	0.0003	>138
Polybutene capacitor oil	2200 (SUS at 100°C)	2.22 (at 1 MHz)	0.0005	>138
Silicone fluid	50	2.7	0.00015	>138
Castor oil	98 (100°C)	3.74	0.06	>138
$C_8F_{16}O$ fluorocarbon	0.64	1.86	<0.0005	>138

because of their gas absorption and oxidation characteristics. Mineral oils used for cable and transformer applications have low polar molecule contents and are characterized by dielectric constants extending from about 2.10 to 2.25 with dissipation factors generally between 2×10^{-5} and 6×10^{-5} at room temperature, depending upon their viscosity and molecular weight. Their dissipation factors increase appreciably at higher temperatures when the viscosities are reduced. Oils may deteriorate in service due to oxidation and moisture absorption. Filtering and treatment with Fullers' earth may improve their properties, but special care must be taken to ensure that the treatment process does not remove the aromatic constituents which are essential to maintaining the gas-absorption characteristics of the oil.

Alkyl benzenes are used as impregnants in high-voltage cables, often as substitutes of the low-viscosity mineral oils in self-contained oil-filled cables. They consist of alkyl chains attached to a benzene ring having the general formula $C_6H_5(CH_2)_nCH_3$; however, branched alkyl benzenes are also employed. Their electrical properties are comparable to those of mineral oils, and they exhibit good gas inhibition characteristics. Due to their detergent character, alkyl benzenes tend to be more susceptible to contamination than mineral oils.

Polybutenes are synthetic oils that are derived from the polymerization of olefins. Their long chains, with isobutene as the base unit, have methyl group side chains with molecular weights in the range from 300 to 1350. Their electrical properties are comparable to those of mineral oils; due to their low cost, they have been used as pipe cable filling oils. Higher viscosity polybutenes have been used as capacitor impregnants. Mixtures of polybutenes and alkyl benzenes have been used to obtain higher ac breakdown strength with impregnated paper systems. They are also compatible and miscible with mineral oils.

Since the discontinued use of the nonflammable polychlorinated biphenyls (PCBs), a number of unsaturated synthetic liquids have been developed for application to capacitors, where due to high stresses evolved gases may readily undergo partial discharge. Most of these new synthetic capacitor fluids are thus gas-absorbing low-molecular-weight derivatives of benzene, with permittivities ranging from 2.66 to 5.25 at room temperature (compared to 3.5 for PCBs). None of these fluids have the nonflammable characteristics of the PCBs; however, they do have high boiling points [Bartnikas, 1994].

Halogenated aliphatic hydrocarbons are derived by replacing the hydrogens by either chlorine or fluorine or both; they may also contain nitrogen and oxygen in their molecular structure. Their dielectric constants range from 1.8 to 3.0, the higher value reflecting some polarity due to molecular asymmetry as a result of branching. They have superior thermal properties to mineral oils and are highly flame-resistant. Fluorocarbons have been used in large power transformers, where both flammability and heat removal are of prime concern.

Silicone liquids consist of polymeric chains of silicon atoms alternating with oxygen atoms, with methyl side groups. For electrical applications, polydimethylsiloxane fluids are used, primarily for transformers as substitutes for the PCBs due to their inherently high flash and flammability points and reduced environmental concerns. They have lower tanδ values than mineral oils but somewhat higher dielectric constants because of their moderately polar nature. The viscosity of silicone fluids exhibits relatively little change with temperature, which is attributed to the ease of rotation about the Si–O–Si bond, thereby overcoming close packing of molecules and reducing intermolecular forces.

There are a large number of organic esters, but only a few are suitable for electrical applications. Their properties are adversely affected by hydrolysis, oxidation, and water content. Due to their reduced dielectric losses at elevated frequencies, they have been used in high-frequency capacitors. Castor oil has found specialized application in energy storage capacitors due to its exceptional resistance to partial discharges. The dielectric constants of esters are substantially higher than those for mineral oils.

Solid Insulating Materials

Solid insulating materials may be classified into two main categories, organic and inorganic. There are an extremely large number of solid insulants available, but in this section only the more commonly representative solid insulants will be considered.

Inorganic Solids

Below are described a number of the more prevalent inorganic dielectrics in use; their electrical and physical properties are listed in Table 55.2.

Alumina (Al_2O_3) is produced by heating aluminum hydroxide or oxyhydroxide; it is widely used as a filler for ceramic insulators. Further heating yields the corundum structure, which in its sapphire form is used for dielectric substrates in microcircuit applications.

Barium titanate ($BaTiO_3$) is an extraordinary dielectric in that below 120°C it behaves as a ferroelectric. That is, the electric displacement is both a function of the field as well as its previous history. Due to spontaneous polarization of the crystal, a dielectric hysteresis loop is generated. The dielectric constant is different in the x and z axis of the crystal (e.g., at 20°C, $\varepsilon' > 4000$ perpendicular to the z axis and $\varepsilon' < 300$ in the x-axis direction.

Porcelain is a multiphase ceramic material that is obtained by heating aluminum silicates until a mullite ($3Al_2O_3 \cdot 2SiO_2$) phase is formed. Since mullite is porous, its surface must be glazed with a high-melting-point glass to render it smooth and impervious and thus applicable for use in overhead line insulators. For high-frequency applications, low-loss single-phase ceramics, such as steatite ($3MgO \cdot 4SiO_2 \cdot H_2O$), are preferred.

Magnesium oxide (MgO) is a common inorganic insulating material, which due to its relatively high thermal conductivity is utilized for insulating heating elements in ovens. The resistance wire elements are placed concentrically within stainless steel tubes, with magnesium oxide packed around them to provide the insulation.

Electrical-grade glasses consist principally of SiO_2, B_2O_3, and P_2O_5 structures that are relatively open to permit ionic diffusion and migration. Consequently, glasses tend to be relatively lossy at high temperatures, though at low temperatures they are suitable for use in overhead line insulators and in transformer, capacitor, and circuit breaker bushings. At high temperatures, their main application lies with incandescent and fluorescent lamps as well as electronic tube envelopes.

Most of the mica used in electrical applications is of the muscovite [$KAl_2(OH)_2Si_3AlO_{10}$] type. Mica is a layer-type dielectric, and mica films are obtained by the splitting of mica blocks. The extended two-dimensionally layered strata of mica prevent the formation of conductive pathways across the mica, resulting in a high dielectric strength. It has excellent thermal stability and due to its inorganic nature is highly resistant to partial discharges. It is used in sheet, plate, and tape form in rotating machines and transformer coils. For example, a mica-epoxy composite is employed in stator bar insulation of rotating machines.

In metal-oxide-silicon (MOS) devices, the semiconductor surface is passivated by thermally growing a silicon dioxide, SiO_2, film (about 5000 Å) with the semiconductor silicon wafer exposed to an oxygen ambient at 1200°C. The resulting SiO_2 dielectric film has good adhesion properties, but due to its open glassy structure is not impervious to ionic impurities (primarily sodium). Accordingly, a denser film structure of silicon nitride, Si_3N_4, is formed in a reaction between silane and ammonia and is pyrolytically deposited on the SiO_2 layer.

TABLE 55.2 Electrical and Physical Properties of Some Common Solid Insulating Materials [Bartnikas and Eichhorn, 1983; Insulation Circuits, 1972]

Material	Specific Gravity	Maximum Operating Temperature (°C)	Dielectric Constant 20°C 60 Hz	Dielectric Constant 20°C 1 kHz	Dielectric Constant 20°C 1 MHz	Dissipation Factor 20°C 60 Hz	Dissipation Factor 20°C 1 kHz	Dissipation Factor 20°C 1 MHz	AC Dielectric Strength (kV cm–1)
Alumina (Al_2O_3)	3.1–3.9	1950	8.5	8.5	8.5	1×10^{-3}	1×10^{-3}	1×10^{-3}	98–157
Porcelain (mullite)	2.3–2.5	1000	8.2	8.2	8.2	1.4×10^{-3}	5.7×10^{-4}	2×10^{-4}	94–157
Steatite $3MgO \cdot 4SiO_2 \cdot H_2O$	2.7–2.9	1000–1100	5.5	5.0	5.0	1.3×10^{-3}	4.5×10^{-4}	3.7×10^{-4}	200
Magnesium oxide (MgO)	3.57	<2800	9.65	9.65	9.69	$<3 \times 10^{-4}$	$<3 \times 10^{-4}$	$<3 \times 10^{-4}$	>2000
Glass (soda lime)	2.47	110–460	6.25	6.16	6.00	5.0×10^{-3}	4.2×10^{-3}	2.7×10^{-3}	4500
Mica ($KAl_2(OH)_2Si_3AlO_{10}$)	2.7–3.1	550	6.9	6.9	5.4	1.5×10^{-3}	2.0×10^{-4}	3.5×10^{-4}	3000–8200
SiO_2 film		<900		3.9			7×10^{-4}		1000–10,000
Si_3N_4		<1000		12.7			$<1 \times 10^{-4}$		1000–10,000
Ta_2O_5	8.2	<1800		28			1×10^{-2}		
HfO_2		4700°F		35			1×10^{-2}		
Low-density PE	(density: 0.910–0.925 g cm^{-3})	70	2.3	2.3	2.3	2×10^{-4}	2×10^{-4}	2×10^{-4}	181–276
Medium-density PE	(density: 0.926–0.940 g cm^{-3})	70	2.3	2.3	2.3	2×10^{-4}	2×10^{-4}	2×10^{-4}	197–295
High-density PE	(density: 0.941–0.965 g cm^{-3})	70	2.35	2.35	2.35	2×10^{-4}	2×10^{-4}	2×10^{-4}	177–197
XLPE	(density: 0.92 g cm^{-3})	90	2.3		2.28	3×10^{-4}		4×10^{-4}	217

TABLE 55.2 (continued) Electrical and Physical Properties of Some Common Solid Insulating Materials [Bartnikas and Eichhorn, 1983; Insulation Circuits, 1972]

Material	Specific Gravity	Maximum Operating Temperature (°C)	Dielectric Constant 20°C 60 Hz	Dielectric Constant 20°C 1 kHz	Dielectric Constant 20°C 1 MHz	Dissipation Factor 20°C 60 Hz	Dissipation Factor 20°C 1 kHz	Dissipation Factor 20°C 1 MHz	AC Dielectric Strength (kV cm–1)
EPR	0.86	300–350°F		3.0–3.5		4×10^{-3}			354–413
Polypropylene	0.90	128–186	2.22–2.28	2.22–2.28	2.22–2.28	$2–3 \times 10^{-4}$	$2.5–3.0 \times 10^{-4}$	4.6×10^{-4}	295–314
PTFE	2.13–2.20	<327	2.0	2.0	2.0	$<2 \times 10^{-4}$	$<2 \times 10^{-4}$	$<2 \times 10^{-4}$	189
Glass-reinforced polyester premix	1.8–2.3	265	5.3–7.3		5.0−6.4	$1–4 \times 10^{-2}$		$0.8–2.2 \times 10^{-2}$	90.6–158
Thermoplastic polyester	1.31–1.58	250	3.3–3.8 (100 Hz)			$1.5–2.0 \times 10^{-3}$			232–295
Polyimide polyester	1.43–1.49	480°F		3.4 (100 kHz)			$1–5 \times 10^{-3}$ (100 kHz)		220
Polycarbonate	1.20	215	3.17		2.96	9×10^{-4}		1×10^{-2}	157
Epoxy (with mineral filler)	1.6–1.9	200 (decomposition temperature)	4.4–5.6	4.2–4.9	4.1–4.6	$1.1–8.3 \times 10^{-2}$	$0.19–1.4 \times 10^{-1}$	$0.13–1.4 \times 10^{-1}$	98.4–158
Epoxy (with silica filler)	1.6–2.0	200 (decomposition temperature)	3.2–4.5	3.2–4.0	3.0–3.8	$0.8–3.0 \times 10^{-2}$	$0.8–3.0 \times 10^{-2}$	$2–4 \times 10^{-2}$	158–217
Silicone rubber	1.1–1.5	700°F	3.3–4.0		3.1–3.7	$1.5–3.0 \times 10^{-2}$		$3.0–5.0 \times 10^{-3}$	158–197

The thin film of Si_3N_4 is characterized by extremely low losses, and its relatively closed structure does not provide any latitude for free sodium movement, thereby providing complete passivation of the semiconductor device. The high dielectric strength of the double film layer of SiO_2 and Si_3N_4 renders it dielectrically effective in field-effect transistor (FET) applications.

In integrated circuit devices, a number of materials are suitable for thin-film capacitor applications. In addition to Al_2O_3, tantalum pentoxide, Ta_2O_5, has been extensively utilized. It is characterized by high-temperature stability and is resistant to acids with the exception of hydrofluoric acid (HF). The high dielectric constant material hafnia (HfO_2) has also been used in thin-film capacitors.

Organic Solids

Solid organic dielectrics consist of large polymer molecules, which generally have molecular weights in excess of 600. Primarily, with the exception of paper, which consists of cellulose that is comprised of a series of glucose units, organic dielectric materials are synthetically derived.

Polyethylene (PE) is perhaps one of the most common solid dielectrics, which is extensively used as a solid dielectric extruded insulant in power and communication cables. Linear PE is classified as a low- (0.910–0.925), medium- (0.926–0.940), or high- (0.941–0.965) density polymer (cf. Table 55.2). Since PE is essentially a long-chain hydrocarbon material in which the repeat unit is $–CH_2–CH_2–$, a low-density PE necessarily implies a high degree of branching. Decreased branching increases the crystallinity as molecules undergo internal folding, which leads to improved stiffness, tear strength, hardness, and chemical resistance. Cross linking of PE produces a thermosetting polymer with a superior temperature rating, improved tensile strength, and an enhanced resistance to partial discharges. Most of the PE used on extruded cables is of the cross-linked polyethylene (XLPE) type.

Ethylene-propylene rubber (EPR) is an amorphous elastomer, which is synthesized from ethylene and propylene. As an extrudent on cables its composition has filler contents up to 50%, comprising primarily clay, with smaller amounts of added silicate and carbon black. The dielectric losses are appreciably enhanced by the fillers, and, consequently, EPR is not suitable for extra-high-voltage applications, with its use being usually confined to lower and intermediate voltages (≤138 kV) and also where high cable flexibility due to its rubber properties may be additionally desired.

Polypropylene has a structure related to that of ethylene with one added methyl group. It is a thermoplastic material having properties similar to high-density PE, though due to its lower density it has also a lower dielectric constant. It has many electrical applications both in bulk form as in molded and extruded insulations as well as in film form in taped capacitor, transformer, and cable insulations.

Polytetrafluoroethylene (PTFE) or Teflon is a fully fluorinated version of PE, having a repeat unit of $[–CF_2–CF_2–]$. It is characterized by a low dielectric constant, extremely low losses, and has excellent temperature stability and is resistive to chemical degradation. It has been extensively used in specialized applications on insulators, wires and cables, transformers, motors, and generators. Its relatively high cost is attributable to both the higher cost of the fluorinated monomers as well as the specialized fabrication techniques required.

Polyesters are obtained most commonly by reacting a dialcohol with a diester; they may be either thermosetting or thermoplastic. The former are usually employed in glass laminates and glass-fiber-reinforced moldings, while thermoplastic polyesters are used for injection-molding applications. They are used in small and large electrical apparatus as well as in electronic applications.

Polyimides (kaptons), as nylons (polyamides), have nitrogen in their molecular structure. They constitute a class of high-temperature thermoplastics that may be exposed to continuous operation at 480°F. When glass-reinforced, they may be exposed to temperatures as high as 700°F; they are used in molded, extruded wire, and film form.

Polycarbonates are thermoplastics that are closely related to polyesters. They are primarily employed in the insulation of electrically powered tools and in the casings of electrical appliances. Polycarbonates may be either compression- or injection-molded and extruded as films or sheets.

Epoxy resins are prepared from an epoxide monomer. The first step involves a reaction between two comonomers, and in the subsequent step the prepolymer is cured by means of a cross-linking agent. Epoxy resins are characterized by low shrinkage and high mechanical strength; they may be reinforced with glass fibers and mixed with mica flakes. Epoxy resins have many applications such as, for example, for insulation of bars in the stators of rotating machines, solid-type transformers, and spacers for compressed-gas-insulated busbars and cables.

TABLE 55.3 Electrical Properties of Taped Solid-Liquid Insulations

Tape	Impregnating Liquid	Average Voltage Stress (kV cm^{-1})	tanδ at Room Temperature	tanδ at Operating Temperature
Kraft paper	Mineral oil	180	3.8×10^{-3} at 23°C	5.7×10^{-3} at 85°C
Kraft paper	Silicone liquid	180	2.7×10^{-3} at 23°C	3.1×10^{-3} at 85°C
Paper-polypropylene-paper (PPP)	Dodecyl benzene	180	9.8×10^{-4} at 18°C	9.9×10^{-4} at 100°C
Kraft paper	Polybutene	180	2.0×10^{-3} at 25°C	2.0×10^{-3} at 85°C

Silicone rubber is classified as an organic-inorganic elastomer, which is obtained from the polymerization of organic siloxanes. They are composed of dimethyl-siloxane repeat units, $(CH_3)_2SiO$–, with the side groups being methyl units. Fillers are added to obtain the desired silicone rubber compounds; cross linking is carried out with peroxides. Since no softeners and plasticizers are required, silicone rubbers are resistant to embrittlement and may be employed for low- temperature applications down to –120°F. Continuous operation is possible up to 500°F, with intermittent usage as high as 700°F.

Solid-Liquid Insulating Systems

Impregnated-paper insulation constitutes one of the earliest insulating systems employed in electrical power apparatus and cables. Although in some applications alternate solid- or compressed-gas insulating systems are now being used, the impregnated-paper system still constitutes one of the most reliable insulating systems available. Proper impregnation of the paper results in a cavity-free insulating system, thereby eliminating the occurrence of partial discharges that inevitably lead to deterioration and breakdown of the insulating system. The cellulose structure of paper has a finite acidity content as well as a residual colloidal or bound water, which is held by hydrogen bonds. Consequently, impregnated cellulose base papers are characterized by somewhat more elevated tanδ values in the order of 2×10^{-3} at 30 kV cm^{-1}. The liquid impregnants employed are either mineral oils or synthetic fluids. Since the dielectric constant of these fluids is normally about 2.2 and that of dried cellulose about 6.5–10, the resulting dielectric constant of the impregnated paper is approximately 3.1–3.5.

Lower-density cellulose papers have slightly lower dielectric losses, but the dielectric breakdown strength is also reduced. The converse is true for impregnated systems utilizing higher-density papers. The general chemical formula of cellulose paper is $C_{12}H_{20}O_{10}$. If the paper is heated beyond 200°C, the chemical structure of the paper breaks down even in the absence of external oxygen, since the latter is readily available from within the cellulose molecule. To avert this process from occurring, cellulose papers are ordinarily not used beyond 100°C.

In an attempt to reduce the dielectric losses in solid-liquid systems, cellulose papers have been substituted in some applications by synthetic papers (cf. Table 55.3). For example in extra-high-voltage cables, cellulose paper–polypropylene composite tapes have been employed. A partial paper content in the composite tapes is necessary both to retain some of the impregnation capability of a porous cellulose paper medium and to maintain the relative ease of cellulose-to-cellulose tape sliding capability upon bending. In transformers the synthetic nylon or polyamide paper (nomex) has been used both in film and board form. It may be continuously operated at temperatures up to 220°C.

Defining Terms

Conductivity σ: Represents the ratio of the leakage current density to the applied electric field density. In general its ac and dc values differ because the mechanisms for establishing the leakage current in the two cases are not necessarily identical.

Dielectric: A material in which nearly all or a large portion of the energy required for its charging can be recovered when the external electric field is removed.

Dielectric constant ε': A quantity that determines the amount of electrostatic energy which can be stored per unit volume per unit potential gradient. It is a real quantity and in Gaussian–CGS units is numerically equal to the ratio of the measured capacitance of the specimen, C, to the equivalent geometrical capacitance *in vacuo,* C_o. It is also commonly referred to as the real value of the permittivity. Note that when the SI system of units is employed, the ratio C/C_o defines the real value of the *relative permittivity* ε'_r.

Dielectric loss: The rate at which the electrical energy supplied to a dielectric material by an alternating electrical field is changed to heat.

Dielectric strength: Represents the value of the externally applied electric field at which breakdown or failure of the dielectric takes place. Unless a completely uniform field gradient can be assured across the dielectric specimen, the resulting breakdown value will be a function of the specimen thickness and the test electrode geometry; this value will be substantially below that of the intrinsic breakdown strength.

Dissipation factor (**tanδ**): Equal to the tangent of the loss angle δ, which is the phase angle between the external electric field vector $\overline{E}$ and the resulting displacement vector $\overline{D}$. It is numerically equal to the ratio of the imaginary permittivity ε'' to the real permittivity, ε'; alternatively, it is defined by the ratio of the leakage current to the displacement (charging or capacitive) current.

Related Topics

54.1 Introduction • 58.4 Material Properties Conducive for Smart Material Applications

References

R. Bartnikas and R. M. Eichhorn (eds.), *Engineering Dielectrics,* Vol. II A, *Electrical Properties of Solid Insulating Materials: Molecular Structure and Electrical Behavior,* STP 783, Philadelphia: ASTM, 1983.

R. Bartnikas (ed.), *Engineering, Dielectrics,* Vol. III, *Electrical Insulating Liquids,* Monograph 2, Philadelphia: ASTM, 1994.

A. von Hippel, *Dielectrics and Waves,* New York: Wiley, 1956.

P. K. Watson, *IEEE Trans. on Dielectrics and Electrical Insulation,* 2, 915–924, 1995.

C. LeGressus and G. Blaise, *IEEE Trans. on Electrical Insulation,* 27, 472–481, 1992.

J. B. Bernstein, Ibid., pp. 152–161.

H. J. Wintle, *IEEE Trans. on Electrical Insulation,* 25, 27–44, 1990.

A. von Hippel and G.M. Lee. *Phys. Rev.,* 59, 824–826, 1941.

W.G. Oakes, *Proc. IEE,* 90(I), 37–43, 1949.

R. Bartnikas (ed.), *Engineering Dielectrics,* Vol. II B, *Measurement Techniques,* STP 926, Philadelphia: ASTM, 1987.

H. C. Miller, *IEEE Trans. on Electrical Insulation,* 28, 512–527, 1993.

R. Bartnikas and E.J. McMahon (eds.), *Engineering Dielectrics,* Vol. I, *Corona Measurement and Interpretation,* STP 669, Philadelphia: ASTM, 1979.

J. C. Fothergill, L. A. Dissado, and P. J. J. Sweeny, *IEEE Trans. on Dielectrics and Electrical Insulation,* 1, 474–486, 1994.

Encyclopedia Issue, *Insul. Circuits,* June/July 1972.

Further Information

The IEEE Dielectrics and Electrical Insulation Society publishes regularly its *IEEE Transactions on Dielectrics and Electrical Insulation,* wherein many new developments in the field of dielectrics are recorded in permanent form. It also sponsors on either an annual or a biennial basis a number of conferences, which provide a forum for rapid dissemination of both the applied and fundamental work carried out on dielectrics and electrical insulating systems. The reader may wish to consult the IEEE Conference Records on the Annual Report, Conference on Electrical Insulation and Dielectric Phenomena, the IEEE International Symposium on Electrical Insulation, and the Electrical/Electronics Insulation Conference. Also, a description of the different test methods on dielectric materials may be found in the *ASTM Book on Standards.*

Smith, R.L. “Sensors”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

56

Sensors

56.1 Introduction
56.2 Physical Sensors
Temperature Sensors • Displacement and Force • Optical Radiation
56.3 Chemical Sensors
Ion-Selective Electrode • Gas Chromatograph
56.4 Biosensors
Immunosensor • Enzyme Sensor
56.5 Microsensors

Rosemary L. Smith
University of California, Davis

56.1 Introduction

Sensors are critical components in all measurement and control systems. The need for computer-compatible sensors closely followed the advent of the microprocessor. Together with the always-present need for sensors in science and medicine, the demand for sensors in automated manufacturing and processing is rapidly growing. In addition, small, inexpensive sensors are finding their way into all sorts of consumer products, from childrens' toys to dishwashers to automobiles. Because of the vast variety of useful things to be sensed and sensor applications, sensor engineering is a multidisciplinary and interdisciplinary field of endeavor. This chapter introduces some basic definitions, concepts, and features of sensors and illustrates them with several examples. The reader is directed to the references and the sources listed under "Further Information" for more details and examples.

There are many terms which are often used synonymously for sensor, including transducer, meter, detector, and gage. Defining the term sensor is not an easy task; however the most widely used definition is that which has been applied to electrical transducers by the Instrument Society of America (ANSI MC6.1, 1975): *Transducer—A device which provides a usable output in response to a specified measurand.* A transducer is more generally defined as a device which converts energy from one form to another. A usable ouput refers to an optical, electrical, or mechanical signal. In the context of electrical engineering, however, a usable output refers to an electrical output signal. The measurand can be a physical, chemical, or biological property or condition to be measured.

Most, but not all, sensors are transducers, employing one or more transduction mechanisms to produce an electrical output signal. Sometimes sensors are classified as direct and indirect sensors according to how many transduction mechanisms are used. For example, a mercury thermometer produces a change in volume of mercury in response to a temperature change via thermal expansion, but the output is a mechanical displacement and not an electrical signal. Another transduction mechanism is required. A thermometer is still a useful sensor since humans can read the change in mercury height using their eyes as the second transducing element. However, in order to produce an electrical output for use in a control loop, the height of the mercury would have to be converted to an electrical signal. This could be accomplished using capacitive effects. However, there are more direct temperature sensing methods, i.e., one where an electrical output is produced in response to a change in temperature. An example is given in the next section on physical sensors. Figure 56.1 depicts a

TABLE 56.1 Physical and Chemical Transduction Principles

Primary Signal	Secondary Signal					
	Mechanical	Thermal	Electrical	Magnetic	Radiant	Chemical
Mechanical	(Fluid) mechanical and acoustic effects (e.g., diaphragm, gravity balance, echo sounder)	Friction effects (e.g., friction calorimeter) Cooling effects (e.g., thermal flow meters)	Piezoelectricity Piezoresistivity Resistive, capacitive, and inductive effects	Magneto-mechanical effects (e.g., piezo-magnetic effect)	Photoelastic systems (stress-induced birefringence) Interferometers Sagnac effect Doppler effect	
Thermal	Thermal expansion (bimetal strip, liquid-in-glass and gas thermometers, resonant frequency) Radiometer effect (light mill)		Seebeck effect Thermoresistance Pyroelectricity Thermal (Johnson) noise		Thermooptical effects (e.g., in liquid crystals) Radiant emission	Reaction activation (e.g., thermal dissociation)
Electrical	Electrokinetic and electro-mechanical effects (e.g., piezoelectricity, electro-meter, Ampere's law)	Joule (resistive) heating Peltier effect	Charge collectors Langmuir probe	Biot-Savart's law	Electrooptical effects (e.g., Kerr effect) Pockel's effect Electroluminescence	Electrolysis Electromigration
Magnetic	Magnetomechanical effects (e.g., magnetorestriction, magnetometer)	Thermomagnetic effects (e.g., Righi-Leduc effect) Galvanomagnetic effects (e.g., Ettingshausen effect)	Thermomagnetic effects (e.g., Ettingshausen-Nernst effect) Galvanomagnetic effects (e.g., Hall effect, magnetoresistance)		Magnetooptical effects (e.g., Faraday effect) Cotton-Mouton effect	
Radiant	Radiation pressure	Bolometer thermopile	Photoelectric effects (e.g., photovoltaic effect, photoconductive effect)		Photorefractive effects Optical bistability	Photosynthesis, -dissociation
Chemical	Hygrometer Electrodeposition cell Photoacoustic effect	Calorimeter Thermal conductivity cell	Potentiometry Conductimetry Amperometry Flame ionization Volta effect Gas-sensitive field effect	Nuclear magnetic resonance	(Emission and absorption) spectroscopy Chemiluminiscence	

Source: T. Grandke and J. Hesse, Introduction, Vol. 1: *Fundamentals and General Aspects, Sensors: A Comprehensive Survey,* W. Gopel, J. Hesse, and J. H. Zemel, Eds., Weinheim, Germany: VCH, 1989. With permission.

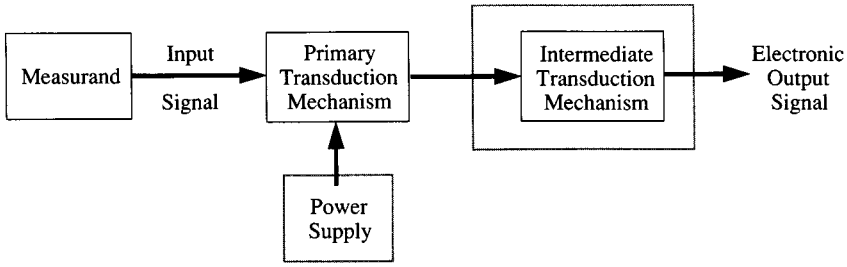


FIGURE 56.1 Sensor block diagram. Active sensors require input power to accomplish transduction. Many sensors employ multiple transduction mechanisms in order to produce an electronic output in response to the measurand.

sensor block diagram identifying the measurand and associated input signal, the primary and intermediate transduction mechanisms, and the electronic output signal. Active sensors require an external power source in order to produce a usable output signal, e.g., the piezoresistor. Table 56.1 is a 6×6 matrix of the more commonly employed physical and chemical transduction mechanisms. Many of the effects listed are described in more detail in this handbook (see Chapters 53–58).

In choosing a particular sensor for a given application, there are many factors to be considered. These deciding factors or specifications can be divided into three major categories: environmental factors, economic factors, and the sensor characteristics. The most commonly encountered factors are listed in Table 56.2, although not all of these factors may be pertinent to a particular application. Most of the environmental factors determine the packaging of the sensor, with packaging meaning the encapsulation or insulation which provides protection and isolation and the input/output leads or connections and cabling. The economic factors determine the type of manufacturing and materials used in the sensor and to some extent the quality of the materials (with respect to lifetime). For example, a very expensive sensor may be cost effective if it is used repeatedly or for very long periods of time. On the other hand, a disposable sensor, such as is desired in many medical applications, should be inexpensive. The sensor characteristics of the sensor are usually the specifications of primary concern. The most important parameters are **sensitivity, stability,** and **repeatability.** Normally, a sensor is only useful if all three of these parameters are tightly specified for a given range of measurand and time of operation. For example, a highly sensitive device is not useful if its output signal drifts greatly during the measurement time and the data obtained is not reliable if the measurement is not repeatable. Other output characteristics, such as selectivity and linearity, can often be compensated for by using additional, independent sensor input or with signal conditioning circuits. In fact, most sensors have a response to temperature, since most tranducing effects are temperature dependent.

Sensors are most often classified by the type of measurand, i.e., physical, chemical, or biological. This is a much simpler means of classification than by transduction mechanism or output signal (e.g., digital or analog), since many sensors use multiple transduction mechanisms and the output signal can always be processed, conditioned, or converted by a circuit so as to cloud the definition of output. A description of each class and examples are given in the following sections. The last section introduces microsensors and gives some examples.

TABLE 56.2

Environmental Factors	Economic Factors	Sensor Characteristics
Temperature range	Cost	Sensitivity
Humidity effects	Availability	Range
Corrosion	Lifetime	Stability
Size		Repeatability
Overrange protection		Linearity
Susceptibility to EM interferences		Error
Ruggedness		Response time
Power consumption		Frequency response
Self-test capability		

56.2 Physical Sensors

Physical measurands include temperature, strain, force, pressure, displacement, position, velocity, acceleration, optical radiation, sound, flow rate, viscosity, and electromagnetic fields. Referring to Table 56.1, all but those transduction mechanisms listed in the chemical column are used in the design of physical sensors. Clearly, they comprise a very large proportion of all sensors. It is impossible to illustrate all of them, but three measurands stand out in terms of their widespread application: temperature, displacement (or associated force), and optical radiation.

Temperature Sensors

Temperature is an important parameter in many control systems, most familiarly in environmental control systems. Several distinctly different transduction mechanisms have been employed. The mercury thermometer was mentioned in the Introduction as a nonelectrical sensor. The most commonly used electrical temperature sensors are thermocouples, thermistors, and resistance thermometers. Thermocouples employ the Seebeck effect, which occurs at the junction of two dissimilar metal wires. A voltage difference is generated at the hot junction due to the difference in the energy distribution of thermally energized electrons in each metal. This voltage is measured across the cool ends of the two wires and changes linearly with temperature over a given range, depending on the choice of metals. To minimize measurement error the cool end of the couple must be kept at a constant temperature, and the voltmeter must have a high input impedance.

The resistance thermometer relies on the increase in resistance of a metal wire with increasing temperature. As the electrons in the metal gain thermal energy, they move about more rapidly and undergo more frequent collisions with each other and the atomic nuclei. These scattering events reduce the mobility of the electrons, and since resistance is inversely proportional to mobility, the resistance increases. Resistance thermometers consist of a coil of fine metal wire. Platinum wire gives the largest linear range of operation. To determine the resistance indirectly, a constant current is supplied and the voltage is measured. A direct measurement can be made by placing the resistor in the sensing arm of a Wheatstone bridge and adjusting the opposing resistor to "balance" the bridge, which produces a null output. A measure of the sensitivity of a resistance thermometer is its temperature coefficient of resistance: TCR = $(\Delta R/R)(1/\Delta T)$ in units of % resistance per degree of temperature.

Thermistors are resistive elements made of semiconductor materials and have a negative coefficient of resistance. The mechanism governing the resistance change of a thermistor is the increase in the number of conducting electrons with an increase in temperature due to thermal generation, i.e., the electrons which are the least tightly bound to the nucleus (valence electrons) gain sufficient thermal energy to break away and become influenced by external fields. Thermistors can be measured in the same manner as resistance thermometers, but thermistors have up to 100 times higher TCR values.

Displacement and Force

Many types of forces are sensed by the displacements they create. For example, the force due to acceleration of a mass at the end of a spring will cause the spring to stretch and the mass to move. Its displacement from the zero acceleration position is governed by the force generated by the acceleration ($F = m \cdot a$) and the restoring force of the spring. Another example is the displacement of the center of a deformable membrane due to a difference in pressure across it. Both of these examples use multiple transduction mechanisms to produce an electronic output: a primary mechanism which converts force to displacement (mechanical to mechanical) and then an intermediate mechanism to convert displacement to an electrical signal (mechanical to electrical).

Displacement can be measured by an associated capacitance. For example, the capacitance associated with a gap which is changing in length is given by C = area × dielectric constant/gap length. The gap must be very small compared to the surface area of the capacitor, since most dielectric constants are of the order of 1×10^{-13} farads/cm and with present methods, capacitance is readily resolvable to only about 10^{-12} farads. This is because measurement leads and contacts create parasitic capacitances the same order of magnitude. If the capacitance is measured at the generated site by an integrated circuit (see Section III), capacitances as small as 10^{-15} farads

can be measured. Displacement is also commonly measured by the movement of a ferromagnetic core inside of an inductor coil. The displacement produces a change in inductance which can be measured by placing the inductor in an oscillator circuit and measuring the change in frequency of oscillation.

The most commonly used force sensor is the strain gage. It consists of metal wires which are stretched in response to a force. The resistance of the wire changes as it undergoes strain, i.e., a change in length, since the resistance of a wire is R = resistivity × length/cross-sectional area. The wire's resistivity is a bulk property of the metal which is a constant for constant temperature. For example, a strain gage can be used to measure acceleration by attaching both ends of the wire to a cantilever beam, with one end of the wire at the attached beam end and the other at the free end. The cantilever beam free end moves in response to an applied force, such as the force due to acceleration which produces strain in the wire and a subsequent change in resistance. The sensitivity of a strain gage is described by the unitless gage factor, $G = (\Delta R/R)/(\Delta L/L)$. For metal wires, gage factors typically range from 2 to 3. Semiconductors are known to exhibit piezoresistivity, which is a change in resistance in response to strain which involves a large change in resistivity in addition to the change in linear dimension. Piezoresistors have gage factors as high as 130. Piezoresistive strain gages are frequently used in **microsensors,** described in Section 56.5.

Optical Radiation

The intensity and frequency of optical radiation are parameters of growing interest and utility in consumer products such as the video camera and home security systems and in optical communications systems. The conversion of optical energy to electronic signals can be accomplished by several mechanisms (see radiant to electronic transduction in Table 56.1); however, the most commonly used is the photogeneration of carriers in semiconductors. The most often-used device is the *p-n* junction photodiode (Section III). The construction of this device is very similar to the diodes used in electronic circuits as rectifiers. The diode is operated in reverse bias, where very little current normally flows. When light is incident on the structure and is absorbed in the semiconductor, energetic electrons are produced. These electrons flow in response to the electric field sustained internally across the junction, producing an externally measurable current. The current magnitude is proportional to the light intensity and also depends on the frequency of the light. Figure 56.2 shows the effects of varying incident optical intensity on the terminal current versus voltage behavior of a *p-n* junction. Note that for zero applied voltage, a net negative current flows when the junction is illuminated. This device can therefore also be a source of power (a solar cell).

56.3 Chemical Sensors

Chemical measurands include ion concentration, chemical composition, rate of reactions, reduction-oxidation potentials, and gas concentration. The last column of Table 56.1 lists some of the transduction mechanisms that have been, or could be, employed in chemical sensing. Two examples of chemical sensors are described

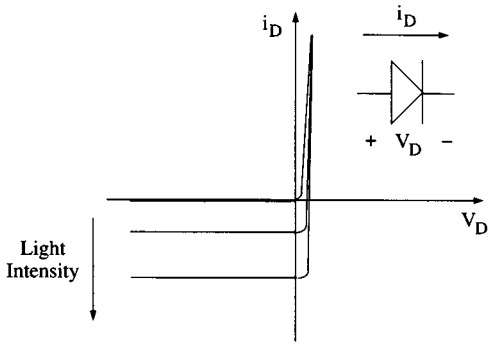


FIGURE 56.2 Sketch of the variation of current versus voltage characteristics of a *p-n* photodiode with incident light intensity.

here: the ion-selective electrode (ISE) and the gas chromatograph. They were chosen because of their general use and availability and because they illustrate the use of a primary (ISE) versus a primary plus intermediate (gas chromatograph) transduction mechanism.

Ion-Selective Electrode (ISE)

As the name implies, ISEs are used to measure the concentration of a specific ion concentration in a solution of many ions. To accomplish this, a membrane which selectively generates a potential which is dependent on the concentration of the ion of interest is used. The generated potential is usually an equilibrium potential, called the Nernst potential, and develops across the interface of the membrane with the solution. This potential is generated by the initial net flow of ions (charge) across the membrane in response to a concentration gradient, and from thence forth the diffusional force is balanced by the generated electric force and equilibrium is established. This is very similar to the so-called built-in potential of a *p-n* junction diode. The ion-selective membrane acts in such a way as to ensure that the generated potential is dependent mostly on the ion of interest and negligibly on any other ions in solution. This is done by enhancing the exchange rate of the ion of interest across the membrane, so it is the fastest moving and, therefore, the species which generates and maintains the potential.

The most familiar ISE is the pH electrode. In this device the membrane is a sodium glass which possesses a high exchange rate for H^+. The generated Nernst potential, E, is given by the expression: $E = E_0 + (RT/F) \ln[H^+]$, where E_0 is a constant for constant temperature, R is the gas constant, and F is the Faraday constant. pH is defined as the negative of the $\log[H^+]$; therefore $\text{pH} = (E_0 - E)(\log e)F/RT$. One pH unit change corresponds to a tenfold change in the molar concentration of H^+ and a 59 mV change in the Nernst potential at room temperature. Other ISEs have the same type of response, but specific to a different ion, depending on the choice of membrane. Many ISEs employ ionophores trapped inside of a polymeric membrane. An ionophore is a molecule which selectively and reversibly binds with an ion and thereby creates a high exchange rate for that particular ion.

The ISE consists of a glass tube with the ion-selective membrane closing that end of the tube which is immersed into the test solution. The Nernst potential is measured by making electrical contact to each side of the membrane. This is done by placing a fixed concentration of conductive filling solution inside of the tube and placing a wire into the solution. The other side of the membrane is contacted by a reference electrode placed inside of the same solution under test. The reference electrode is constructed in the same manner as the ISE but it has a porous membrane which creates a liquid junction between its inner filling solution and the test solution. That junction is designed to have a potential which is invariant with changes in concentration of any ion in the test solution. The reference electrode, solution under test, and the ISE form an electrochemical cell. The reference electrode potential acts like the ground reference in electric circuits, and the ISE potential is measured between the two wires emerging from the respective two electrodes. The details of the mechanisms of transduction in ISEs are beyond the scope of this chapter. The reader is referred to Bard and Faulkner [1980] and Janata [1989].

Gas Chromatograph

Molecules in gases have thermal conductivities which are dependent on their masses; therefore, a pure gas can be identified by its thermal conductivity. One way to determine the composition of a gas is to first separate it into its components and then measure the thermal conductivity of each. A gas chromatograph does exactly that. The gas flows through a long narrow column, which is packed with an adsorbant solid (for gas–solid chromatography) wherein the gases are separated according to the retentive properties of the packing material for each gas. As the individual gases exit the end of the tube one at a time, they flow over a heated wire. The amount of heat transferred to the gas depends on its thermal conductivity. The gas temperature is measured a short distance downstream and compared to a known gas flowing in a separate sensing tube. The temperature is related to the amount of heat transferred and can be used to derive the thermal conductivity according to thermodynamic theory and empirical data. This sensor required two transductions: a chemical to thermal energy transduction followed by a thermal to electrical transduction.

56.4 Biosensors

Biological measurands are biologically produced substances, such as antibodies, glucose, hormones, and enzymes. Biosensors are not the same as biomedical sensors, which are any sensors used in biomedical applications, such as blood pressure sensors, or electrocardiogram electrodes. Many biosensors are biomedical sensors; however, they are also used in industrial applications, e.g., the monitoring and control of fermentation reactions. Table 56.1 does not include biological signals as a primary signal because they can be classified as either chemical or physical in nature. Biosensors are of special interest because of the very high selectivity of biological reactions and binding. However, the detection of that reaction or binding is often elusive. A very familiar commercial biosensor is the in-home pregnancy test sensor, which detects the presence of human growth factor in urine. That device is a nonelectrical sensor since the output is a color change which the eye senses. In fact, most biosensors require multiple transduction mechanisms to arrive at an electrical output signal. Two examples are given below: an immunosensor and an enzyme sensor. Rather than examine a specific species, the examples describe a general type of sensor and transduction mechanism, since the same principles can be applied to a very large number of biological species of the same type.

Immunosensor

Commercial techniques for detecting antibody-antigen binding utilize optical or x-radiation detection. An optically fluorescent molecule or radioisotope is nonspecifically attached to the species of interest in solution. The complementary binding species is chemically attached to a glass substrate or glass beads which are packed into a column. The tagged solution containing the species of interest, say the antibody, is passed over the antigen-coated surface, where the two selectively bind. After the specific binding occurs, the nonbound fluorescent molecules or radioisotopes are washed away, and the antibody concentration is determined by fluorescence spectroscopy or with a scintillation counter, respectively. These sensing techniques are quite costly and bulky, and therefore other biosensing mechanisms are rapidly being developed. One experimental technique uses the change in the mechanical properties of the bound antibody-antigen complex in comparison to an unbound surface layer of antigen. It uses a shear mode, surface acoustic wave (SAW) device (see Chapter 51 and [Ballentine et al., 1997]) to sense this change as a change in the propagation time of the wave between the generating electrodes and the pick-up electrodes some distance away on the same piezoelectric substrate. The substrate surface is coated with the antigen and it is theorized that upon selectively binding with the antibody, this layer stiffens, changing the mechanical properties of the interface and therefore the velocity of the wave. The advantages of this device are that the SAW device produces an electrical signal (a change in oscillation frequency when the device is used in the feedback loop of an oscillator circuit) which is dependent on the amount of bound antibody; it requires only a very small amount of the antigen which can be very costly; the entire device is small, robust and portable; and the detection and readout method is inexpensive. However, there are numerous problems which currently preclude its commercial use, specifically a large temperature sensitivity and responses to nonspecific adsorption, i.e., by species other than the desired antibody.

Enzyme Sensor

Enzymes selectively react with a chemical substance to modify it, usually as the first step in a chain of reactions to release energy (metabolism). A well-known example is the selective reaction of glucose oxidase (enzyme) with glucose to produce gluconic acid and peroxide, according to

$$C_6H_{12}O_6 + O_2 \xrightarrow{\text{glucose oxidase}} \text{gluconic acid} + H_2O_2 + 80 \text{ kilojoules heat}$$

An enzymatic reaction can be sensed by measuring the rise in temperature associated with the heat of reaction or by the detection and measurement of byproducts. In the glucose example, the reaction can be sensed by measuring the local dissolved peroxide concentration. This is done via an electrochemical analysis technique called amperometry [Bard and Faulkner, 1980]. In this method, a potential is placed across two inert metal

wire electrodes immersed in the test solution and the current which is generated by the reduction/oxidation reaction of the species of interest is measured. The current is proportional to the concentration of the reducing/oxidizing species. A selective response is obtained if no other available species has a lower redox potential. Because the selectivity of peroxide over oxygen is poor, some glucose sensing schemes employ a second enzyme called catalase which converts peroxide to oxygen and hydroxyl ions. The latter produces a change in the local pH. As described earlier, an ISE can then be used to convert the pH to a measurable voltage. In this latter example, glucose sensing involves two chemical-to-chemical transductions followed by a chemical-to-electrical transduction mechanism.

56.5 Microsensors

Microsensors are sensors that are manufactured using integrated circuit fabrication technologies and/or **micromachining.** Integrated circuits are fabricated using a series of process steps which are done in batch fashion, meaning that thousands of circuits are processed together at the same time in the same way. The patterns which define the components of the circuit are photolithographically transferred from a template to a semiconducting substrate using a photosensitive organic coating. The coating pattern is then transferred into the substrate or into a solid-state thin film coating through an etching or deposition process. Each template, called a mask, can contain thousands of identical sets of patterns, with each set representing a circuit. This "batch" method of manufacturing is what makes integrated circuits so reproducible and inexpensive. In addition, photoreduction enables one to make extremely small features, on the order of microns, which is why this collection of process steps is referred to as microfabrication. The resulting integrated circuit is contained in only the top few microns of the semiconductor substrate and the submicron thin films on its surface. Hence, integrated circuit technology is said to consist of a set of planar, microfabrication processes. Micromachining refers to the set of processes which produce three-dimensional microstructures using the same photolithographic techniques and batch processing as for integrated circuits. Here, the third dimension refers to the height above the substrate of the deposited layer or the depth into the substrate of an etched structure. Micromachining produces third dimensions in the range of 1–500 μm (typically). The use of microfabrication to manufacture sensors produces the same benefits as it does for circuits: low cost per sensor, small size, and highly reproducible behavior. It also enables the integration of signal conditioning, compensation circuits and actuators, i.e., entire sensing and control systems, which can dramatically improve sensor performance for very little increase in cost. For these reasons, there is a great deal of research and development activity in microsensors.

The first microsensors were integrated circuit components, such as semiconductor resistors and *p-n* junction diodes. The piezoresistivity of semiconductors and optical sensing by the photodiode were already discussed. Diodes are also used as temperature-sensing devices. When forward-biased with a constant diode current, the resulting diode voltage increases approximately linearly with increasing temperature. The first micromachined microsensor to be commercially produced was the silicon pressure sensor. It was invented in the mid-to-late 1950s at Bell Labs and commercialized in the 1960s. This device contains a thin silicon diaphragm (≈10 μm) which is produced by chemical etching. The diaphragm deforms in response to a pressure difference across it (Fig. 56.3). The deformation produces two effects: a position-dependent displacement which is maximum at the diaphragm center and position-dependent strain which is maximum near the diaphragm edge. Both of these effects have been used in microsensors to produce an electrical output which is proportional to differential pressure. The membrane displacement is sensed capacitively as previously described in one type of pressure sensor. The strain is sensed in another by placing a piezoresistor, fabricated in the same silicon substrate, along one edge of the diaphragm. The two leads of the piezoresistor are connected to a Wheatstone bridge. The latter type of sensor is called a piezoresistive pressure sensor and is the commercially more common type of pressure microsensor. Pressure microsensors constituted about 5% of the total U.S. consumption of pressure sensors in 1991. Most of them are used in the medical industry as disposables due to their low cost and small, rugged construction. Many other types of microsensors are commercially under development, including accelerometers, mass flow rate sensors, and biosensors.

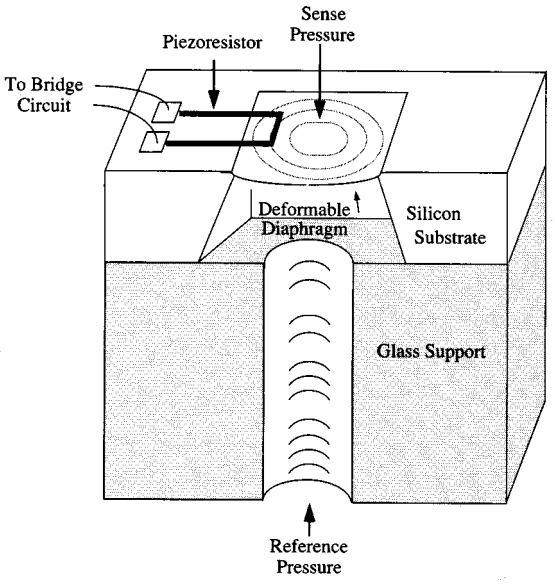


FIGURE 56.3 Schematic cross section of a silicon piezoresistive pressure sensor. A differential pressure deforms the silicon diaphragm, producing strain in the integrated piezoresistor. The change in resistance is measured via a Wheatstone bridge.

Defining Terms

Micromachining: The set of processes which produce three-dimensional microstructures using sequential photolithographic pattern transfer and etching or deposition in a batch processing method.

Microsensor: A sensor which is fabricated using integrated circuit and micromachining technologies.

Repeatability: The ability of a sensor to reproduce output readings for the same value of measurand, when applied consecutively and under the same conditions.

Sensitivity: The ratio of the change in sensor output to a change in the value of the measurand.

Sensor: A device which produces a usable output in response to a specified measurand.

Stability: The ability of a sensor to retain its characterisctics over a relatively long period of time.

Related Topics

58.6 Smart Sensors • 114.1 Introduction • 114.2 Physical Sensors • 114.3 Chemical Sensors • 114.4 Bioanalytical Sensors • 114.5 Applications

References

ANSI, "Electrical Transducer Nomenclature and Terminology," ANSI Standard MC6.1-1975 (ISA S37.1), Research Triangle Park, N.C.: Instrument Society of America, 1975.

D. S. Ballentine, Jr. et al., *Acoustic Wave Sensors: Theory, Design, and Physico-Chemical Applications,* San Diego, Calif.: Academic Press, 1997.

A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications,* New York: John Wiley & Sons, 1980.

R. S. C. Cobbold, *Transducers for Biomedical Measurements: Principles and Applications,* New York: John Wiley & Sons, 1974.

W. Göpel, J. Hesse, and J. N. Zemel, Eds., *Sensors: A Comprehensive Survey,* vol. 1, *Fundamentals and General Aspects,* T. Grandke and W. H. Ko, Eds., Weinheim, Germany: VCH, 1989.

J. Janata, *Principles of Chemical Sensors,* New York, Plenum Press, 1989.

Further Information

Sensors: A Comprehensive Survey, W. Gopel, J. Hesse, and J. N. Zemel, editors. Weinheim, F.R.G. VCH, 1989–1994.

Vol. 1: *Fundamentals and General Aspects,* T. Grandke and W. H. Ko, Eds.

Vol. 2, 3, pt. 1–2: *Chemical and Biochemical Sensors,* W. Gopel et al., Eds.

Vol. 4: *Thermal Sensors,* T. Ricolfi and J. Scholz, Eds.

Vol. 5: *Magnetic Sensors,* R. Boll and K. J. Overshott, Eds.

Vol. 6: *Optical Sensors,* E. Wagner, R. Dandliker, and K. Spenner, Eds.

Vol. 7: *Mechanical Sensors,* H. H. Bau, N. F. deRooij, and B. Kloeck, Eds.

J. Carr, *Sensors and Circuits: Sensors, Transducers, and Supporting Circuits for Electronic Instrumentation, Measurement, and Control,* Englewood Cliffs, N.J.: Prentice-Hall, 1993.

J. R. Carstens, *Electrical Sensors and Transducers,* Englewood Cliffs, N.J.: Regents/Prentice-Hall, 1993.

M. J. Usher and D. A. Keating, *Sensors and Transducers: Characteristics, Applications, Instrumentation, Interfacing,* 2nd ed., New York: Macmillan, 1996.

S. M. Sze, Ed., *Semiconductor Sensors,* New York: John Wiley & Sons, 1994.

D. Tandeske, *Pressure Sensors: Selection and Application,* New York: Marcel Dekker, 1991.

Sensors and Actuators is a technical journal devoted to solid-state sensors and actuators, which is published bimonthly by Elsevier Press in two volumes: Vol. A: *Physical Sensors* and Vol. B: *Chemical Sensors.*

The International Conference on Solid-State Sensors and Actuators is held every 2 years, hosted in rotation by the U.S., Japan, and Europe. It is sponsored in part by IEEE in the U.S. and a digest of technical papers is published and available through IEEE.

Young, D., Pu, Y. “Magnetooptics”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

57

Magnetooptics

57.1 Introduction

57.2 Classification of Magnetooptic Effects
Faraday Rotation or Magnetic Circular Birefringence • Cotton-Mouton Effect or Magnetic Linear Birefringence • Kerr Effects

57.3 Applications of Magnetooptic Effects
Optical Isolator and Circulator • MSW-Based Guided-Wave Magnetooptic Bragg Cell • Magnetooptic Recording

David Young
Rockwell Semiconductor Systems

Yuan Pu
Applied Materials

57.1 Introduction

When a magnetic field **H** is applied to a magnetic medium (crystal), a change in the magnetization **M** within the medium will occur as described by the constitution relation of the Maxwell equations $\mathbf{M} = \overleftrightarrow{\chi} \cdot \mathbf{H}$ where $\overleftrightarrow{\chi}$ is the magnetic susceptibility tensor of the medium. The change in magnetization can in turn induce a perturbation in the complex optical permittivity tensor $\overleftrightarrow{\varepsilon}$. This phenomenon is called the magnetooptic effect. Mathematically, the magnetooptic effect can be described by expanding the permittivity tensor as a series in increasing powers of the magnetization [Torfeh et al., 1977] as follows:

$$\overleftrightarrow{\varepsilon} = \varepsilon_0[\varepsilon_{ij}] \tag{57.1}$$

where

$$\varepsilon_{ij}(\mathbf{M}) = \varepsilon_r\delta_{ij} + jf_1\ e_{ijk}M_k + f_{ijkl}M_kM_l$$

Here j is the imaginary number. M_1, M_2, and M_3 are the magnetization components along the principal crystal axes X, Y, and Z, respectively. ε_0 is the permittivity of free space. ε_r is the relative permittivity of the medium in the paramagnetic state (i.e., $\mathbf{M} = 0$), f_1 is the first-order magnetooptic scalar factor, $f_{ijk\ l}$ is the second-order magnetooptic tensor factor, δ_{ij} is the Kronecker delta, and e_{ijk} is the antisymmetric alternate index of the third order. Here we have used Einstein notation of repeated indices and have assumed that the medium is quasi-transparent so that $\overleftrightarrow{\varepsilon}$ is a Hermitian tensor. Moreover, we have also invoked the Onsager relation in thermodynamical statistics, i.e., $\varepsilon_{ij}(\mathbf{M}) = \varepsilon_{ji}(-\mathbf{M})$. The consequences of Hermiticity and Onsager relation are that the real part of the permittivity tensor is an even function of **M** whereas the imaginary part is an odd function of **M**. For a cubic crystal, such as YIG (yttrium-iron-garnet), the tensor f_{ijkl} reduces to only three independent terms. In terms of Voigt notation, they are f_{11}, f_{12}, and f_{44}. In a principal coordinate system, the tensor can be expressed as

$$f_{ijkl} = f_{12}\delta_{ij}\delta_{kl} + f_{44}\ (\delta_{il}\delta_{kj} + \delta_{ik}\delta_{lj}) + \Delta f\,\delta_{kl}\delta_{ij}\delta_{jk} \tag{57.2}$$

where $\Delta f = f_{11} - f_{12} - 2f_{44}$.

In the principal crystal axes [100] coordinate system, the magnetooptic permittivity reduces to the following forms:

$$\overleftrightarrow{\varepsilon} = \varepsilon_0 \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{12}^{\star} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{13}^{\star} & \varepsilon_{23}^{\star} & \varepsilon_{33} \end{bmatrix}$$

where * denotes complex conjugate operation. The elements are given by

paramagnetic state

$$\overleftrightarrow{\varepsilon} = \varepsilon_0 \begin{bmatrix} \varepsilon_r & 0 & 0 \\ 0 & \varepsilon_r & 0 \\ 0 & 0 & \varepsilon_r \end{bmatrix}$$

Faraday rotation

$$+ \quad \varepsilon_0 \begin{bmatrix} 0 & +jf_1M_3 & -jf_1M_2 \\ -jf_1M_3 & 0 & +jf_1M_1 \\ +jf_1M_2 & -jf_1M_1 & 0 \end{bmatrix}$$

Cotton-Mouton effect

$$+ \quad \varepsilon_0 \begin{bmatrix} f_{11}M_1^2 + f_{12}M_2^2 + f_{12}M_3^2 & 2f_{44}M_1M_2 & 2f_{44}M_1M_3 \\ 2f_{44}M_1M_2 & f_{12}M_1^2 + f_{11}M_2^2 + f_{12}M_3^2 & 2f_{44}M_2M_3 \\ 2f_{44}M_1M_3 & 2f_{44}M_2M_3 & f_{12}M_1^2 + f_{12}M_2^2 + f_{11}M_3^2 \end{bmatrix} \tag{57.3}$$

In order to keep the discussion simple, analytic complexities due to optical absorption of the magnetic medium have been ignored. Such absorption can give rise to magnetic circular dichroism (MCD) and magnetic linear dichroism (MLD). Interested readers can refer to Hellwege [1978] and Arecchi and Schulz-DuBois [1972] for more in-depth discussions on MCD and MLD.

57.2 Classification of Magnetooptic Effects

Faraday Rotation or Magnetic Circular Birefringence

The classic Faraday rotation takes place in a cubic or isotropic transparent medium where the propagation direction of transmitted light is parallel to the direction of applied magnetization within the medium. For example, if the direction of magnetization and the propagation of light is taken as *Z*, the permittivity tensor becomes (assuming second-order effect is insignificantly small):

$$\overleftrightarrow{\varepsilon} \cong \varepsilon_0 \begin{bmatrix} \varepsilon_r & jf_1M_3 & 0 \\ -jf_1M_3 & \varepsilon_r & 0 \\ 0 & 0 & \varepsilon_r \end{bmatrix} \tag{57.4}$$

The two eigenmodes of light propagation through the magnetooptic medium can be expressed as a right circular polarized (RCP) light wave

$$\tilde{E}_1(Z) = \begin{bmatrix} 1 \\ j \\ 0 \end{bmatrix} \exp\left[j\left(\omega t - \frac{2\pi n_+}{\lambda_0} Z \right)\right] \tag{57.5a}$$

and a left circular polarized (LCP) light wave

$$\tilde{E}_2(Z) = \begin{bmatrix} 1 \\ -j \\ 0 \end{bmatrix} \exp\left[j\left(\omega t - \frac{2\pi n_-}{\lambda_0} Z \right)\right] \tag{57.5b}$$

where $n_\pm^2 \cong \varepsilon_r \pm f_1 M_3$; ω and λ_0 are the angular frequency and the wavelength of the incident light, respectively. n_+ and n_- are the refractive indices of the RCP and LCP modes, respectively. These modes correspond to two counterrotating circularly polarized light waves. The superposition of these two waves produces a linearly polarized wave. The plane of polarization of the resultant wave rotates as one circular wave overtakes the other. The rate of rotation is given by

$$\begin{aligned} \theta_F &\cong \frac{\pi f_1 M_3}{\lambda_0 \sqrt{\varepsilon_r}} \quad \text{rad/m} \\ &= \frac{1.8 f_1 M_3}{\lambda_0 \sqrt{\varepsilon_r}} \quad \text{degree/cm} \end{aligned} \tag{57.6}$$

θ_F is known as the Faraday rotation (FR) coefficient. When the direction of the magnetization is reversed, the angle of rotation changes its sign. Since two counterrotating circular polarized optical waves are used to explain FR, the effect is thus also known as optical magnetic circular birefringence (MCB). Furthermore, since the senses of polarization rotation of forward traveling and backward traveling light waves are opposite, FR is a nonreciprocal optical effect. Optical devices such as **optical isolators** and **optical circulators** use the Faraday effect to achieve their nonreciprocal functions. For ferromagnetic and ferrimagnetic media, the FR is characterized under a magnetically saturated condition, i.e., $M_3 = M_S$, the saturation magnetization of the medium. For paramagnetic or diamagnetic materials, the magnetization is proportional to the external applied magnetic field H_0. Therefore, the FR is proportional to the external field or $\theta_F = VH_0$ where $V = \chi_0 f_1 \pi/(\lambda_0 \sqrt{\varepsilon_r})$ is called the Verdet constant and χ_0 is the magnetic susceptibility of free space.

Cotton-Mouton Effect or Magnetic Linear Birefringence

When transmitted light is propagating perpendicular to the magnetization direction, the first-order isotropic FR effect will vanish and the second-order anisotropic Cotton-Mouton (CM) effect will dominate. For example, if the direction of magnetization is along the Z axis and the light wave is propagating along the X axis, the permittivity tensor becomes

$$\overleftrightarrow{\varepsilon} = \varepsilon_0 \begin{bmatrix} \varepsilon_r + f_{12} M_3^2 & 0 & 0 \\ 0 & \varepsilon_r + f_{12} M_3^2 & 0 \\ 0 & 0 & \varepsilon_r + f_{11} M_3^2 \end{bmatrix} \tag{57.7}$$

The eigenmodes are two linearly polarized light waves polarized along and perpendicular to the magnetization direction:

$$\tilde{E}_{//}(x) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \exp\left[j\left(\omega t - \frac{2\pi}{\lambda_0} n_{//} x \right) \right] \tag{57.8a}$$

$$\tilde{E}_{\perp}(x) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \exp\left[j\left(\omega t - \frac{2\pi}{\lambda_0} n_{\perp} x \right) \right] \tag{57.8b}$$

with $n_{//}^2 = \varepsilon_r + f_{11}M_3^2$ and $n_{\perp}^2 = \varepsilon_r + f_{12}M_3^2$; $n_{//}$ and $n_{\perp}$ are the refractive indices of the parallel and perpendicular linearly polarized modes, respectively. The difference in phase velocities between these two waves gives rise to a magnetic linear birefringence (MLB) of light which is also known as the CM or Voigt effect. In this case, the light transmitted through the crystal has elliptic polarization. The degree of ellipticity depends on the difference $n_{//} - n_{\perp}$. The phase shift or retardation can be found by the following expression:

$$\psi_{cm} \cong \frac{\pi(f_{11} - f_{12})M_3^2}{\lambda_0\sqrt{\varepsilon_r}} \quad \text{rad/m}$$

or

$$\frac{1.8(f_{11} - f_{12})M_3^2}{\lambda_0\sqrt{\varepsilon_r}} \quad \text{degree/cm} \tag{57.9}$$

Since the sense of this phase shift is unchanged when the direction of light propagation is reversed, the CM effect is a reciprocal effect.

Kerr Effects

Kerr effects occur when a light beam is reflected from a magnetooptic medium. There are three distinct types of Kerr effects, namely, polar, longitudinal (or meridional), and transverse (or equatorial). Figure 57.1 shows the configurations of these Kerr effects. A reflectivity tensor relation between the incident light and the reflected light can be used to describe the phenomena as follows:

$$\begin{bmatrix} E_{r\perp} \\ E_{r//} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} E_{i\perp} \\ E_{i//} \end{bmatrix} \tag{57.10}$$

where r_{ij} is the reflectance matrix. $E_{i\perp}$ and $E_{i//}$ are, respectively, the perpendicular (TE) and parallel (TM) electric field components of the incident light waves (with respect to the plane of incidence). $E_{r\perp}$ and $E_{r//}$ are, respectively, the perpendicular and parallel electric field components of the reflected light waves.

The diagonal elements r_{11} and r_{22} can be calculated by Fresnel reflection coefficients and Snell's law. The off-diagonal elements r_{12} and r_{21} can be derived from the magnetooptic permittivity tensor, the applied magnetization and Maxwell equations with the use of appropriate boundary conditions [Arecchi and Schulz-DuBois, 1972]. It is important to note that all the elements of the reflectance matrix r_{ij} are dependent on the angle of incidence between the incident light and the magnetooptic film surface.

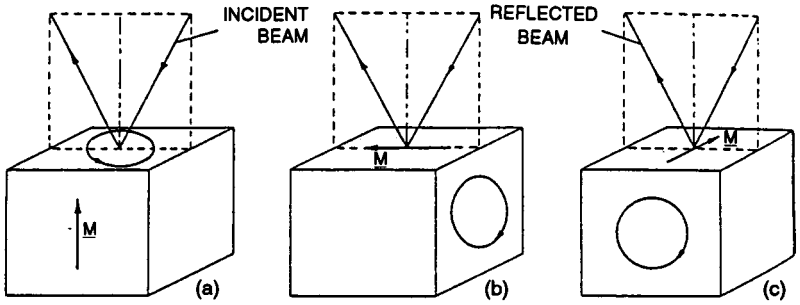


FIGURE 57.1 Kerr magnetooptic effect. The magnetization vector is represented by M while the plane of incidence is shown dotted. (a) Polar; (b) longitudinal; (c) transverse. (*Source:* A.V. Sokolov, *Optical Properties of Metals*, London: Blackie, 1967. With permission.)

Polar Kerr Effect

The polar Kerr effect takes place when the magnetization is perpendicular to the plane of the material. A pair of orthogonal linearly polarized reflected light modes will be induced and the total reflected light becomes elliptically polarized. The orientation of the major axis of the elliptic polarization of the reflected light is the same for both TE ($E_{i\perp}$) or TM ($E_{i//}$) linearly polarized incident lights since $r_{12} = r_{21}$.

Longitudinal or Meridional Kerr Effect

The longitudinal Kerr effect takes place when the magnetization is in the plane of the material and parallel to the plane of incidence. Again, an elliptically polarized reflected light beam will be induced, but the orientation of the major axis of the elliptic polarization of the reflected light is opposite to each other for TE ($E_{i\perp}$) and TM ($E_{i//}$) linearly polarized incident lights since $r_{12} = -r_{21}$.

Transverse or Equatorial Kerr Effect

This effect is also known as the equatorial Kerr effect. The magnetization in this case is in the plane of the material and perpendicular to the plane of incidence. The reflected light does not undergo a change in its polarization since $r_{12} = r_{21} = 0$. However, the intensity of the TM ($E_{r//}$) reflected light will be changed if the direction of the magnetic field is suddenly reversed. For TE ($E_{r\perp}$) reflected light, this modulation effect is at least two orders of magnitude smaller and is usually ignored.

57.3 Applications of Magnetooptic Effects

Optical Isolator and Circulator

In fiber-optic-based communication systems with gigahertz bandwidth or coherent detection, it is often essential to eliminate back reflections from the fiber ends and other surfaces or discontinuities because they can cause amplitude fluctuations, frequency instabilities, limitation on modulation bandwidth, noise or even damage to the lasers. An optical isolator permits the forward transmission of light while simultaneously preventing reverse transmission with a high degree of extinction. The schematic configuration of a conventional optical isolator utilizing bulk rotator and permanent magnet [Johnson, 1966] is shown in Fig. 57.2. It consists of a 45-degree polarization rotator which is nonreciprocal so that back-reflected light is rotated by exactly 90 degrees and can therefore be excluded from the laser. The nonreciprocity is furnished by the Faraday effect. The basic operation principle is as follows: A Faraday isolator consists of rotator material immersed in a longitudinal magnetic field between two polarizers. Light emitted by the laser passes through the second polarizer being oriented at 45 degrees relative to the transmission axis of the first polarizer. Any subsequently reflected light is then returned through the second polarizer, rotated by another 45 degrees before being extinguished by the first polarizer—thus optical isolation is achieved.

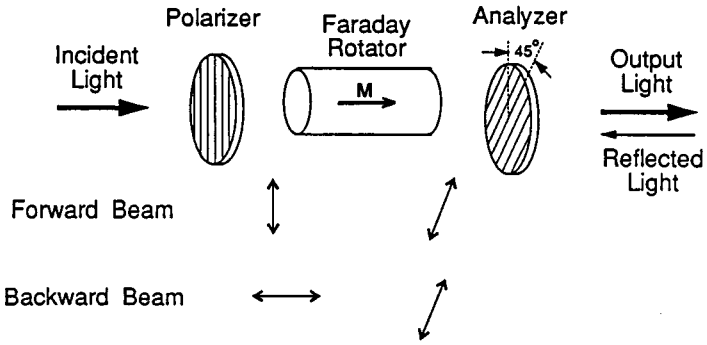


FIGURE 57.2 Schematic of an optical isolator. The polarization directions of forward and backward beams are shown below the schematic.

TABLE 57.1 Characteristics of YIG and BIG Faraday Rotators

	YIG	BIG
Verdet constant (min/cm-Gauss)		
1300 nm	10.5[a]	–806
1550 nm	9.2	–600
Saturated magnetooptic rotation (degree/mm)		
1300 nm	20.6	–136.4
1550 nm	18.5	–93.8
Thickness for 45-degree rotation (mm)		
1300 nm	2.14	0.33
1550 nm	2.43	0.48
Typical insertion loss (dB)	>0.4	<0.1
Typical reverse isolation (dB)	30–35	40
Required magnetic field (Gauss)	>1600	120
Magnetically tunable	No	Yes[b]

[a] Variable.
[b] Some BIG not tunable.
Source: D.K. Wilson, "Optical isolators adapt to communication needs," *Laser Focus World*, p. 175, April 1991. ©PennWell Publishing Company. With permission.

The major characteristics of an optical isolator include isolation level, insertion loss, temperature dependence, and size of the device. These characteristics are mainly determined by the material used in the rotator. Rotating materials generally fall into three categories: the paramagnetics (such as terbium-doped borosilicate glass), the diamagnetic (such as zinc selenide), and the ferromagnetic (such as rare-earth garnets). The first two kinds have small Verdet constants and mostly work in the visible or shorter optical wavelength range. Isolators for use with the InGaAsP semiconductor diode lasers ($\lambda_0 = 1100$–1600 nm), which serve as the essential light source in optical communication, utilize the third kind, especially the yttrium-iron-garnet (YIG) crystal. A newly available ferromagnetic crystal, epitaxially grown bismuth-substituted yttrium-iron-garnet (BIG), has an order-of-magnitude stronger Faraday rotation than pure YIG, and its magnetic saturation occurs at a smaller field [Matsuda et al., 1987]. The typical parameters with YIG and BIG are shown in Table 57.1. As the major user of optical isolators, fiber optic communication systems require different input-output packaging for the isolators. Table 57.2 lists the characteristics of the isolators according to specific applications [Wilson, 1991].

For the purpose of integrating the optical isolator component into the same substrate with the semiconductor laser to facilitate monolithic fabrication, integrated waveguide optical isolators become one of the most exciting areas for research and development. In a waveguide isolator, the rotation of the polarization is accomplished in a planar or channel waveguide. The waveguide is usually made of a magnetooptic thin film, such as YIG or BIG film, liquid phase epitaxially grown on a substrate, typically gadolinium-gallium-garnet (GGG) crystals. Among the many approaches in achieving the polarization rotation, such as the 45-degree rotation type or the

TABLE 57.2 Applications of Optical Isolators

Application	Type	Wavelength Tunable	Isolation (dB)	Insertion Loss (dB)	Return Loss (dB)
Fiber to fiber	PI	Yes/no	30–40	1.0–2.0	> 60
Fiber to fiber	PS	Normally no	33–42	1.0–2.0	> 60
Single fiber	PS	No	38–42	Complex	Complex
Bulk optics	PS	No	38–42	0.1–0.2	

PI = polarization insensitive.
PS = polarization sensitive.
Source: D.K. Wilson, "Optical isolators adapt to communication needs," *Laser Focus World,* p. 175, April 1991. With permission.

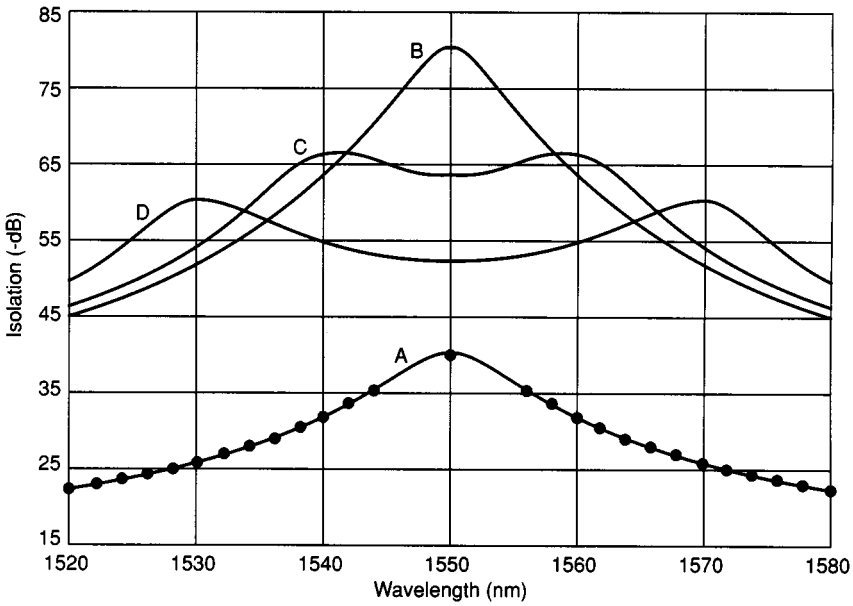


FIGURE 57.3 Isolation performance of four isolators centered around 1550 nm shows the effects of different configurations. A single-stage isolator (curve A) reaches about –40 dB isolation, and two cascaded single-wavelength isolators (curve B) hit –80 dB. Wavelength broadening (curves C and D) can be tailored by cascading isolators tuned to different wavelengths. (*Source:* D.K. Wilson, "Optical isolators adapt to communication needs," *Laser Focus World*, April 1991. With permission.)

unidirectional TE-TM mode converter type, the common point is the conversion or coupling process between the TE and TM modes of the waveguide. Although very good results have been obtained in some specific characteristics, for example, 60-dB isolation [Wolfe et al., 1990], the waveguide optical isolator is still very much in the research and development stage.

Usually, the precise wavelength of any given semiconductor diode is uncertain. Deviation from a specified wavelength could degrade isolator performance by 1 dB/nm, and an uncertainty of 10 nm can reduce isolation by 10 dB. Therefore, a tunable optical isolator is highly desirable. A typical approach is to simply place two isolators, tuned to different wavelengths, in tandem to provide a broadband response. Curves C and D of Fig. 57.3 show that isolation and bandwidth are a function of the proximity of the wavelength peak positions. This combination of nontunable isolators has sufficiently wide spectral bandwidth to accommodate normal wavelength variations found in typical diode lasers. In addition, because the laser wavelength depends on its operating temperature, this broadened spectral bandwidth widens the operating temperature range without decreasing isolation.

The factors that limit isolation are found in both the polarizers and the Faraday rotator materials. Intrinsic strain, inclusions, and surface reflections contribute to a reduction in the purity of polarization which affects

isolation. About 40 dB is the average isolation for today's materials in a single-isolator stage. If two isolators are cascaded in tandem, it is possible to double the isolation value.

Finally, an optical circulator [Fletcher and Weisman, 1965] can be designed by replacing the polarizers in a conventional isolator configuration with a pair of calcite polarizing prisms. A laser beam is directed through a calcite prism, then through a Faraday rotator material which rotates the polarization plane by 45 degrees, then through a second calcite prism set to pass polarization at 45 degrees. Any reflection beyond this second calcite prism returns through the second prism, is rotated by another 45 degrees through the Faraday material, and, because its polarization is now 90 degrees from the incident beam, is deflected by the first calcite prism. The four ports of the circulator then are found as follows: (1) the incident beam, (2) the exit beam, (3) the deflected beam from the first calcite prism, and (4) the deflected beam from the second calcite prism.

MSW-Based Guided-Wave Magnetooptic Bragg Cell

When a ferrimagnet is placed in a sufficiently large externally applied dc magnetic field, H_0, the ferrimagnetic materials become magnetically saturated to produce a saturation magnetization $4\pi M_S$. Under this condition, each individual magnetic dipole will precess in resonance with frequency $f_{res} = \gamma H_0$ where γ is the gyromagnetic ratio (γ = 2.8 MHz/Oe). However, due to the dipole–dipole coupling and quantum mechanical exchange coupling, the collective interactions among neighboring magnetic dipole moments produce a continuum spectrum of precession modes or spin waves at frequency bands near f_{res}. Exchange-free spin wave spectra obtained under the magnetostatic approximation are known as magnetostatic waves (MSWs) [Ishak, 1988]. In essence, MSWs are relatively slow propagating, dispersive, magnetically dominated electromagnetic (EM) waves which exist in biased ferrites at microwave frequencies (2–20 GHz). In a ferrimagnetic film with a finite thickness, such as a YIG thin film epitaxially grown on a nonmagnetic substrate such as GGG, MSW modes are classified into three types: magnetostatic surface wave (MSSW), magnetostatic forward volume wave (MSFVW), and magnetostatic backward volume wave (MSBVW), depending on the orientation of the dc magnetic field with respect to the film plane and the propagation direction of the MSW. At a constant dc magnetic field, each type of mode only exists in a certain frequency band. An important feature of MSW is that these frequency bands can be simply tuned by changing the dc magnetic field.

As a result of the Faraday rotation effect and Cotton-Mouton effect, the magnetization associated with MSWs will induce a perturbation in the dielectric tensor. When MSW propagates in the YIG film, it induces a moving optical grating which facilitates the diffraction of an incident guided light beam. If the so-called Bragg condition is satisfied between the incident guided light and the MSW-induced optical grating, Bragg diffraction takes place. An optical device built based on this principle is called the **magnetooptic Bragg cell** [Tsai and Young, 1990].

A typical MSFVW-based noncollinear coplanar guided-wave magnetooptic Bragg cell is schematically shown in Fig. 57.4. Here a homogeneous dc bias magnetic field is applied along the Z axis to excite a Y-propagating MSFVW generated by a microstrip line transducer. With a guided lightwave coupled into the YIG waveguide and propagating along the X axis, a portion of the lightwave is Bragg-diffracted and mode-converted (TE to TM mode and vice versa). The Bragg-diffracted light is scanned in the waveguide plane as the frequency of the MSFVW is tuned. Figure 57.5 shows the scanned light spots by tuning the frequency at a constant dc magnetic field.

MSW-based guided-wave magnetooptic Bragg cell is analogous to surface acoustic wave (SAW)-based guided-wave acoustooptic (AO) Bragg cell and has the potential to significantly enhance a wide variety of integrated optical applications which had previously been implemented with SAW. These include TE-TM mode converter, spectrum analyzer, convolvers/correlators, optical frequency shifters, tunable narrowband optical filters, and optical beam scanners/switches [Young, 1989]. In comparison to their AO counterparts, the MSW-based magnetooptic Bragg cell modules may possess the following unique advantages: (1) A much larger range of tunable carrier frequencies (2–20 GHz, for example) may be readily obtained by varying a dc magnetic field. Such high and tunable carrier frequencies with the magnetooptic device modules allow direct processing at the carrier frequency of wide-band RF signals rather than indirect processing via frequency down-conversion, as is required with the AO device modules. (2) A large magnetooptic bandwidth may be realized by means of a simpler transducer geometry. (3) Much higher and electronically tunable modulation/switching and scanning speeds are possible as the velocity of propagation for the MSW is higher than that of SAW by one to two orders of magnitude, depending upon the dc magnetic field and the carrier frequency.

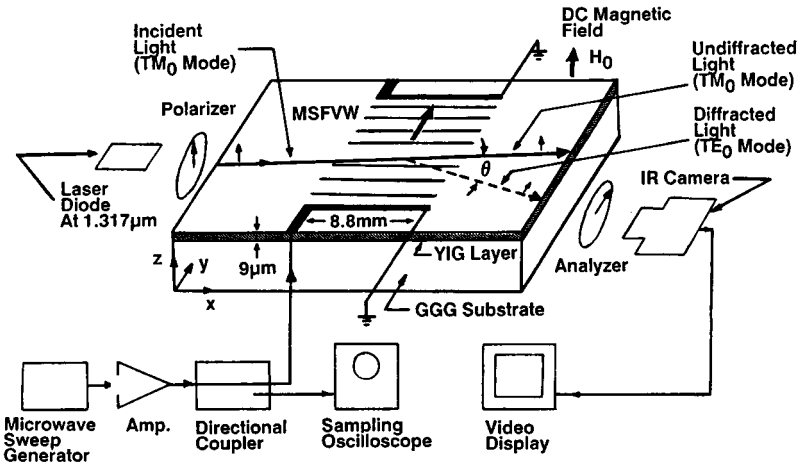


FIGURE 57.4 Experimental arrangement for scanning of guided-light beam in YIG-GGG waveguide using magnetostatic forward waves.

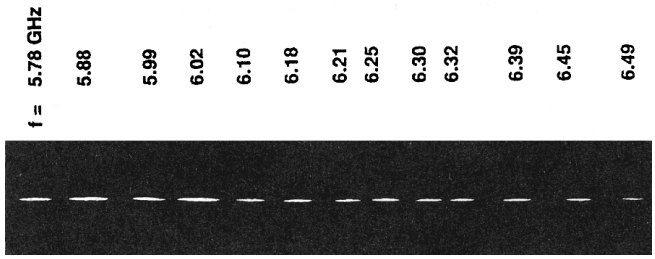


FIGURE 57.5 Deflected light spots obtained by varying the carrier frequency of MSFVW around 6 GHz.

Magnetooptic Recording

The write/erase mechanism of the **magnetooptical (MO) recording system** is based on a thermomagnetic process in a perpendicularly magnetized magnetooptic film. A high-power pulsed laser is focused to heat up a small area on the magnetooptic medium. The coercive force of the MO layer at room temperature is much greater than that of a conventional non-MO magnetic recording medium. However, this coercive force is greatly reduced when the heated spot is raised to a critical temperature. Application of a bias magnetic field can then easily reverse the polarization direction of the MO layer within the heated spot. As a result, a very small magnetic domain with magnetization opposite to that of the surrounding area is generated. This opposite magnetic domain will persist when the temperature of the medium is lowered. The magnetization-reversed spot represents one bit of stored data. To erase data, the same thermal process can be applied while reversing the direction of the bias magnetic field.

To read the stored information optically, the Kerr effect is used to detect the presence of these very small magnetic domains within the MO layer. When a low-power polarized laser beam is reflected by the perpendicularly oriented MO medium, the polarization angle is twisted through a small angle θ_k, the Kerr rotation. Furthermore, the direction of this twist is either clockwise or counterclockwise, depending on the orientation of the perpendicular magnetic moment, which is either upward or downward. Therefore, as the read beam scans across an oppositely magnetized domain from the surrounding medium, there is a total change of $2\theta_k$ in the polarization directions from the reflected beam coming from the two distinct regions. Reading is done by detecting this phase difference.

The MO recording medium is one of the most important elements in a high-performance MO data-storage system. In order to achieve fast writing and erasing functions, a large Kerr rotation is required to produce an

acceptable carrier-to-noise (C/N) ratio. In general, a high-speed MO layer has a read signal with a poor C/N ratio, while a good read-performance MO layer is slow in write/erase sensitivity.

Defining Terms

Cotton-Mouton effect: Second-order anisotropic reciprocal magnetooptic effect which causes a linearly polarized incident light to transmit through as an elliptically polarized output light wave when the propagation direction of the incident light is perpendicular to the direction of the applied magnetization of the magnetooptic medium. It is also known as magnetic linear birefringence (MLB).

Faraday rotation: First-order isotropic nonreciprocal magnetooptic effect which causes the polarization direction of a linearly polarized transmitted light to rotate when the propagation direction of the incident light wave is parallel to the direction of the applied magnetization of the magnetooptic medium. It is also known as magnetic circular birefringence (MCB).

Kerr effects: Reflected light from a magnetooptic medium can be described by the optical Kerr effects. There are three types of Kerr effects: polar, longitudinal, and transverse, depending on the directions of the magnetization with respect to the plane of incidence and the reflecting film surface.

Magnetooptic Bragg cell: A magnetically tunable microwave signal processing device which uses optical Bragg diffraction of light from a moving magnetooptic grating generated by the propagation of magnetostatic waves within the magnetic medium.

Magnetooptic recording system: A read/write data recording system based on a thermomagnetic process to write oppositely magnetized domains onto a magnetooptic medium by means of high-power laser heating. Magnetooptic Kerr effect is then employed to read the data by using a low-power laser as a probe beam to sense the presence of these domains.

Optical circulator: A four-port optical device that can be used to monitor or sample incident light (input port) as well as reflected light (output port) with the two other unidirectional coupling ports.

Optical isolator: A unidirectional optical device which only permits the transmission of light in the forward direction. Any reflected light from the output port is blocked by the device from returning to the input port with a very high extinction ratio.

Related Topics

35.1 Maxwell Equations • 39.1 Passive Microwave Devices

References

F.T. Arecchi and E.O. Schulz-DuBois, *Laser Handbook*, D4, Amsterdam: North-Holland, 1972, pp. 1009–1027.

P.C. Fletcher and D.L. Weisman, "Circulators for optical radar systems," *Applied Optics*, vol. 4, pp. 867–873, 1965.

K.H. Hellwege, *Landolt-Bornstein Numerical Data and Functional Relationships in Science and Technology, New Series*, vols. 4 and 12, New York: Springer-Verlag, 1978.

W. S. Ishak, "Magnetostatic wave technology: A review," *Proc. IEEE*, vol. 76, pp. 171–187, 1988.

B. Johnson, "The Faraday effect at near infrared wavelength in rare-earth garnet," *Brit. J. Appl. Phys.*, vol. 17, p. 1441, 1966.

K. Matsuda, H. Minemoto, O. Kamada, and S. Isbizuka, "Bi-substituted, rare-earth iron garnet composite film with temperature independent Faraday rotation for optical isolators," *IEEE Trans. Mag.*, vol. MAG-23, p. 3479, 1987.

M. Torfeh, L. Courtois, L. Smoczynski, H. Le Gall, and J.M. Desvignes, "Coupling and phase matching coefficients in a magneto-optical TE-TM mode converter," *Physica*, vol. 89B, pp. 255–259, 1977.

C.S. Tsai and D. Young, "Magnetostatic-forward-volume-wave-based guided-waveless magneto-optic Bragg cells and applications to communications and signal processing," *IEEE Trans. on Microwave Theory and Technology*, vol. MTT-38(5), pp. 560–570, 1990.

D.K. Wilson, "Optical isolators adapt to communication needs," *Laser Focus World*, p. 175, April 1991.

R. Wolfe et al., "Edge tuned ridge waveguide magneto-optic isolator," *Applied Physics Letters*, vol. 56, p. 426, 1990.

D. Young, "Guided Wave Magneto-Optic Bragg Cells Using Yttrium Iron Garnet-Gadolinium Gallium Garnet (YIG-GGG) Thin Film at Microwave Frequencies," Ph.D. Dissertation, University of California, Irvine, 1989.

Further Information

Current publications on magnetooptics can be found in the following journals: *Intermag Conference Proceedings* published by the IEEE and *International Symposium Proceedings on Magneto-Optics.*

For in-depth discussion on magnetic bubble devices, please see, for example, *Magnetic-Bubble Memory Technology* by Hsu Chang, published by Marcel Dekker, 1978.

An excellent source of information on garnet materials can be found in *Magnetic Garnet* by Gerhard Winkler, published by Friedr. Vieweg & Sohn, 1981.

Numerous excellent reviews on the subject of magnetooptics have been published over the years, for example, J.F. Dillon, Jr., "Magnetooptics and its uses," *Journal of Magnetism and Magnetic Materials*, vol. 31–34, pp. 1–9, 1983; M.J. Freiser, "A survey of magnetooptic effects," *IEEE Transactions on Magnetics*, vol. MAG-4, pp. 152–161, 1968; G. A. Smolenskii, R.V. Pisarev, and I.G. Sinii, "Birefringence of light in magnetically ordered crystals," *Sov. Phys. Usp.*, vol. 18, pp. 410–429, 1976; and A.M. Prokhorov, G.A. Smolenskii, and A.N. Ageev, "Optical phenomena in thin-film magnetic waveguides and their technical application," *Sov. Phys. Usp.*, vol. 27, pp. 339–362, 1984.

Neelakanta, P.S. “Smart Materials”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

58

Smart Materials

58.1 Introduction

58.2 Smart/Intelligent Structures

58.3 Objective-Based Classification of Smart/Intelligent Materials
Smart Structural Materials • Smart Thermal Materials • Smart Acoustical Materials • Smart Electromagnetic Materials • Pyrosensitive Smart Materials

58.4 Material Properties Conducive for Smart Material Applications
Piezoelectric Effect • Magnetostrictive Effect • Electroplastic Effect • Shape-Memory Effects • Electrorheological Property • Nonlinear Electro-optic Properties • Nonlinear Electroacoustic Properties • Pyrosensitive Properties • Nonlinear Electromagnetic Properties

58.5 State-of-the-Art Smart Materials
Piezoelectric Smart Materials • Magnetostrictive Smart Materials • Electroplastic Smart Materials • Shape-Memory Smart Materials • Electrorheological Smart Fluids • Electro-optic Smart Materials • Electroacoustic Smart Materials • Electromagnetic Smart Materials • Pyrosensitive Smart Materials

58.6 Smart Sensors
Fiber-Optic-Based Sensors • Piezoelectric-Based Sensors • Magnetostriction-Based Sensors • Shape-Memory Effects-Based Sensors • Electromagnetics-Based Sensors • Electroacoustic Smart Sensors

58.7 Examples of Smart/Intelligent Systems
Structural Engineering Applications • Electromagnetic Applications

58.8 High-Tech Application Potentials

58.9 Conclusions

P. S. Neelakanta
Florida Atlantic University

58.1 Introduction

Smart materials are a class of materials and/or composite media having inherent intelligence together with self-adaptive capabilities to external stimuli. Also known as **intelligent materials,** they constitute a few subsets of the material family that "manifest their own functions intelligently depending on environmental changes" [Rogers and Rogers, 1992].

Classically, such intelligent material systems have been conceived in the development of mechanical structures that contain their own sensors, actuators and self-assessing computational feasibilities in order to modify their structural (elastic) behavior via feedback control capabilities. The relevant concepts have stemmed from intelligent forms of natural (material) systems, namely, living organisms; hence, in modern concepts smart or

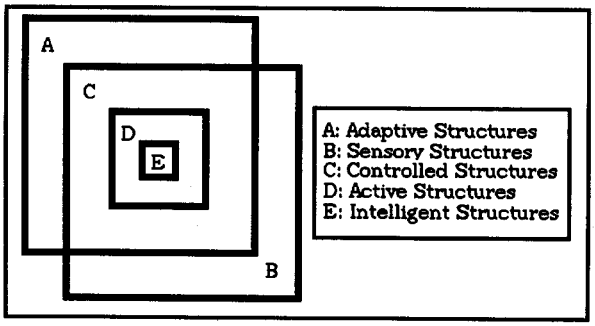


FIGURE 58.1 Set of structures. (Adapted from B. K. Wada, J. L. Fanson, and E.F. Crawley, "Adaptive structures," *J. Intell. Mat. Syst. Struct.*, 1, 1990.)

intelligent materials and systems are conceived as those that mimic the life functions of sensing, actuation, control, and intelligence.

The inherent intelligence and self-adaptable control of artificial smart materials should be programmable in terms of the constituent processing, microstructural characteristics, and defects to permit the self-conditionings to adapt in a controlled manner to various types of stimuli. The dividing line between smart materials and the so-called **intelligent structures** is not, however, distinct. In simple terms, intelligent material systems are constructed of smart materials with a dedicated, discrete set of integrated actuators, sensors, and so on, and smart materials contain largely a built-in or embedded set of distributed sensors. In general, the term *smart materials* usually connotes the structural constituent in which the discrete functions of sensing, actuation, signal processing and control are tangibly integrated. Intelligent structures, as an extension, are constructed with smart materials to respond to the environment around them in a predetermined, desired manner.

Intelligent or smart materials that manifest their own functions intelligently vis-à-vis the changes in their surroundings are capable of performing, in general (Chong et al., 1990):

- Primary functions specifying the adaptive roles of the sensor, the effector and processor capabilities (including the memory functions)
- Macroscopic functions that enclave the extensive or global aspects of the intelligence inherent in the materials
- Built-in social utility aspects with an instilled human-like intelligence with hyper-performance capabilities

58.2 Smart/Intelligent Structures

The framework of intelligent structures as a subset in the gamut of conventional material-based systems is illustrated in Fig. 58.1. This general classification of material structures refer to [Chong et al., 1990]:

- Sensory structures, "which possess sensors that enable the determination or monitoring of system states or characteristics" [Chong et al., 1990]
- Adaptive structures, which possess actuators that facilitate the alteration of system-states or characteristics in a controlled manner
- Sensory systems, which may contain sensors, but no actuators
- Adaptive systems, which contain actuators, but no sensors

Referring to Fig. 58.1, the intersection of sensory versus adaptive structures depicts the controlled structures with a feedback architecture. That is, the active structure has an integrated controlled unit with sensors and/or actuators that have structural as well as control functionality. Hence, the logical subset that defines an intelligent structure is a highly integrated unit (with controlled logic, electronics, etc.) that provides the cognitive element of a distributed or a hierarchic controlled structure.

58.3 Objective-Based Classification of Smart/Intelligent Materials

Smart Structural Materials

Intelligent structural engineering materials are the classical versions of smart systems in which the mechanical (elastic) properties of a structure can be modified adaptively by means of an imbedded distribution of smart material(s), and an associated (integral) set of sensors and actuators together with an external control system to facilitate adaptive changes in the elastic behavior of structures so that motion, vibration, strength, stiffness, redistribution of load path in response to damage, etc. are controlled.

Smart Thermal Materials

A **smart thermal material,** in response to environmental demands, can self-adaptively influence its thermal states (temperature or such thermal properties as conductivity, diffusivity, absorptivity), by means of an integrated conglomeration of thermal sensors, heaters, or actuators with an associated control system.

Smart Acoustical Materials

Smart acoustical materials can be classified as those that have self-adaptive characteristics on their acoustical behavior (such as transmission, reflection, and absorption of acoustical energy) by means of sensors that assess the acoustical states (intensity, frequency, response, etc.), along with a set of actuators (dampers, exciters) with an associated control system. Again, the self-adaptive behavior of these materials is in response to ambient acoustical changes.

Smart Electromagnetic Materials

Smart Magnetic Shielding Materials

As warranted by the surroundings, the self-adaptive shielding effectiveness to magnetic fields at low frequencies (power frequencies such as 60 or 50 Hz) can be achieved by means of an integrated set of magnetic field sensors and actuators (magnetic biasing, current elements, etc.) plus a control system arrangement [Neelakanta and Subramaniam, 1992].

High-Frequency Smart Shielding Materials

Corresponding to radio and higher frequency environments, the shielding requirement warrants curtailing both electric and magnetic fields. Hence, the relevant self-adaptive intelligent shielding system would consist of an array of distributed electromagnetic sensors with appropriate elements (actuators) and a control system.

Smart Radar-Absorbing Materials

Absorption of microwave/millimeter wave energy at radar frequency is useful in radar stealth applications. Adaptively controllable smart radar-absorbing materials (smart RAMs) can be synthesized with integrated distribution of electromagnetic detectors (sensors) with appropriate actuators and control system [Neelakanta et al., 1992].

Smart Optical Surface Materials

Smart optical surface materials can be envisioned as those in which the surface optical properties (hue, intensity, etc.) can be adaptively controlled by means of an intelligent sensor/actuator combinational control system.

Pyrosensitive Smart Materials

Electromagnetic active surfaces constituted by pyrosensitive inclusions have been successfully developed to manage the electromagnetic reflection and/or absorption characteristics from the active surface by means of thermal actuation of the pyrosensitive nodes imbedded in the medium [Neelakanta et al., 1992]. With the inclusion of a feedback systems, smart operation in adaptively manipulating the active surface characteristics can be achieved.

58.4 Material Properties Conducive for Smart Material Applications

Certain specific characteristics of materials make them suitable for smart material applications. These properties are:

1. Piezoelectric effect
2. Magnetostrictive effect
3. Electroplastic effect
4. Shape-memory effects
5. Electrorheological properties
6. Nonlinear electro-optic properties
7. Nonlinear electroacoustic properties
8. Nonlinear electromagnetic properties
9. Pyrosensitive properties

Piezoelectric Effect

Piezoelectric property of a material refers to the ability to induce opposite charges at two faces (correspondingly, to exhibit a voltage difference between the faces) of the material as a result of the strain due to mechanical force (either tension or compression) applied across the surfaces. This process is also reversible in the sense that a mechanical strain would be experienced in the material when subjected to opposite electric charging at the two faces by means of an applied potential.

In the event of such an applied voltage being alternating, the material specimen will experience vibrations. Likewise, an applied vibration on the specimen would induce an alternating potential change between the two faces. The most commonly known materials that exhibit piezoelectric properties are natural materials like quartz and a number of crystalline and polycrystalline compounds.

The strain versus the electric phenomenon perceived in piezoelectric materials is dictated by a coefficient that has components referred to a set of orthogonal coordinate axes (which are correlated to standard crystallographic axes). For example, denoting the piezoelectric coefficient (ratio between piezoelectric strain component to applied electric field component at a constant mechanical stress or vice versa) as d_{mn}, the subscript n (1 to 3) refers to the three euclidian orthogonal axes, and $m = 1$ to 6 specifies the mechanical stress-strain components. The unit for d_{mn} is meter/volt which is the same as coulomb/newton.

In the piezoelectric phenomenon, there is an electromechanical synergism expressed as a coupling factor K defined by K^2, which quantifies the ratio of mechanical energy converted into electric charges to the mechanical energy impressed on the material. Being a reversible process, a relevant inverse ratio is also applicable.

Magnetostrictive Effect

Magnetostrictive effect refers to the structural strain experienced in a material subjected to a polarizing magnetic flux. A static strain of $\Delta l/l$ is produced by a dc polarizing magnetic flux density B_o such that $\Delta l/l = CB_o^2$, where C is a material constant expressed in (meter^4/weber^2) taking the units for B_o as weber/meter^2.

The magnetic stress constant (Λ) in (newton/weber) is given by $\Lambda = 2CB_oY_o$ where Y_o refers to the Young's modulus of a linearly strained free bar. The coefficient (Λ) could be both positive or negative. For example, nickel contracts with increasing B_o, whereas magnetic alloys such as 45 Permalloy (45% Ni + 55% Fe), Alfer (13% Al, 87% Fe) exhibit positive magnetostrictive coefficient [Reed, 1988].

Electroplastic Effect

The **electroplastic effect** (EPE) refers to the plastic deformation of metals with the application of high-density electric current with an enhanced deformation rate (that persists in addition to that caused by the side effects of the current such as joule-heating and the magnetic pinch effect). The plastic strain rate resulting from a current pulse is given by $\varepsilon_I/\varepsilon_A = \alpha J^2 \exp(\beta J)$ where ε_I is the strain rate occurring during the current pulse, ε_A is the strain rate in the absence of the current pulse, J is the current density and α and β are material constants. Typically the EPE has been observed in zinc, niobium, titanium, etc.

Shape-Memory Effects

The mechanism by which a plastically deformed object in the low-temperature martensitic condition regains its original shape when the external stress is removed and heat is applied is referred to as the shape-memory

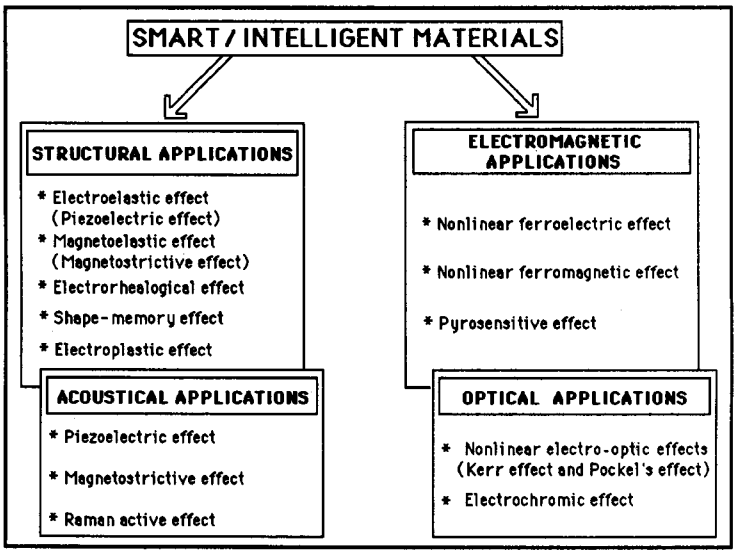


FIGURE 58.2 Application-specific classification of smart/intelligent materials.

effect (SME) [Jackson et al., 1972]. It is a memory mechanism that is the result of a martensitic transformation taking place during heating.

Alhough the exact mechanism by which the shape-memory effect occurs is still under study, the process by which the original shape is regained is associated with a reverse transformation of the deformed martensitic phase to the higher temperature austenite phase. A group of nickel-titanium alloys (referred to as Nitinol) of proper composition exhibit the shape-memory property and are widely used in smart material applications [Jackson et al., 1972].

Electrorheological Property

Electrorheological property is the property exhibited by certain fluids that are capable of altering their flow characteristics depending on an external applied electric field. These fluids have a fast response time, only a few milliseconds. Once the external field is applied, there is a form of progressive gelling of the fluid proportional to the applied field strength. Without the applied field, the fluid flows freely. If the electrified electrorheological (ER) fluid is sheared by an applied force larger than a certain critical value, it flows. Below this critical value of applied shear force, the electrified fluid remains in the gel phase [Gandhi and Thompson, 1989].

An electrorheological fluid requires particles (1 to 100 mm in diameter) dispersed in a carrier fluid. Sometimes a surfactant is also added to help the dispersion of particles in the fluid. The surfactant is used to prevent particle interaction that could otherwise result in a tendency for the particulates to clump together when the fluid is allowed to stand still over a stretch of time. The tendency of the particles to clump together is referred to as settling.

The applied electric field to perceive the electrorheological phenomenon is usually in the order of 4 kV/mm. When the electric field is applied, the positive and negative charges on the suspended particles are separated, forming a dipole of charges. These dipoles then align (polarize) themselves by mutual forces of attraction and repulsion to other similar dipoles, resulting in unique flow characteristics. In the absence of an electric field, there is no dipole separation of charges, and hence the fluid returns to its normal flow.

An ideal electrorheological fluid is one that has a low viscosity in the absence of an applied field and that which transforms into a high-viscosity gel capable of withstanding high shear stresses when the field is on. Further, it must also have a low power consumption. The first reported ER fluid consisted of finely dispersed suspensions of starch or silica gel in mineral oil nearly 40 years ago.

Nonlinear Electro-optic Properties

In certain materials that are optically transparent when subjected to an external electric field, the refractive index of the material changes. Invariably the electric field versus optical effect thus experienced is nonlinear,

with the result that a time-varying electric field will modulate the refractive index, and hence a phase shift is experienced by the light passing through the medium. In materials that have a central symmetry, this phenomenon is called the Kerr effect; in noncentrosymmetric materials, it is referred to as Pockel's effect [Kaminow, 1965].

Nonlinear Electroacoustic Properties

Electroacoustic synergism is experienced in certain classes of materials in which the mechanical atomic vibrations are influenced by the electronic polarizability, with the result that nonlinear interaction between the atomic displacements versus the electric field causes modulation effects resulting in the generation of new sideband frequencies. Such sidebands (labeled Raman frequencies) and the response function of a Raman active medium have the form

$$H(\omega) = A_1E(\omega) + A_2E^2(\omega) + A_3E^3(\omega) + \cdots$$

Pyrosensitive Properties

The **pyrosensitive property** is governed by a class of materials known as solid electrolytes. On thermally energizing such materials, they exhibit superionic electric conduction (also known as fast ion conduction), with the result that the medium, which is dielectric under cold conditions, becomes conducting at elevated temperatures. Correspondingly, the media that are embedded with solid electrolytes show different extents of electromagnetic reflection/transmission characteristics at low and high temperatures and hence can be manipulated thermally [Neelakanta et al., 1992].

Typical solid electrolytes that can be adopted for such pyrosensitive applications are, for example, AgI and $RbAg_4I_5$. The materials like β-AgI and β-alumina show increasing conductivity with increasing temperature. The compound β-AgI exhibits superionic conductivity, with an abrupt transition at a temperature close to 147°C. This transition is known as the β- to α-phase transition, and there are a host of other materials that exhibit this phenomenon. For example, the material $RbAg_4I_5$ has a high electrical conductivity even at room temperature. It has also been observed that solid electrolytes provide sufficiently high electrical conductivity in the α-phase even when included in low volume fractions in a mixture with a nonsolid-electrolyte host [Neelakanta et al., 1992].

Nonlinear Electromagnetic Properties

Basically, the nonlinear electromagnetic properties can manifest as two subsets of material characteristics, namely, **nonlinear dielectric properties** and **nonlinear magnetic properties.**

Nonlinear Dielectric Properties

Dielectric materials whose permittivity has a distinct dependence on the intensity of the applied electric field are referred to as active or nonlinear dielectrics. Such materials demonstrate very high values of permittivity (in the order of several thousand), pronounced dependence of dielectric parameters on the temperature, and a loop of electric hysteresis under the action of an alternating voltage.

Ferroelectrics are the most typical example of nonlinear dielectrics. Rochelle's salt (potassium sodium tartrate) was the first substance in which nonlinearity was discovered. All ferroelectrics, however, possess nonlinear properties only within a definite temperature range. The temperature transition points over which the ferroelectric materials gain or lose their ferroelectric properties are referred to as Curie points. The arsenates and dihydrogen phosphates of alkali metals are also examples of ferroelectric materials.

Piezoelectrics also fall under the category of active dielectrics. Electrets, which are capable of preserving an electric charge for a long period of time (hence regarded analogous to permanent magnets), exhibit highly nonlinear dielectric properties.

Nonlinear Magnetic Properties

Ferromagnetic materials are materials in which the permanent magnetic dipoles align themselves parallel to each other. These materials have a characteristic temperature below and above which their properties differ greatly. This temperature is referred to as the Curie temperature. Above the Curie temperature they behave as paramagnetic materials, while below it they exhibit the well known hysteresis B versus H curves. Examples of such ferromagnetic materials are iron, Mu-metal, and Supermalloy. Ferrimagnetic materials are similar in their hysteresis properties to ferromagnetic materials but differ from them in that their magnetic dipoles align

themselves antiparallel to each other. Ferrites are the most popular ferrimagnetic materials, and they are of the greatest interest in electrical engineering applications.

58.5 State-of-the-Art Smart Materials

Piezoelectric Smart Materials

Piezoelectric smart materials find applications primarily in intelligent structures deploying electroelastic synergism, and a class of ceramics (popularly known as ferroelectric ceramics) have emerged in recent times for such applications. Typically, such ceramics include the base polycrystalline piezoelectrics such as $BaTiO_3$, $CdTiO_3$, $PbZrO_3$, and $PbTiO_3$, formulated with various stoichiometric proportions. Another class of piezoelectric flexible composite that has the potential for smart applications is a compound consisting of $PbTiO_3$ and chloroprene rubber. A set of glass ceramic composites containing the crystalline phases of Li_2SiO_3, $Li_2Si_2O_5$, $Ba_2TiSi_2O_8$, $Ba_2TiGe_2O_8$, $Li_2B_4O_7$, etc. are also emerging samples in smart material engineering [Chong et al., 1990].

Piezoelectric smart materials can also be made from the family of polymers, namely, polyvinylidene fluoride (PVDF). The main advantages of using this polymer are that it can be formed into very thin sheets and has excellent mechanical strength combined with high sensitivity to pressure changes.

Another piezoelectric material recently developed in the NTK Research facility in Japan is a kind of rubber-based material referred to as piezoelectric rubber. This material is composed of a base material of synthetic rubber, namely, chloroben, dispersed with fine particles of a popular piezoelectric ceramic, called PZT (lead zirconium titanate). Piezoelectric rubber combines the favorable properties of PZT, namely, high sensitivity, chemical inertness, linearity, and simplicity, with that of the rubber base, namely, flexibility. The main drawback with the piezoelectric rubber is in making an electrical contact with it. This problem has been circumvented by the development of a coaxial cable connection that is easier to use [Ting, 1990].

Magnetostrictive Smart Materials

Materials with a high degree of magnetostriction are deployed in modern intelligent structures. Typically, the amount of strain inducible with intelligent materials in the current state of the art is 2000 ppm. These are alloys made with iron and rare earth materials such as terbium (Te), dysprosium (Dy), and niobium (Nb). A commercially known material of this category is Terfenol [Reed, 1988]. Magnetostrictive transducers for smart applications have also been developed with a certain class of metallic glass materials.

Electroplastic Smart Materials

Electroplastic materials are useful as smart elastic media inasmuch as the stimulus that modifies the elastic deformation is the electric current that can be controlled externally. The usefulness of these materials for smart systems under room temperature conditions is still under investigation.

Shape-Memory Smart Materials

Shape-memory smart materials include three categories, namely shape-memory alloys (SMA), shape-memory hybrid composites (SMHC), and shape-memory polymers (SMP).

Nickel–titanium (Nitinol) alloys of proper composition exhibit unique memory, or shape-restoration force characteristics, and are the most popular shape-memory alloys. When the material is plastically deformed in its low-temperature phase and then heated above its characteristic transition temperature, the original configuration or shape is restored. Deformations up to 6–8% can be completely restored by heating the material. It is this property that is used in smart electromechanical actuations.

Shape-memory hybrid composites are composite materials that contain SMA fibers or films in such a way that they can be mechanically controlled by heat. These materials can be heated by passing a current through the fibers. SMHCs offer a wide scope of applications in material–structure interaction. The fibers used in these composites are also made of Nitinol alloys.

The third form of shape-memory materials are the shape-memory polymers. These materials have an elastic memory, meaning that a large reversible change in the elastic modulus exists across the glass-transition

temperature. In other words, across the glass-transition temperature, the material can change from a glass to rubbery state, allowing significant deformation in response to temperature changes. Shape-memory polymers, in general, are durable, lightweight, and transparent. Nippon Zeon Company and Mitsubishi Company have developed high-performance SMPs in the recent past [Chong et al., 1990]. While the SMP of Nippon Zeon Company is polynorborene based, Mitsubishi's SMP is polyurethane based, which overcomes crucial weaknesses such as poor processability and limited-temperature operating range. In their applications SMPs can be used either as an elastic memory material or a shape-memory material. Depending on which of these possibilities are used, the range of applications differs.

Electrorheological Smart Fluids

Current research on electrorheological fluids is focused toward development of carrier–particle combinations that result in the desirable characteristics to achieve smart elastic behavior [Gandhi and Thompson, 1989]. The earlier versions of electrorheological fluids contained adsorbed water, which limited their operating temperature change (up to 80°C). Particles in the newer electrorheological fluids are, however, based on polymers, minerals, and ceramics, which have a higher operating range (200°C). Also, the increase in power consumption is less with temperature increments in the recent anhydrous systems. The most commonly used carrier fluids are silicone oil, mineral oil, and chlorinated paraffin, which offer good insulation and compatibility for particulate dispersion.

Electro-optic Smart Materials

Typically potassium dihydrogen phosphate (KDP) exhibits electro-optic behavior. Synthetic materials that have the ability to alter their refractive index (and hence the optical transmission and reflection characteristics) in the presence of an electric stimulus can be comprehended as viable **smart sensor** applications.

Electroacoustic Smart Materials

Although classically the nonlinear interaction of a vibrational (acoustic) wave and an electromagnetic wave has been studied in reference to Raman active media, relevant concepts can be exercised for smart engineering applications using those materials that exhibit strong vibrational versus piezoelectric characteristics. The NTK piezorubber, PZT ceramics, $LiNBO_3$, PZT with donor additives, insolvent additives, etc. are viable candidates for smart applications in addition to piezoelectric polymers.

Electromagnetic Smart Materials

In recent times a number of materials that possess ferroelectric properties have been discovered, the most popular of which is barium titanate ($BaTiO_3$). Barium titanate has an excellent prospect as a smart material because of the several advantages it offers, such as high mechanical strength, resistance to heat and moisture, and ease of manufacturing. $BaTiO_3$ and other similar materials are frequently referred to as ferroelectric ceramics. Also, electrets such as polymethylmethacrylate offer promise for smart applications.

Among the nonlinear magnetic materials, ferromagnetic materials such as Alnico V, platinum– cobalt, and a variety of ferrites are possible smart materials.

Pyrosensitive Smart Materials

Pyrosensitive smart materials are useful in realizing intelligent electromagnetic active surfaces, radar-absorbing materials, electromagnetic shielding, and so on. For example, it has been demonstrated [Neelakanta et al., 1992] that the microwave reflection characteristics at a surface of a composite medium comprised of thermally controllable, solid-electrolytic zones (made of AgI pellets) show broadband microwave absorption/reflection characteristics under elevated temperatures. This principle can be adopted in conjunction with an electromagnetic sensor to provide a controllable feedback for thermal activation of fast-ion zones reconfigurably in order to acheive smart active-surface characteristics. Exclusive for this application, depending on the temperature limited conditions, the solid electrolyte can be chosen on the basis of its α- to β-phase transition characteristics. In order to keep the cost of the system low, a mixture phase can also be adopted, in which, commensurate with the elevated temperature operation, the host medium of the mixture could be a ceramic (dielectric).

58.6 Smart Sensors

Fiber-Optic-Based Sensors

The field of sensing technology has been revolutionized in the past decade by the entry of fiber optics. The properties of fiber optics that have made the technology suitable for communications are responsible for it being successful as a sensor as well. Fiber-optic sensors are of two types, namely, extrinsic and intrinsic. In the extrinsic type, the fiber itself acts only as a transmitter and does no part of the sensing. In an intrinsic type, however, the fiber acts as a sensor by using one of its intrinsic properties, such as induced birefringence or electrochromatism, to detect a phenomenon or quantify a measurement. Relevant to smart systems, the use of fiber optics in conjunction with optical (sensors) is based on changes in optical effects such as refractive index, optical absorption, luminescence, and chromic properties due to alterations in the environment in which the fiber is imbedded. Such alterations refer to strain or other elastic characteristics and thermal and/or electromagnetic properties [Claus, 1991]. Surfaces located with smart fiber sensors are known as smart skins.

Piezoelectric-Based Sensors

The most conventional form of sensing technology is that of piezoelectric materials, which generate an electrical response to a stimulus. In recent times piezoelectric materials have been greatly improved in mechanical strength and sensitivity. Pressure and vibration can be directly sensed as a one-to-one transduction effect resulting from the elastic-to-piezoelectric effect. Bending, on the other hand, can be sensed via piezoabsorption characteristics.

Magnetostriction-Based Sensors

The use of metallic glass as a distributive magnetostrictive sensor has been studied. Typically, in the imbedded smart sensing applications using the magnetostrictive property, the magnetic field is in the submicrogauss regime, and the nonlinearity associated with the hysteresis of magnetostriction provides a detectable sensor signal. Pressure and force, which cause static or quasi-static magnetic fields, as well as vibrations, which induce alternating magnetic fields, can be regarded as direct magnetostrictive sensor responses. In the bending mode, corresponding magnetostrictive absorption can also be sensed via reduction in the Q-factor due to absorption losses in a magnetostrictively tunable system.

Shape-Memory Effects-Based Sensors

The latest form of sensing technology utilizes shape-memory materials, namely, Nitinol alloys. The Nitinol sensors are used to measure strain and consist of superelastic Nitinol wires. The basic concept is to measure the change in resistance of a Nitinol wire used as an unbalanced arm of a Wheatstone bridge as a function of the strain. The desirable properties of Nitinol in such a sensing application are its high sensitivity and superelastic nature (which permits strains up to 6% to be accurately and repeatedly measured). The piezoelectric and Nitinol sensing materials can also be used for actuation applications.

Electromagnetics-Based Sensors

Smart electromagnetic sensors are simple deviations of classic electric/magnetic probes, more properly known as antennas or pickups. Depending on changes in the surroundings vis-à-vis the electromagnetic characteristics, these sensors respond and yield a corresponding signal. Again, the environmental changes refer to possible alterations caused by elastic, thermal, optical, magnetic, electric, and/or chemical influences.

Electroacoustic Smart Sensors

Electroacoustic smart sensors are embedded acoustic (vibration) sensors (similiar to a microphone) that adaptively yield a signal proportional to the acoustic input. Such inputs could result from changes in the alterations in the surroundings caused by elastic or thermal effects.

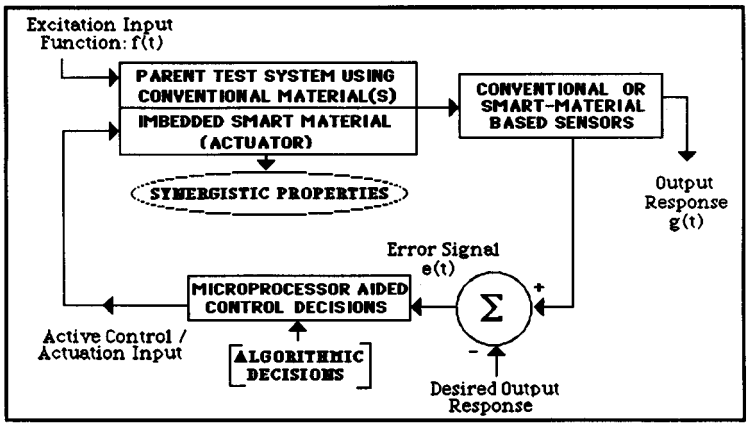


FIGURE 58.3 Schematic of a smart system.

As far as smart sensor technology is concerned, in fact, all the synergistic responses and effects between the electric and nonelectric phenomena just discussed can be judiciously adopted. Considering the state-of-the-art technology and practical considerations, however, the existing smart sensors are limited to the aforesaid versions. Future trends could, however, include other possible electric to nonelectric synergistic responses.

58.7 Examples of Smart/Intelligent Systems

The method of synthesizing a smart/intelligent system is illustrated in Fig. 58.3. The output response under a given set of input condition(s) of a parent test system is normally decided by the properties of the constituent (conventional) materials. If the system-states (changes) under the influence of external inputs are sensed, however, an appropriate feedback control can be used to actuate an imbedded smart material in the parent unit, so that output will track adaptively a desired response. The feedback path may include relevant electronic hardware (such as microprocessors) for on-line processing of the feedback signal to optimize the system performance.

Essentially, the smart materials can be adopted in two regimes of the system shown in Fig. 58.3. The *sensing unit* can be zones of an integrated set of smart material that senses the response of the parent system on a real time basis. (Sometimes, conventional sensors/tranducers can serve this purpose, as well.) The *actuating unit,* built-in as a part of the parent structure, consists of a smart material, which upon receiving the electric signal from the feedback loop modifies the response of the parent system, as dictated by the input signal. Thus, the actuation is based on the synergism between the electric input to the corresponding material property of the parent structure being altered.

The feedback control unit may consist of decision logic, which can relatively modify the error signal being fed to the actuator. The decision logic refers to, for example, response linearization, time-averaged smoothing, amplitude-limiting, and bandwidth control. On the basis of the general schematic depicted in Fig. 58.3, the following discusses a few examples of application-specific intelligent systems using smart materials.

Structural Engineering Applications

Figure 58.4 illustrates a smart vibration control strategy in structural beams. Normally, the parent beam is made of conventional materials, and its vibrational characteristics are decided by the elastic behavior of the constituent materials. Suppose a smart material is imbedded in the test beam. This material could be one of the types indicated in Fig. 58.2. A vibration sensor yields an electric output proportional to the vibration. Suppose the dynamic response of the beam (as observed at the output of the sensor) deviates from the desired characteristics. Then an error signal can be generated, which in turn can be used to develop an optimal control signal and this control signal can be fed back to the smart material whose elastic behavior is then altered as a function of the control input. As a result, the vibration characteristics of the entire (parent) structure are modified, or the system is dynamically tuned in an adaptive manner.

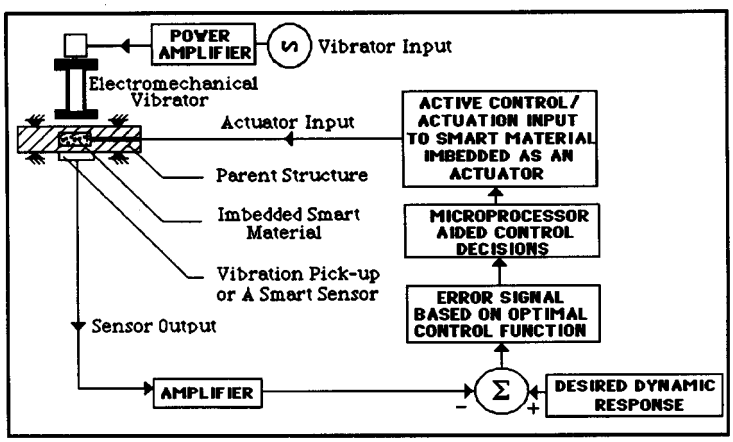


FIGURE 58.4 Active control of vibrating beams.

The vibration sensor used can be either a conventional transducer (such as resistive, capacitive, inductive, or optical displacement versions) or it can be a smart sensor by itself. For example, an optical fiber with a leaky sheath (which permits the light energy to leak from the core to the outside surface) can be imbedded in the parent structure. When the structure is deformed, the extent of light leakage from the fiber to the surrounding material will modify proportionately. Hence, the detected light signal from the fiber optics, when detected, delivers information on the deformation or the dynamic structural characteristics of the test beam. This sensor can be made smart by integrating a distributed set of fibers that can sense strain, vibration, temperature, if needed, and so on, so that the network implemented with appropriate algorithms will provide exhaustive data for a comprehensive adaptive feedback control strategy.

Although the scheme illustrated in Fig. 58.4 refers to vibration control (or damping) in structures, judicious choice of subsystems and materials will permit adaptive control over other structural aspects also, namely, strain, bending moment, and redistribution of load path in response to failures.

Electromagnetic Applications

Smart material/structural techniques can be adopted in electromagnetic systems. The following are possible applications:

- Smart low-frequency magnetic shields
- Smart high-frequency electromagnetic shields
- Smart electrostatic dissipative/conductive surfaces
- Smart radar-absorbing materials (smart RAMs)
- Smart linear and aperture antennas

In all the preceding applications, the basic consideration is that the relevant structure can smartly and adaptively change its electromagnetic properties (normally specified via dielectric permittivity, magnetic permeability, and electrical conductivity parameters) so that the desired electromagnetic performance is acheived. Two typical systems are detailed next.

Electromagnetic Active Surface Embedded with Ferroelectric Inclusions

Figure 58.5 illustrates the concept of a smart electromagnetic active surface. The surface is made of a mixture of polyacrylamide, ferrite, and barium titanate on a ceramic substrate. This skin material, which represents a lossy, nonlinear electromagnetic medium with anisotropic ferroelectric and ferromagnetic properties, offers different extents of surface impedence, in the presence and absence of an electric voltage stimulus applied to

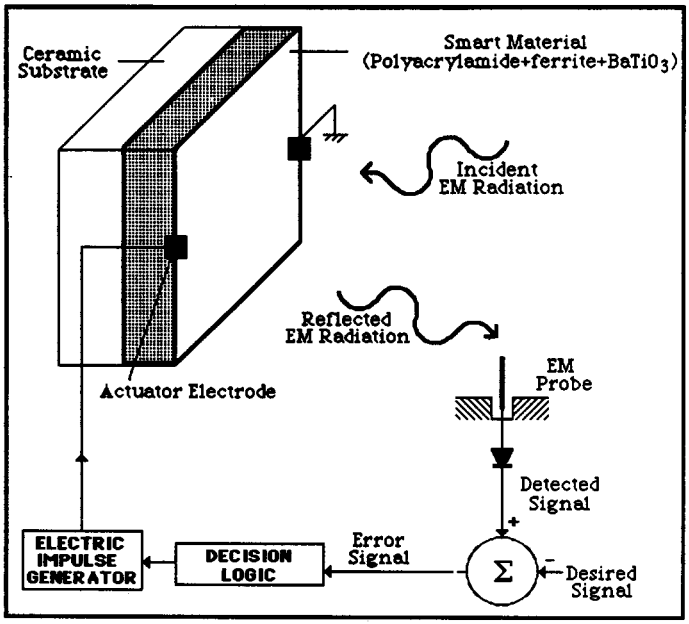


FIGURE 58.5 Smart electromagnetic active surface.

it. Hence, the reflection coefficient of this material to electromagnetic energy can be altered via electric stimulus. Relevant feedback can facilitate adaptive smart responsiveness of the system as illustrated [Neelakanta et al., 1992].

Smart Electromagnetic Aperture

The aperture-radiation of microwaves can be smartly controlled by using a pyrosensitive material as illustrated in Fig. 58.6. A set of solid-electrolyte (AgI) pellets interconnected via nichrome heating elements is placed at the aperture of a microwave horn. At room temperature, the pellets behave as dielectrics (β-phase AgI). When heated, however, the β-phase AgI changes to a highly conducting medium (α-phase), which would mask a part of the aperture, thus modifying the radiation pattern of the horn antenna. Again, an appropriate feedback loop would render the functioning of this system intelligent [Neelakanta et al., 1992].

58.8 High-Tech Application Potentials

Although smart material technology is in its infancy pending significant efforts to make it usable on a wide scale, the existing results and ongoing research have confirmed the usability of these materials in several avenues of modern high-technology systems.

Currently imaginable enclaves for the use of intelligent materials not only include structural engineering but also such areas as electromagnetics and biomedical, optical, and biological techniques. Relevant research has also been focused heavily in aerospace, aeronautics, marine vessel, and robotic applications.

Adaptive, self-monitoring of well-being by a system that has an integrated set of smart devices to self-assess its performance, diagnosing any malfunctions and failures and able to change its system characteristics vis-à-vis the environment, has been the objective of the relevant seed research pursued until now. For example, self-health checks by aircraft via a network of smart-skin sensors offer real-time monitoring of the structural well-being of tomorrow's aircraft [Claus, 1991]. The protocols in such efforts include self-diagnosis, prediction and notification, and self-repair strategies relevant to mechanical structures (such as aircraft bodies).

Another domain of smart material application is in self-induced morphologies in the infrastructure of the material with self-adaptive adjustments to the surroundings. Examples of this category include materials usable over a wide range of temperatures (as in space shuttles), with a smart adaptibility to transform according to the environment. Similarly, in radar stealth applications, the target skin could offer variable electromagnetic absorption over a broad band of radar frequencies.

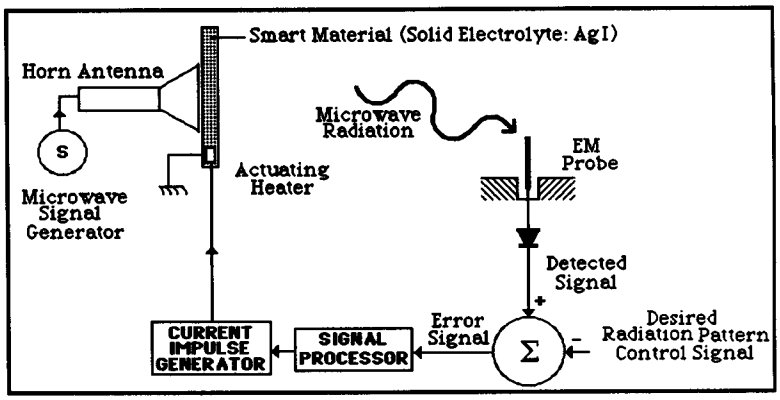


FIGURE 58.6 Smart electromagnetic aperture radiation control.

Extensions of smart material concepts can cover selective acoustical absorptions and adaptive chromic controls in glasses, mirrors, etc. In short, viable smart systems can be conceived with various combinations of material characteristics discussed earlier together with the advent of new conventional materials, innovative sensors, advances in microcomputers, artificial intelligence, neural networking, and other upcoming technologies. Currently imaginable outlets for smart materials are summarized in the following list.

1. Structural/mechanical engineering
   - Airborne/space-borne systems with smart skins for adaptive self-health check feasibilities
   - Earthquake-resistant intelligent buildings
   - Large deployable space structures
   - Nondestructive evaluation of large structures
2. Thermal engineering
   - Adaptive heat transfer and heat-resistant structures (space shuttles, etc.)
3. Optical engineering
   - Adaptive hue, optical transparency, reflection, opaqueness control in glasses and mirrors
4. Electromagnetic engineering
   - Magnetic and electrostatic shielding
   - High-frequency shielding
   - Radar-absorbing materials
   - Active surfaces
   - Adaptive scattering/radiation control
5. Acoustical engineering
   - Active absorption/reflection of sonar radiation
   - Adaptive anechoic chambers
6. Chemical engineering
   - Materials with adaptive adsorption characteristics
   - Adaptive corrosion-resistant materials
7. Biomedical engineering
   - Materials with smart structural properties usable as artificial limbs
   - Materials with adaptive biochemical properties
8. Warfare systems
   - Smart shelters
   - Shock-resistant structures

58.9 Conclusions

The quest for new materials in scientific endeavors and engineering applications is everlasting. The emergence of the smart material concept has set a trend that science and technology in the coming years will rely to a large extent on the development of exotic materials, with intelligent materials being the leading candidates. Such materials will be hyper-functional with unstereotyped purposiveness responses to novel and changing situations.

Defining Terms

Electroacoustic smart materials: Materials that have self-adaptive characteristics in their acoustical behavior (such as transmission, reflection, and absorption of acoustical energy) in response to an external stimulus applied as a function of the sensed acoustical response.

Electromagnetic smart materials: Materials such as shielding materials, radar-absorbing materials (RAMs), and electromagnetic surface materials, in all of which some electromagnetic properties can be adaptively controlled by means of an external stimulus dictated by the sensed electromagnetic response.

Electro-optic smart materials: Materials in which optical properties are changed self-adaptively with an external electric stimulus proportional to the sensed optical characteristics.

Electroplastic effect: Plastic deformation of metals with the application of high-density electric current.

Electroplastic smart materials: Materials with smart properties of elastic deformation changes proportional to a controlled electric current applied in proportion to the sensed deformation.

Electrorheological property: Property exhibited by some fluids that are capable of altering their flow characteristics depending on an externally applied electric field.

Electrorheological smart fluids: Fluids with smart flow characteristics dictated to change self-adaptively by means of an electric field applied in proportion to the sensed flow parameters.

Intelligent structures: Structures constructed of smart materials with a dedicated, discrete set of integrated actuators, sensors, etc., in order to respond to the environment around them in a predetermined (desired) manner.

Magnetostrictive effect: Structural strain experienced in a material subjected to a polarizing magnetic flux, or reversibly, experiencing magnetic property changes to external mechanical stresses.

Magnetostrictive smart materials: A class of materials with self-adaptively modifiable elastic properties in response to a magnetic field applied in proportion to sensed stress–strain information.

Nonlinear dielectric property: The distinct dependence of the electric permittivity of certain dielectric materials on the intensity of an applied electric field.

Nonlinear electroacoustic property: Nonlinear interaction between the atomic displacement and the electric field experienced in certain materials that would cause modulation effects resulting in the generation of new sideband frequencies (called Raman frequencies).

Nonlinear electro-optic property: Nonlinear changes in the refractive index of certain optically transparent materials with change(s) in the externally applied electric field.

Nonlinear magnetic property: Nonlinear dependence of the magnetic susceptibility of certain materials on the intensity of an applied magnetic field.

Piezoelectric property: Ability of a material to induce opposite charges at two faces (correspondingly to exhibit a voltage difference between the faces) of the material as a result of the strain due to a mechanical force applied across the faces; reversibly, application of a potential across the faces would induce a mechanical strain.

Piezoelectric smart materials: Materials capable of changing their elastic characteristics (by virtue of their piezoelectric property) self-adaptively in response to an externally applied electric potential proportional to the observed elastic behavior.

Pyrosensitive properties: Exhibited by materials known as solid electrolytes whose electromagnetic properties can be altered by temperature.

Pyrosensitive smart materials: Materials that self-adaptively (smartly) manage the electromagnetic surface characteristics of active surfaces constituted by pyrosensitive inclusions, in response to an external temperature-inducing stimulus applied per the feedback information on electromagnetic characteristics.

Shape-memory effects: Mechanism by which a plastically deformed object in the low-temperature martensitic condition regains its original shape when the external stress is removed and heat is applied.

Shape-memory smart materials: Materials that smartly change their elastic characteristics by virtue of their shape-restoration characteristics achieved by means of an external stimulus in proportion to the magnitude of sensed shape changes.

Smart, or intelligent, materials: A class of materials and/or composite media having inherent intelligence together with self-adaptive capabilities to external stimuli applied in proportion to a sensed material response.

Smart sensors: Sensors with inherent intelligence via bulit-in electronics.

Smart structural materials: Materials in which the mechanical (elastic) properties can be modified adaptively through the application of external stimuli.

Smart thermal materials: Materials that can influence their thermal states (temperature or thermal properties such as conductivity) self-adaptively by means of an external control in response to environmental demands.

Related Topics

49.1 Introduction • 49.2 Mechanical Characteristics • 50.2 Equation of State • 55.5 Dielectric Materials • 56.2 Physical Sensors

References

K.P. Chong, S.C. Liu, and J.C. Li (Eds.), *Intelligent Structures,* London and New York: Elsevier, 1990.

R.O. Claus, "Fiber sensors as nerves for smart materials," *Photonics Spectra,* vol. 25, no. 4, p. 75, 1991.

B. Culshaw, *Smart Structures and Materials,* Boston, Mass: Artech House, 1996.

M.V. Gandhi and B.S. Thompson, "A new generation of revolutionary ultra-advanced intelligent materials featuring electrorheological fluids," in *Smart Materials, Structures, and Mathematical Issues,* Lancaster, Pa.: Technomic Publishing, 1989, pp. 63–68.

C.M. Jackson, H.J. Wagner, and R.J. Wasilewski, *55-Nitinol—The Alloy with a Memory: Its Physical Metallurgy, Properties and Application,* NASA-SP-5110, 1972.

I.P. Kaminow, "Parametric principles in optics," *IEEE Spectrum,* vol. 2, p. 40, 1965.

P.S. Neelakanta, J. Abello, and C. Gu, "Microwave reflection at an active surface imbedded with fast-ion conductors," *IEEE Trans. Microwave Theory Tech.,* vol. MTT-40, no. 5, pp. 28–30, 1992.

P.S. Neelakanta and K. Subramaniam, "Controlling the properties of electromagnetic composites," *Adv. Materials and Process,* vol. 141, no. 3, pp. 20–25, 1992.

R.S. Reed, "Shock isolation using an active magnetostrictive element," in *Proc. 59th Shock and Vibration Symp.* vol. I, Albuquerque, New Mex., Oct. 18–20, 1988.

C.A. Rogers and R.C. Rogers, *Recent Advances in Adaptive and Sensory Materials and Their Applications,* Lancaster, Pa.: Technomic Publishing, 1992.

R. Ting, "The hydroacoustic behavior of piezoelectric composite materials," *Ferroelectrics,* vol. 102, pp. 215–224, 1990.

Further Information

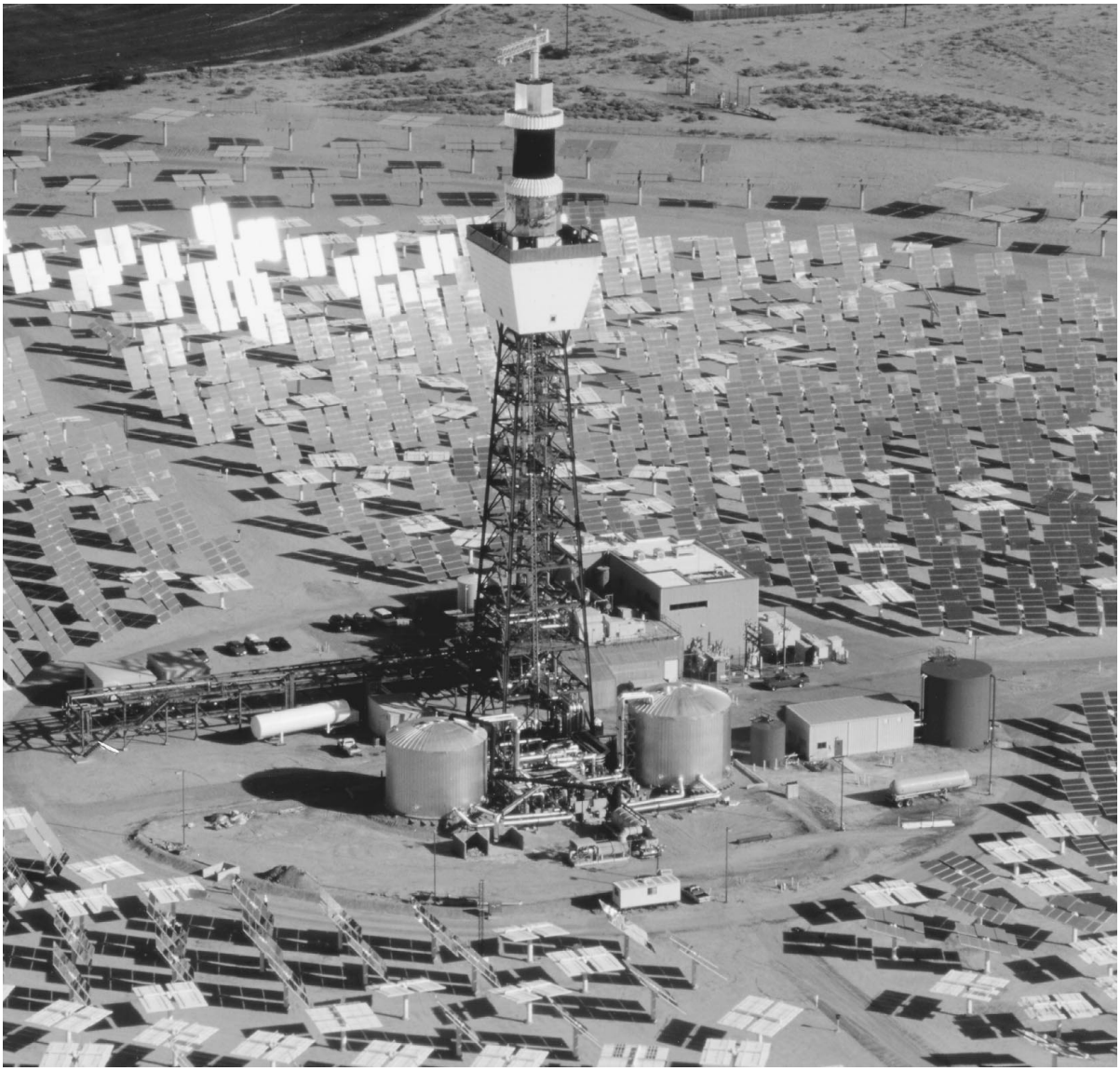
Intelligent Structures, edited by K.P Chong, S.C. Liu, and J.C. Li, contains the papers presented in an international workshop on intelligent structures held on 23–26 July 1990 in Taipei, Taiwan, Elsevier Science Publishers, 1990.

Another source is the *Proceedings of the International Workshop on Intelligent Materials,* The Society of Non-Traditional Technology, 1989.

Recent Advances in Adaptive and Sensory Materials and Their Applications, by C.A. Rogers and R.C. Rogers, Lancaster, Pa.: Technomic Publishing Co., Inc., 1992.

The author's publication with K. Subramaniam, "Controlling the Properties of Electromagnetic Composites" in *Advanced Materials and Processes,* vol. 141, no. 3, pp. 20–25, 1992.

Kersting, W.H. "Section VI – Energy"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



California's Mojave Desert is the site of Solar Two, the world's most technically advanced solar power plant. Solar Two uses an innovative molten salt technology to capture and store the sun's energy. This technology vastly differs from other solar technologies because, for the first time, it allows the practical storage of solar energy during the day to generate electricity at night.

Solar Two uses 1,926 heliostats (mirrors) located on a field in circular formation around a 300-ft tower to focus sunlight on a central receiver which generates a clean, inexhaustible supply of energy. The heliostats are equipped with two dual-axis electrical motors that position the mirrors to reflect the most sunlight onto the receiver via a computerized program.

Scheduled for testing until 1998, Solar Two will demonstrate the technical and economical viability of an emission-free solar thermal power plant to produce virtually unlimited supplies of electricity. (Photo courtesy of Southern California Edison.)

VI

Energy

59 Conventional Power Generation *G.G. Karady*
Fossil Power Plants • Nuclear Power Plants • Geothermal Power Plants • Hydroelectric Power Plants

60 Power Systems and Generation *R. Ramakumar, A.M. Barnett, L.L. Kazmerski, J.P. Benner, T.J. Coutts*
Distributed Power Generation • Photovoltaic Solar Cells • Thermophotovoltaics

61 Transmission *M.S. Chen, K.C. Lai, R.S. Thallam, M.E. El-Hawary, C. Gross, A.G. Phadke, R.B. Gungor, J.D. Glover*
Alternating Current Overhead: Line Parameters, Models, Standard Voltages, Insulators • Alternating Current Underground: Line Parameters, Models, Standard Voltages, Cables • High-Voltage Direct-Current Transmission • Compensation • Fault Analysis in Power Systems • Protection • Transient Operation of Power Systems • Planning

62 Power Quality *J. Arrillaga*
Power Quality Disturbances • Power Quality Monitoring • Power Quality Conditioning

63 Power Systems *L.L. Grigsby, A.P. Hanson, R.A. Schlueter, N. Alemadi*
Power System Analysis • Voltage Instability

64 Power Transformers *C.A. Gross*
Transformer Construction • Power Transformer Modeling • Transformer Performance • Transformers in Three-Phase Connections • Autotransformers

65 Energy Distribution *G.G. Karady*
Primary Distribution System • Secondary Distribution System • Radial Distribution System • Secondary Networks • Load Characteristics • Voltage Regulation • Capacitors and Voltage Regulators

66 Electrical Machines *C.C. Liu, K.T. Vu, Y. Yu, D. Galler, E.G. Strange, Chee-Mun Ong*
Generators • Motors • Small Electric Motors • Simulation of Electric Machinery

67 Energy Management *K.N. Stanton, J.C. Giri, A.J. Bose*
Power System Data Acquisition and Control • Automatic Generation Control • Load Management • Energy Management • Security Control • Operator Training Simulator

68 Power System Analysis Software *C.P. Arnold, N.R. Watson*
Early Analysis Programs • The Second Generation of Programs • Further Development of Programs

William H. Kersting
New Mexico State University

THE GENERATION, TRANSMISSION, AND DISTRIBUTION of electrical energy remains one of the most exciting and challenging areas of electrical engineering. Without a safe, reliable, and economic supply of electrical energy, all industry would come to a grinding halt.

This section will present chapters discussing the theory and methods for the generation, transmission, and distribution of electrical energy. While the fundamentals have been around a long time, the application of the fundamentals will continue to take on many new forms.

The great majority of electrical energy continues to be generated by large conventional power plants, which are discussed in Chapter 61. As Chapter 62 will explain, fuel cells, wind, and solar are becoming important components of the overall generation picture.

The transmission of electrical energy over long distances and at increasingly higher voltages has become an ever more important component as "open access" of these facilities becomes a reality. Chapter 63 will present the theory of transmission, including alternating current and direct current transmission, both overhead and underground. In addition, the discussion will include the protection of these facilities, transient operation, and planning.

A modern power system operates at many different voltage levels. Because of this, transformers play a key role. The theory is the same for all voltage levels and will be presented in Chapter 66. The chapter will also discuss application of the theory for different types of transformers commonly used.

The final component in bringing electrical energy to the ultimate user is the distribution system. For many years the stepchild to the more costly generation and transmission components, distribution systems are now playing an important role in increasing reliability and service at a reduced cost. Chapter 67 presents an overview of this key component.

Generators and motors are still the primary devices for converting energy from mechanical to electrical and vice versa. Chapter 68 is devoted to the theory of ac/dc motors and generators.

The automatic control of the total power system is presented in Chapter 69. This is one area of power systems that has changed dramatically and continues to change on almost a daily basis. In many ways, the total field of electrical engineering is applied in the control of a modern power system.

Nomenclature

Symbol	Quantity	Unit
D	damping coefficient	
DF	demand factor	
δ	power angle	degree
δ	torque angle	degree
f_s	slip frequency	Hz
ϕ	core flux	Wb
ϕ_f	magnetic field flux	Wb
I_{ac}	rms ac current	A
I_f	field current	A
I_r	rotor current	A
J	moment of inertia of rotor	kg-m^2/rad
K_a	armature constant of a dc machine	
K_t	torque constant of a dc machine	
LSF	loss factor	
λ	magnetic flux linkages	Wb/m
n	rotor speed	rpm
n	Steinmetz constant	
N	number of turns	
N_p	number of stator pole pairs	
ω_s	slip frequency radian	
P	real power	W
P_e	eddy current loss	W
P_h	hysteresis power loss	W
Q	reactive power	
r	radius of conductor	m
ρ	resistivity of conductor	Ωm
s	slip of an induction motor	
$\overline{S}$	complex power	
T	shaft torque	N-m
θ	rotor angle	degree
θ	power factor angle	degree
V_a	armature back emf	V
w	shaft speed	rad/s

Karady, G.G. “Conventional Power Generation”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

59

Conventional Power Generation

59.1 Introduction

59.2 Fossil Power Plants
Fuel Handling • Boiler • Turbine • Generator • Electric System • Condenser • Stack and Ash Handling • Cooling and Feedwater System

59.3 Nuclear Power Plants
Pressurized Water Reactor • Boiling-Water Reactor

59.4 Geothermal Power Plants

59.5 Hydroelectric Power Plants

George G. Karady
Arizona State University

59.1 Introduction

The electric energy demand of the world is continuously increasing, and most of the energy is generated by conventional power plants, which remain the only cost-effective method for generating large quantities of energy.

Power plants utilize energy stored in the earth and convert it to electrical energy that is distributed and used by customers. This process converts most of the energy into heat, which increases the entropy of the earth. In this sense, power plants deplete the earth's energy supply. Efficient operation becomes increasingly important to conserve energy.

Typical energy sources used by power plants include fossil fuel (gas, oil, and coal), nuclear fuel (uranium), geothermal energy (hot water, steam), and hydro energy (water falling through a head).

Around the turn of the century, the first fossil power plants used steam engines as the prime mover. These plants were evolved to an 8- to 10-MW capacity, but increasing power demands resulted in the replacement by a more efficient steam boiler–turbine arrangement. The first commercial steam turbine was introduced by DeLaval in 1882. The boilers were developed from heating furnaces. Oil was the preferred and most widely used fuel in the beginning. The oil shortage promoted coal-fired plants, but the adverse environmental effects (sulfur dioxide generation, acid rain, dust pollution, etc.) curtailed their use in the late seventies. Presently the most acceptable fuel is natural gas, which minimizes pollution and is available in large quantities. During the next two decades, gas-fired power plants will dominate the electric industry.

The hydro plants' ancestors are water wheels used for pumping stations, mill driving, etc. Water-driven turbines were developed in the last century and used for generation of electricity since the beginning of their commercial use. However, most of the sites that can be developed economically are currently being utilized. No significant new development is expected in the United States in the near future.

Nuclear power plants appeared after the Second World War. The major development occurred during the sixties; however, by the eighties environmental considerations stopped plant development in the United States and slowed it down all over the world. Presently, the future of nuclear power generation is unclear, but the abundance of nuclear fuel and the expected energy shortage in the early part of the next century may rejuvenate nuclear development if safety issues can be resolved.

Geothermal power plants are the product of the clean energy concept, although the small-scale, local application of geothermal energy has a long history. Presently only a few plants are in operation. The potential for further development is limited because of the unavailability of geothermal energy sites that can be developed economically.

Typical technical data for different power plants is shown in Table 59.1.

59.2 Fossil Power Plants

The operational concept and major components of a fossil power plant are shown in Fig. 59.1.

Fuel Handling

The most frequently used **fuels** are oil, natural gas, and coal. Oil and gas are transported by rail, on ships, or through pipelines. In the former case the gas is liquefied. Coal is transported by rail or ships if the plant is near a river or the sea. The power plant requires several days of fuel reserve. Oil and gas are stored in large metal tanks, and coal is kept in open yards. The temperature of the coal layer must be monitored to avoid self-ignition.

Oil is pumped and gas is fed to the burners of the **boiler.** Coal is pulverized in large mills, and the powder is mixed with air and transported by air pressure, through pipes, to the burners. The coal transport from the yard to the mills requires automated transporter belts, hoppers, and sometimes manually operated bulldozers.

Boiler

Two types of boilers are used in modern power plants: subcritical water-tube drum-type and supercritical once-through type. The former operates around 2500 psi, which is under the water critical pressure of 3208.2 psi. The latter operates above that pressure, at around 3500 psi. The superheated steam temperature is about 1000°F (540°C) because of turbine temperature limitations.

A typical subcritical water-tube drum-type boiler has an inverted-U shape. On the bottom of the rising part is the furnace where the fuel is burned. The walls of the furnace are covered by water pipes. The drum and the **superheater** are at the top of the boiler. The falling part of the U houses the reheaters, **economizer** (water heater), and air preheater, which is supplied by the forced-draft fan. The induced-draft fan forces the flue gases out of the system and sends them up the stack, which is located behind the boiler. A flow diagram of the drum-type boiler is shown in Fig. 59.2. The steam generator has three major systems: fuel, air-flue gas, and water-steam.

Fuel System. Fuel is mixed with air and injected into the furnace through burners. The burners are equipped with nozzles, which are supplied by preheated air and carefully designed to assure the optimum air-fuel mix. The fuel mix is ignited by oil or gas torches. The furnace temperature is around 3000°F.

Air-Flue Gas System. Ambient air is driven by the forced-draft fan through the air preheater, which is heated by the high-temperature (600°F) flue gases. The air is mixed with fuel in the burners and enters into the furnace, where it supports the fuel burning. The hot combustion flue gas generates steam and flows through the boiler

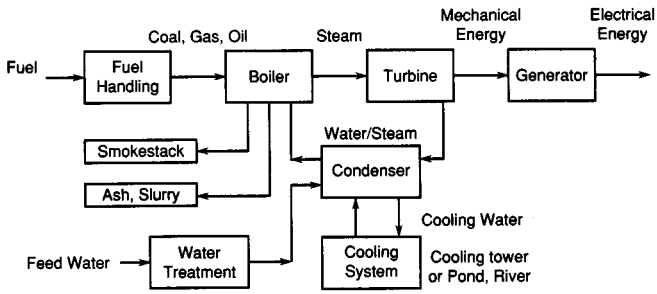


FIGURE 59.1 Major components of a fossil power plant.

TABLE 59.1 Power Plant Technical Data

Generation Type	Typical MW Size	Capitalized Plant Cost, $/kW	Construction Lead Time, Years	Heat Rate, Btu/kWh	Fuel Cost, $/MBtu	Fuel Type	Equivalent Forced Outage Rate	Equivalent Scheduled Outage Rate	O&M Fixed, $/kW/year	Cost Variable $/MWh
Nuclear	1200	2400	10	10,400	1.25	Uranium	20	15	25	8
Pulverized coal steam	500	1400	6	9,900	2.25	Coal	12	12	20	5
Atmospheric fluidized bed	400	1400	6	9,800	2.25	Coal	14	12	17	6
Gas turbine	100	350	2	11,200	4.00	Nat. gas	7	7	1	5
Combined-cycle	300	600	4	7,800	4.00	Nat. gas	8	8	9	3
Coal-gasification combined-cycle	300	1500	6	9,500	2.25	Coal	12	10	25	4
Pumped storage hydro	300	1200	6	—	—		5	5	5	2
Conventional hydro	300	1700	6	—	—		3	4	5	2

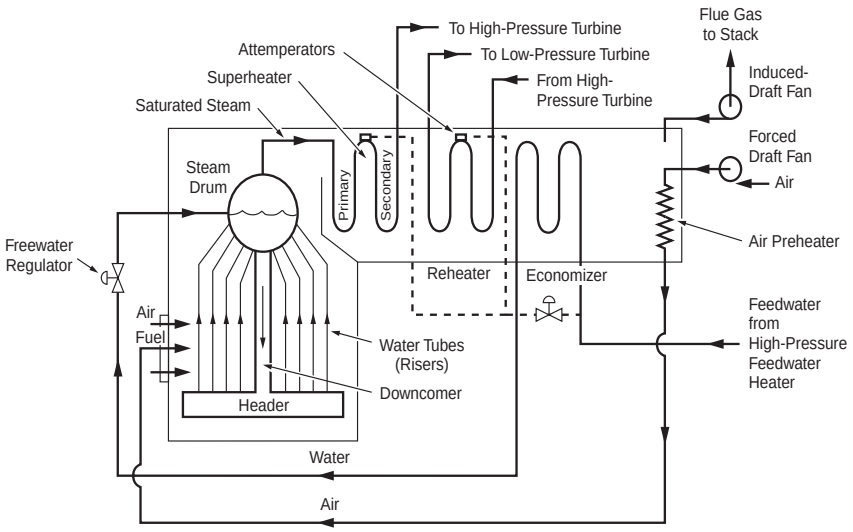


FIGURE 59.2 Flow diagram of a typical drum-type steam boiler.

to heat the superheater, reheaters, economizer, etc. Induced-draft fans, located between the boiler and the stack, increase the flow and send the 300°F flue gases to the atmosphere through the stack.

Water-Steam System. Large pumps drive the feedwater through the high-pressure heaters and the economizer, which further increases the water temperature (400–500°F). The former is heated by steam removed from the turbine; the latter is heated by the flue gases. The preheated water is fed to the steam drum. Insulated tubes, called downcomers, are located outside the furnace and lead the water to a header. The header distributes the hot water among the risers. These are water tubes that line the furnace walls. The water tubes are heated by the combustion gases through both convection and radiation. The steam generated in these tubes flows to the drum, where it is separated from the water. Circulation is maintained by the density difference between the water in the downcomer and the water tubes. Saturated steam, collected in the drum, flows through the superheater. The superheater increases the steam temperature to about 1000°F. Dry superheated steam drives the high-pressure turbine. The exhaust from the high-pressure turbine goes to the reheater, which again increases the steam temperature. The reheated steam drives the low-pressure turbine.

The typical supercritical once-through-type boiler concept is shown in Fig. 59.3.

The feedwater enters through the economizer to the boiler, which consists of riser tubes that line the furnace wall. All the water is converted to steam and fed directly to the superheater. The latter increases the steam temperature above the critical temperature of the water and drives the turbine. The construction of these steam generators is more expensive than the drum-type units but has a higher operating efficiency.

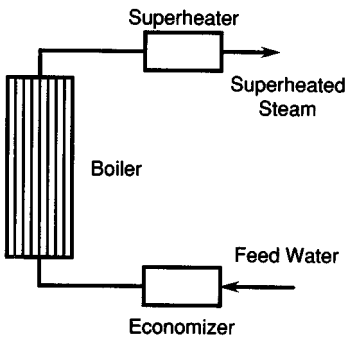


FIGURE 59.3 Concept of once-through-type steam generator.

Turbine

The turbine converts the heat energy of the steam into mechanical energy. Modern power plants usually use one high-pressure and one or two lower-pressure turbines. A typical turbine arrangement is shown in Fig. 59.4.

The figure shows that only one bearing is between each of the machines. The shafts are connected to form a tandem compound steam turbine unit. High-pressure steam enters the high-pressure turbine to flow through and drive the turbine. The exhaust is reheated in the boiler and returned to the lower-pressure units. Both the rotor and the stationary part of the turbine have blades. The length of the blades increases from the steam entrance to the exhaust.

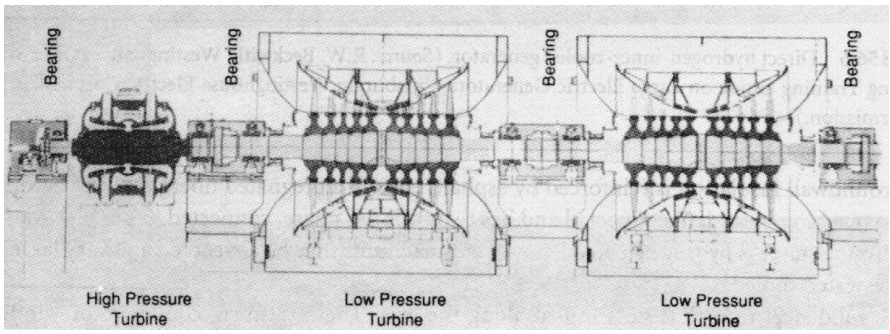


FIGURE 59.4 Large tandem compound steam turbine. (*Source:* M.M. El-Wakil, *Power Plant Technology,* New York: McGraw-Hill, 1984, p. 210. With permission.)

Figure 59.5 shows the blade arrangement of an *impulse*-type turbine. Steam enters through nozzles and flows through the first set of moving rotor blades. The following stationary blades change the direction of the flow and direct the steam into the next set of moving blades. The nozzles increase the steam speed and reduce pressure, as shown in the figure. The impact of the high-speed steam, generated by the change of direction and speed in the moving blades, drives the turbine.

The *reaction*-type turbine has nonsymmetrical blades arranged like those shown in Fig. 59.5. The blade shape assures that the pressure continually drops through all rows of blades, but steam velocity decreases in the moving blades and increases in the stationary blades.

Generator

The generator converts mechanical energy from the turbines into electrical energy. The major components of the generator are the frame, stator core and winding, rotor and winding, bearings, and cooling system. Figure 59.6 shows the cross section of a modern hydrogen-cooled generator.

The stator has a laminated and slotted silicon steel iron core. The stacked core is clamped and held together by insulated axial through bolts. The stator winding is placed in the slots and consists of a copper-strand configuration with woven glass insulation between the strands and mica flakes, mica mat, or mica paper ground-wall insulation. To avoid insulation damage caused by vibration, the groundwall insulation is reinforced by asphalt, epoxy-impregnated fiberglass, or Dacron. The largest machine stator is Y-connected and has two coils per phase, connected in parallel. Most frequently, the stator is hydrogen-cooled; however, small units may be air-cooled and very large units may be water-cooled.

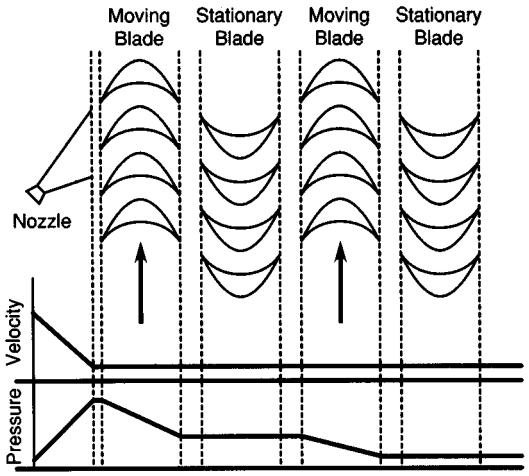


FIGURE 59.5 Velocity and pressure variation in an impulse turbine.

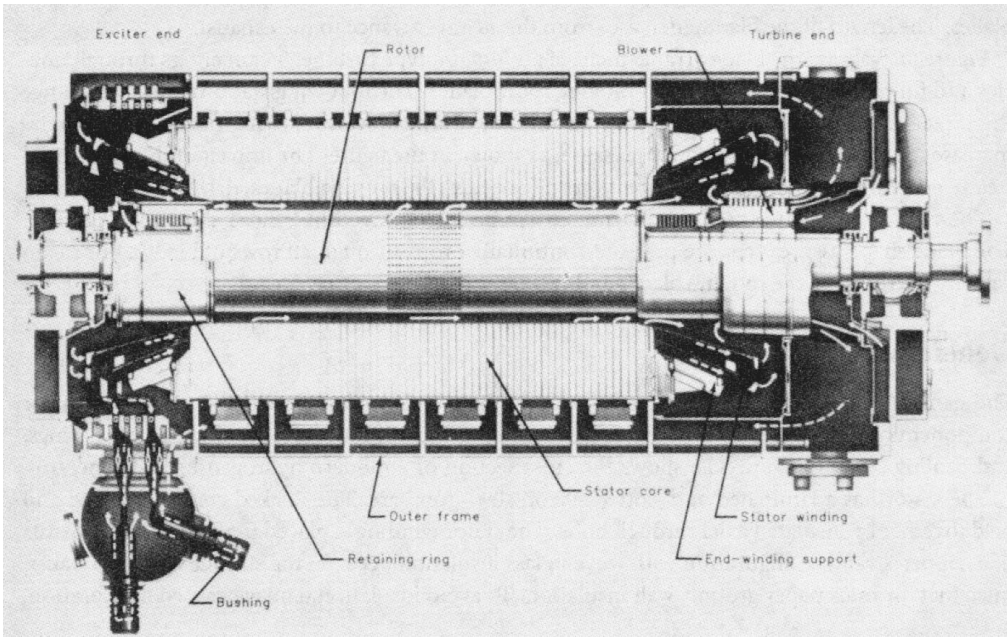


FIGURE 59.6 Direct hydrogen-inner-cooled generator. (*Source:* R.W. Beckwith, Westinghouse Power Systems Marketing Training Guide on Large Electric Generators, Pittsburgh: Westinghouse Electric Corp. 1979, p. 54. With permission.)

The solid steel rotor has slots milled along the axis. The multiturn, copper rotor winding is placed in the slots and cooled by hydrogen. Cooling is enhanced by subslots and axial cooling passages. The rotor winding is restrained by wedges inserted in the slots.

The rotor winding is supplied by dc current, either directly by a brushless excitation system or through collector rings. The rotor is supported by bearings at both ends. The non-drive-end bearing is insulated to avoid shaft current generated by stray magnetic fields. The hydrogen is cooled by a hydrogen-to-water heat exchanger mounted on the generator or installed in a closed-loop cooling system.

The dc current of the rotor generates a rotating magnetic field that induces an ac voltage in the stator winding. This voltage drives current through the load and supplies the electrical energy.

Electric System

Energy generated by the power plant supplies the electric network through transmission lines. The power plant operation requires auxiliary power to operate mills, pumps, etc. The auxiliary power requirement is approximately 10 to 15%.

Smaller generators are directly connected in parallel using a busbar. Each generator is protected by a circuit breaker. The power plant auxiliary system is supplied from the same busbar. The transmission lines are connected to the generator bus, either directly or through a transformer.

The larger generators are unit-connected. In this arrangement the generator is directly connected, without a circuit breaker, to the main transformer. A conceptual one-line diagram is shown in Fig. 59.7. The generator supplies main and auxiliary transformers without circuit breakers. The units are connected in parallel at the high-voltage side of the main transformers by a busbar. The transmission lines are also supplied from this bus. Circuit breakers are installed at the secondary side of the main and auxiliary transformer. The application of generator circuit breaker is not economical in the case of large generators. Because of the generator's large short-circuit current, special expensive circuit breakers are required. However, the transformers reduce the short-circuit current and permit the use of standard circuit breakers at the secondary side. The disconnect switches permit visual observation of the off state and are needed for maintenance of the circuit breakers.

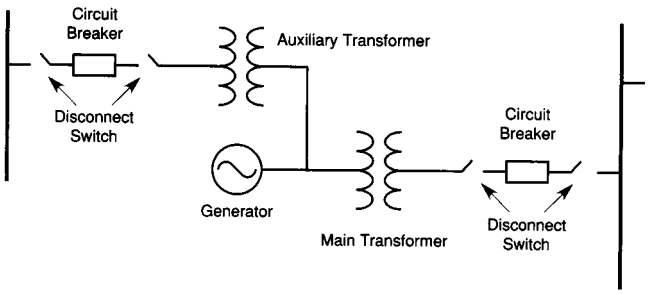


FIGURE 59.7 Conceptual one-line diagram for a unit-connected generator.

Condenser

The condenser condenses turbine exhaust steam to water, which is pumped back to the steam generator through various water heaters. The condensation produces a vacuum, which is necessary to exhaust the steam from the turbine. The condenser is a shell-and-tube heat exchanger, where steam condenses on water-cooled tubes. Cold water is obtained from the cooling towers or other cooling systems. The condensed water is fed through a deaerator, which removes absorbed gases from the water. Next, the gas-free water is mixed with the feedwater and returned to the boiler. The gases absorbed in the water may cause corrosion (oxygen) and increase condenser pressure, adversely affecting efficiency. Older plants use a separate deaerator heater, while deaerators in modern plants are usually integrated in the condensor, where injected steam jets produce pressure drop and remove absorbed gases.

Stack and Ash Handling

The stack is designed to disperse gases into the atmosphere without disturbing the environment. This requires sufficient stack height, which assists the fans in removing gases from the boiler through natural convection. The gases contain both solid particles and harmful chemicals. Solid particles, like dust, are removed from the flue gas by electrostatic precipitators or bag-house filters. Harmful sulfur dioxide is eliminated by scrubbers. The most common is the lime/limestone scrubbing process.

Coal-fired power plants generate a significant amount of ash. The disposition of the ash causes environmental problems. Several systems have been developed in past decades. Large ash particles are collected by a water-filled ash hopper, located at the bottom of the furnace. Fly ash is removed by filters, then mixed with water. Both systems produce sludge that is pumped to a clay-lined pond where water evaporates and the ash fills disposal sites. The clay lining prevents intrusion of groundwater into the pond.

Cooling and Feedwater System

The condenser is cooled by cold water. The open-loop system obtains the water from a river or sea, if the power plant location permits it. The closed-loop system utilizes cooling towers, spray ponds, or spray canals. In the case of spray ponds or canals, the water is pumped through nozzles, which generate fine sprays. Evaporation cools the water sprays as they fall back into the pond. Several different types of cooling towers have been developed. The most frequently used is the wet cooling tower, where the hot water is sprayed on top of a latticework of horizontal bars. The water drifts downward and is cooled, through evaporation, by the air, which is forced by fans or natural draft upward.

The power plant loses a small fraction of the water through leakage. The feedwater system replaces this lost water. Replacement water has to be free from absorbed gases, chemicals, etc., because the impurities cause severe corrosion in the turbines and boiler. The water treatment system purifies replacement water by pretreatment, which includes filtering, chlorination, demineralization, condensation, polishing. These complicated chemical processes result in a corrosion-free high-quality feedwater.

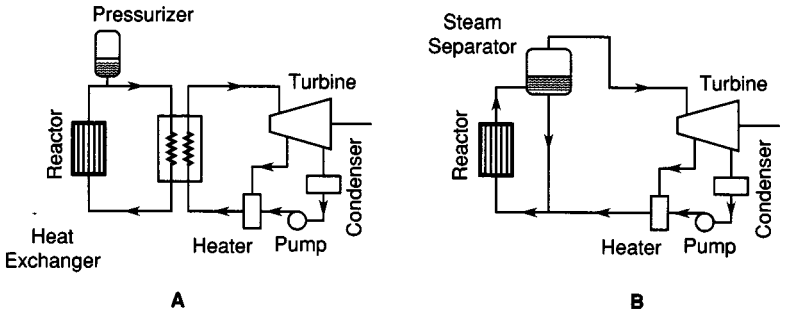


FIGURE 59.8 (A) Power plant with PWR; (B) power plant with BWR.

59.3 Nuclear Power Plants

More than 500 nuclear power plants operate around the world. Close to 300 operate pressurized water reactors (PWRs), more than 100 are built with boiling-water reactors (BWRs), about 50 use gas-cooled reactors, and the rest are heavy-water reactors. In addition a few fast breeder reactors are in operation. These reactors are built for better utilization of uranium fuel. The modern nuclear plant size varies from 100 to 1200 MW.

Pressurized Water Reactor

The general arrangement of a power plant with a PWR is shown in Fig. 59.8(A).

The **reactor** heats the water from about 550 to about 650°F. High pressure, at about 2235 psi, prevents boiling. Pressure is maintained by a pressurizer, and the water is circulated by a pump through a heat exchanger. The heat exchanger evaporates the feedwater and generates steam, which supplies a system similar to a conventional power plant. The advantage of this two-loop system is the separation of the potentially radioactive reactor cooling fluid from the water-steam system.

The reactor core consists of fuel and control rods. Grids hold both the control and fuel rods. The fuel rods are inserted in the grid following a predetermined pattern. The fuel elements are Zircaloy-clad rods filled with UO_2 pellets. The control rods are made of a silver (80%), cadmium (5%), and indium (15%) alloy protected by stainless steel. The reactor operation is controlled by the position of the rods. In addition, control rods are used to shut down the reactor. The rods are released and fall in the core when emergency shutdown is required. Cooling water enters the reactor from the bottom, flows through the core, and is heated by nuclear fission.

Boiling-Water Reactor

In the BWR shown in Fig. 59.8(B), the pressure is low, about 1000 psi. The nuclear reaction heats the water directly to evaporate it and produce wet steam at about 545°F. The remaining water is recirculated and mixed with feedwater. The steam drives a turbine that typically rotates at 1800 rpm. The rest of the plant is similar to a conventional power plant. A typical reactor arrangement is shown in Fig. 59.9. The figure shows all the major components of a reactor. The fuel and control rod assembly is located in the lower part. The steam separators are above the core, and the steam dryers are at the top of the reactor. The reactor is enclosed by a concrete dome.

59.4 Geothermal Power Plants

The solid crust of the earth is an average of 20 mil (32 km) deep. Under the solid crust is the molten mass, the magma. The heat stored in the magma is the source of geothermal energy. The hot molten magma comes close to the surface at certain points in the earth and produces volcanoes, hot springs, and geysers. These are the signs of a possible geothermal site. Three forms of geothermal energy are considered for development.

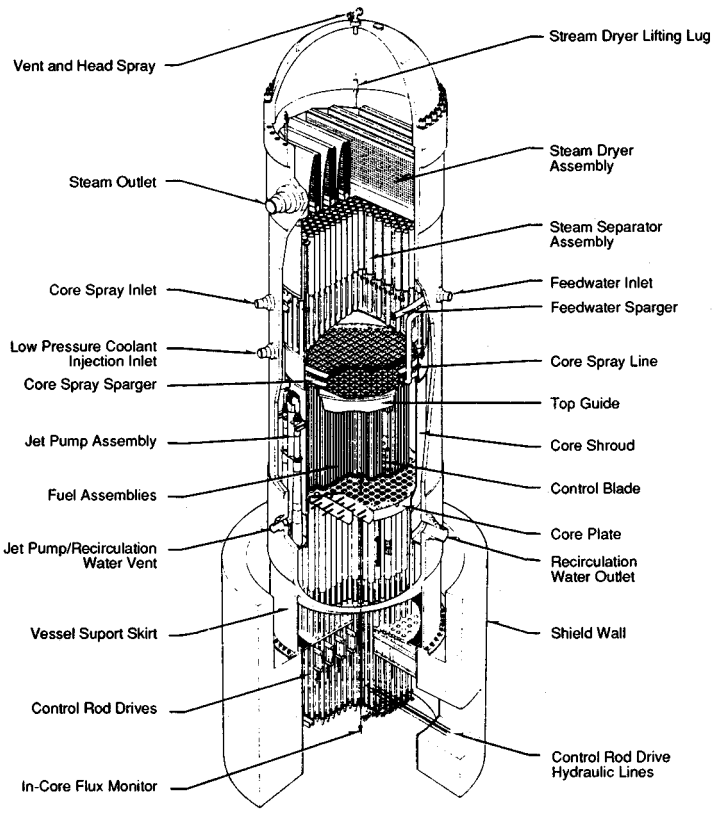


FIGURE 59.9 Typical BWR reactor arrangement. (*Source:* Courtesy of General Electric Company.)

Hydrothermal Source. This is the most developed source. Power plants, up to a capacity of 2000 MW, are in operation worldwide. Heat from the magma is conducted upward by the rocks. The groundwater drifts down through the cracks and fissures to form reservoirs when water-impermeable solid rock bed is present. The water in this reservoir is heated by the heat from the magma. Depending on the distance from the magma and rock configuration, steam, hot pressurized water, or the mixture of the two are generated. Signs of these underwater reservoirs include hot springs and geysers. The reservoir is tapped by a well, which brings the steam-water mixture to the surface to produce energy. The geothermal power plant concept is illustrated in Fig. 59.10.

The hot water and steam mixture is fed into a separator. If the steam content is high, a centrifugal separator is used to remove the water and other particles. The obtained steam drives a turbine. The typical pressure is

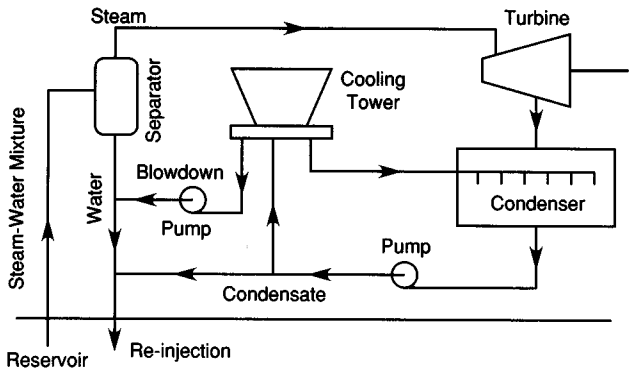


FIGURE 59.10 Concept of a geothermal power plant.

around 100 psi and the temperature is around 400°F (200°C). If the water content is high, the water-steam mixture is led through a flashed-steam system where the expansion generates a better quality of steam and separates the steam from the water. The water is returned to the ground, the steam drives the turbine. Typically the steam entering the turbine has a temperature of 120 to 150°C and a pressure of 30 to 40 psi.

The turbine drives a conventional generator. The typical rating is in the 20- to 100-MW range. The exhaust steam is condensed in a direct-contact condenser. A part of the obtained water is reinjected into the ground. The rest of the water is fed into a cooling tower to provide cold water to the condenser.

Major problems with geothermal power plants are the minerals and noncondensable gases in the water. The minerals make the water highly corrosive, and the separated gases cause air pollution. An additional problem is noise pollution. The centrifugal separator and blowdowns require noise dampers and silencers.

Petrothermal Source. Some fields have only hot rocks under the surface. Utilization of this petrothermal source requires pumping surface water through a well in a constructed hole to a reservoir. The hot water is then recovered through another well. The problem is the formation of a reservoir. The U.S. government is studying practical uses of petrothermal sources.

Geopressured Source. In deep underground holes (8000 to 30,000 ft) a mixture of pressurized water and natural gas, like methane, may sometimes be found. These geopressured sources promise power generation through the combustion of methane and the direct recovery of heat from the water. The geopressured method is currently in an experimental stage, with operating pilot plants.

59.5 Hydroelectric Power Plants

Hydroelectric power plants convert energy produced by a water head into electric energy. A typical hydroelectric power plant arrangement is shown in Fig. 59.11.

The head is produced by building a dam across a river, which forms the upper-level reservoir. In the case of low head, the water forming the reservoir is fed to the turbine through the intake channel or the turbine is integrated in the dam. The latter arrangement is shown in Fig. 59.11(A). **Penstock** tubes or tunnels are used for medium- [Fig. 59.11(B)] and high-head plants (Fig. 59.12). The spillway regulates the excess water by opening gates at the bottom of the dam or permitting overflow on the spillway section of the dam. The water discharged from the turbine flows to the lower or tail water reservoir, which is usually a continuation of the original water channel.

High-Head Plants. High-head plants (Fig. 59.12) are built with impulse turbines, where the head-generated water pressure is converted into velocity by nozzles and the high-velocity water jets drive the turbine runner.

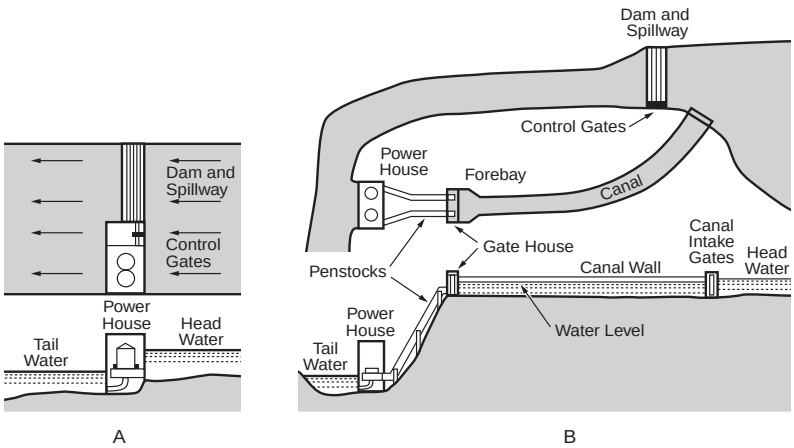


FIGURE 59.11 Hydroelectric power plant arrangement. (A) Low-head plant, (B) medium-head plant.

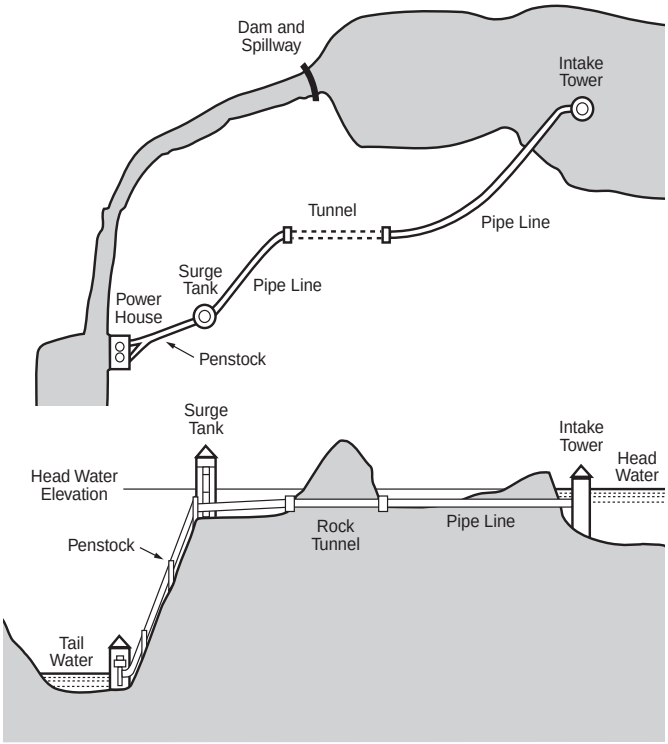


FIGURE 59.12 Hydroelectric power plant arrangement, high-head plant.

Low- and Medium-Head Plants. Low- and medium-head installations (Fig. 59.11) are built with reaction-type turbines, where the water pressure is mostly converted to velocity in the turbine. The two basic classes of reaction turbines are the propeller or Kaplan type, mostly used for low-head plants, and the Francis type, mostly used for medium-head plants. The cross section of a typical low-head Kaplan turbine is shown in Fig. 59.13.

The vertical shaft turbine and generator are supported by a thrust bearing immersed in oil. The generator is in the upper, watertight chamber. The turbine runner has 4 to 10 propeller types, and adjustable pitch blades. The blades are regulated from 5 to 35 degrees by an oil-pressure-operated servo mechanism. The water is evenly distributed along the periphery of the runner by a concrete spiral case and regulated by adjustable wicket blades. The water is discharged from the turbine through an elbow-shaped draft tube. The conical profile of the tube reduces the water speed from the discharge speed of 10–30 ft/s to 1 ft/s to increase turbine efficiency.

Hydrogenerators. The hydrogenerator is a low-speed (100 to 360 rpm) salient-pole machine with a vertical shaft. A typical number of poles is from 20 to 72. They are mounted on a pole spider, which is a welded, spoked wheel. The spider is mounted on the forged steel shaft. The poles are built with a laminated iron core and stranded copper winding. Damper bars are built in the pole faces. The stator is built with slotted, laminated iron core that is supported by a welded steel frame. Windings are made of stranded conductors insulated between the turns by fiberglass or Dacron-glass. The ground insulation is multiple layers of mica tape impregnated by epoxy or polyester resins. The older machines use asphalt and mica tape insulation, which is sensitive to corona-discharge-caused insulation deterioration. Direct water cooling is used for very large machines, while the smaller ones are air- or hydrogen-cooled. Some machines use forced-air cooling with an air-to-water heat exchanger. A braking system is installed in larger machines to stop the generator rapidly and to avoid damage to the thrust.

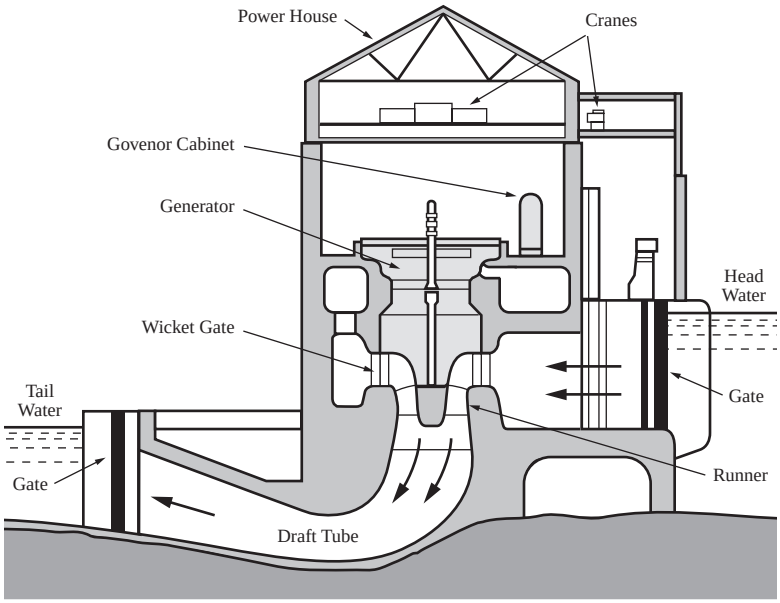


FIGURE 59.13 Typical low-head hydroplant with Kaplan turbine.

Defining Terms

Boiler: A steam generator which converts the chemical energy stored in the fuel (coal, gas, etc.) to thermal energy by burning. The heat evaporates the feedwater and generates high-pressure steam.

Economizer: A heat exchanger which increases the feedwater temperature. It is heated by the flue gases.

Fuel: Thermal power plants use coal, natural gas, and oil as a fuel, which is burned in the boiler. Nuclear power plants use uranium as a fuel.

Penstock: A water tube which feeds the turbine. It is used when the slope is too steep for using an open canal.

Reactor: A container where the nuclear reaction takes place. The reactor converts the nuclear energy to heat.

Superheater: A heat exchanger which increases the steam temperature to about 1000°F. It is heated by the flue gases.

Surge tank: An empty vessel which is located at the top of the penstock. It is used to store water surge when the turbine valve is suddenly closed.

References

A.J. Ellis, "Using geothermal energy for power," *Power*, 123(10), October 1979.

M.M. El-Wakil, *Power Plant Technology*, New York: McGraw-Hill, 1984.

A.V. Nero, *A Guidebook to Nuclear Reactors*, Berkeley: University of California Press Ltd., 1979.

J. Weisman and L.E. Eckart, *Modern Power Plant Engineering*, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

Further Information

Other recommended publications include the "Power Plant Electrical References Series," published by EPRI, which consists of several books dealing with power plant electrical system design. A good source of information on the latest developments is *Power* magazine, which regularly publishes articles on power plants.

Additional books include the following:

S. Glasstone and M.C. Edlund, *The Elements of Nuclear Reactor Theory*, New York: Van Nostrand, 1952, p. 416.

G. Murphy, *Elements of Nuclear Engineering*, New York: Wiley, 1961.

M.A. Schultz, *Control of Nuclear Reactors and Power Plants,* New York: McGraw-Hill, 1955.
R.H. Shannon, *Handbook of Coal-Based Electric Power Generation,* Park Ridge, N.J.: Noyes, 1982, p. 372.
E.J.G. Singer, *Combustion: Fossil Power Systems,* Windsor, Conn.: Combustion Engineering, Inc., 1981.
B.G.A. Skrotzki and W.A. Vopat, *Power Station Engineering and Economy,* New York: McGraw-Hill, 1960.
M.J. Steinberg and T.H. Smith, *Economy Loading of Power Plants and Electric Systems,* New York: Wiley, 1943, p. 203.
Various, *Electric Generation: Steam Stations,* B.G.A. Skrotzki, Ed., New York: McGraw-Hill, 1970, p. 403.
Various, *Steam,* New York: Babcock & Wilcox, 1972.
Various, *Steam: Its Generation and Use,* New York: Babcock & Wilcox, 1978.

Ramakumar, R., Barnett, A.M., Kazmerski, L.L., Benner, J.P., Coutts, T.J. “Power Systems and Generation”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

60

Power Systems and Generation

R. Ramakumar
Oklahoma State University

Allen M. Barnett
AstroPower, Inc.

Lawrence L. Kazmerski
National Renewable Energy Laboratory

John P. Benner
National Renewable Energy Laboratory

Timothy J. Coutts
National Renewable Energy Laboratory

60.1 Distributed Power Generation
Photovoltaics • Wind-Electric Conversion • Hydro • Geothermal • Tidal Energy • Fuel Cells • Solar-Thermal-Electric Conversion • Biomass Energy • Thermoelectrics • Thermionics • Integrated System Concepts • System Impacts

60.2 Photovoltaic Solar Cells
Solar Cell Operation and Characteristics • Solar Cell Types and Their Optimization • Crystalline Silicon • III-V Semiconductors • Thin-Film Solar Cells • Dye-Sensitized Cells • Module Technologies • Photovoltaic Power Systems

60.3 Thermophotovoltaics
Background • Design Considerations of a TPV System • Optical Control of Sub-bandgap Energies • Development of PV Cells • Status of System Development • Systems and Applications

60.1 Distributed Power Generation

Distributed generation (DG) refers to small (a few watts up to 1 MW) power plants at or near the loads, operating in a stand-alone mode or connected to a grid at the distribution or subtransmission level, and geographically scattered throughout the service area. Typically they harness unconventional energy resources such as **insolation,** wind, **biomass,** tides and waves, and **geothermal.** Small plants powered by site-specific conventional energy resources such as low-head and small hydro and natural gas are also included in this general group.

Interest in DG has been growing steadily since the dramatic oil embargo of 1973. In addition to the obvious advantages realized by the development of renewable energy sources, DG is ideally suited to power small remote loads, located far from the grid. An entire family of small power sources has been developed and employed for space, underwater, and biomedical applications. Another niche for these systems is in energizing remote rural areas of developing countries. It is estimated that there are more than one million remote villages in the world with no grid connection and minimally sustained by locally available energy sources. Integrated renewable energy systems (**IRES**), a special subset of DG, are ideally suited for these situations.

General Features

DG will have one or more of the following features:

- Small size
- Intermittent input resource
- Stand-alone or interface at the distribution or subtransmission level
- Extremely site-specific inputs
- Located near the loads

- Remoteness from conventional grid supply
- Availability of energy storage and reconversion for later use

Potential and Future

Globally, the potential for DG is vast. Even extremely site-specific resources such as tides, geothermal, and small hydro are available in significant quantities. Assessments of the future for various DG technologies vary, depending on the enthusiasm of the estimator. However, in almost all cases, the limitations are economic rather than technical. Concerns over the unrestricted use of depletable energy resources and the ensuing environmental problems such as the greenhouse effect and global warming are providing the impetus necessary for the continued development of technologies for DG.

Motivation

Among the powerful motivations for the entry of DG are:

- Less capital investment and less capital at risk in the case of smaller installations
- Easier to site smaller plants under the ever-increasing restrictions
- Likely to result in improved reliability and availability
- Location near load centers decreases delivery costs and lowers transmission and distribution losses
- In terms of the cost of power *delivered,* DG is becoming competitive with large central-station plants, especially with the advent of open access and competition in the electric utility industry

DG Technologies

Many technologies have been proposed and employed for DG. Power ratings of DG systems vary from milliwatts to megawatts, depending on the application. A listing of the technologies is given below.

- **Photovoltaics** (PV)
- **Wind-electric conversion** systems
- Mini and micro hydro
- Geothermal plants
- Tidal and wave energy conversion
- **Fuel cells**
- **Solar-thermal-electric conversion**
- Biomass utilization
- **Thermoelectrics**
- **Thermionics**
- Small cogeneration plants powered by natural gas and supplying electrical and thermal energies

The technology involved in the last item above is mature and very similar to that of conventional thermal power plants and therefore will not be considered in this section.

Photovoltaics

PV refers to the direct conversion of insolation (incident solar radiation) to electricity. A PV cell (also known as a solar cell) is simply a large-area semiconductor *pn* junction diode with the junction positioned very close to the top surface. Typically, a metallic grid structure on the top and a sheet structure in the bottom collect the minority carriers crossing the junction and serve as terminals. The minority carriers are generated by the incident photons with energies greater than or equal to the energy gap of the semiconductor material.

Since the output of an individual cell is rather low (1 or 2 W at a fraction of a volt), several (30 to 60) cells are combined to form a module. Typical module ratings range from 40 to 50 W at 15 to 17 V. PV modules are progressively put together to form panels, arrays (strings or trackers), groups, segments (subfields), and ultimately a PV plant consisting of several segments. Plants rated at several MW have been built and operated successfully.

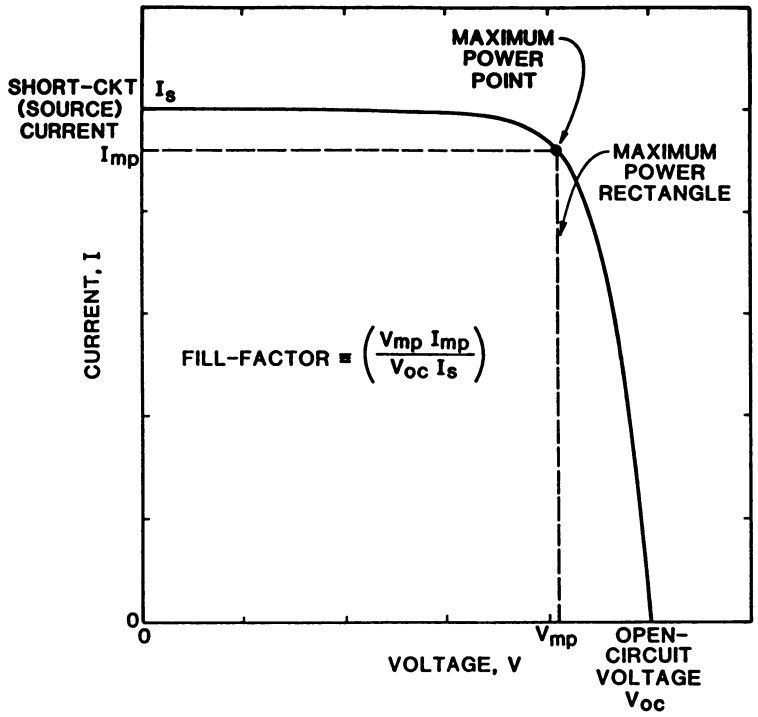


FIGURE 60.1 Typical current-voltage characteristic of an illuminated solar cell.

Advantages of PV include demonstrated low operation and maintenance costs, no moving parts, silent and simple operation, almost unlimited lifetime if properly cared for, no recurring fuel costs, modularity, and minimal environmental effects. The disadvantages are its cost, need for large collector areas due to the diluteness of insolation, and the diurnal and seasonal variability of the output.

PV systems can be flat-plate or concentrating type. While flat-plate systems utilize the global (direct and diffuse) radiation, concentrator systems harness only the direct or beam radiation. As such, concentrating systems must track (one axis or two axis) the sun. Flat-plate systems may or may not be mounted on trackers.

By 1990, efficiencies of flat-plate crystalline and thin-film cells had reached 23 and 15%, respectively. Efficiencies as high as 34% were recorded for concentrator cells. Single-crystal and amorphous PV module efficiencies of 12 and 5% were achieved by the early '90s. For an average module efficiency of 10% and an insolation of 1 kW/m^2 on a clear afternoon, 10 m^2 of collector area is required for each kW of output.

The output of a PV system is dc and inversion is required for supplying ac loads or for utility-interactive operation. While the required fuel input to a conventional power plant depends on its output, the input to a PV system is determined by external factors such as cloud cover, time of day, season of the year, geographic location, orientation, and geometry of the collector. Therefore, PV systems are operated, as far as possible, at or near their maximum outputs. Also, PV plants have inertialess generation and are subject to rapid changes in their outputs due to moving clouds.

The current-voltage (*IV*) characteristic of an illuminated solar cell is shown in Figure 60.1. It is given as

$$I = I_s - I_o\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

where I_o and I_s are the dark and source currents, respectively, k is the Boltzmann constant (1.38×10^{-23} J/K), T is the temperature in K, and e is the electronic charge. Under ideal conditions (identical cells), for a PV module with a series-parallel arrangement of cells, the *IV* characteristic will be similar, except that the current scale should be multiplied by the number of parallel branches and the voltage scale by the number of cells in series in the module. The source current varies linearly with insolation. The dark current increases as the cell

operating temperature increases. Also, the larger the energy gap of the material, the smaller the dark current. The ratio of source current to dark current should be made as large as possible for improved operation.

Single-crystal silicon is still the dominant technology for fabricating PV devices. Polycrystalline, semicrystalline, and amorphous silicon technologies are developing rapidly to challenge this. Highly innovative technologies such as spheral cells are being introduced to reduce costs. Concentrator systems typically employ gallium arsenide or multiple junction cells. Many other materials and thin-film technologies are under investigation as potential candidates.

PV applications range from milliwatts (consumer electronics) to megawatts (central station plants). They are suitable for portable, remote, stand-alone, and utility-interactive applications. PV systems should be considered as energy sources and their design should maximize the conversion of insolation into useable electrical form. Power requirements of practical loads are met using an energy storage and reconversion system or utility interconnection. Concentrating systems have been designed and operated to provide both electrical and low-grade thermal outputs with combined peak utilization efficiencies approaching 60%.

The vigorous growth of PV technology is manifested by a doubling of world PV module shipments in six years — from 42 MW in 1989 to 84 MW in 1995. Tens of thousands of small (<1 kW) systems are in operation around the world. Thousands of kilowatt-size systems (1 to 10s of kW) also have been installed and are in operation. Many intermediate-scale systems (10 to 100s of kW) and large-scale systems (1 MW or larger) are being installed by utility- and government-sponsored programs as proof-of-concept experiments and to glean valuable operational data.

By 1988, nearly 11 MW of PV was interconnected to the utility system in the United States alone. Most were in the 1- to 5-kW range. The two major exceptions are the 1-MW Hesperia-Lugo project installed in 1982 and the 6.5-MW Carrisa Plains project installed in 1984, both in California. In Germany, a 340-kW system began operation in 1988 as part of a large program. Switzerland had a plan to install 1 MW of PV in 333 roof-mounted units of 3 kW each. By 1990, the installed capacity of PV in Italy exceeded 3 MW. Many nations have recognized the vast potential of PV and have established their own PV programs within the past decade. A view of the 300 kW flat-plate grid-connected PV system installed and operated by the city of Austin electric utility department in Austin, Texas is shown in Figure 60.2.

From a capital cost of $7000/kW in 1988 with an associated levelized energy cost of 32¢/kWh, even with a business-as-usual scenario, a twofold reduction to $3500/kW by 2000 and an additional 3-to-1 reduction to $1175/kW by 2030 are being projected. The corresponding energy costs are 15 and 5¢/kWh, respectively. These

FIGURE 60.2 A view of the city of Austin PV-300 flat-plate grid-connected photovoltaic system. (Courtesy of the city of Austin electric utility department.)

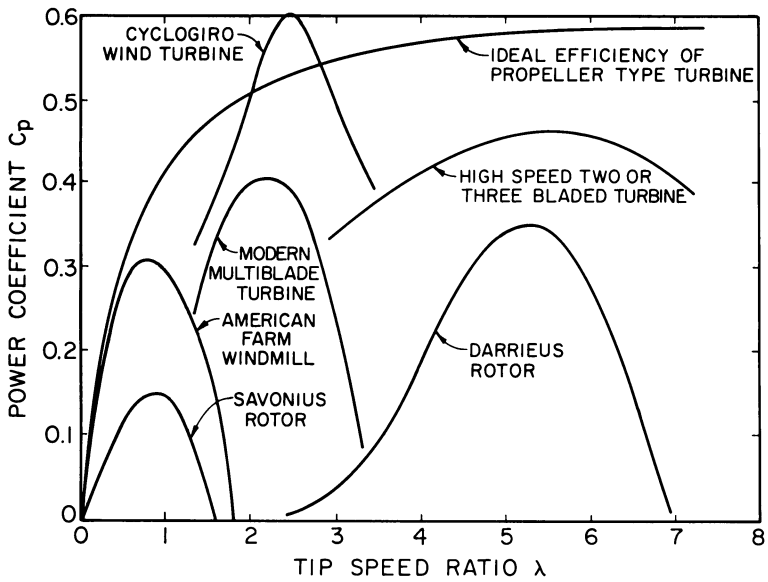


FIGURE 60.3 Typical aeroturbine characteristics.

estimates put the cost of energy from PV in par with the cost of energy from conventional plants in the early part of the twenty-first century.

Wind-Electric Conversion

Wind energy is intermittent, highly variable, and site-specific, exists in three dimensions, and is the least dependent upon latitude among all renewable resources. The power density (in W/unit area) in moving air (wind) is a cubic function of wind speed and therefore even small increases in average wind speeds can lead to significant increases in the capturable energy. Wind sites are typically classified as good, excellent, or outstanding, with associated mean wind speeds of 13, 16, and 19 mph, respectively.

Aeroturbines employ lift and/or drag forces to convert wind energy to rotary mechanical energy, which is then converted to electrical energy by coupling a suitable generator. The power coefficient C_p of an aeroturbine is the fraction of the incident power converted to mechanical shaft power, and it is a function of the tip speed-to-wind speed ratio λ as shown in Figure 60.3. For a given propeller configuration, at any given wind speed, there is an optimum tip speed that maximizes C_p.

Several types of aeroturbines are available. They can have horizontal or vertical axes, number of blades ranging from one to several, mounted upwind or downwind, and fixed- or variable-pitch blades with full blade control or tip control. Vertical-axis (Darrieus) turbines are not self-starting and require a starting mechanism. Today, horizontal-axis turbines with two or more blades are the most prevalent, and considerable work is underway to develop advanced versions of these.

The electrical output P_e of a wind-electric conversion system (WECS) is given as

$$P_e = \eta_g \eta_m A C_p K v^3$$

where η_g and η_m are the efficiencies of the electrical generator and mechanical interface, respectively, A is the swept area, K is a constant, and v is the wind speed incident on the aeroturbine.

There are two basic options for wind-electric conversion. With varying wind speeds, the aeroturbine can be operated at a constant speed by blade-pitch control, and a conventional synchronous machine is then employed to generate constant-frequency ac. More commonly, an induction generator is used with or without an adjustable var supply. In this case, the aeroturbine will operate at a nearly constant speed. Alternatively, the aeroturbine rotational speed can be allowed to vary with wind to maintain a constant and optimum tip speed ratio, and then a combination of special energy converters and power electronics is employed to obtain utility-grade ac.

The variable-speed option allows optimum efficiency operation of the turbine over a wide range of wind speeds, resulting in increased outputs with lower structural loads and stresses. All future utility-grade advanced turbines are expected to operate in the variable-speed mode and use power electronics to convert the variable-frequency output to constant frequency with minimal harmonic distortion.

Large-scale harnessing of wind energy will require hundreds or even thousands of WECS arranged in a wind farm with spacings of about 2 to 3 diameters crosswind and about 10 diameters apart downwind. The power output of an individual WECS will fluctuate over a wide range, and its statistics strongly depend on the wind statistics. When many WECS are used in a wind farm, some smoothing of the total power output will result, depending on the statistical independence of the outputs of individual WECS. This is desirable, especially with high (>20%) penetration of WECS in the generation mix. While the output of WECS is not dispatchable, with large wind farms the possibility of assigning some capacity credit to the overall output significantly improves.

Although wind-electric conversion has overall minimum environmental impacts, the large rotating structures involved do generate some noise and introduce visual aesthetics problems. By locating wind energy systems sufficiently far from centers of population, these effects can be minimized. The envisaged potential for bird kills turned out to be not a serious problem. Wind energy systems occupy only a very small fraction of the land. However, the area surrounding them can be used only for activities such as farming and livestock grazing. Thus, there is some negative impact on land use.

Today, the cost of energy delivered by wind plants rivals those obtained from some nonrenewable sources. By 1990, wind became the most utilized and competitive option among all the solar energy technologies for the bulk power market at a cost of generation of about 8¢/kWh (or roughly 7¢/kWh in 1987 dollars). Ongoing research and development work in new design tools, advanced airfoils, site tailoring, operating strategies, array spacing, and improved reliability and manufacturability is expected to bring the cost of energy further down by a factor of 2 to 3.

At around 1600 MW, nearly 90% of all the WECS installed in the world are in California. They are expected to generate nearly 3 billion kWh of electricity per year to the state's utilities to which they are interconnected. Although their lack of control and the intermittent nature of wind-derived energy are not embraced enthusiastically by electric utilities, this gap is expected to be bridged very soon with appropriate computer controls and operating strategies. Wind energy is already an economical option for remote areas endowed with good wind regimes. The modularity of WECS, coupled with the associated environmental benefits, potential for providing jobs, and economic viability point to a major role for wind energy in the generation mix of the world in the decades to come.

Hydro

Hydropower is a mature but neglected and one of the most promising renewable energy technologies. In the context of DG, small (less than 15 MW), mini (less than 1 MW), and micro (less than 100 kW) hydroelectric plants are of interest. The source of hydropower is the hydrologic cycle driven by the energy from the sun. Most of the sites for DG hydro are either low-head (2 to 20 m) or medium-head (20 to 150 m). The global hydroelectric potential is vast. One estimate puts it at 31 GW for Indonesia alone! The installed capacity of small hydro in the People's Republic of China was exceeding 7 GW by 1980.

Both impulse and reaction turbines have been employed for small-scale hydro for DG. Several standardized units are available in the market. Most of the units are operated at constant speed with governor control and are coupled to synchronous machines to generate utility-grade ac. If the water source is highly variable, it may be necessary to employ variable-speed operation. If the speed variations are not large, induction generators can be used. Special variable-speed constant-frequency (VSCF) generation schemes may be needed if the range of speed variations is large ($> \pm 10\%$). Permanent magnet generators provide another alternative, especially if the output is to be rectified and stored for later use in the case of very small units.

Geothermal

Geothermal plants exploit the heat stored in the form of hot water and steam in the earth's crust at depths of 2000 to 8000 ft. By nature, these resources are extremely site-specific and slowly run down (depletable) over a period of years. For electric power generation, the resource should be at least around 250°C. Depending on the temperature and makeup, dry steam, flash steam, or binary technology can be employed. Of these, dry natural steam is the best since it eliminates the need for a boiler.

The three basic components of a geothermal plant are (1) a production well to bring the resource to the surface, (2) a turbine generator system for energy conversion, and (3) an injection well to recycle the spent geothermal fluids back into the reservoir.

Worldwide deployment of geothermal plants reached 5000 MW by 1987 in 17 countries. Nearly one-half of this was in the United States. The Geysers plant north of San Francisco is the largest in the world with an installed capacity of 516 MW. In some developing countries, the Philippines for example, geothermal plants supply nearly 20% of their electrical needs.

Tidal Energy

The origin of **tidal energy** is the upward-acting gravitational force of the moon, which results in a cyclic variation in the potential energy of water at a point on the earth's surface. These variations are amplified by topographical features such as the shape and size of estuaries. The ratio between maximum spring tide and minimum at neap can be as much as 3 to 1. In estuaries, the tidal range can be as large as 10 to 15 m.

Power can be generated from a tidal estuary in two basic ways. A single basin can be used with a barrage at a strategic point along the estuary. By installing turbines at this point, electricity can be generated both when the tide is ebbing or flooding. In the two-basin scheme, generation can be time-shifted to coincide with hours of peak demand by using the basins alternately.

As can be expected, tidal energy conversion is very site-specific. The largest tidal power plant is the single-basin scheme at La Rance in Brittany, France. It is rated at 240 MW and employs 24 vane-type horizontal turbines and alternator motors, each rated at 10 MVA. The plant has been in operation since 1966 with good technical and economic results. It has generated, on the average, around 500 GWh of net energy per year. The Severn estuary in the southwest of England and the Bay of Fundy in the border between the United States and Canada with the highest known tidal range of 17 m have been extensively studied for tidal power generation. There are several other possible sites around the world, but the massive capital costs required have delayed their exploitation.

Fuel Cells

A fuel cell is a simple static device that converts the chemical energy in a fuel directly, isothermally, and continuously into electrical energy. Fuel and oxidant (typically oxygen in air) are fed to the device in which an electrochemical reaction takes place that oxidizes the fuel, reduces the oxidant, and releases energy. The energy released is in both electrical and thermal forms. The electrical part provides the required output. Since a fuel cell completely bypasses the thermal-to-mechanical conversion involved in a conventional power plant and since its operation is isothermal, fuel cells are not Carnot-limited. Efficiencies in the range of 43 to 55% are forecasted for modular dispersed generators featuring fuel cells.

The low (< 0.05 lb/MWh) airborne emissions of fuel cell plants make them prime candidates for siting in urban areas. The possibility of using fuel cells in combined heat and power (CHP) units provides the cleanest and most efficient energy system option utilizing valuable (or imported) natural gas resources.

Hydrocarbon fuel (natural gas or LNG) or gasified coal is reformed first to produce hydrogen-rich (and sulphur-free) gas that enters the fuel cell stack where it is electrochemically "burned" to produce electrical and thermal outputs. The electrical output of a fuel cell is low-voltage high-current dc. By utilizing a properly organized stack of cells and an inverter, utility-grade ac output is obtained.

Early MW-scale demonstration plants employed phosphoric acid fuel cells. Molten carbonate fuel cell systems have shown considerable promise in recent years with demonstrated efficiencies in the 50 to 55% range based on the higher heating value. Another competitor in the long range is the solid oxide fuel cell that can be intergrated with a coal gasifier and a steam bottoming cycle.

Solar-Thermal-Electric Conversion

The quality of thermal energy needed for DG employing solar-thermal-electric conversion necessitates concentrated sunlight. Parabolic troughs, parabolic dishes, and central receivers are used to generate temperatures in the range of 400 to 500, 800 to 900, and >500°C, respectively.

Technical feasibility of the central receiver system was demonstrated in the early '80s by the 10-MWe Solar One system in Barstow, California. Over a six-year period, this system delivered 37 GWh of net energy to the

Southern California Edison's grid with an overall system efficiency in the range of 7 to 8%. With improvements in heliostat and receiver technologies, annual system efficiencies of 14 to 15% and generation cost of 8 to 12¢/kWh have been projected.

Parabolic-dish electric-transport technology for DG was under active development at the Jet Propulsion Laboratory (JPL) in Pasadena, California, in the late '70s and early '80s. Prototype modules with Stirling engines reached a record 29% overall efficiency of conversion from insolation to electrical output. Earlier parabolic-dish designs collected and transported thermal energy to a central location for conversion to electricity. Advanced designs such as the one developed at JPL employed engine driven generators at the focal points of the dishes, and energy was collected and transported in electrical form.

By far the largest installed capacity (nearly 400 MW) of solar-thermal-electric DG employs parabolic-trough collectors and oil to transport the thermal energy to a central location for conversion to electricity via a steam-Rankine cycle. With the addition of a natural gas burner for hybrid operation, this technology, developed by LUZ under the code name SEGS (solar electric generating system), accounts for more than 90% of the world's solar electric capacity, all located in Daggett, Kramer Junction, and Harper Lake in California. Generation costs of around 8 to 9¢/kWh have been realized with SEGS. This technology uses natural gas to compensate for the temporal variations of insolation and firms up the power delivered by the system. This compensation may come during 7 to 11 P.M. in summer and during 8 A.M. to 5 P.M. in winter. SEGS will require about 5 acres/MW or can deliver 130 MW/mi^2 of land area.

Biomass Energy

Biological sources provide a wide array of materials that have been and continue to be used as energy sources. Wood, wood wastes, and residue from wood processing industries, sewage or municipal solid waste, cultivated herbaceous and other energy crops, waste from food processing industries, and animal wastes are lumped together by the term ***biomass.*** The most compelling argument for the use of biomass technologies is the inherent recycling of the carbon by photosynthesis. In addition to the obvious method of burning biomass, conversion to liquid and gaseous fuels is possible, thus expanding the application possibilities.

In the context of electric power generation, the role of biomass is expected to be for repowering old units and for use in small (20 to 50 MW) new plants. Several new high-efficiency conversion technologies are either already available or under development for the utilization of biomass. The technologies and their overall conversion efficiencies are listed below.

- FBC (fluidized-bed combustor), 36–38%
- EPS (energy performance system) combustor, 34–36%
- BIG/STIG (biomass-integrated gasifier/steam-injected gas turbine), 38–47%

Acid or enzymatic hydrolysis, gasification, and aqueous pyrolysis are some of the other technology options available for biomass utilization.

Anaerobic digestion of animal wastes is being used extensively in developing countries to produce biogas, which is utilized directly as a fuel in burners and for lighting. An 80–20 mixture of biogas and diesel has been used effectively in biogas engines to generate electricity in small quantities.

Biomass-fueled power plants are best suited in small (<100 MW) sizes for DG to serve base load and intermediate loads in the eastern United States and in many other parts of the world. This contribution is clean, renewable, and reduces CO_2 emissions. Since biomass fuels are sulphur-free, these plants can be used to offset CO_2 and SO_2 emissions from new fossil power plants. Ash from biomass plants can be recycled and used as fertilizer. A carefully planned and well-managed SRWC (short-rotation woody crop) plantation program with yields in the range of 6 to 12 dry tons/acre/year can be effectively used to mitigate greenhouse gases and contribute thousands of MW of DG to the U.S. grid by the turn of the century.

Thermoelectrics

Thermal energy can be directly converted to electrical energy by using the thermoelectric effects in materials. Semiconductors offer the best option as thermocouples since thermojunctions can be constructed using a *p*-type

and an *n*-type material to cumulate the effects around a thermoelectric circuit. Moreover, by using solid solutions of tellurides and selenides doped to result in a low density of charge carriers, relatively moderate thermal conductivities and reasonably good electrical conductivities can be achieved.

In a thermoelectric generator, the Seebeck voltage generated under a temperature difference drives a dc current through the load circuit. Even though there is no mechanical conversion, the process is still Carnot-limited since it operates over a temperature difference. In practice, several couples are assembled in a series-parallel configuration to provide dc output power at the required voltage.

Typical thermoelectric generators employ radioisotope or nuclear reactor or hydrocarbon burner as the heat source. They are custom-made for space missions as exemplified by the SNAP (systems for nuclear auxiliary power) series and the RTG (radioisotope thermoelectric generator) used by the Apollo astronauts. Maximum performance over a large temperature range is achieved by cascading stages. Each stage consists of thermocouples electrically in series and thermally in parallel. The stages themselves are thermally in series and electrically in parallel.

Tellurides and selenides are used for power generation up to 600°C. Silicon germanium alloys turn out better performance above this up to 1000°C. With the materials available at present, conversion efficiencies in the 5 to 10% range can be expected. Whenever small amounts of silent reliable power is needed for long periods of time, thermoelectrics offer a viable option. Space, underwater, biomedical, and remote terrestrial power such as cathodic protection of pipelines fall into this category.

Thermionics

Direct conversion of thermal energy into electrical energy can be achieved by employing the Edison effect—the release of electrons from a hot body, also known as thermionic emission. The thermal input imparts sufficient energy (≥ work function) to a few electrons in the emitter (cathode), which helps them escape. If these electrons are collected using a collector (anode) and a closed path through a load is established for them to complete the circuit back to the cathode, then electrical output is obtained. Thermionic converters are heat engines with electrons as the working fluid and, as such, are subject to Carnot limitations.

Converters filled with ionizable gases such as cesium vapor in the interelectrode space yield higher power densities due to space charge neutralization. Barrier index is a parameter that signifies the closeness to ideal performance with no space charge effects. As this index is reduced, more applications become feasible.

A typical example of developments in thermionics is the TFE (thermionic fuel element) that integrates the converter and nuclear fuel for space nuclear power in the kW to MW level for very long (7 to 10 years) duration missions. Another niche is the thermionic cogeneration burner module, a high-temperature burner equipped with thermionic converters. Electrical outputs of 50 kW/MW of thermal output have been achieved. High (600 to 650°C) heat rejection temperatures of thermionic converters are ideally suited for producing flue gas in the 500 to 550°C range for industrial processes. A long-range goal is to use thermionic converters as toppers for conventional power plants. Such concepts are not economical at present.

Integrated System Concepts

DG technologies offer many possibilities for integrated operation. Integrated systems may be stand-alone with energy storage and reconversion or include grid connection. Also, both renewable and conventional systems can be integrated to achieve the required operational characteristics. Integrated renewable energy systems (**IRES**) that harness several manifestations of solar energy to supply a variety of energy and other needs have many advantages and applications worldwide. The complementary nature of some of the resources (insolation and wind, for example) over the annual cycle can be exploited by IRES to decrease the amount of energy storage necessary and lower the overall cost of energy.

System Impacts

Response of distribution systems to high penetrations of DG is not yet fully understood. Also, the nature of the response will depend on the DG technology involved. However, there are some general areas of potential impacts common to most of the technologies: (i) voltage flicker, imbalance, regulation, etc.; (ii) power quality;

(iii) real and reactive power flow modifications; (iv) islanding; (v) synchronization during system restoration; (vi) transients; (vii) protection issues; (viii) load following capability; and (ix) dynamic interation with the rest of the system. Since there are very few systems with high penetration of DG, studies based on detailed models should be undertaken to forecast potential problems and arrive at suitable solutions.

Defining Terms

Biomass: General term used for wood, wood wastes, sewage, cultivated herbaceous and other energy crops, and animal wastes.

Distributed generation: Small power plants at or near loads and scattered throughout the service area.

Fuel cell: Device that converts the chemical energy in a fuel directly and isothermally into electrical energy.

Geothermal energy: Thermal energy in the form of hot water and steam in the earth's crust.

Hydropower: Conversion of potential energy of water into electricity using generators coupled to impulse or reaction water turbines.

Insolation: Incident solar radiation.

IRES: Acronym for integrated renewable energy system, a collection of devices that harness several manifestations of solar energy to supply a variety of energy and other needs.

Photovoltaics: Conversion of insolation into dc electricity by means of solid state *pn* junction diodes.

Solar-thermal-electric conversion: Collection of solar energy in thermal form using flat-plate or concentrating collectors and its conversion to electrical form.

Thermionics: Direct conversion of thermal energy into electrical energy by using the Edison effect (thermionic emission).

Thermoelectrics: Direct conversion of thermal energy into electrical energy using the thermoelectric effects in materials, typically semiconductors.

Tidal energy: The energy contained in the varying water level in oceans and estuaries, originated by lunar gravitational force.

Wind-electric conversion: The generation of electrical energy using electromechanical energy converters driven by aeroturbines.

Related Topic

22.1 Physical Properties

References

S.W. Angrist, *Direct Energy Conversion,* 4th ed., Boston, Mass.: Allyn and Bacon, 1982.

R. C. Dorf, *Energy, Resources, & Policy,* Reading, Mass.: Addison-Wesley, 1978.

J. J. Fritz, *Small and Mini Hydropower Systems,* New York: McGraw-Hill, 1984.

J. F. Kreider and F. Kreith (eds.), *Solar Energy Handbook,* New York: McGraw-Hill, 1981.

T. Moore, "On-site utility applications for photovoltaics," *EPRI J.,* p. 27, 1991.

R. Ramakumar and J. E. Bigger, "Photovoltaic Systems," *Proceedings of the IEEE,* vol. 81, no. 3, pp. 365–377, 1993.

R. Ramakumar, "Renewable energy sources and developing countries," *IEEE Transactions on Power Apparatus and Systems,* vol. PAS-102, no. 2, pp. 502–510, 1983.

R. Ramakumar, "Wind-electric conversion utilizing field modulated generator systems," *Solar Energy,* vol. 20, no. 1, pp. 109–117, 1978.

R. Ramakumar, I. Abouzahr, and K. Ashenayi, "A knowledge-based approach to the design of integrated renewable energy systems," *IEEE Transactions on Energy Conversion,* vol. EC-7, no. 4, pp. 648–659, 1992.

R. Ramakumar, H. J. Allison, and W. L. Hughes, "Solar energy conversion and storage systems for the future," *IEEE Transactions on Power Apparatus and Systems,* vol. PAS-94, no. 6, pp. 1926–1934, 1975.

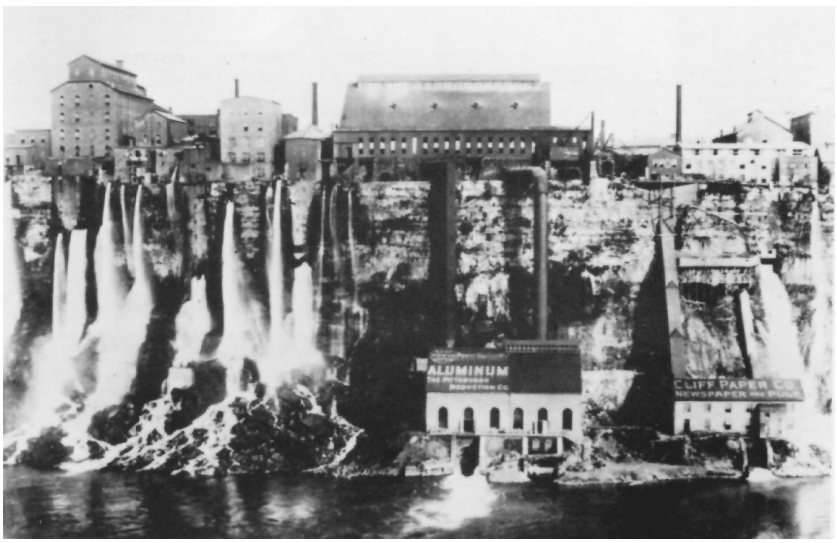
R. H. Taylor, *Alternative Energy Sources for the Centralised Generation of Electricity,* Bristol, U.K.: Hilger, 1983.

The Potential of Renewable Energy, An interlaboratory white paper, prepared for the U.S. Department of Energy, Solar Energy Research Institute, Golden, Colo., 1990.

Niagara Falls Electrical Technology Showplace

At the close of the 19th century, Niagara Falls, New York, represented a showplace for displaying how far the electrical engineering profession had come in one short decade. Here, electrical engineers were confronted with one of the great technical challenges of the age — how to harness the enormous power latent in Niagara's thundering waters and make it available for useful work.

Years of study and heated debate preceded the start-up of the first Niagara Falls Power Station in the summer of 1895, as engineers and financiers argued about whether electricity could be relied on to transmit large amounts of power the 20 miles to Buffalo and, if so, whether it should be direct or alternating current. The success of the giant polyphase alternating current generators made clear the directions that electric power technology would take in the new century, and the attraction of novel industries that consumed great amounts of electricity, such as aluminum and other electrochemical manufacturers, showed the vast potential for growth and change that electricity held for the future. (Courtesy of IEEE Center for the History of Electrical Engineering.)



The discovery of how to use electricity to make aluminum in 1886 gave Niagara Falls its first major consumer of power — the Pittsburgh Reduction Company, known today as the Aluminum Company of America (ALCOA). (Photo courtesy of IEEE Center for the History of Electrical Engineering.)

60.2 Photovoltaic Solar Cells

Allen M. Barnett and Lawrence L. Kazmerski

Photovoltaic solar cells are semiconductor diodes that are designed to absorb sunlight and convert it into electricity. The absorption of sunlight creates free minority carriers, which determine the solar cell current. These carriers are collected and separated by the junction of the diode, which determines the voltage. Photovoltaic solar cells have been the power supply of choice for satellites since 1958; 350 kilowatts of solar cells were sold for space applications in 1998. The widespread use of photovoltaic solar cells for terrestrial applications began during the oil crisis of 1973. The market for these solar cells has grown from 240 kilowatts in 1976 to 160 megawatts in 1998. Space solar cells cost approximately 100 times as much as terrestrial solar cells, so the revenue difference between the two markets is not as great as the power generation difference.

Solar Cell Operation and Characteristics

The basic operation of a **solar cell** is shown in Fig. 60.4. Photons of light are absorbed by the semiconductor material and each photon that is absorbed generates an electron-hole pair. The generated minority carriers diffuse to the junction where they are collected. The number of collected carriers determines the current. The voltage is determined by the junction characteristics. The equivalent circuit is shown in Fig. 60.5. The characteristic curve of the photovoltaic solar cell can be determined by first calculating the collected minority carriers and then separately calculating the current–voltage characteristic of the diode. Superposition can be used to combine them.

The maximum current for any solar cell is dependent on the bandgap of the absorbing semiconductor and the solar spectrum. Any photon with an energy greater than the bandgap can be expected to generate one electron-hole pair, which will lead to one collectable minority carrier. The absorption coefficient of the semiconductor material determines the thickness required to absorb the sunlight with energies greater than the bandgap. As examples, a silicon thickness of 0.5 mm will absorb 93% of the sunlight with an energy above its

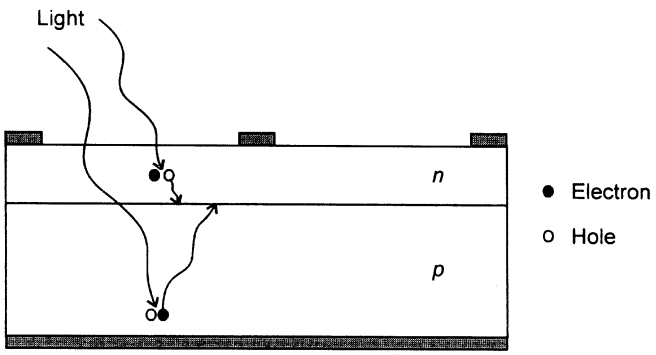


FIGURE 60.4 Operation of a solar cell.

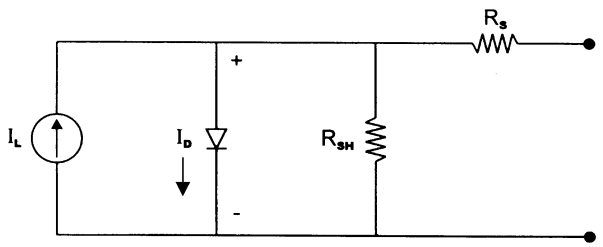


FIGURE 60.5 Equivalent circuit of a solar cell.

bandgap, while a gallium arsenide thickness of 5 μm will absorb 96% of the sunlight with an energy above it bandgap.

The photogenerated minority carriers must diffuse to the junction to be collected. These carriers will recombine as a function of the distance that they travel in accordance with the following formula $n = n_0 e^{x/L}$, where x is the distance traveled and L is the diffusion length for the minority carrier.

$$L = \sqrt{kT\mu\tau/q} = \sqrt{D\tau} \tag{60.1}$$

The other determinant of current is the spectrum of the light. The two most commonly used spectra in modeling are AM1.5 for terrestrial applications and AM0 for space applications. AM1.5 is the spectrum of the sunlight after it has passed through an equivalent thickness of 1.5 atmospheres. The integrated power of this spectrum is 970 W m^{-2}. AM0 is the spectrum of the sunlight before it passes through any of the atmosphere (0 atmospheres) and has an integrated power of 1353 W m^{-2}. AM1.5D is used to describe direct sunlight for applications that use optics to focus the sunlight onto the solar cell, these solar cells are called concentrators and the available power to the optical surface is 752 W m^{-2}.

The solar cell voltage is determined by the diode junction. The most common and most efficient junction is the *pn* junction. The diode characteristic is determined by $J = J_0(e^{qV/kT} - 1)$. J_0 is the reverse saturation current of the solar cell and is most commonly described as

$$J_0 = q\frac{D_p}{L_p}\frac{n_1^2}{N_D} + q\frac{D_n}{L_n}\frac{n_1^2}{N_A} \tag{60.2}$$

This is the case where surface recombination can be ignored and the device thickness is much greater than the minority carrier diffusion length. Unfortunately, surface recombination is an important factor in a number of semiconductor materials, including silicon. Also, the minority carrier diffusion length can be greater than the solar cell thickness. The equation for J_0 then becomes the more formidable (see [Hovel, 1976] or [Fahrenbruch and Bube, 1983]).

Superposition of the theoretical maximum generated current as a function of energy gap and the diode characteristic leads to the solar cell curve shown in Fig. 60.6. This is the diode curve offset by the light-generated current (I_L). The maximum current is called the short circuit current, I_{sc}, while the maximum voltage is called the open circuit voltage, V_{oc}. The maximum power point is shown in Fig. 60.6 as the product of V_{mp} and I_{mp} as the IV curve is traced by varying the load.

The maximum power is often described by $V_{oc} \times I_{sc} \times FF$, where FF is called the **fill factor** and is defined as $V_{mp} \times I_{mp}/V_{oc} \times I_{sc}$.

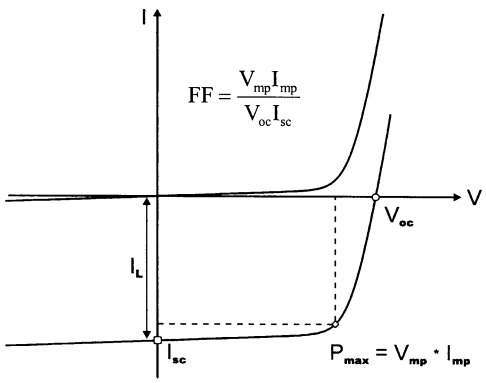


FIGURE 60.6 Current voltage characteristic of solar cell showing superposition.

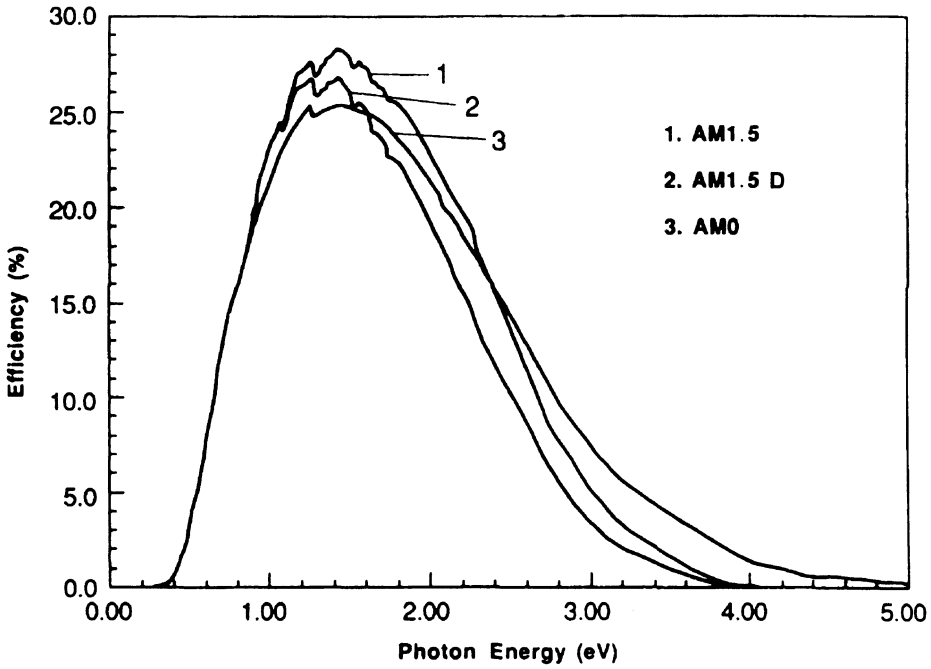


FIGURE 60.7 Single-junction solar cell efficiencies for the standard spectra.

Efficiency is defined as the maximum power divided by the power in the solar spectrum (for the solar cell area). The theoretical maximum efficiency as a function of photon energy and solar spectrum can be calculated and is shown in Fig. 60.7. The effect of energy gap is embedded in term n_i in the equation for J_0.

Solar Cell Types and Their Optimization

When one reviews the defining equation for an ideal solar cell,

$$I = I_0 \left[e^{(qV/kT)} - 1 \right] - I_L \tag{60.3}$$

it becomes apparent that there are two types of solar cell design improvements:

1. Those that improve the light-generated current, I_L
2. Those that improve the open-circuit voltage by decreasing the diode saturation current, I_0

The full impact of new solar cell designs is often hard to separate from the three parameters used to characterize the performance of any solar cell: short-circuit current, open-circuit voltage, fill factor and their product, P_{max}. However, these parameters merely summarize, and often obscure, the potential performance of any particular design. Light-generated current, together with all the various recombination processes, are the fundamental determinants of solar cell performance. A high efficiency solar cell design must maximize light-generated current and minimize losses in the bulk base and emitter, within the collecting junction, and at the surfaces. Because of these requirements, a particular device design for any material is typically optimized to obtain, first, light-generated current, then open-circuit voltage, and finally fill factor. Many examples of this approach can be found for a variety of solar cell materials, including silicon, GaAs, amorphous silicon, $CuInSe_2$, $CuInGaSe_2$, CdTe, and InP.

More recently, however, an approach that emphasizes open-circuit voltage, rather than short-circuit current, has been followed. Notable examples are present high-efficiency crystalline silicon and GaAs solar cells. This

philosophy is due, in part, to the fact that for any photovoltaic material, optimizing the light-generated current is, in general, an easier task to accomplish than optimizing the voltage-sensitive parameters. Therefore, a better measure of the potential performance of a material or design is the actual achieved open-circuit voltage, relative to the theoretical open-circuit voltage, rather than short-circuit current, which often has a value approaching the theoretically expected current.

Although the open-circuit voltage of a *pn* junction solar cell is influenced by the operating temperature, the light-generated current and the minority carrier losses**,** the minority carrier loss term visually dominates since it can change by orders of magnitude for different solar cell designs, materials, and fabrication processes. Minority carriers can recombine in each of the various regions of the solar cell — in the base, within the junction depletion region, in the emitter, and at either the front or the rear surface. For completeness, it should be mentioned that recombination at each surface consists of two parts: recombination at the ohmic contact and recombination at the free surface (if the metallization does not cover the whole surface, such as the front of the solar cell). In the limiting thick-base case, recombination at the surfaces and within the junction and emitter of the solar cell will be minimized in comparison to the base recombination term, and minority carrier losses will be controlled solely by the minority carrier lifetime in the base "bulk" region. For the thin-base device, however, base recombination will be reduced with volume, and minority carrier losses will be increasingly influenced by the surfaces and contacts.

Solar cell design innovations that significantly impact high-efficiency performance include:

1. Increased photon absorption with surface texturing and antireflection coatings to reduce top surface reflection
2. Surface passivation and low-recombination emitters that decrease both bulk and surface recombination near the top surface
3. High-low junctions and heterojunctions for reducing surface recombination, particularly at the back surface
4. Reduced-area and heterojunction contacts to reduce contact recombination losses
5. High-injection conditions that can lead to reduced J_0
6. And for polycrystalline materials, grain-boundary passivation to reduce recombination at the defects at the grain boundaries

The first design innovation contributes to improved performance primarily by increasing the light-generated current. The second innovation leads to increases in light-generated current for a direct bandgap material and to increased voltage in all materials. The next four lead to decreases in minority carrier recombination, and increased open-circuit voltage.

In addition to the *pn* junction, there are a number of other ways to make semiconductor diode junctions. These approaches, in addition to the *pn* or homojunction, are heterojunction, metal/semiconductor junction, metal/insulator/semiconductor (MIS), semiconductor/insulator/semiconductor (SIS), and electrolyte/semiconductor.

Band diagrams for these six structures are shown on Fig. 60.8.
For a detailed discussion of the physics of these junction types, see [Fonash, 1981; Green, 1982; and Hovel, 1976]. The junction is designed to maximize the voltage at any current; accordingly the objective is to minimize J_0. The *pn* junction leads to the minimum values for I_0, in part due to the lack of interface states and the ability to tune the doping. Other junction types can approximate, but usually fall short of the values achieved by the *pn* junction.

The status of the technologies — including an examination of the problems, opportunities, and issues with the various technical approaches — is summarized in this chapter. Bulk crystalline (Si, GaAs) and thin-film Cu-ternaries and multinaries, CdTe, hydrogenated amorphous silicon and dye-sensitized, $CuInSe_2$, CdTe technologies are reviewed. Applications for flat-plate and concentrator modes of operation are considered. This chapter can only present a brief overview, and the reader is directed to several other sources for more detailed evaluations and technical discussions [Kazmerski, 1997]. It can also only provide a "snapshot" of technologies that are changing with advancing R&D and manufacturing.

Following is a discussion of specific solar cells.

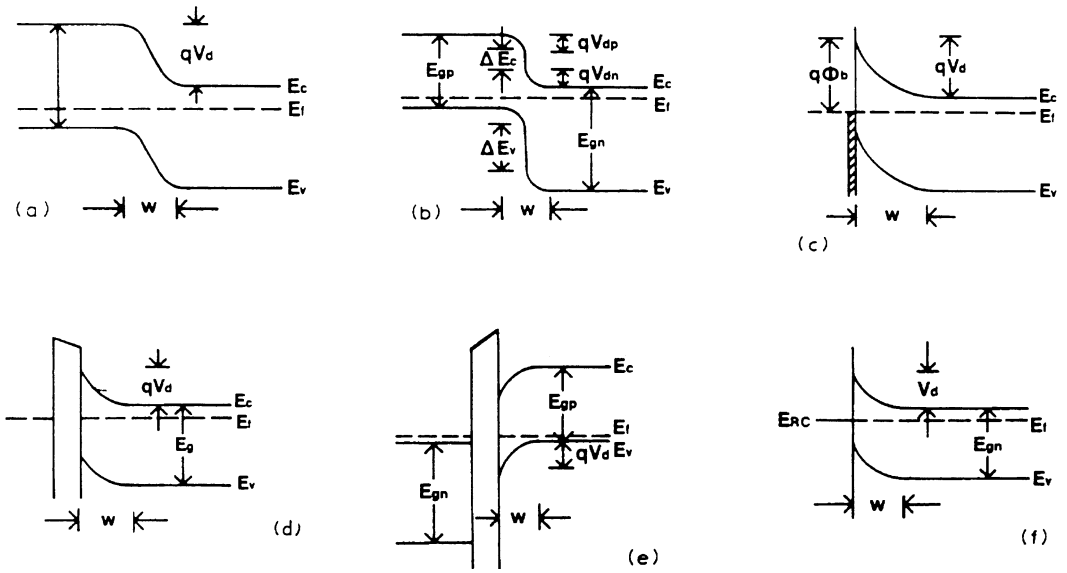


FIGURE 60.8 Band diagrams for various solar cell junctions: (a) homojunction; (b) heterojunction; (c) metal/semiconductor; (d) metal/insulator/semiconductor (MIS); (e) semiconductor/insulator/semiconductor (SIS); and, (f) electrolyte/semiconductor.

Crystalline Silicon

Silicon continues to be the foundation of the PV industry. The material is abundant, and Si solar cells have demonstrated reliability in both space and terrestrial environments [Green, 1996]. Although the properties of this semiconductor might make it non-ideal from a theoretical view, Si solar cells have reached among the highest efficiencies among the PV material options. The evolution of the high-efficiency Si solar cell is illustrated in the device cross sections shown in Fig. 60.9. The relatively simple, conventional *p-n* junction has given way to more complicated designs and structures — all aimed at capturing every incident photon, maximizing electron-hole generation, and generating maximum currents. This has called for improving antireflection approaches on the front surfaces, providing for multiple passes for the light, incorporating back surface electric field to reflect noncollected carriers, and minimizing obscuration of the incident light on the front surfaces. The evolution of these designs has included metal/insulator/*n*-type/*p*-type (MINP), passivated-emitter solar cells (PESC), single-sided and doubled-sided buried contact (SSBC and DSBC), point contact, and bifacial cells (see [Green, 1996]). Efficiencies up to 24% have been verified with monocrystalline Si cells. It should be noted that although performance enhancement is attained through these more complex designs, there is also a potentially significant increase in manufacturing cost.

A major cost factor for the Si solar cell is that associated with a high-perfection wafer. This has directed attention toward less energy-intensive process, which sacrifice the crystalline order and higher device efficiency for the benefits of lower energy production and perhaps the utilization of lower purity feedstock Si. Casting (and some sister technologies) has become conventional in the current Si manufacturing industry. Bulk and ribbon approaches for sheet-Si material have undergone extensive investigations and development over the past 20 years. Among the more developed technologies is the Edge Film-Fed Growth (EFG) process, which involves the shaping of Si through a special die and forming a connected octagon of flat sheets. The cells are cut from the connected structure by lasers, and large area solar cells (100 cm^2) with efficiencies exceeding 14% have been produced.

Treatments of the multi- and polycrystalline Si with hydrogen, lithium, aluminum, arsenic, and phosphorus have been used with varying degrees of success to minimize the effects of active defects and boundary surfaces. There is considerable effort on identifying treatment processes that benefit commercially produced polycrystalline Si products and to implement them into manufacturing. The best bulk multicrystalline Si cells have reported 18.6% efficiency.

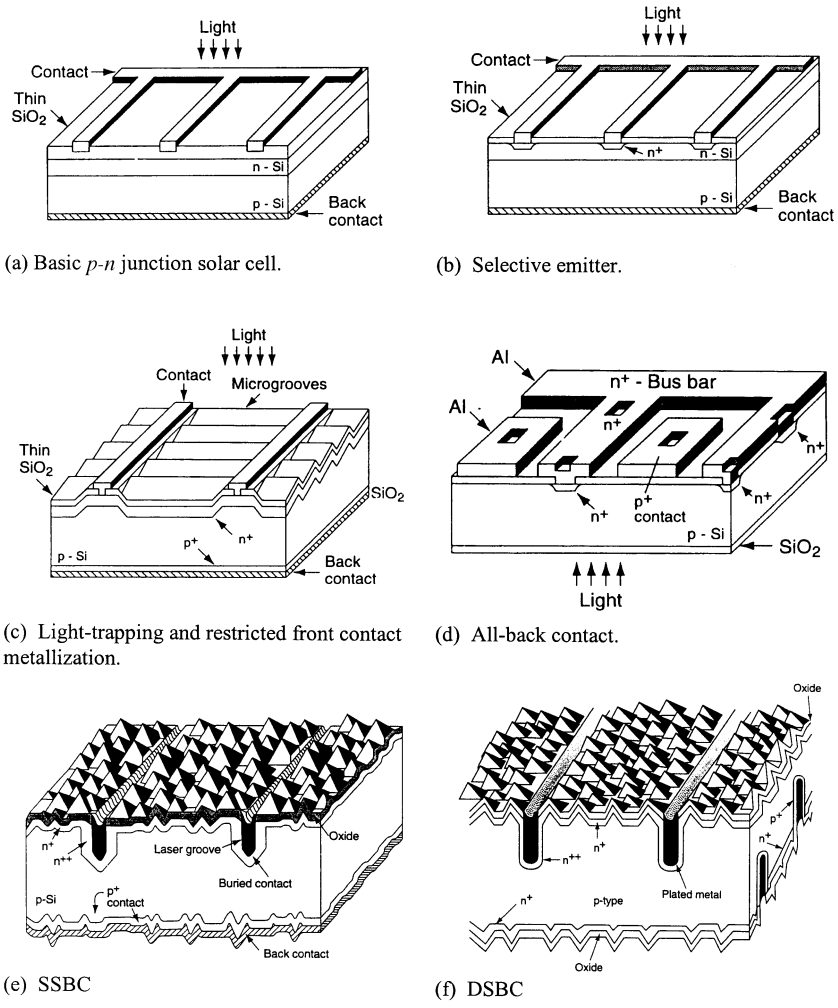


FIGURE 60.9 Evolution of silicon solar cell designs.

For better materials utilization, thin-film Si has always been of keen interest to the photovoltaics community. Early work in this area was limited to cells having efficiencies in the 5% regime, much below expectations. With the evolution of better cell engineering and understanding of Si deposition, higher efficiencies have been realized. Among these are a 20.6% (CVD of Si on Si) 47 μm thick active layer, a 47 μm thinned Si cell (CVD of Si on Si) with efficiency of 21.5%, and a PERL structure. However, the economy of these cells is not expected to meet large-scale production requirements in their current technologies. (Figure 60.10 presents some structures of advanced thin-film Si designs.) They serve as indicators that high efficiencies can be reached with thin-Si layers. The economy of films of Si requires high-performance and low-cost materials, process, and production methods. AstroPower Silicon-Film™ solar cells have reached 16.6%. This thin-Si solar cell represents a first-phase commercial design. Other more advanced designs are also illustrated in Fig. 60.10.

Si-based solar cells continue to lead the industry in performance, reliability, and availability. Major issues and concerns for Si include:

- Silicon feedstock (materials availability, solar-grade Si, competitiveness with electronic technologies)
- Manufacturing costs (yields, complexity)
- Manufacturing capacity (current output, demands, plans for increased capacity)
- Research (materials production, processing, solar cell design, thin films)

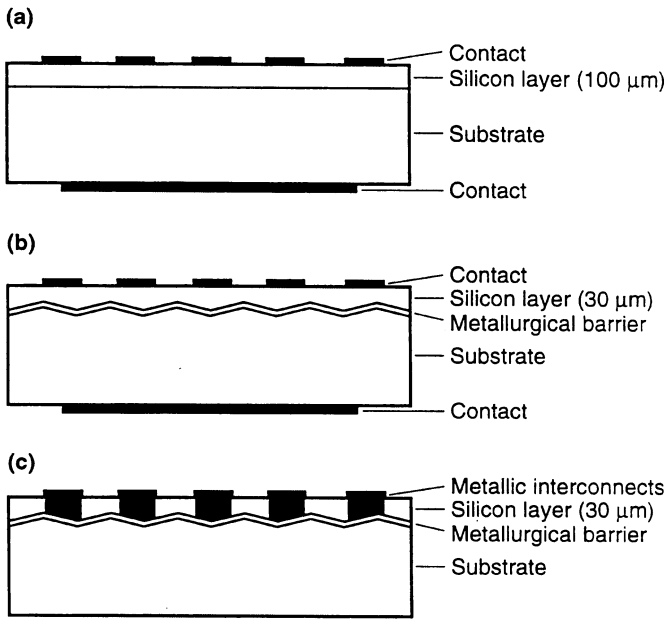


FIGURE 60.10 Silicon-Film™ products: (a) current design; (b) thinner solar cell design; and (c) thin, interconnected solar cell design.

III-V Semiconductors

Semiconductors such as GaAs, GaAlAs, GaInAsP, InSb, and InP have been receiving attention for photovoltaics because they have exceptional characteristics that offer energy conversions exceeding one third of the sun's energy. Although cost is the overriding consideration for terrestrial applications, some compelling arguments can be made for their use in concentrators. Space markets have expanded for these materials, since solar cell cost has been less a factor compared to performance. Because of the ability to adjust the electro-optical properties (e.g., bandgap), these materials lend themselves extremely well to multiple bandgap cell designs. Fig. 60.11 presents structures for two-, three-, and four-terminal approaches. Solar cells in the 30 to 34% efficiency range have been realized for these structures, and research continues in order to bring about better performances with lower complexities [Friedman, 1998]. There has been some recent attention directed toward the two-terminal, two-junction tandem — with the best efficiency of 30.3% under non-concentrator conditions. While these robust performing solar cells boast the highest efficiencies, the cost demands for terrestrial markets are still impeding their acceptance in competition with other approaches. Dominant issues include:

- Cost (materials, manufacturing, and processing)
- Industry (primarily directed toward space applications currently)
- Research (materials, processing, cell engineering)

Emerging from the development of lower-bandgap III-V cells is the resurgent *thermophotovoltaic* (TPV) technology [Coutts and Ward, 1998]. These are photovoltaic devices designed to work with infrared or thermal sources rather than sunlight. Early considerations of this technology used semiconductors with bandgaps in the range of 0.9 to 1.1 eV (primarily Ge and Si), that are matched to black-body temperatures above 2000K, which is difficult to engineer real systems. The development of devices in the 0.5 eV range correspond to temperatures of approximately 1500K, more suitable for system realization.

The complete TPV system includes: (1) a fuel and a burner that is non-direct solar, (2) a radiator using either selective or broadband emitters, (3) a long-wave photon recovery mechanism, (4) a PV cell or converter, and (5) a waste-heat recuperation system. TPV represents a growing and intensive research area for photovoltaics.

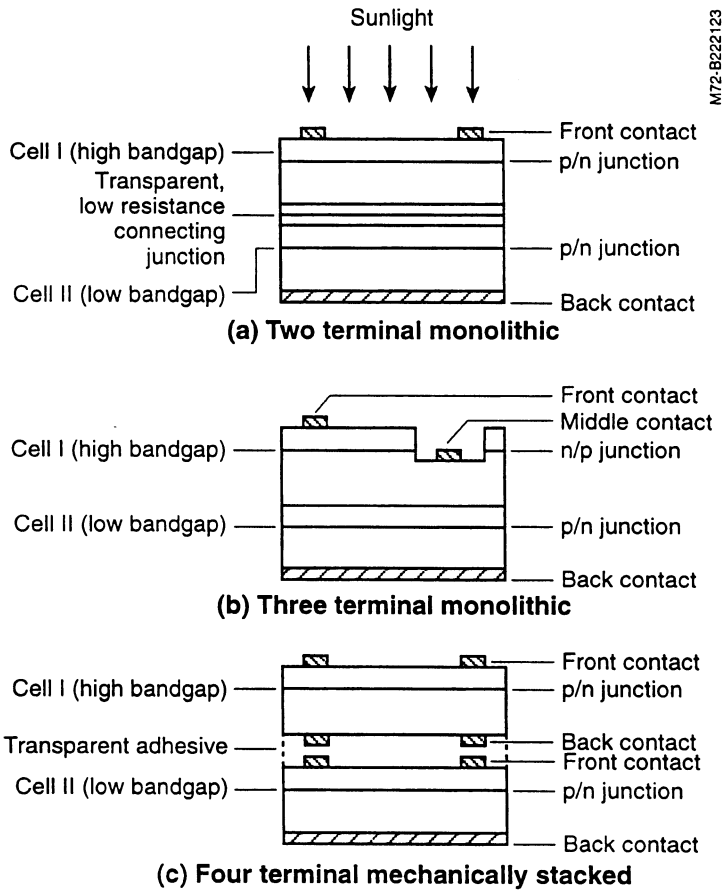


FIGURE 60.11 Multijunction structures for high-efficiency cells: (a) two terminal; (b) three terminal; and, (c) four terminal.

System efficiencies have promise of better than 40%. Certainly, an attractive system quality is that it offers "24-hour power" potential, directly from a fueled burner. Selective emitters, primarily rare-earth oxides (e.g., Yb_2O_3 for Si, Ho_2O_3 for III-V alloys) and broadband emitters, black and gray-body types (e.g., SiC-based for 1300 to 1700K), are under consideration. The TPV cells are central to the system, and include InGaAs, GaSb, and InGaAsSb. The GaSb cells have been introduced into commercial products. Thin-film semiconductor cells based on Ge and chalcopyrites are research projects. The radiation sources for terrestrial applications remain the most intriguing. These vary from biomass related (e.g., burning of wood-powder) to industrial waste heat (e.g., large furnaces used in float-glass production). In any case, TPV is in the area of next-generation technologies that could make a very large contribution to energy generation.

Thin-Film Solar Cells

The arguments for thin-film solar cells for terrestrial PV applications are primarily based on materials utilization, large-scale manufacturing potential, and better energy economy for production. The focus of this section is on the major approaches based on copper indium selenide, cadmium telluride, hydrogenated amorphous silicon (α Si:H), and dye-sensitized cells (see [Kazmerski, 1997]). Solar cells made from these materials generally have lower voltages than predicted by the energy gap because the diode technology is not an ideal *pn* junction. Additionally, there may be grain boundary losses due to additional minority carrier recombination, parallel diodes, and parasitic resistance, as shown in Fig. 60.12.

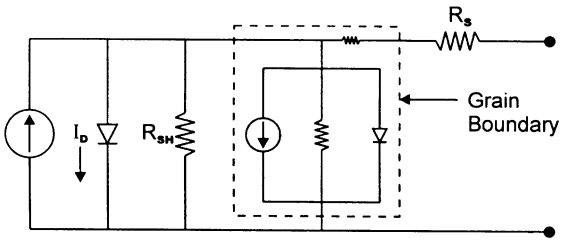


FIGURE 60.12 Equivalent circuit of a solar cell with grain boundaries.

Cu-Ternaries and Multinaries

Interest in Cu-ternary semiconductors began in the early 1970s for solar cells. $CuInSe_2$ and $CuInS_2$ (and their various alloys) have dominated R&D. Advantageous properties include: (1) suitable bandgaps for homojunctions or heterojunctions; (2) direct bandgap transitions minimizing requirements for absorber-layer thickness; (3) *n*- and *p*-type semiconductor types achievable; (4) lattice and electron affinity matches with window layer partners; (5) high optical absorption coefficients; and (6) stable electro-optical properties.

Most of the emphasis has been on modifications of CdS/Cu(In,Ga)Se_2 (or CIGS) heterojunction devices. The cross-sectional representations in Fig. 60.13 indicate the relative complexity in structure. Each of the layers, thicknesses, interfaces, and compositions are ascribed to the engineering of the cell for optimal performance and reliability. The best research cells have been demonstrated as high as a remarkable 18.8% for these true, polycrystalline thin films. Certainly, the positive and perhaps unique factors that favor this thin-film technology are stability and large-area production potential — with performance characteristics similar to those for laboratory devices. (Commercial modules with better than 12% efficiency and 4 ft^2 areas are available, using Cu(In,Ga)$(S,Se)_2$ active layers.) The variety of techniques (vacuum and non-vacuum deposition) used to make the cells speaks to the potential of low-cost manufacturing. Recently, a 15% "Cd-free" ZnO/CIGS research device has been reported. This and other Cu ternaries are undergoing research: $CuGaSe_2$ and $CuInS_2$ are prime contenders. The issues and concerns with $CuInSe_2$ and alloys include:

- Research (chemical paths to materials realization, window heteropartner, process development, minority carrier properties, contacts, role of sodium, alloy compositions)
- Complexity (manufacturing costs, control)
- Stability (increase in efficiency with light exposure)
- Device issues (low open-circuit voltage, high short-circuit current)
- Scale-up
- Manufacturing base (enhancing the current embryonic commercial base and products)

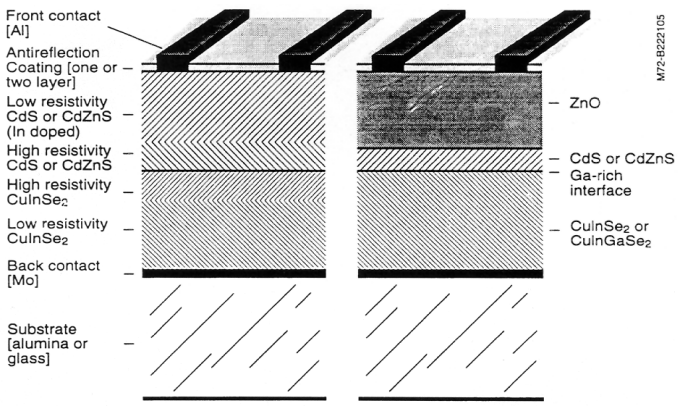


FIGURE 60.13 Device cross sections for Cu(In,Ga)Se_2 solar cell designs.

Cadmium Telluride

Since the 1960s, CdTe has been a candidate photovoltaic material. Evaporation, spraying, screen printing/sintering, and electrodeposition have been used to produce efficient solar cells. Inherent to most solar cell processing is a chemical treatment in $CdCl_2$: methanol solutions at high temperature (~400°C). The beneficial effects of this process have been attributed to enhanced grain size, identification, evolution of a *p*-i-*n* or heterojunction, surface alteration/passivation, alteration of shallow and/or deep electronic levels, improvement in morphology, and the formation of an interfacial CdSTe layer. The best solar cell efficiency has been measured at 15.8%. There is some concern that this record efficiency has not been exceeded since its report more than 6 years ago.

There is a small manufacturing base. Modules in excess of 9% have been confirmed. The devices are relatively stable, although there are concerns with humidity. Concerns for the environmental effects of Cd surround this (and other technologies). A great deal of validation (fire testing, leaching) testing has been undertaken by the industry and research laboratories. The reader is directed to the literature about this sensitive area for Cd-based cells. Issues and concern for CdTe include:

- Research (process development, modeling, interface optimization, contacts, chemical treatments, role of oxygen)
- Substrates (cost of borosilicate glass and use of sodium glasses)
- Cadmium (environmental concerns and availability)
- Stability (Cu diffusion, contact oxidation, contact degradation, humidity)
- Scale-up (cell vs. module performance levels)
- Manufacturing base

Hydrogenated-Amorphous Silicon

In contrast to more perfect crystalline materials, amorphous semiconductors have neither short- nor long-range structural order. At its introduction, a-Si:H seemed to be the ideal photovoltaic candidate. Its bandgap can be varied over tenths of eVs by changing the hydrogen content. Because its physics are considerably different from its single-crystal relative, its absorption characteristics make it about 100 times more effective in absorbing the sun's irradiance. It also has benefited technologically because there are other electronic technologies (transistors, flat-panel displays) that have enhanced knowledge and understanding of its properties.

The evolution of the a-Si:H cell is illustrated in Fig. 60.14. The development of various device structures has been an integral part of improving solar cell performance. The inherent light instabilities [Staebler and Wronski, 1977] have been minimized by engineering of the layer thicknesses, and by the use of multiple or tandem structures. The origins and cure for the light instabilities have not been completely identified, but "stabilized" cells and modules having less than 10% change in output characteristics have been produced. Many solar cells and modules with efficiencies exceeding 10% have been reported and confirmed.

Amorphous silicon-based solar cells and modules have serious problems with several stability mechanisms. The stabilized efficiencies of research solar cells are about 33% less than other thin-film options (i.e., 12% vs. 18%). Although solar cell and module engineering have minimized the Staebler-Wronski effect, the stability issue remains a major research, manufacturing, and consumer acceptance issue. Module design has requirements beyond those for crystalline Si and the ingress of any environmental entity has major influence because of the large surface-to-volume ratios involved. Stability and reliability of a device is the major issue relating to amorphous Si technology.

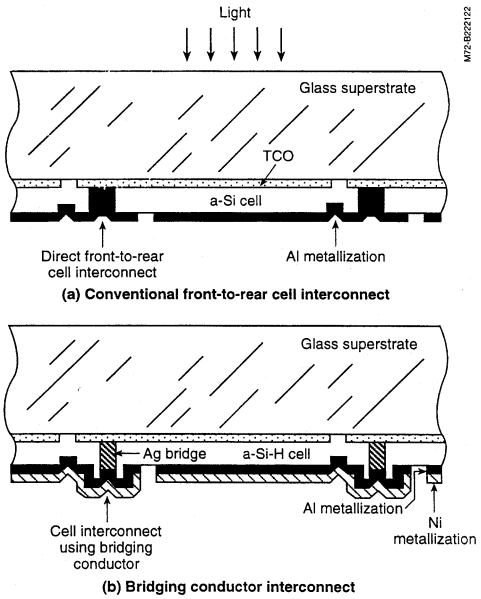


FIGURE 60.14 Amorphous Si:H module designs, showing interconnections.

These issues may require the development of new methods of producing the films, new film structures (e.g., nanocrystalline Si or multilayers), or a new a-Si technology.

Issues and concerns with the amorphous silicon technology primarily relate to the overriding stability issues. However, a variety of other concerns do accompany this technology:

- Stability and reliability
- Research (modeling, characterization, analysis, deposition research)
- Production costs (capital costs)
- Manufacturing capacity (vacuum based costs, yields, production volume)

Dye-Sensitized Cells

Among the thin-film technologies on the near horizon is the dye-sensitized solar cell. The device, illustrated in Fig. 60.15, utilizes an oxide semiconductor (TiO_2) having a bandgap in the 3-eV range that is insensitive to the solar spectrum. The extension of the photoresponse across the visible portion of the spectrum is achieved by separation of the two steps of the photovoltaic process. The oxide semiconductor in the electrochemical system is sensitized by a monolayer of an electroactive dye having an optical absorption band extending across the width of the visible spectrum. Charge separation occurs by electron loss from the photoexcited dye to the semiconductor substrate. Following absorption of a photon, the excited state of the dye is such that relaxation

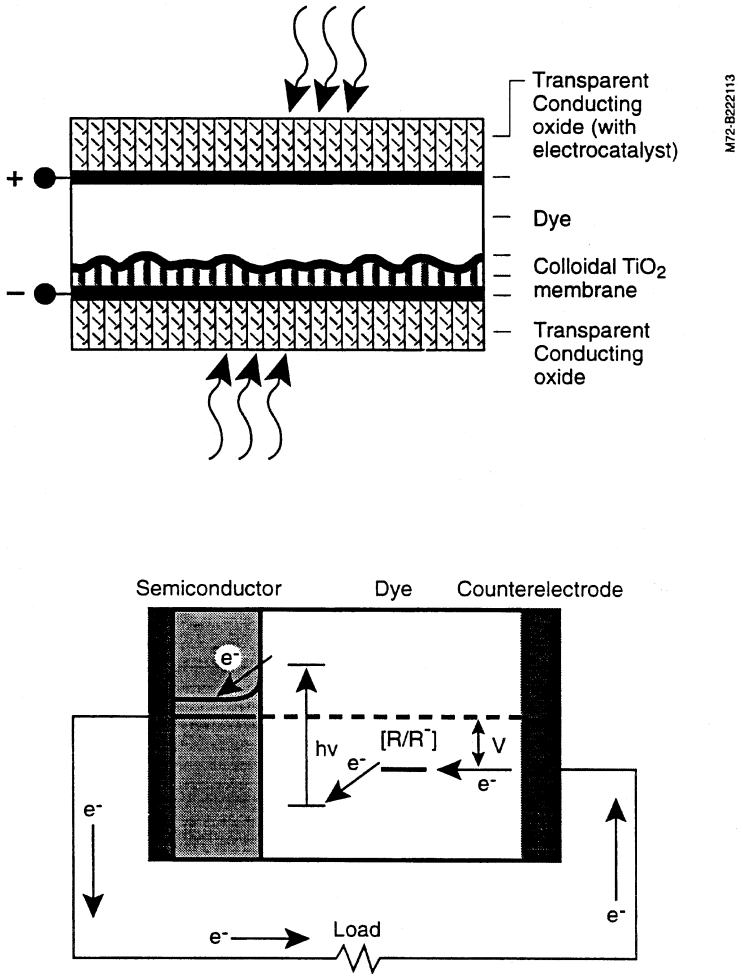


FIGURE 60.15 Device cross section for dye-sensitized solar cell.

by electron loss to the semiconductor substrate is possible, leaving the dye molecule as an oxidized species. The original ground state of the dye is restored by charge transfer reactions with the redox electrolyte. The solar cell circuit is completed by a metallized counterelectrode at which a reduction reaction with a redox system takes place. Research cells with better than 10% efficiency have been confirmed. Such cells have substantially better conversion efficiencies at lower light intensity levels. The technology is promising because it involves a low-cost processing of large areas. The solar cell continues to show growth in performance and in stability, a parameter that has been of concern to research and manufacturing groups. This technology is still in its infancy, and represents one area of thin films that is an alternative to conventional junction solar cell approaches.

Module Technologies

Solar cells are typically electrically connected into series and parallel strings to deliver a desired voltage and current and encapsulated into a supporting structure for environmental protection and strength. The module is a composite structure involving mechanical, optical, and electrical optimization that has required the collaborating and overlapping knowledge of physics, chemistry, materials science, and engineering to ensure viability [Lasnier and Ang, 1990]. The materials used for support and encapsulation depend on both the solar cell type and the application/installation. The module construction determines not only its cost, but also its lifetime. Module design has occupied a significant portion of the development efforts in photovoltaics, and the complexities and details for all photovoltaic materials are beyond the scope of this review. However, module issues are almost as important as solar cell issues because they directly impact the performance, lifetime, and cost of the photovoltaic technology [Wohlgemuth, 1995]. The module is central to meeting not only the efficiency goals (e.g., 15 to 25% for modules), but also the system cost (e.g., $1.00 to $1.50/W) and system lifetime (e.g., >30 years) in the 2010 to 2030 timeframe. (See Fig. 60.16.)

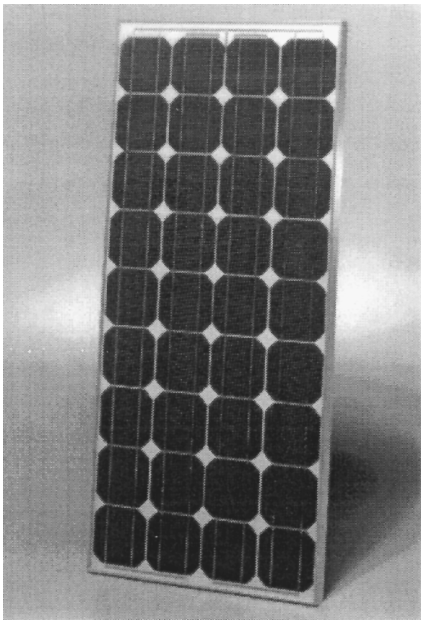


FIGURE 60.16 75 W, 36 solar cell module.

Even the most developed and commercialized of the photovoltaics approaches, crystalline Si, has required many redirections of its module construction to meet either operational or cost limitations over the past 25 years. Two examples highlight this ability and necessity for change. First, the module has been redesigned to eliminate framing to decrease materials cost, and improve loading (weight) and aerodynamics when configured into the array. In this same area, some modules, traditionally only the DC delivery system, have integrated the inverter into its construction to meet AC energy requirements. The second example has to do with the encapsulation of the Si cells in a polymer — commonly ethylene vinyl acetate (EVA). This encapsulant was identified through several development programs as a cost-effective and environment-resistant material for PV modules in the late 1970s. In the mid-1980s, the EVA in some modules was observed to turn yellow and brown. Subsequently, new and improved polymer encapsulants have also been developed to replace the original EVA formulations and UV-absorption glass has been implemented.

The evolving thin-film and concentrator technologies have additional complexities and, likely, as yet unknown problems. These advanced technologies are much different from their flat-plate silicon cousins. Consider the thin-film solar cells. Because the surface-to-volume ratios in these solar cell types are extremely high compared with bulk counterparts, materials and environmental interactions are not only enhanced, but affect relatively larger portions of the structures. These structures are also more complex, beyond the numerous interfaces that are inherent to the device itself. The cross sections for a-Si:H integrated module designs, shown in Fig. 60.14, illustrate areas of concern for shunting (bridges), contact openings, electromigration, interdiffusion, delamination, and microdefects that affect macroscale electrical behavior. Moreover, concentrators present module designs and complexities that have little relationship to their "one-sun" relatives.

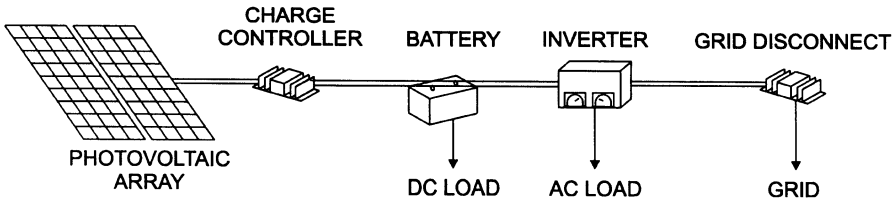


FIGURE 60.17 Grid-connected photovoltaic system.

Photovoltaic Power Systems

The photovoltaic module is the power generating component of a power system. Systems can be generically divided into stand-alone and grid-connected. A stand-alone system provides power directly to a load and usually includes storage. Stand-alone systems can vary greatly in complexity. A relatively simple system is a lighting system that includes a battery, battery charge regulator, the photovoltaic module(s), and the lights. More complex stand-alone systems include a range of battery charging systems, provide power for a whole house, provide power to a telecommunications repeater, or provide power to a satellite. All of these systems have battery and the battery charge control electronics in common. The simplest stand-alone system is a water pump that can be directly connected to the photovoltaic module. In this case, no electricity storage is required because the pumped water can be stored if it is not immediately used.

Electricity grid-connected systems generally exist in two basic forms. Systems that immediately convert the generated electricity from direct current (dc) to alternating current (ac) and synchronize this power with the electricity grid are the most fundamental for safety reasons. These systems are designed to not supply electricity to the grid when there is a grid power outage. The most complex photovoltaic systems are grid-connected systems with storage — sometimes called uninterruptible power supplies. These systems disconnect from the utility grid when there is a power outage but continue to provide electricity to the load. The load can often be a home, commercial building, or a sensitive load, such as a computer. The systems include the ability to synchronize with the grid, the ability to provide electricity to the grid when the amount generated is greater than the load, battery charging, and the conversion of dc to ac. A schematic of one of these systems is shown in Fig. 60.17. All other systems can be derived from this complex system by removing components.

Defining Terms

Photovoltaic effect: Conversion of photons to electricity.
Solar cell: Diode that converts sunlight to electricity using photovoltaic effect.
Fill factor: A measure of the relative squareness of the solar cell diode curve.
Electricity grid: Transmission and distribution system for centrally generated electricity.

References

T.J. Coutts and S. Ward, Thermophotovoltaic Solar Cells, *Scientific American,* 1998.
A.L. Fahrenbruch and R.H. Bube, *Fundamentals of Solar Cells,* Academic Press, New York, 1983.
S.J. Fonash, *Solar Cell Device Physics,* New York, Academic Press, 1981.
D. Friedman, *Proc. 2nd World Conference on Photovoltaic Solar Energy Conversion,* Vienna, Austria, IEEE Press, 1998.
M.A. Green, *Silicon Solar Cells,* University of New South Wales Press, Sydney, Australia, 1996.
M.A. Green, *Solar Cells: Operating Principles, Technology and System Applications,* Prentice-Hall, Englewood Cliffs, NJ, 1982.
H.J. Hovel, *Semiconductors and Semimetals,* Vol. 2, Academic Press, 1976.
L.L. Kazmerski, *Photovoltaics: A Review of Cell and Module Technologies, Renewable and Sustainable Energy Reviews,* chap. 1, pp. 71, 1997.

F. Lasnier and T.G. Ang, *Photovoltaic Engineering Handbook,* Adam Hilger Publishing, Bristol, UK, 1990.
D.L. Staebler and C.R. Wronski, *Applied Physics Letter,* Vol. 31, 292–294, 1977.
J. Wohlgemuth, *Proc. 24th IEEE Photovoltaic Specialists Conference,* New York, IEEE Press, pp. 901–904, 1995.

Further Information

Chopra, K.L. and S.R. Das, *Thin Film Solar Cells,* Plenum Press, New York, 1983.
R. Hill, *Applications of Photovoltaics,* Adam Hilger Publishing, Bristol, UK, 1989.
A. Luque and G.L. Araujo (Eds.), *Physical Limitations to Photovoltaic Energy Conversion,* Adam Hilger Publishing, Bristol, UK, 1990.
J.A. Mazer, *Solar Cells — An Introduction to Crystalline Photovoltaic Technology,* Kluwer Academic Publishers, 1997.
A. Ricaud, *Photopiles Solaires,* PPUR, EPFL, Lausanne, Switzerland, 1997.
R.J. Van Overstraeten and R.P. Mertens, *Physics, Technology and Use of Photovoltaics,* Adam Hilger, Bristol, UK, 1986.
S.R. Wenham, M.A. Green, and M.E. Watt, *Applied Photovoltaics,* Centre for Photovoltaics and Systems, Australia, 1994.

60.3 Thermophotovoltaics

John P. Benner and Timothy J. Coutts

Thermophotovoltaic (TPV) is the title given to a process of generating electric power by photovoltaic conversion of the energy radiated from a thermal source. The source of energy can be a hydrocarbon or hydrogen burner, concentrated solar energy, radioisotope or nuclear reactor heating the emitter to an operating temperature in the range of 1000°C to 1800°C. Two significant attributes of TPV systems, relative to photovoltaic solar energy conversion, are that the source is in close proximity to the cell and thus delivers a high power density and that the emission spectrum lies in the infrared. TPV cells share many design features with concentrator solar cells, but are generally made from semiconductors with smaller bandgaps in order to achieve peak response better matched to the spectrum of the emitter. Silicon devices, however, are still viable for systems with emitter temperatures in the upper end of the above range. The final component essential for high conversion efficiency is some mechanism for controlling the flux of sub-bandgap photons such that their energy is not lost. This can be achieved with either selective emission or optical systems to return unused energy to the emitter. Theoretically, TPV systems could achieve conversion efficiencies approaching 30%, far higher than any other candidates for energy conversion in this temperature range. In the near term, a system efficiency of 10% is a practical goal. System design must balance the conversion efficiency and power density from the PV converter. This important trade-off will be described later.

TPV systems are now commercially available operating at power levels in the range of several tens of watts to about 500 W. Market analyses indicate that TPV may be competitive for a variety of distributed generation applications of up to about 100 kW. In addition, many industrial processes operate in a range of temperatures suitable for TPV conversion. As the technology matures, these may provide the opportunity for co-generation of heat, as well as electricity. In this type of application, TPV units of several megawatt electric generating capacity would be reached to correspond with the scale of the waste heat from associated manufacturing processes.

Background

The first reference to the concept of TPV appeared in 1961 and attributes the original idea to Professor Pierre Agrain [1]. In a series of lectures at MIT in the early 1960s, Agrain assumed that radiation from the emitter that is not useful for conversion in the photocell could be returned to the emitter. Photocells were available in several materials systems, but only silicon had reached conversion efficiencies of a level for consideration in power conversion. Silicon's limited utilization of the emission spectrum from early systems was addressed in

1964 with proposed use of selective emitters, development of rare-earth oxide systems for selective emission, and with use of germanium photocells [2–4]. By the mid-1970s, system designs of up to 1 kW with projected efficiencies of 6 to 7% were brought to the stage of working prototypes [5]. Interest waned as the benefits did not appear to outweigh the challenges in thermal management and structural stability through thermal cycling to the required emitter temperatures.

More than a decade later, the advances in high-efficiency solar photovoltaics and innovations in selective emitter development prompted renewed interest in TPV systems. Solar-to-electric energy conversion efficiencies greater than 30% were achieved in multiple-junction structures. These devices exploit the range of properties that can be obtained from III–V semiconductors. The ability to tune the photocell to a desired response characteristic provided the degree of freedom necessary to address the remaining system design issues effectively.

Design Considerations of a TPV System

Figure 60.18 is a cross-sectional drawing of a prototype TPV unit under development by McDermott Technologies for use as a portable power source and battery charger [6]. One of the major advantages of TPV is the wide variety of fuels that can be selected for powering the system. This system will use diesel fuel, greatly simplifying logistics for the planned use by the Army. The combination of burner, radiator, gas flow channels, and emitter is designed to optimize the temperature uniformity over the emitter surface and also to isolate the PV cells from any of the hot combustion products. At a flame temperature of 1700K, the TPV system can be designed for very low NO_x emission. The steady burn contributes to quiet and reliable operation. Burner efficiency is improved by thermal **recuperation.** This consists of a heat exchanger that recirculates a high percentage of the heat of the combustion products to raise the temperature of the incoming combustion air and fuel. Note that the exhaust is mixed with the air flowing over the cooling fins for the PV cells, dropping the final outlet temperature to only 30 to 50°C over ambient.

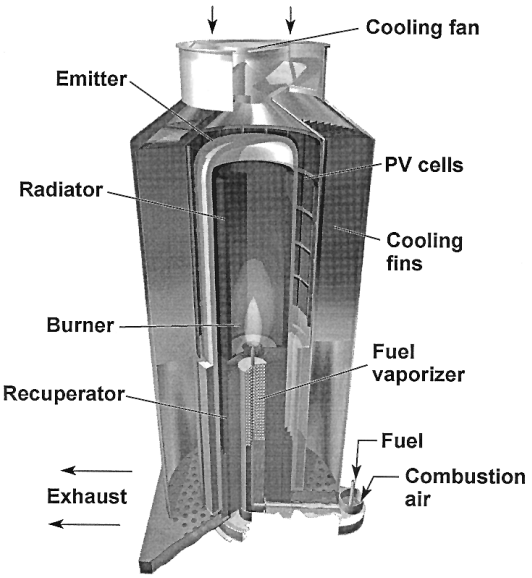


FIGURE 60.18 Schematic of the McDermott Technologies portable TPV generator. (Used with permission of the American Institute of Physics.)

The cylindrical geometry and flow from the burner, up through the radiator, then down through the channel between the radiator and the emitter, are key features in the thermal design. The energy emitted depends on emitter temperature in proportion to T^4. System performance will clearly require a high degree of uniformity of emitter surface temperature. In this particular system, the emitter produces a black-body spectrum. The front surface of the cell is coated with a selective filter that reflects long-wavelength light back to the emitter. This not only minimizes the cooling load on the PV array, but also returns useful energy to the emitter, maintaining more efficient energy conversion. The solar cells are made of GaSb. These usefully absorb energies to a maximum wavelength of about 1.8 μ.

Figure 60.19 shows the emission spectrum for several temperatures, overlaid by the cut-off energies for silicon and GaSb. For an emitter temperature of 1700K, the silicon device is able to use only about 5% of the available flux, while the device with a bandgap at about 0.7 eV can use 30%. When corrected for obscuration of the front contact grid, this portion of the available spectrum that the cell can use is called the **spectral utilization factor.** A silicon photovoltaic cell can achieve efficiencies of about 25% under illumination by the solar spectrum. The major losses are in relaxation of carriers generated by high-energy photons back to the bandgap and nonabsorbing sub-bandgap energies. The photovoltaic cell in a TPV system can be quite efficient — 40 to 50% — in converting the usable energy into electricity, since most of this energy arrives at close to the bandgap energy for a selective radiator. For this reason, the silicon device will be more efficient than the 0.7 eV device.

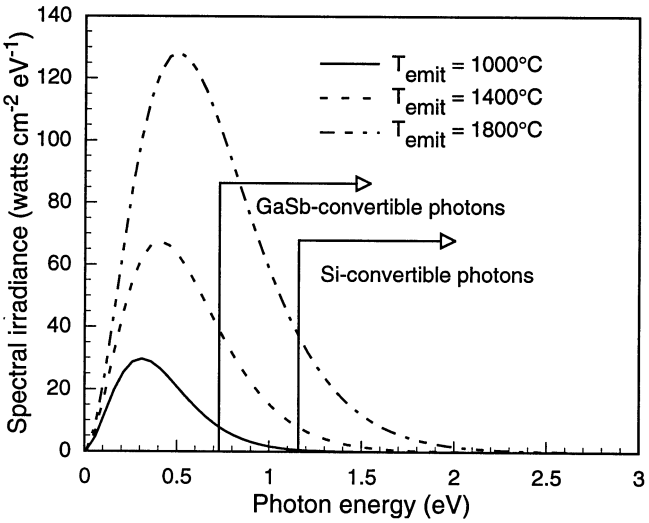


FIGURE 60.19 Black-body emission spectrum and spectral utilization for GaSb and Si photovoltaic cells.

And, as mentioned, some proportion of the sub-bandgap photons are not lost, but recirculated. However, the power density produced by the silicon-based system will be quite small. The overall system efficiency is more critically dependent on the efficiency of the spectral control components — the efficiency of recirculating the unused energy. This challenge is further compounded by the broad bandwidth needed for spectral control. Figure 60.20 plots the trade-off between conversion efficiency and power density for PV cells of various cut-off wavelengths and a black-body emitter at 1700K. The calculations assume that 100% of the sub-bandgap energy is usefully returned to the emitter. If this figure is reduced to 50%, the theoretical silicon-based system efficiency will drop to under 10%, while the 0.7 eV system efficiency will fall to 25%. The power density is, however, unaffected. The practical challenges in efficiently recirculating the sub-bandgap energies highlight the importance of the development of low-bandgap TPV cells.

For high-efficiency III-V-based PV cells, such as the GaSb device, economic considerations demand that TPV systems operate a PV cell at power densities of about 1 W cm^{-2}. The geometry of the TPV system also has a major impact on power density for the converter. For example, it would be technically desirable to evacuate the region between the emitter and cell to reduce convection losses or to place a filter in this region, but economic considerations may render these improvements impracticable. In the cylindrical geometry, if the

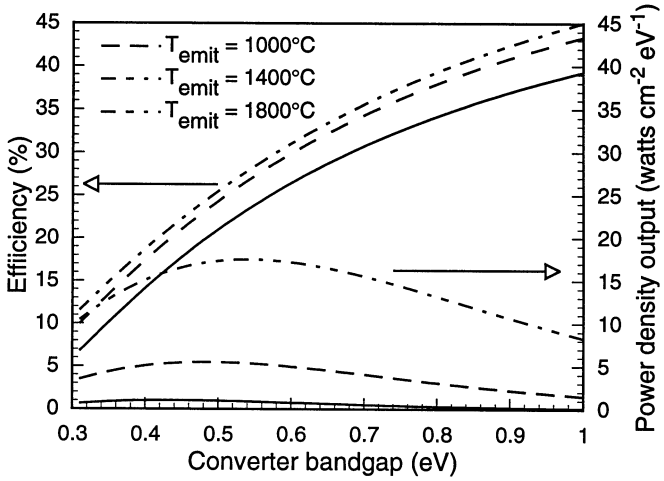


FIGURE 60.20 Calculated power density and converter efficiency for TPV operating at three different emitter temperatures.

converter array is separated far from the emitter, then the **view factor** is reduced. The view factor is the ratio of the photon flux per unit area impinging on the converter to that emanating from the emitter.

The TPV system has several potential advantages, including:

- Versatile fuel usage
- High efficiency
- High power density
- Quiet
- Clean, low NO_x emission
- No moving parts

Optical Control of Sub-bandgap Energies

Advances in selective emitters provided a major influence in rekindling interest in TPV development. Several investigators have shown the utility of rare-earth oxides such at holmia, ytterbia, erbia in modifying the emission spectrum from a broad band to one that selectively emits a substantial portion of the energy in a narrow band around a characteristic resonant frequency. Ytterbia emits with good selectivity at a peak wavelength of 980 nm — close to the bandgap of silicon (1070 nm). Given the relative maturity, efficiency, and low cost of silicon solar cells, development of an effective ytterbia emitter and surrounding system for use of silicon converters presents a pathway with potential for good performance and quick development for TPV. Unfortunately, the bulk properties of the substrate for these materials also contribute to the total radiation. Even a very low value of out-of-band emissivity integrated over a broad spectrum presents unacceptable losses. Mantles, similar to those used in camping lanterns, alleviate this loss by eliminating most of the bulk of the support structure. Nelson was the first to observe that fibrous emitters for TPV systems yield the same improvement [7]. However, it is difficult to scale the mantle structure to larger sizes and maintain acceptable mechanical performance. Several solutions for design of mechanically robust emitters are in development. One example uses selectively emitting fibers embedded in a ported ceramic block that allows the radiating fibers to work within the flame while the ceramic substrate remains relatively cool.

The alternative to selective emission is to reflect sub-bandgap energy back to the emitter. Several types of filters show potential for this task, including dielectric layers, plasma filters, metallic reflectors on the back surface of the cell, or some combination of these. An ideal filter will have 100% transmission up to the band edge of the semiconductor and 100% reflection for lower energies. Even very small absorption losses in the filter system can produce unacceptable cooling loads at typical TPV system power densities.

A new class of systems for **optical control** is under development using geometrical feature sizes on the order of the wavelength of the light. Patterning a metal film to produce a high-density array of antenna elements can achieve an inductive resonance that produces a bandpass filter. In a somewhat similar way, producing a fine periodic surface structure on the emitter material can produce wavelength-selective behavior.

Development of PV Cells

High-performance silicon solar cells are more widely available and lower in cost than other types of photovoltaic devices. For this reason, they remain in consideration for a number of TPV prototype systems. As discussed, choosing silicon for the PV converter places stringent requirements on other subsystems. TPV system design can take one of two paths to improve on Si-based converters; namely, either raise the emitter temperature (which will worsen thermal management problems) or seek PV devices with smaller bandgaps and commensurately longer wavelength response. During the early phase of TPV research, only Ge cells offered a longer wavelength response. However, the intrinsic carrier concentration of germanium is too high for this device technology ever to reach a high efficiency [8].

Development of very high efficiency tandem PV cells for use with concentrated solar power provided the other major stimulus for renewed interest in TPV. Two designs were particularly important in that they took the approach of developing low bandgap booster cells of GaSb or $Ga_xIn_{1-x}As$ for use under existing high-performance devices [9–11]. Both GaSb and $Ga_{0.47}In_{0.53}As$ have bandgaps of about 0.7 eV, corresponding to a

cut-off wavelength near 1.8 μm. This response enabled new designs and applications for TPV using lower emitter temperatures and less-demanding photon recuperation.

Ongoing research in TPV converters is advancing on two fronts. First, materials are in development with even lower bandgaps in the $Ga_xIn_{1-x}As_ySb_{1-y}$ system and in $Ga_xIn_{1-x}As$. As shown in Figure 60.20, maximum power densities should be achieved for devices at about 0.5 eV or less. The availability of devices over a narrow range of bandgaps opens the possibility of further system efficiency gains through use of cascade multiple-junction converters. A two-cell stack with bandgaps of 0.7 eV and 0.53 eV should be current-matched for a black-body emitter operating at about 1500°C. The use of such tandem cells will likely place an even greater premium on achieving exceedingly uniform temperature profile over the emitter area.

The other avenue of converter development is in the creation of monolithic interconnected modules. These devices are fabricated in semiconductor layers grown on semi-insulating substrates (or an isolation layer on a conducting substrate), using processing to isolate mesas into individual cells and then to connect the cells in series to produce a higher-voltage, lower-current circuit. This greatly simplifies the balance-of-system design and assembly because the converter size is limited only by dimensions of available substrates rather than current handling capability for the single high-flux device. The use of a semi-insulating substrate in the design may also provide the simplest method for optical and thermal control by incorporating a back surface reflector. This option is essentially precluded in other converter designs because of excessive free-carrier absorption losses in the relatively thick substrate.

All of the TPV converters described above present some challenge in reaching a cost target of about $1 per watt. In the case of silicon, the power densities are quite small. For the GaSb or GaInAs devices, the cost of the substrate (GaSb or InP) and device processing leaves little margin. Thus, as in solar photovoltaics, there is growing research interest in thin-film technologies and in low-cost substrate alternatives.

Status of System Development

Despite the elegance of the physics of the emitters, optical control, and PV converters, the engineering of components and system designs to minimize optical and thermal losses is probably the major near-term barrier. Monte Carlo analyses of photon paths have provided insights into the mechanisms of photon loss for characteristic Lambertian emission profiles. These analyses yield some surprising results. Off-normal angle of incidence and internal reflection can multiply the effective cross section of absorbing or reflecting materials. As mentioned, the emitter operates at relatively high temperatures and must maintain a highly uniform temperature profile. Considerations of **view factor** and power density require close spacing between hot and cool components. Thus, some designs include evacuation of the region between emitter and converter. Small oversights in estimating conductive losses have also broadened the gap between predicted and achieved performance.

Systems and Applications

One class of systems, used for explanation here, is intended for military use. The Army consumes large quantities of batteries for communications and a range of other applications. In these kinds of applications, systems of less than 5% conversion efficiency using diesel fuel compare favorably to batteries.

Remote communication systems also benefit from new approaches for power delivery. For example, many are currently powered by hybrid systems that may contain combinations of small diesel engines, photovoltaics, thermoelectrics, and batteries. In many remote applications, risk of environmental damage from spilled liquid fuels limits choices for power generation. TPV systems offer efficient conversion of energy from gaseous sources. This offers greatly reduced maintenance, reduced delivery cost of maintaining the fuel supply, and improved environmental acceptance.

In the commercial sector, one of the early applications might well be for recreational vehicles and boats. Quiet operation carries premium value. One of the next early opportunities may well be for self-powered appliances. A furnace can be designed to extract the electricity needed to power the blower motor by TPV conversion of radiant emission from the burner. More than a million households in the U.S. experience power outages annually, leaving their homes without electric power to drive circulation circuits on heating units. The

Gas Research Institute estimates that a unit costing about $500 added to furnaces could capture much of this market to make furnaces self-powered units. The power demands for these will be in the range of 200 to 500 W. This path leads to larger energy units that could provide electricity, heat, and hot water for a home.

The PV cell array for TPV systems are likely to represent about half of the system cost for units using crystalline silicon or III–V devices. How might the applications for TPV systems change if thin-film photovoltaics or other advances significantly drop projected system costs? One possibility may be in co-generation in industrial processes. For example, float-glass manufacturing processes 600 tons of material each day for a typical line. These factories melt feedstock in crucibles 100 ft long by 30 ft wide operating at 1500°C. Projected technology advances both in improving energy efficiency of glass manufacturing as well as in TPV create the potential for covering the top of this melt unit with an umbrella of TPV converters. A large part of the electric demand for the factory might be met by the co-generation unit. The engineering problems may be challenging, however, even if the benefits are potentially great.

Defining Terms

Recuperator: A heat exchanger that extracts energy from the combustion products to heat the incoming fuel and air. In TPV systems, optical control is sometimes called photon recuperation.

Optical control: Technology such as a selective emitter or filter used to minimize loss of unusable sub-bandgap photons by the photovoltaic converter.

View factor: Ratio of the photon flux per unit area impinging on the converter to that emanating from the emitter.

Spectral utilization factor: The fraction of the incident energy that the photovoltaic converter can use to generate electricity.

References

1. D. C. White, B. D. Wedlock, and J. Blair, Recent advances in thermal energy conversion, *15th Annual Proceedings, Power Sources Conference,* May 1961, pp. 125–132.
2. D. C. White and R. J. Schwartz, P-I-N structures for controlled spectrum photovoltaic converters, *Proceedings NATO AGARD Conference,* Cannes, France, March 1964.
3. E. Kittl, Thermophotovoltaic energy conversion, *Proceedings 20th Annual Power Sources Conference,* May 1966, pp. 178–182.
4. R. W. Beck and E. N. Sayers, Study of Germanium Devices for Use in a Thermophotovoltaic Converter, Progress Report No. 2 (Final Report) DA28-043-AMC-02543(E), General Motors Corporation, 1967.
5. E. Kittl and G. Guazzoni, Design analysis of TPV-generator system, *Proceedings 25th Power Sources Symp.,* May 1972, pp. 106–109.
6. C. L. DeBellis, M. V. Scotto, L. Fraas, J. Samaras, R. C. Watson, and S. W. Scoles, Component development for 500 watt diesel fueled portable thermophotovoltaic (TPV) power supply, *Thermophotovoltaic Generation of Electricity: Fourth NREL Conference, AIP Confrence Proceedings 460,* Woodbury NY.
7. R. E. Nelson, U.S. Patent No. 4,584,426, filed July 1984, issued April 1986.
8. J. L. Gray and A. El-Husseini, A simple parametric study of TPV system efficiency and output power density including a comparison of several TPV materials, *Thermophotovoltaic Generation of Electricity: Second NREL Conference, AIP Confrence Proceedings 358,* Woodbury NY, pp. 3–15.
9. L.M. Fraas, J. E. Avery, P. E. Gruenbaum, and V. S. Sundarum, Fundamental characterization studies of GaSb solar cells, *Conference Record of the 22nd IEEE Photovoltaic Specialists Conference,* IEEE, New York, 1991, pp. 80–89.
10. M. W. Wanlass, T. A. Gessert, G. S. Horner, K. A. Emery, and T. J. Coutts, InP/$Ga_{0.47}As_{0.53}$ monolithic, two-junction, three-terminal tandem solar cells, *Proc. 10th Space Photovoltaic Research and Technology Conference,* NASA, 1989, pp. 102–116.
11. M. W. Wanlass, J. S. Ward, K. A. Emery, T. A. Gessert, C. R. Osterwald, and T. J. Coutts, High-performance concentrator tandem cells based on IR-sensitive bottom cells, *Solar Cells,* 30, 363, 1991.

Further Information

Proceedings from the NREL Conferences 1, 2, 3, and 4 on Thermophotovoltaic Generation of Electricity; AIP Conference Proceedings Volumes 321, 358, 401, and 460, American Institute of Physics, Woodbury, NY; T. J. Coutts and M. C. Fitzgerald, Thermophotovoltaics, *Scientific American,* September, 1998, pp. 90–95.

Chen, M.S., Lai, K.C., Thallam, R.S., El-Hawary, M.E., Gross, C., Phadke, A.G., Gungor, R.B., Glover, J.D. “Transmission”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

61

Transmission

Mo-Shing Chen
University of Texas at Arlington

K.C. Lai
University of Texas at Arlington

Rao S. Thallam
Salt River Project, Phoenix

Mohamed E. El-Hawary
Technical University of Nova Scotia

Charles Gross
Auburn University

Arun G. Phadke
Virginia Polytechnic Institute and State University

R.B. Gungor
University of South Alabama

J. Duncan Glover
FaAAElectrical Corporation

61.1 Alternating Current Overhead: Line Parameters, Models, Standard Voltages, Insulators
Line Parameters • Models • Standard Voltages • Insulators

61.2 Alternating Current Underground: Line Parameters, Models, Standard Voltages, Cables
Cable Parameters • Models • Standard Voltages • Cable Standards

61.3 High-Voltage Direct-Current Transmission
Configurations of DCTransmission • Economic Comparison of AC and DC Transmission • Principles of Converter Operation • Converter Control • Developments

61.4 Compensation
Series Capacitors • Synchronous Compensators • Shunt Capacitors • Shunt Reactors • Static VAR Compensators (SVC)

61.5 Fault Analysis in Power Systems
Simplifications in the System Model • The Four Basic Fault Types • An Example Fault Study • Further Considerations

61.6 Protection
Fundamental Principles of Protection • Overcurrent Protection • Distance Protection • Pilot Protection • Computer Relaying

61.7 Transient Operation of Power Systems
Stable Operation of Power Systems

61.8 Planning
Planning Tools • Basic Planning Principles • Equipment Ratings • Planning Criteria • Value-Based Transmission Planning

61.1 Alternating Current Overhead: Line Parameters, Models, Standard Voltages, Insulators

Mo-Shing Chen

The most common element of a three-phase power system is the overhead transmission line. The interconnection of these elements forms the major part of the power system network. The basic overhead transmission lines consist of a group of phase conductors that transmit the electrical energy, the earth return, and usually one or more neutral conductors (Fig. 61.1).

Line Parameters

The transmission line parameters can be divided into two parts: series impedance and shunt admittance. Since these values are subject to installation and utilization, e.g., operation frequency and distance between cables, the manufacturers are often unable to provide these data. The most accurate values are obtained through measuring in the field, but it has been done only occasionally.

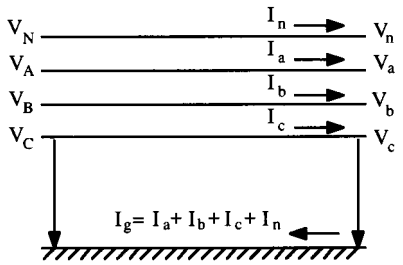


FIGURE 61.1 A three-phase transmission line with one neutral wire.

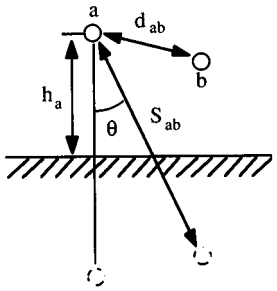


FIGURE 61.2 Geometric diagram of conductors *a* and *b*.

Though the symmetrical component method has been used to simplify many of the problems in power system analysis, the following paragraphs, which describe the formulas in the calculation of the line parameters, are much more general and are not limited to the application of symmetrical components. The sequence impedances and admittances used in the symmetrical components method can be easily calculated by a matrix transformation [Chen and Dillon, 1974]. A detailed discussion of symmetrical components can be found in Clarke [1943].

Series Impedance

The network equation of a three-phase transmission line with one neutral wire (as given in Fig. 61.1) in which only series impedances are considered is given as follows:

$$\begin{bmatrix} V_A \\ V_B \\ V_C \\ V_N \end{bmatrix} = \begin{bmatrix} Z_{aa-g} & Z_{ab-g} & Z_{ac-g} & Z_{an-g} \\ Z_{ba-g} & Z_{bb-g} & Z_{bc-g} & Z_{bn-g} \\ Z_{ca-g} & Z_{cb-g} & Z_{cc-g} & Z_{cn-g} \\ Z_{na-g} & Z_{nb-g} & Z_{nc-g} & Z_{nn-g} \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \\ I_n \end{bmatrix} + \begin{bmatrix} V_a \\ V_b \\ V_c \\ V_n \end{bmatrix} \tag{61.1}$$

where Z_{ii-g} = self-impedance of phase *i* conductor and Z_{ij-g} = mutual impedance between phase *i* conductor and phase *j* conductor.

The subscript *g* indicates a ground return. Formulas for calculating Z_{ii-g} and Z_{ij-g} were developed by J. R. Carson based on an earth of uniform conductivity and semi-infinite in extent [Carson, 1926]. For two conductors *a* and *b* with earth return, as shown in Fig. 61.2, the self- and mutual impedances in ohms per mile are

$$Z'_{aa-g} = z_a + j\omega \frac{\mu_0}{2\pi} \ln \frac{2h_a}{r_a} + \omega \frac{\mu_0}{\pi} (p + jq) \tag{61.2}$$

$$Z'_{ab-g} = j\omega \frac{\mu_0}{2\pi} \ln \frac{S_{ab}}{d_{ab}} + \omega \frac{\mu_0}{\pi} (p + jq) \tag{61.3}$$

where the "prime" is used to indicate distributed parameters in per-unit length; $z_a = r_c + jx_i$ = conductor *a* internal impedance, Ω/mi; h_a = height of conductor *a*, ft; r_a = radius of conductor *a*, ft; d_{ab} = distance between conductors *a* and *b*, ft; S_{ab} = distance from one conductor to image of other, ft; $\omega = 2\pi f$; *f* = frequency, cycles/s; μ_0 = the magnetic permeability of free space, $\mu_0 = 4\pi \times 10^{-7} \times 1609.34$ H/mi; and *p, q* are the correction terms for earth return effect and are given later.

The conductor internal impedance consists of the effective resistance and the internal reactance. The effective resistance is affected by three factors: temperature, frequency, and current density. In coping with the temperature effect on the resistance, a correction can be applied.

TABLE 61.1 Electrical Properties of Metals Used in Transmission Lines

Metal	Relative Conductivity (Copper = 100)	Electrical Resistivity at 20°C, Ω · m (10^{-8})	Temperature Coefficient of Resistance (per °C)
Copper (HC, annealed)	100	1.724	0.0039
Copper (HC, hard-drawn)	97	1.777	0.0039
Aluminum (EC grade, 1/2 H-H)	61	2.826	0.0040
Mild steel	12	13.80	0.0045
Lead	8	21.4	0.0040

$$R_{\text{new}} = R_{20^\circ}[1 + \alpha\ (T_{\text{new}} - 20)] \tag{61.4}$$

where R_{new} = resistance at new temperature, T_{new} = new temperature in °C, R_{20° = resistance at 20°C (Table 61.1), and α = temperature coefficient of resistance (Table 61.1).

An increase in frequency causes nonuniform current density. This phenomenon is called *skin effect*. Skin effect increases the effective ac resistance of a conductor and decreases its internal inductance. The internal impedance of a solid round conductor in ohms per meter considering the skin effect is calculated by

$$z = \frac{\rho m}{2\pi r}\frac{I_0(mr)}{I_1(mr)} \tag{61.5}$$

where ρ = resistivity of conductor, Ω · m; r = radius of conductor, m; I_0 = modified Bessel function of the first kind of order 0; I_1 = modified Bessel function of the first kind of order 1; and $m = \sqrt{j\omega\mu/\rho}$ = reciprocal of complex depth of penetration.

The ratios of effective ac resistance to dc resistance for commonly used conductors are given in many handbooks [such as *Electrical Transmission and Distribution Reference Book* and *Aluminum Electrical Conductor Handbook*]. A simplified formula is also given in Clarke [1943].

p and q are the correction terms for earth return effect. For perfectly conducting ground, they are zero. The determination of p and q requires the evaluation of an infinite integral. Since the series converge fast at power frequency or less, they can be calculated by the following equations:

$$p = \frac{\pi}{8} - \frac{1}{3\sqrt{2}}k\cos\theta + \frac{k^2}{16}\left[\left(0.6728 + \ln\frac{2}{k}\right)\cos 2\theta + \theta\sin 2\theta\right] + \frac{k^3\cos 3\theta}{45\sqrt{2}} - \frac{\pi k^4\cos 4\theta}{1536} \tag{61.6}$$

$$q = -0.0386 + \frac{1}{2}\ln\frac{2}{k} + \frac{1}{3\sqrt{2}}k\cos\theta - \frac{\pi k^2\cos 2\theta}{64} + \frac{k^3\cos 3\theta}{45\sqrt{2}} - \frac{k^4}{384}\left[\left(\ln\frac{2}{k} + 1.0895\right)\cos 4\theta + \theta\sin 4\theta\right] \tag{61.7}$$

with

$$k = 8.565 \times 10^{-4} D\sqrt{\frac{f}{\rho}}$$

where $D = 2h_i$ (ft), $\theta = 0$, for self-impedance; $D = S_{ij}$ (ft), for mutual impedance (see Fig. 61.2 for θ); and ρ = earth resistivity, Ω/m^3.

Shunt Admittance

The shunt admittance consists of the conductance and the capacitive susceptance. The conductance of a transmission line is usually very small and is neglected in steady-state studies. A capacitance matrix related to phase voltages and charges of a three-phase transmission line is

$$Qabc = Cabc \cdot Vabc \quad \text{or} \quad \begin{bmatrix} Q_a \\ Q_b \\ Q_c \end{bmatrix} = \begin{bmatrix} C_{aa} & -C_{ab} & -C_{ac} \\ -C_{ba} & C_{bb} & -C_{bc} \\ -C_{ca} & -C_{cb} & C_{cc} \end{bmatrix} \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} \tag{61.8}$$

The capacitance matrix can be calculated by inverting a potential coefficient matrix.

$$Qabc = Pabc^{-1} \cdot Vabc \quad \text{or} \quad Vabc = Pabc \cdot Qabc$$

or

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} P_{aa} & P_{ab} & P_{ac} \\ P_{ba} & P_{bb} & P_{bc} \\ P_{ca} & P_{cb} & P_{cc} \end{bmatrix} \begin{bmatrix} Q_a \\ Q_b \\ Q_c \end{bmatrix} \tag{61.9}$$

$$P_{ii} = \frac{l}{2\pi\varepsilon} \ln \frac{2h_i}{r_i} \tag{61.10}$$

$$P_{ij} = \frac{l}{2\pi\varepsilon} \ln \frac{S_{ij}}{d_{ij}} \tag{61.11}$$

where d_{ij} = distance between conductors i and j, h_i = height of conductor i, S_{ij} = distance from one conductor to the image of the other, r_i = radius of conductor i, ε = permittivity of the medium surrounding the conductor, and l = length of conductor.

Though most of the overhead lines are bare conductors, aerial cables may consist of cable with shielding tape or sheath. For a single-core conductor with its sheath grounded, the capacitance C_{ii} in per-unit length can be easily calculated by Eq. (61.12), and all C_{ij}'s are equal to zero.

$$C = \frac{2\pi\varepsilon_0\varepsilon_r}{\ln(r_2/r_1)} \tag{61.12}$$

where ε_0 = absolute permittivity (dielectric constant of free space), ε_r = relative permittivity of cable insulation, r_1 = outside radius of conductor core, and r_2 = inside radius of conductor sheath.

Models

In steady-state problems, three-phase transmission lines are represented by lumped-π equivalent networks, series resistances and inductances between buses are lumped in the middle, and shunt capacitances of the

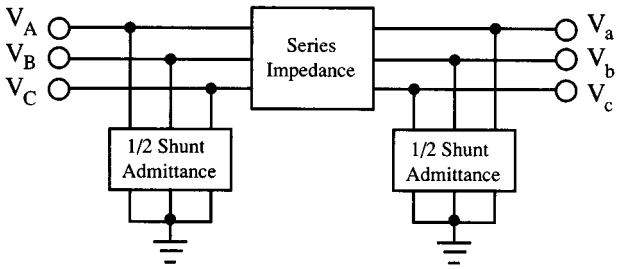


FIGURE 61.3 Generalized conductor model.

TABLE 61.2 Standard System Voltage, kV

	Rating	
Category	Nominal	Maximum
	34.5	36.5
	46	48.3
	69	72.5
	115	121
	138	145
	161	169
	230	242
Extra-high voltage (EHV)	345	362
	400 (principally in Europe)	
	500	550
	765	800
Ultra-high voltage (UHV)	1100	1200

transmission lines are divided into two halves and lumped at buses connecting the lines (Fig. 61.3). More discussion on the transmission line models can be found in El-Hawary [1995].

Standard Voltages

Standard transmission voltages are established in the United States by the American National Standards Institute (ANSI). There is no clear delineation between distribution, subtransmission, and transmission voltage levels. Table 61.2 shows the standard voltages listed in ANSI Standard C84 and C92.2, all of which are in use at present.

Insulators

The electrical operating performance of a transmission line depends primarily on the insulation. Insulators not only must have sufficient mechanical strength to support the greatest loads of ice and wind that may be reasonably expected, with an ample margin, but must be so designed to withstand severe mechanical abuse, lightning, and power arcs without mechanically failing. They must prevent a flashover for practically any power-frequency operation condition and many transient voltage conditions, under any conditions of humidity, temperature, rain, or snow, and with accumulations of dirt, salt, and other contaminants which are not periodically washed off by rains.

The majority of present insulators are made of glazed porcelain. Porcelain is a ceramic product obtained by the high-temperature vitrification of clay, finely ground feldspar, and silica. Porcelain insulators for transmission may be disks, posts, or long-rod types. Glass insulators have been used on a significant proportion of transmission lines. These are made from toughened glass and are usually clear and colorless or light green. For transmission voltages they are available only as disk types. Synthetic insulators are usually manufactured as long-rod or post types. Use of synthetic insulators on transmission lines is relatively recent, and a few questions

TABLE 61.3 Typical Line Insulation

Line Voltage, kV	No. of Standard Disks	Controlling Parameter (Typical)
115	7–9	Lightning or contamination
138	7–10	Lightning or contamination
230	11–12	Lightning or contamination
345	16–18	Lightning, switching surge, or contamination
500	24–26	Switching surge or contamination
765	30–37	Switching surge or contamination

about their use are still under study. Improvements in design and manufacture in recent years have made synthetic insulators increasingly attractive since the strength-to-weight ratio is significantly higher than that of porcelain and can result in reduced tower costs, especially on EHV and UHV transmission lines.

NEMA Publication "High Voltage Insulator Standard" and AIEE Standard 41 have been combined in ANSI Standards C29.1 through C29.9. Standard C29.1 covers all electrical and mechanical tests for all types of insulators. The standards for the various insulators covering flashover voltages (wet, dry, and impulse; radio influence; leakage distance; standard dimensions; and mechanical-strength characteristics) are addressed. These standards should be consulted when specifying or purchasing insulators.

The electrical strength of line insulation may be determined by power frequency, switching surge, or lightning performance requirements. At different line voltages, different parameters tend to dominate. Table 61.3 shows typical line insulation levels and the controlling parameter.

Defining Term

Surge impedance loading (SIL): The surge impedance of a transmission line is the characteristic impedance with resistance set to zero (resistance is assumed small compared to reactance). The power that flows in a lossless transmission line terminated in a resistive load equal to the line's surge impedance is denoted as the surge impedance loading of the line.

Related Topics

3.5 Three-Phase Circuits • 55.2 Dielectric Losses

References

Aluminum Electrical Conductor Handbook, 2nd ed., Aluminum Association, 1982.

J. R. Carson, "Wave propagation in overhead wires with ground return," *Bell System Tech. J.*, vol. 5, pp. 539–554, 1926.

M. S. Chen and W. E. Dillon, "Power system modeling," *Proc. IEEE*, vol. 93, no. 7, pp. 901–915, 1974.

E. Clarke, *Circuit Analysis of A-C Power Systems,* vols. 1 and 2, New York: Wiley, 1943.

Electrical Transmission and Distribution Reference Book, Central Station Engineers of the Westinghouse Electric Corporation, East Pittsburgh, Pa.

M. E. El-Hawary, *Electric Power Systems: Design and Analysis*, revised edition, Piscataway, N.J.: IEEE Press, 1995.

Further Information

Other recommended publications regarding EHV transmission lines include *Transmission Line Reference Book, 345 kV and Above,* 2nd ed., 1982, from Electric Power Research Institute, Palo Alto, Calif., and the IEEE Working Group on Insulator Contamination publication "Application guide for insulators in a contaminated environment," *IEEE Trans. Power Appar. Syst.,* September/October 1979.

Research on higher voltage levels has been conducted by several organizations: Electric Power Research Institute, Bonneville Power Administration, and others. The use of more than three phases for electric power transmission has been studied intensively by sponsors such as the U.S. Department of Energy.

61.2 Alternating Current Underground: Line Parameters, Models, Standard Voltages, Cables

Mo-Shing Chen and K.C. Lai

Although the capital costs of an underground power cable are usually several times those of an overhead line of equal capacity, installation of underground cable is continuously increasing for reasons of safety, security, reliability, aesthetics, or availability of right-of-way. In heavily populated urban areas, underground cable systems are mostly preferred.

Two types of cables are commonly used at the transmission voltage level: pipe-type cables and self-contained oil-filled cables. The selection depends on voltage, power requirements, length, cost, and reliability. In the United States, over 90% of underground cables are pipe-type design.

Cable Parameters

A general formulation of impedance and admittance of single-core coaxial and pipe-type cables was proposed by Prof. Akihiro Ametani of Doshisha University in Kyoto, Japan [Ametani, 1980]. The impedance and admittance of a cable system are defined in the two matrix equations

$$\frac{d(V)}{dx} = -[Z] \cdot (I) \tag{61.13}$$

$$\frac{d(I)}{dx} = -[Y] \cdot (V) \tag{61.14}$$

where (V) and (I) are vectors of the voltages and currents at a distance x along the cable and $[Z]$ and $[Y]$ are square matrices of the impedance and admittance. For a pipe-type cable, shown in Fig. 61.4, the impedance and admittance matrices can be written as Eqs. (61.15) and (61.16) by assuming:

1. The displacement currents and dielectric losses are negligible.
2. Each conducting medium of a cable has constant permeability.
3. The pipe thickness is greater than the penetration depth of the pipe wall.

$$[Z] = [Z_i] + [Z_p] \tag{61.15}$$

$$[Y] = j\omega[P]^{-1} \tag{61.16}$$

$$[P] = [P_i] + [P_p]$$

where $[P]$ is a potential coefficient matrix.

$$[Z_i] = \text{single-core cable internal impedance matrix}$$

$$= \begin{bmatrix} [Z_{i1}] & [0] & \cdots & [0] \\ [0] & [Z_{i2}] & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & [Z_{in}] \end{bmatrix} \tag{61.17}$$

$$[Z_p] = \text{pipe internal impedance matrix}$$

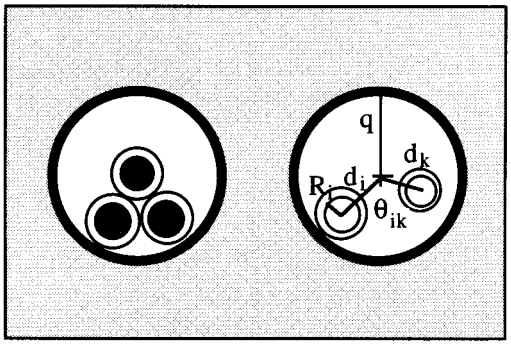


FIGURE 61.4 A pipe-type cable system.

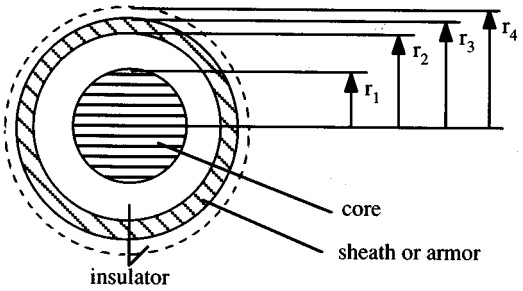


FIGURE 61.5 A single-core cable cross section.

$$= \begin{bmatrix} [Z_{p11}] & [Z_{p12}] & \cdots & [Z_{p1n}] \\ [Z_{p12}] & [Z_{p22}] & \cdots & [Z_{p2n}] \\ \vdots & \vdots & \ddots & \vdots \\ [Z_{p1n}] & [Z_{p2n}] & \cdots & [Z_{pnn}] \end{bmatrix} \tag{61.18}$$

The diagonal submatrix in $[Z_i]$ expresses the self-impedance matrix of a single-core cable. When a single-core cable consists of a core and sheath (Fig. 61.5), the self-impedance matrix is given by

$$[Z_{ij}] = \begin{bmatrix} Z_{ccj} & Z_{csj} \\ Z_{csj} & Z_{ssj} \end{bmatrix} \tag{61.19}$$

where

$$Z_{ssj} = \text{sheath self-impedance}$$

$$= Z_{\text{sheath-out}} + Z_{\text{sheath/pipe-insulation}} \tag{61.20}$$

$$Z_{csj} = \text{mutual impedance between the core and sheath}$$

$$= Z_{ssj} - Z_{\text{sheath-mutual}} \tag{61.21}$$

$$Z_{ccj} = \text{core self-impedance}$$

$$= (Z_{\text{core}} + Z_{\text{core/sheath-insulation}} + Z_{\text{sheath-in}}) + Z_{csj} - Z_{\text{sheath-mutual}} \tag{61.22}$$

where

$$Z_{\text{core}} = \frac{\rho m}{2\pi r_1} \frac{I_0(mr_1)}{I_1(mr_1)} \tag{61.23}$$

$$Z_{\text{core/sheath-insulation}} = \frac{j\omega\mu_1}{2\pi} \ln \frac{r_2}{r_1} \tag{61.24}$$

$$Z_{\text{sheath-in}} = \frac{\rho m}{2\pi r_2 D}\left[I_0(mr_2)K_1(mr_3) + K_0(mr_2)I_1(mr_3)\right] \tag{61.25}$$

$$Z_{\text{sheath-mutual}} = \frac{\rho}{2\pi r_2 r_3 D} \tag{61.26}$$

$$Z_{\text{sheath-out}} = \frac{\rho m}{2\pi r_3 D}\left[I_0(mr_3)K_1(mr_2) + K_0(mr_3)I_1(mr_2)\right] \tag{61.27}$$

$$Z_{\text{sheath/pipe-insulation}} = \frac{j\omega\mu_0}{2\pi}\cosh^{-1}\left(\frac{q^2 + R_i^2 - d_i^2}{2qR_i}\right) \tag{61.28}$$

where ρ = resistivity of conductor, $D = I_1(mr_3)K_1(mr_2) - I_1(mr_2)K_1(mr_3)$, γ = Euler's constant = 1.7811, I_i = modified Bessel function of the first kind of order i, K_i = modified Bessel function of the second kind of order i, and $m = \sqrt{j\omega\mu/\rho}$ = reciprocal of the complex depth of penetration.

A submatrix of $[Z_p]$ is given in the following form:

$$[Z_{pjk}] = \begin{bmatrix} Z_{pjk} & Z_{pjk} \\ Z_{pjk} & Z_{pjk} \end{bmatrix} \tag{61.29}$$

Z_{pjk} in Eq. (61.29) is the impedance between the jth and kth inner conductors with respect to the pipe inner surface. When $j = k$, $Z_{pjk} = Z_{\text{pipe-in}}$; otherwise Z_{pjk} is given in Eq. (61.31).

$$Z_{\text{pipe-in}} = \frac{\rho m}{2\pi q}\frac{K_0(mq)}{K_1(mq)} + \frac{j\omega\mu}{\pi}\sum_{n=1}^{\infty}\left[\left(\frac{d_i}{q}\right)^{2n}\frac{K_n(mq)}{n\mu_r K_n(mq) - mqK_n'(mq)}\right] \tag{61.30}$$

$$Z_{pjk} = \frac{j\omega\mu_0}{2\pi}\left\{\begin{aligned}&\ln\frac{q}{S_{jk}} + \frac{\mu_r}{mq}\frac{K_0(mq)}{K_1(mq)} \\ &+\sum_{n=1}^{\infty}\left(\frac{d_j d_k}{q^2}\right)^n \cos(n\theta_{jk})\left[2\mu_r\frac{K_n(mq)}{n\mu_r K_n(mq) - mqK_n'(mq)} - \frac{1}{n}\right]\end{aligned}\right\} \tag{61.31}$$

where q is the inside radius of the pipe (Fig. 61.4).

The formulation of the potential coefficient matrix of a pipe-type cable is similar to the impedance matrix.

$$[P_i] = \begin{bmatrix} [P_{i1}] & [0] & \cdots & [0] \\ [0] & [P_{i2}] & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & [P_{in}] \end{bmatrix} \tag{61.32}$$

$$[P_p] = \begin{bmatrix} [P_{p11}] & [P_{p12}] & \cdots & [P_{p1n}] \\ [P_{p12}] & [P_{p22}] & \cdots & [P_{p2n}] \\ \vdots & \vdots & \ddots & \vdots \\ [P_{p1n}] & [P_{p2n}] & \cdots & [P_{pnn}] \end{bmatrix} \tag{61.33}$$

The diagonal submatrix in $[P_i]$ expresses the potential coefficient matrix of a single-core cable. When a single-core cable consists of a core and sheath (Fig. 61.5), the submatrix is given by

$$[P_{ij}] = \begin{bmatrix} P_{cj} + P_{sj} & P_{sj} \\ P_{sj} & P_{sj} \end{bmatrix} \tag{61.34}$$

where

$$P_{sj} = \left(\frac{1}{2}\pi\varepsilon_0\varepsilon_{sj}\right) \ln\frac{r_4}{r_3} \tag{61.35}$$

$$P_{cj} = \left(\frac{1}{2}\pi\varepsilon_0\varepsilon_{cj}\right) \ln\frac{r_2}{r_1} \tag{61.36}$$

ε_0 = absolute permittivity of free space, ε_{sj} = relative permittivity of insulation outside sheath, and ε_{cj} = relative permittivity of insulation outside core.

Submatrix $[P_{pjk}]$ of $[P_p]$ is given by

$$[P_{pjk}] = \begin{bmatrix} P_{pjk} & P_{pjk} \\ P_{pjk} & P_{pjk} \end{bmatrix} \tag{61.37}$$

P_{pjk} in Eq. (61.37) is the potential coefficient between the jth and kth inner conductors with respect to the pipe inner surface. When $j = k$, $P_{pjk} = P_{\text{pipe-in}}$; otherwise P_{pjk} is given in Eq. (61.39).

$$P_{\text{pipe-in}} = \frac{\ln\left\{\dfrac{q}{R_i}\left[1 - \left(\dfrac{d_i}{q}\right)^2\right]\right\}}{2\pi\varepsilon_p\varepsilon_0} \tag{61.38}$$

$$p_{pjk} = \frac{1}{2\pi\varepsilon_p\varepsilon_0}\left[\ln\frac{q}{S_{jk}} - \sum_{n=1}^{\infty}\left(\frac{d_j d_k}{q^2}\right)^n \cdot \frac{\cos(n\theta_{jk})}{n}\right] \tag{61.39}$$

where ε_p is the relative permittivity of insulation inside the pipe; R_i is the outer radius of cable i; and d_i, d_j, and d_k are the inner radii of cables i, j, and k.

Models

Refer to "Models" in Section 61.1.

Standard Voltages

In the United States, the underground transmission cables are rated 69 to 345 kV (refer to Table 61.2 in Section 61.1). Cables rated 550 kV are used commercially in Japan. In the United States, cables installed at the 550-kV level are used in relatively short distances, for example, at the Grand Coulee Dam.

Cable Standards

The most universal standardizing authority for cables is the International Electrotechnical Commission (IEC). The IEC standards cater to a large variety of permissible options and serve mainly as a basis for the preparation of national standards. In the United States, in addition to national standards for materials and components, there are cable standards in widespread use by industry issued by four bodies: Underwriter's Laboratories (UL), Association of Edison Illuminating Companies (AEIC), and jointly by the Insulated Power Cables Engineers Association and the National Electrical Manufacturers' Association (IPCEA/NEMA).

Related Topic

55.5 Dielectric Materials

References

A. Ametani, "A general formulation of impedance and admittance of cables," *IEEE Trans. Power Syst.*, vol. PAS-99, no. 3, pp. 902–910, 1980.

P. Graneau, *Underground Power Transmission*, New York: Wiley, 1979.

D. McAllister, *Electric Cables Handbook,* New York: Granada Technical Books, 1982.

B. M. Weedy, *Underground Transmission of Electric Power*, New York: Wiley, 1980.

Further Information

The development of advanced cable systems is continuously supported by government and utilities. Information and reports regarding these activities are available from two principal funding agencies, the Electric Power Research Institute (EPRI) and the U.S. Department of Energy.

61.3 High-Voltage Direct-Current Transmission

Rao S. Thallam

The first commercial high-voltage direct-current (HVDC) power transmission system was commissioned in 1954, with an interconnection between the island of Gotland and the Swedish mainland. It was an undersea cable, 96 km long, with ratings of 100 kV and 20 MW. There are now more than 50 systems operating throughout the world, and several more are in the planning, design, and construction stages. HVDC transmission has become acceptable as an economical and reliable method of power transmission and interconnection. It offers advantages over alternating current (ac) for long-distance power transmission and as asynchronous interconnection between two ac systems and offers the ability to precisely control the power flow without inadvertent loop flows in an interconnected ac system. Table 61.4 lists the HVDC projects to date (1995), their ratings, year commissioned (or the expected year of commissioning), and other details. The largest system in operation, Itaipu HVDC transmission, consists of two ±600-kV, 3150-MW-rated bipoles, transmitting a total of 6300 MW power from the Itaipu generating station to the Ibiuna (formerly Sao Roque) converter station in southeastern Brazil over a distance of 800 km.

TABLE 61.4 HVDC Projects Data

	HVDC Supplier†	Year Commissioned	Power Rating, MW	DC Volts, kV	Line/Cable, km	Location
			Mercury Arc Valves			
Moscow-Kashira[a]	F	1951	30	±100	100	Russia
Gotland I[a]	A	1954	20	±100	96	Sweden
English Channel	A	1961	160	±100	64	England-France
Volgograd-Donbass[b]	F	1965	720	±400	470	Russia
Inter-Island	A	1965	600	±250	609	New Zealand
Konti-Skan I	A	1965	250	250	180	Denmark-Sweden
Sakuma	A	1965	300	2125	B-B[f]	Japan
Sardinia	I	1967	200	200	413	Italy
Vancouver I	A	1968	312	260	69	Canada
Pacific Intertie	JV	1970	1440	±400	1362	USA
		1982	1600			
Nelson River I[c]	I	1972	1620	±450	892	Canada
Kingsnorth	I	1975	640	±266	82	England
			Thyristor Valves			
Gotland Extension	A	1970	30	±150	96	Sweden
Eel River	C	1972	320	2 × 80	B-B	Canada
Skagerrak I	A	1976	250	250	240	Norway-Denmark
Skagerrak II	A	1977	500	±250	240	Norway-Denmark
Skagerrak III	A	1993	440	±350	240	Norway-Denmark
Vancouver II	C	1977	370	–280	77	Canada
Shin-Shinano	D	1977	300	2 × 125	B-B	Japan
		1992	600	3 × 125		
Square Butte	C	1977	500	±250	749	USA
David A. Hamil	C	1977	100	50	B-B	USA
Cahora Bassa	J	1978	1920	±533	1360	Mozambique-S. Africa
Nelson River II	J	1978	900	±250	930	Canada
		1985	1800	±500		
C-U	A	1979	1000	±400	710	USA
Hokkaido-Honshu	E	1979	150	125	168	Japan
	E	1980	300	250		
		1993	600	±250		
Acaray	G	1981	50	25.6	B-B	Paraguay
Vyborg	F	1981	355	1 × 170 (±85)	B-B	Russia (tie with Finland)
	F	1982	710	2 × 170		
			1065	3 × 170		
Duernrohr	J	1983	550	145	B-B	Austria
Gotland II	A	1983	130	150	100	Sweden
Gotland III	A	1987	260	±150	103	Sweden
Eddy County	C	1983	200	82	B-B	USA
Chateauguay	J	1984	1000	2 × 140	B-B	Canada
Oklaunion	C	1984	200	82	B-B	USA
Itaipu	A	1984	1575	±300	785	Brazil
	A	1985	2383			
	A	1986	3150	±600		
	A	1987	6300	2 × ±600		
Inga-Shaba	A	1982	560	±500	1700	Zaire
Pac Intertie Upgrade	A	1984	2000	±500	1362	USA
Blackwater	B	1985	200	57	B-B	USA
Highgate	A	1985	200	±56	B-B	USA
Madawaska	C	1985	350	140	B-B	Canada
Miles City	C	1985	200	±82	B-B	USA
Broken Hill	A	1986	40	2 × 17(±8.33)	B-B	Australia
Intermountain	A	1986	1920	±500	784	USA

Thyristor Valves (continued)

TABLE 61.4 (continued) HVDC Projects Data

	HVDC Supplier†	Year Commissioned	Power Rating, MW	DC Volts, kV	Line/Cable, km	Location
Cross-Channel						
Les Mandarins	H	1986	2000	±270	72	France
Sellindge	I	1986	2000	±270	72	England
Descantons-Comerford	C	1986	690	±450	172	Canada-USA
SACOI[d]	H	1986	200	200	415	Corsica Island
SACOI[e]		1992	300			Italy
Urguaiana Freq. Conv.	D	1987	53.7	17.9	B-B	Brazil (tie with Uruguay)
Virginia Smith (Sidney)	G	1988	200	55.5	B-B	USA
Gezhouba-Shanghai	B+G	1989	600	500	1000	China
		1990	1200	±500		
Konti-Skan II	A	1988	300	285	150	Sweden-Denmark
Vindhyachal	A	1989	500	2 × 69.7	B-B	India
Pac Intertie Expansion	B	1989	1100	±500	1362	USA
McNeill	I	1989	150	42	B-B	Canada
Fenno-Skan	A	1989	500	400	200	Finland-Sweden
Sileru-Barsoor	K	1989	100	+100	196	India
			200	+200		
			400	±200		
Rihand-Delhi	A	1991	750	+500	910	India
		1991	1500	±500		
Hydro Quebec-New Eng.	A	1990	2000[g]	±450	1500	Canada-USA
Welch-Monticello		1995	300		B-B	USA
		1998	600			
Etzenricht		1993	600	160	B-B	Germany (tie with Czech)
Vienna South-East	G	1993	600	160	B-B	Austria (tie with Hungary)
DC Hybrid Link	AB	1993	992	+270/–350	617	New Zealand
Chandrapur-Padghe		1997	1500	±500	900	India
Chandrapur-Ramagundam		1996	1000	2 × 205	B-B	India
Leyte-Luzun		1997	1000	350	440	Philippines
Haenam-Cheju I		1997	300	±180	100	South Korea
Baltic Cable Project		1994	600	450	250	Sweden-Germany
Victoria-Tasmania		1995	300	300		Australia
Kontek HVDC Intercon		1995	600	600		Denmark
Scotland-N. Ireland		1998	250	150	60	United Kingdom
Greece-Italy		1998	500			Italy
Tiang-Guang		1998	1800	500	903	China
Visakhapatnam	I	1998	500	205	B-B	India
Thailand-Malaysia		1998	300	300	110	Malaysia-Thailand
Rivera		1998	70		B-B	Urguay

†A–ASEA; H–CGEE Alsthom;
B–Brown Boveri; I–GEC (formerly Eng. Elec.);
C–General Electric; J–HVDC W.G. (AEG, BBC, Siemens);
D–Toshiba; K–(Independent);
E–Hitachi; AB–ABB Brown Boveri;
F–Russian; JV–Joint Venture (GE and ASEA).
G–Siemens;

[a] Retired from service.
[b] 2 valve groups replaced with thyristors in 1977.
[c] 2 valve groups in Pole 1 replaced with thyristors by GEC in 1991.
[d] 50-MW thyristor tap.
[e] Uprate with thyristor valves.
[f] Back-to-back HVDC system.
[g] Multiterminal system. Largest terminal is rated 2250 MW.

Source: Data compiled by D. J. Melvold, Los Angeles Department of Water and Power.

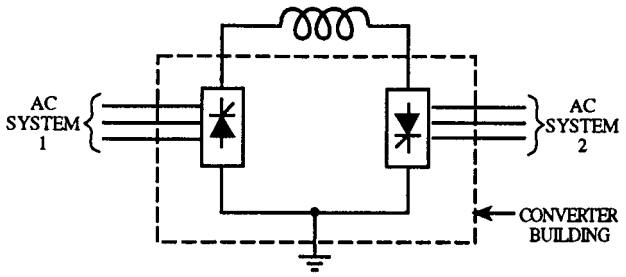


FIGURE 61.6 Back-to-back dc system.

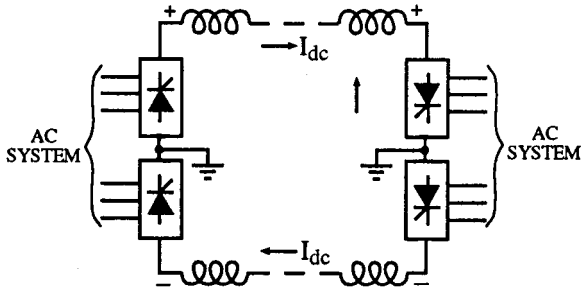


FIGURE 61.7 Bipolar dc system.

Configurations of DC Transmission

HVDC transmission systems can be classified into three categories:

1. Back-to-back systems
2. Two-terminal, or point-to-point, systems
3. Multiterminal systems

These will be briefly described here.

Back-to-Back DC System

In a back-to-back dc system (Fig. 61.6), both rectifier and inverter are located in the same station, usually in the same building. The rectifier and inverter are usually tied with a reactor, which is generally of outdoor, air-core design. A back-to-back dc system is used to tie two **asynchronous ac systems** (systems that are not in synchronism). The two ac systems can be of different operating frequencies, for example, one 50 Hz and the other 60 Hz. Examples are the Sakuma and Shin-Shinano converter stations in Japan. Both are used to link the 50- and 60-Hz ac systems. The Acaray station in Paraguay links the Paraguay system (50 Hz) with the Brazilian system, which is 60 Hz. Back-to-back dc links are also used to interconnect two ac systems that are of the same frequency but are not operating in synchronism. In North America, eastern and western systems are not synchronized, and Quebec and Texas are not synchronized with their neighboring systems. A dc link offers a practical solution as a tie between nonsynchronous systems. Thus to date, there are 10 back-to-back dc links in operation interconnecting such systems in North America. Similarly, in Europe, eastern and western systems are not synchronized, and dc offers the practical choice for interconnection between them.

Two-Terminal, or Point-to-Point, DC Transmission

Two-terminal dc systems can be either **bipolar** or monopolar. Bipolar configuration, shown in Fig. 61.7, is the commonly used arrangement for systems with overhead lines. In this system, there will be two conductors, one for each polarity (positive and negative) carrying nearly equal currents. Only the difference of these currents, which is usually small, flows through ground return.

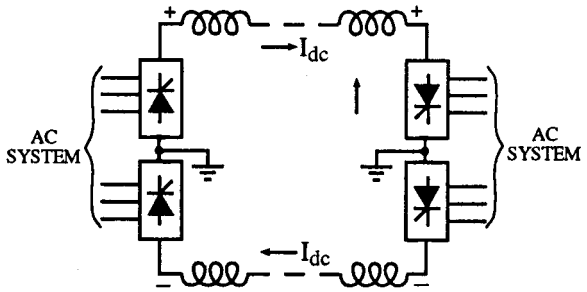


FIGURE 61.8 Monopolar ground return dc system.

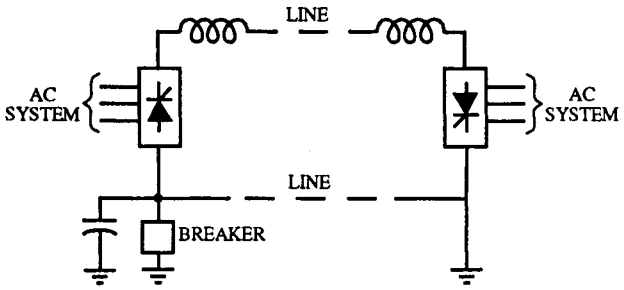


FIGURE 61.9 Monopolar metallic return dc system.

A monopolar system will have one conductor, either positive or negative polarity with current returning through either ground or another metallic return conductor. The monopolar ground return current configuration, shown in Fig. 61.8, has been used for undersea cable systems, where current returns through the sea. This configuration can also be used for short-term emergency operation for a two-terminal dc line system in the event of a pole outage. However, concerns for corrosion of underground metallic structures and interference with telephone and other utilities will restrict the duration of such operation. The total ampere-hour operation per year is usually the restricting criterion.

In a monopolar metallic return system, shown in Fig. 61.9, return current flows through a conductor, thus avoiding problems associated with ground return current. This method is generally used as a contingency mode of operation for a normal bipolar transmission system in the event of a partial converter (one-pole equipment) outage. In the case of outage of a one-pole converter, the conductor of the affected pole will be used as the returning conductor. A metallic return transfer breaker will be opened, diverting the return current from the ground path and into the pole conductor. This conductor will be grounded at one end and will be insulated at the other end. This system can transmit half the power of the normal bipolar system capacity (and can be increased if overload capacity is available). However, the line losses will be doubled compared to the normal bipolar operation for the same power transmitted.

Multiterminal DC Systems

There are two basic configurations in which the dc systems can be operated as multiterminal systems:

1. Parallel configuration
2. Series configuration

Parallel configuration can be either radial-connected [Fig. 61.10(a)] or mesh-connected [Fig. 61.10(b)]. In a parallel-connected multiterminal dc system, all converters operate at the same nominal dc voltage, similar to ac system interconnections. In this operation, one converter determines the operating voltage, and all other terminals operate in a current-controlling mode.

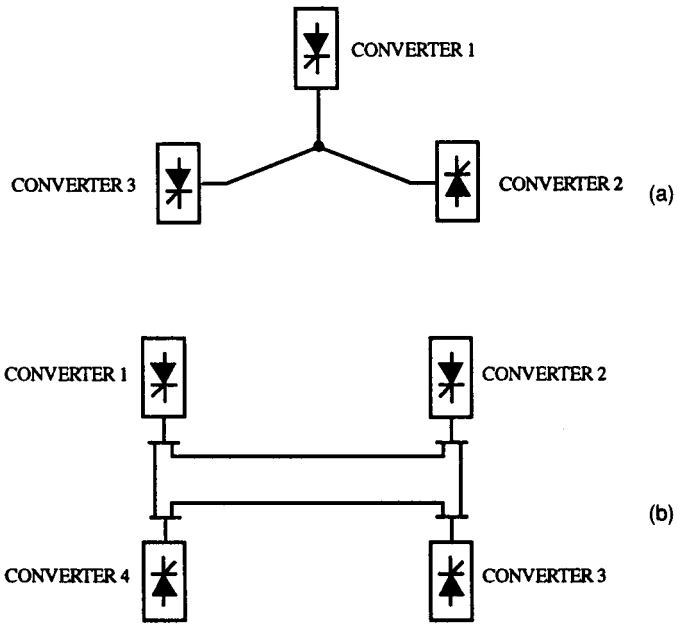


FIGURE 61.10 (a) Parallel-connected radial MTDC system; (b) parallel-connected mesh-type MTDC system.

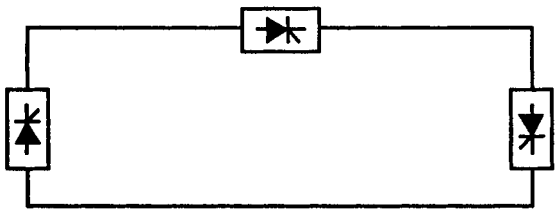


FIGURE 61.11 Series-connected MTDC system.

In a series-connected multiterminal dc system (Fig. 61.11), all converters operate at the same current. One converter sets the current that will be common to all converters in the system. Except for the converter that sets the current, the remaining converters operate in voltage control mode (constant **firing angle** or constant **extinction angle**). The converters operate almost independently without requirement for high-speed communication between them. The power output of a non-current-controlling converter is varied by varying its voltage. At all times, the sum of the voltages across the rectifier stations must be larger than the sum of voltages across the inverter stations. Disadvantages of a series-connected system are (1) reduced efficiency because full line insulation is not used at all times and (2) operation at higher firing angles will lead to high converter losses and higher reactive power requirements from the ac system.

There are now two truly multiterminal dc systems in operation. The Sardinia–Corsica–Italy three-terminal dc system was originally commissioned as a two-terminal (Sardinia–Italy) system in 1967 with a 200-MW rating. In 1986, the Corsica tap was added and the system was upgraded to a 300-MW rating. The two-terminal Hydro Quebec–New England HVDC interconnection (commissioned in 1985) was extended to a five-terminal system and commissioned in 1990 (see Table 61.4). The largest terminal of this system at Radisson station in Quebec is rated at 2250 MW. Two more systems, the Nelson River system in Canada and the Pacific NW-SW Intertie in the United States, also operate as multiterminal systems. Each of these systems has two converters at each end of the line, but the converters at each end are constrained to operate in the same mode, either rectifier or inverter.

Economic Comparison of AC and DC Transmission

In cases where HVDC is selected on technical considerations, it may be the only practical option, as in the case of an asynchronous interconnection. However, for long-distance power transmission, where both ac and HVDC are practical, the final decision is dependent on total costs of each alternative. Total cost of a transmission system includes the line costs (conductors, insulators, and towers) plus the right-of-way (R-o-W) costs. A dc line with two conductors can carry almost the same amount of power as the three-phase ac line with the same size of line conductors. However, dc towers with only two conductors are simpler and cheaper than three-phase ac towers. Hence the per-mile costs of line and R-o-W will be lower for a dc line. Power losses in the dc line are also lower than for ac for the same power transmitted. However, the HVDC system requires converters at the two ends of the line; hence the terminal costs for dc are higher than for ac. Variation of total costs for ac and dc as a function of line length are shown in Fig. 61.12. There is a break-even distance above which the total costs of dc option will be lower than the ac transmission option. This is in the range of 500 to 800 km for overhead lines but much shorter for cables. It is between 20 and 50 km for submarine cables and twice as far for underground cables.

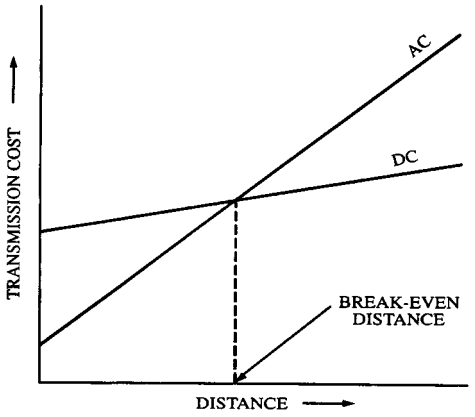


FIGURE 61.12 Transmission cost as function of line length.

Principles of Converter Operation

Converter Circuit

Since the generation and most of the transmission and utilization is alternating current, HVDC transmission requires conversion from ac to dc (called rectification) at the sending end and conversion back from dc to ac (called inversion) at the receiving end. In HVDC transmission, the basic device used for conversion from ac to dc and from dc to ac is a three-phase full-wave bridge converter, which is also known as a Graetz circuit. This is a three-phase six-pulse converter. A three-phase twelve-pulse converter will be composed of two three-phase six-pulse converters, supplied with voltages differing in phase by 30 degrees (Fig. 61.13). The phase difference of 30 degrees is obtained by supplying one six-pulse bridge with a Y/Y transformer and the other by Y/Δ transformer.

Relationships between AC and DC Quantities

Voltages and currents on ac and dc sides of the converter are related and are functions of several converter parameters including the converter transformer. The following equations are provided here for easy reference. Detailed derivations are given in Kimbark [1971].

E_{LL} = rms line-to-line voltage of the converter ac bus
I_1 = rms value of fundamental frequency component of the converter ac current
h = harmonic number
α = valve firing delay angle (from the instant the valve voltage is positive)
u = **overlap angle** (also called **commutation angle**)
ϕ = phase angle between voltage and current
$\cos\phi$ = displacement power factor
V_{d0} = ideal no-load dc voltage (at $\alpha = 0$ and $u = 0$)
L_c = commutating circuit inductance
$\beta = 180 - \alpha$ = angle of advance for inverter
$\gamma = 180 - (\alpha + u)$ = margin angle for inverter

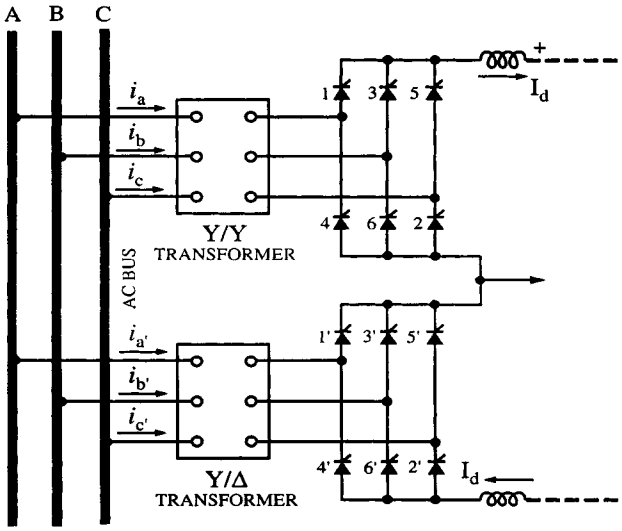


FIGURE 61.13 Basic circuit of a 12-pulse HVDC converter.

with $\alpha = 0$, $u = 0$,

$$V_{d0} = \frac{3\sqrt{2}}{\pi} E_{LL} = 1.35E_{LL} \tag{61.40}$$

With $\alpha > 0$, and $u = 0$

$$V_d = V_{d0} \cos \alpha \tag{61.41}$$

Theoretically α can vary from 0 to 180 degrees (with $u = 0$); hence V_d can vary from $+V_{d0}$ to $-V_{d0}$. Since the valves conduct current in only one direction, variation of dc voltage from V_{d0} to $-V_{d0}$ means reversal of power flow direction and the converter mode of operation changing from rectifier to inverter.

$$I_1 = \frac{\sqrt{6}}{\pi} I_d = 0.78I_d \tag{61.42}$$

$$\cos \phi = \cos \alpha = \frac{V_d}{V_{d0}} \tag{61.43}$$

With $\alpha > 0$ and $0 < u > 60°$,

$$V_d = V_{d0} \frac{\cos \alpha + \cos(\alpha + u)}{2} \tag{61.44}$$

$$V_d = \frac{3\sqrt{2}}{\pi} E_{LL} \frac{\cos \alpha + \cos(\alpha + u)}{2} \tag{61.45}$$

$$I_1 \approx \frac{\sqrt{6}}{\pi} I_d \tag{61.46}$$

The error in Eq. (61.46) is only 4.3% at $u = 60$ degrees (maximum overlap angle for normal steady-state operation), and it will be even lower (1.1%) for most practical cases when u is 30 degrees or less. It can be seen from Eqs. (61.45) and (61.46) that the ratio between ac and dc currents is almost fixed, but the ratio between ac and dc voltages varies as a function of α and u. Hence the HVDC converter can be viewed as a variable-ratio voltage transformer, with almost fixed current ratio.

$$P_{\text{dc}} = V_d I_d \tag{61.47}$$

$$P_{\text{ac}} = \sqrt{-3} E_{LL} I_1 \cos \phi \tag{61.48}$$

Substituting for V_d and I_d in (61.47) and comparing with (61.48),

$$\cos \phi \approx \frac{\cos \alpha + \cos(\alpha + u)}{2} \tag{61.49}$$

From Eqs. (61.44) and (61.49),

$$\cos \phi \approx \frac{V_d}{V_{d0}} \tag{61.50}$$

From Eqs. (61.40), (61.44), and (61.49),

$$V_d \approx 1.35 E_{LL} \cos \phi \tag{61.51}$$

AC Current Harmonics

The HVDC converter is a harmonic current source on the ac side. Fourier analysis of an ac current waveform, shown in Fig. 61.14, shows that it contains the fundamental and harmonics of the order 5, 7, 11, 13, 17, 19, etc. The current for zero degree overlap angle can be expressed as

$$i(t) = \frac{2\sqrt{3}}{\pi} I_d \left(\begin{array}{l} \cos\ \omega t - \frac{1}{5} \cos\ 5\omega t + \frac{1}{7} \cos\ 7\omega t \\ - \frac{1}{11} \cos\ 11\omega t + \frac{1}{13} \cos\ 13\omega t + \cdots \end{array} \right) \tag{61.52}$$

and

$$I_{h0} = \frac{I_{10}}{h} \tag{61.53}$$

where I_{10} and I_{h0} are the fundamental and harmonic currents, respectively, at $\alpha = 0$ and $u = 0$.

Equation (61.53) indicates that the magnitudes of harmonics are inversely proportional to their order.

Converter ac current waveform $i_{a'}$ for phase a with a Y/Δ transformer is also shown in Fig. 61.14. Fourier analysis of this current shows that the fundamental and harmonic components will have the same magnitude

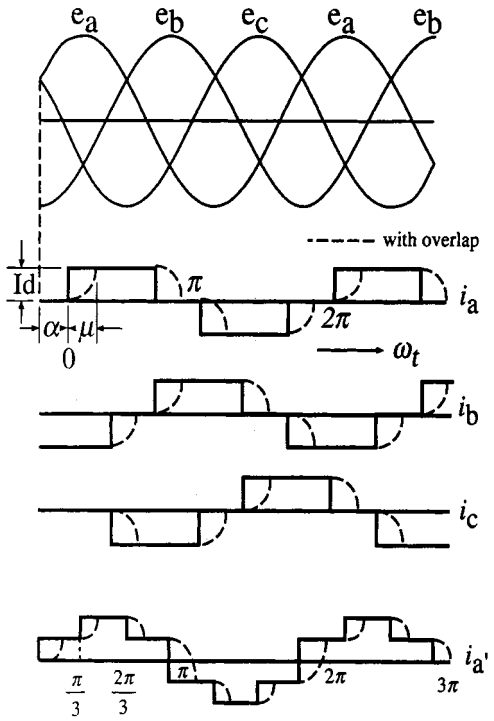


FIGURE 61.14 AC line current waveforms, i_a, i_b, i_c with Y/Y transformer and $i_{a'}$ with Y/Δ transformer.

as in the case of the Y/Y transformer. However, harmonics of order 5, 7, 17, 19, etc. are in phase opposition, whereas harmonics of order 11, 13, 23, 25, etc. are in phase with the Y/Y transformer case. Hence harmonics of order 5, 7, 17, 19, etc. will be canceled in a 12-pulse converter and do not appear in the ac system. In practice they will not be canceled completely because of imbalances in converter and transformer parameters.

Effect of Overlap. The effect of overlap due to commutation angle is to decrease the amplitude of harmonics from the case with zero overlap. Magnitudes of harmonics for a general case with finite firing angle (α) and overlap angle (u) are given by

$$\frac{I_h}{I_{h0}} = \frac{1}{x}\left[A^2 + B^2 - 2AB\cos(2\alpha + u)\right]^{1/2} \tag{61.54}$$

where

$$A = \frac{\sin(h+1)\dfrac{u}{2}}{h+1} \qquad B = \frac{\sin(h-1)\dfrac{u}{2}}{h-1}$$
$$x = \cos\alpha - \cos(\alpha + u)$$

Noncharacteristic Harmonics. In addition to the harmonics described above, converters also generate other harmonics due to "nonideal" conditions of converter operation. Examples of the nonideal conditions are converter ac bus voltage imbalance, perturbation of valve firing pulses, distortion of ac bus voltages, and unbalanced converter transformer impedances. Harmonics generated due to these causes are called *noncharacteristic* harmonics. These are usually smaller in magnitude compared to characteristic harmonics but can create problems if resonances exist in the ac system at these frequencies. In several instances additional filters were installed at the converter ac bus to reduce levels of these harmonics flowing into the ac system.

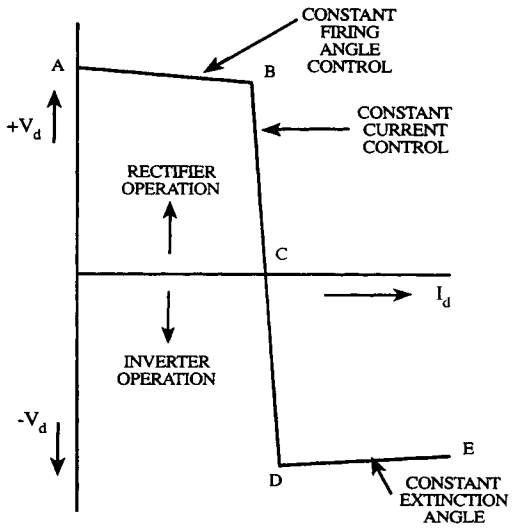


FIGURE 61.15 HVDC converter static characteristic.

Converter Control

The static characteristic of a HVDC converter is shown in Fig. 61.15. There are three distinct features of this characteristic.

***Constant Firing Angle Characteristic* (A–B).** If the converter is operating under constant firing angle control, the converter characteristic can be described by the equation

$$V_d = V_{d0} \cos \alpha - \frac{3\omega L_c}{\pi} I_d \tag{61.55}$$

When the ordered current is too high for the converter to deliver, it will operate at the minimum firing angle (usually 5 degrees). Then the current will be determined by the voltage V_d and the load. This is also referred to as the natural voltage characteristic. The converter in this mode is equivalent to a dc voltage source with internal resistance R_c, where

$$R_c = \frac{3\omega L_c}{\pi} \tag{61.56}$$

Constant Current Control. This is the usual mode of operation of the rectifier. When the converter is operating in constant current control mode, the firing angle is adjusted to maintain dc current at the ordered value. If the load current goes higher than the ordered current for any reason, control increases the firing angle to reduce dc voltage and the converter operation moves in the direction from *B* to *C*. At point *C*, the firing angle reaches 90 degrees (neglecting overlap angle), the voltage changes polarity, and the converter becomes an inverter. From *C* to *D*, the converter works as an inverter.

Constant Extinction Angle Control. At point *D*, the inverter firing angle has increased to a point where further increase can cause commutation failure. The inverter for its safe operation must be operated with sufficient angle of advance β, such that under all operating conditions the extinction angle γ is greater than the valve deionization angle. The deionization angle is defined as the time in electrical degrees from the instant current reaches zero in a particular valve to the time the valve can withstand the application of positive voltage.

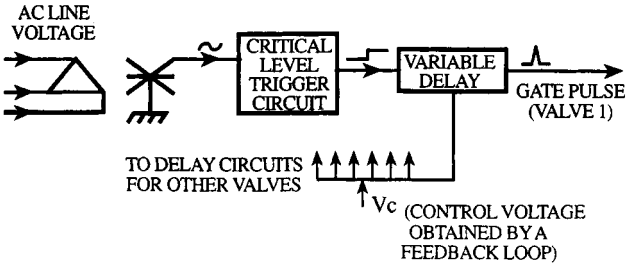


FIGURE 61.16 Constant α control.

Typical minimum values of γ are 15 to 20 degrees for mercury arc valves and slightly less for thyristor valves. During the range *D* to *E*, the increase of load current increases the overlap angle, which reduces the dc voltage. This is the negative resistance characteristic of the inverter.

The functional requirements for HVDC converter control are:

1. Minimize the generation of noncharacteristic harmonics.
2. Safe inverter operation with fewest possible commutation failures even with distorted ac voltages.
3. Lowest possible consumption of reactive power. This requires operation with smallest possible delay angle α and extinction angle γ without increased risk of commutation failures.
4. Smooth transition from current control to extinction angle control.
5. Sufficient stability margins and response time when the ratio of the ac system short-circuit strength and the rated dc power (short-circuit ratio) is low.

Individual Phase Control

In the early HVDC systems individual phase control systems were used. The firing angle of each valve is calculated individually and operated either as constant α or constant γ control.

A schematic of the individual phase control system is shown in Fig. 61.16. Six timing voltages are derived from the ac bus voltage, and the six grid pulses are generated at nominally identical delay times subsequent to the respective voltage-zero crossings. The delays are produced by independent delay circuits and controlled by a common direct voltage V_c, which is derived through a feedback loop to control constant dc line current or constant power. Several variations of this control were used until the late 1960s.

Disadvantages of Individual Phase Control. With distorted ac bus voltages, the firing pulses will be unequally spaced, thus generating noncharacteristic harmonics in ac current. This in turn will further distort the ac bus voltage. This process could lead to harmonic instability, particularly with ac systems of low short-circuit capacity (high-impedance system). Control system filters were tried to solve this problem. However, the filters could increase the potential for commutation failures and also reduce the speed of control system response for faults or disturbances in the ac system.

Equal Pulse Spacing Control

A control system based on the principle of equal spacing of firing pulses at intervals of 60 degrees (electrical) independent of ac bus voltages was developed in the late 1960s. The basic components of this system, shown in Fig. 61.17, consist of a voltage-controlled oscillator and ring counter. The frequency of the oscillator is directly proportional to the dc control voltage V_c. Under steady-state conditions, pulse frequency is precisely 6*f*, where *f* is the ac system frequency. The phase of each grid pulse will have some arbitrary value relative to the ac bus voltage. If the three-phase ac bus voltages are symmetrical sine waves with no distortion, then α is the same for all valves. The oscillator will be phase-locked with the ac system frequency to avoid drifting. The dc control voltage V_c is derived from a feedback loop for constant current, constant power, or constant extinction angle γ.

The control systems used in recent projects are digital-based and much more sophisticated than the earlier versions.

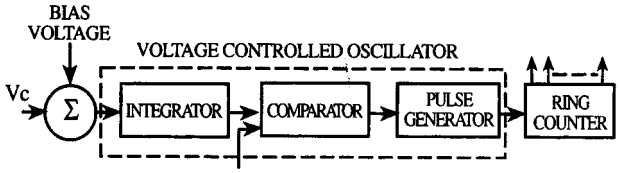


FIGURE 61.17 Equal pulse spacing control.

Developments

During the last two or three decades, several developments in HVDC technology have taken place that improved viability of the HVDC transmission. Prior to 1970 mercury arc valves were used for converting from ac to dc and dc to ac. They had several operational problems including frequent arcbacks. Arcback is a random phenomenon that results in failure of a valve to block conduction in the reverse direction. This is most common in the rectifier mode of operation. In rectifier operation, the valve is exposed to inverse voltage for approximately two-thirds of each cycle. Arcbacks result in line-to-line short circuits, and sometimes in three-phase short circuits, which subject the converter transformer and valves to severe stresses.

Thyristors

Thyristor valves were first used for HVDC transmission in the early 1970s, and since then have completely replaced mercury arc valves. The term thyristor valve, carried over from mercury arc valve, is used to refer to an assembly of series and parallel connection of several thyristors to make up the required voltage and current ratings of one arm of the converter. The first test thyristor valve in a HVDC converter station was installed in 1967, replacing a mercury arc valve in the Ygne converter station on the island of Gotland (see Gotland I in Table 61.4). The Eel River back-to-back station in New Brunswick, Canada, commissioned in 1972, was the first all-thyristor HVDC converter station. The voltage and current ratings of thyristors have increased steadily over the last two decades. Figure 61.18 shows the maximum blocking voltage of thyristors from the late 1960s to date. The current ratings have also increased in this period from 1 to 4 kA. Some of the increased current ratings were achieved with large-diameter silicon wafers (presently 100-mm diameter) and with improved cooling systems. Earlier projects used air-cooled thyristors. Water-cooled thyristors are used for all the recent projects.

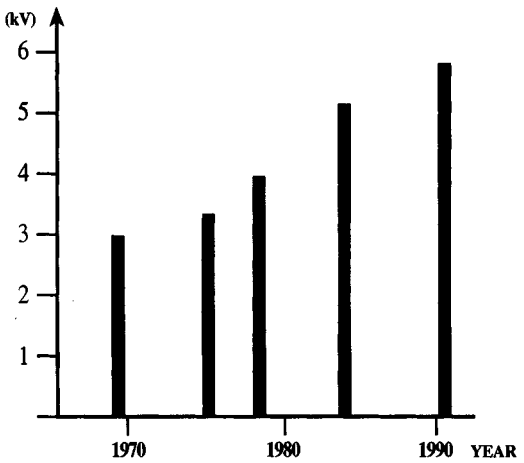


FIGURE 61.18 Maximum blocking voltage development.

Other recent developments include direct light-triggered thyristors and gate turn-off (GTO) thyristors. GTO thyristors have some advantages for HVDC converters connected to weak ac systems. They are now available in ratings up to 4.5 kV and 4 kA but have not yet been applied in HVDC systems.

DC Circuit Breakers

Interrupting the current in ac systems is aided by the fact that ac current goes through zero every half-cycle or approximately every 8 ms in a 60-Hz system. The absence of natural current zero in dc makes it difficult to develop a dc circuit breaker. There are three principal problems in designing a dc circuit breaker:

1. Forcing current zero in the interrupting element
2. Controlling the overvoltages caused by large di/dt in a highly inductive circuit
3. Dissipating large amounts of energy (tens of megajoules)

The second and third problems are solved by the application of zinc oxide varistors connected line to ground and across the breaking element. The first is the major problem, and several different solutions are adopted by different manufacturers. Basically, current zero is achieved by inserting a counter voltage into the circuit.

In the circuit shown in Fig. 61.19, opening CB (air-blast circuit breaker) causes current to be commutated to the parallel LC circuit. The commutating circuit will be oscillatory, which creates current zero in the circuit breaker. The opening of CB increases the voltage across the commutating circuit, which will be limited by the zinc oxide varistor ZnO_1 by entering into conduction. The resistance *R* is the closing resistor in series with switch *S*.

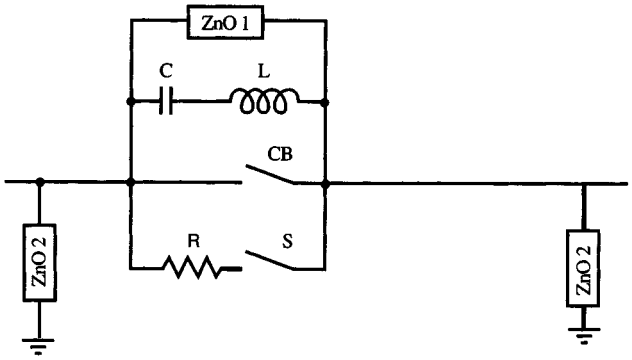


FIGURE 61.19 DC circuit breaker (one module).

It should be noted that a two-terminal dc system does not need a dc breaker since the fast converter control response can bring the current quickly to zero. In multiterminal systems, dc breakers can provide additional flexibility of operation. The multiterminal dc systems commissioned so far have not employed dc breakers.

Defining Terms

Asynchronous ac systems: AC systems with either different operating frequencies or that are not in synchronism.

Bipole: DC system with two conductors, one positive and the other negative polarity. Rated voltage of a bipole is expressed as ±100 kV, for example.

Commutation: Process of transferring current from one valve to another.

Commutation angle (overlap angle): Time in electrical degrees from the start to the completion of the commutation process.

Extinction angle: Time in electrical degrees from the instant the current in a valve reaches zero (end of conduction) to the time the valve voltage changes sign and becomes positive.

Firing angle (delay angle): Time in electrical degrees from the instant the valve voltage is positive to the application of firing pulse to the valve (start of conduction).

Pulse number of a converter: Number of ripples in dc voltage per cycle of ac voltage. A three-phase two-way bridge is a six-pulse converter.

Thyristor valve: Assembly of series and parallel connection of several thyristors to make up the required current and voltage ratings of one arm of the converter.

Related Topic

30.2 Power Conversion

References

J.D. Ainsworth, "The phase locked oscillator: A new control system for controlled static converters," *IEEE Trans. Power Appar. Syst.*, vol. PAS-87, pp. 859–865, March 1968.

A. Ekstrom and L. Eklund, "HVDC thyristor valve development," in Proceedings of the International Conference on DC Power Transmission, Montreal, pp. 220–227, 1984.

A. Ekstrom and G. Liss, "A refined HVDC control system," *IEEE Trans. Power Appar. Syst.*, vol. PAS-89, no. 5, pp. 723–732, May/June 1970.

E. W. Kimbark, *Direct Current Transmission,* vol. I, New York: Wiley-Interscience, 1971.

W.F. Long et al., "Considerations for implementing multiterminal dc systems," *IEEE Trans. Power Appar. Syst.,* pp. 2521–2530, September 1985.
K.R. Padiyar, *HVDC Power Transmission Systems—Technology and System Interactions,* New Delhi: Wiley Eastern Limited, 1990.
J. Reeve and P.C.S. Krishnayya, "Unusual current harmonics arising from high-voltage dc transmission," *IEEE Trans. Power Appar. Syst.,* vol. PAS-87, no. 3, pp. 883–893, March 1968.
R.S. Thallam and J. Reeve, "Dynamic analysis of harmonic interaction between AC and DC power systems," *IEEE Trans. Power Appar. Syst.,* vol. PAS-93, no. 2, pp. 640–646, March/April 1974.
E. Uhlmann, *Power Transmission by Direct Current,* Berlin: Springer-Verlag, 1975.

Further Information

The three textbooks cited under References are excellent for further reading. The IEEE (USA) and IEE (UK) periodically hold conferences on "DC Transmission." The last IEEE conference was held in 1984 in Montreal, and the IEE conference was held in 1991 (conf. publ. no. 345) in London. Proceedings can be ordered from these organizations.

61.4 Compensation

Mohamed E. El-Hawary

The term *compensation* is used to describe the intentional insertion of reactive power-producing devices, either capacitive or inductive, to achieve a desired effect in the electric power system. The effects include improved voltage profiles, enhanced stability performance, and improved transmission capacity. The devices are connected either in series or in **shunt** (parallel) at a particular point in the power circuit.

For illustration purposes, we consider the circuit of Fig. 61.20, where the link has an impedance of $R + jX$, and it is assumed that $V_1 > V_2$ and V_1 leads V_2. The corresponding phasor diagram for zero R and lagging load current I is shown in Fig. 61.21. The approximate relationship between the scalar voltage difference between two nodes in a network and the flow of reactive power Q can be shown to be [Weedy, 1972]

$$\Delta V = \frac{RP_2 + XQ_2}{V_2} \tag{61.57}$$

In most power circuits, $X >> R$ and the voltage difference ΔV determines Q.

The flow of power and reactive power is from A to B when $V_1 > V_2$ and V_1 leads V_2. Q is determined mainly by $V_1 - V_2$. The direction of reactive power can be reversed by making $V_2 > V_1$. It can thus be seen that if a scalar voltage difference exists across a largely reactive link, the reactive power flows toward the node of lower voltage. Looked at from an alternative point of view, if there is a reactive power deficit at a point in an electric network, this deficit has to be supplied from the rest of the circuit and hence the voltage at that point falls. Of course, a surplus of reactive power generated will cause a voltage rise. This can be interpreted as providing voltage support by supplying reactive power at that point.

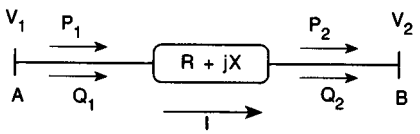


FIGURE 61.20 Two nodes connected by a link.

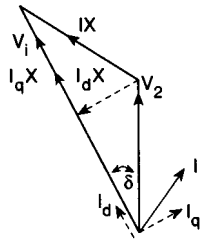


FIGURE 61.21 Phasor diagram for system shown in Fig. 61.20.

Assuming that the link is reactive, i.e., with $R = 0$, then $P_1 = P_2 = P$. In this case, the active power transferred from point A to point B can be shown to be given by [El-Hawary, 1995]

$$P = P_{\max} \sin \delta \tag{61.58}$$

The maximum power transfer $P_{\max}$ is given by

$$P_{\max} = \frac{V_1 V_2}{X} \tag{61.59}$$

It is clear that the power transfer capacity defined by Eq. (61.59) is improved if V_2 is increased.

Series Capacitors

Series capacitors are employed to neutralize part of the inductive reactance of a power circuit, as shown in Fig. 61.22. From the phasor diagram of Fig. 61.23 we see that the load voltage is higher with the capacitor inserted than without the capacitor.

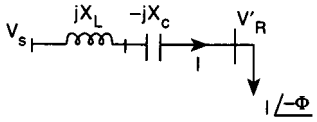


FIGURE 61.22 Line with series capacitor.

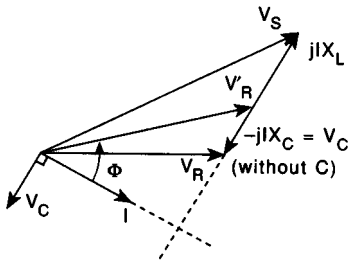


FIGURE 61.23 Phasor diagram corresponding to Fig. 61.22.

Introducing series capacitors is associated with an increase in the circuit's transmission capacity [from (61.59) with a net reduction in X] and enhanced stability performance as well as improved voltage conditions on the circuit. They are also valuable in other aspects such as:

- Controlling reactive power balance
- Load distribution and control of overall transmission losses

Series-capacitor compensation delays investments in additional overhead lines for added transmission capacity, which is advantageous from an environmental point of view.

The first worldwide series-capacitor installation was a 33-kV 1.25-MVAR bank on the New York Power & Light system, which was put in service in 1928. Since then, many higher-capacity, higher-voltage installations have been installed in the United States, Canada, Sweden, Brazil, and other countries.

The reduction in a circuit's inductive reactance increases the short-circuit current levels over those for the noncompensated circuit. Care must be taken to avoid exposing series capacitors to such large short-circuit currents, since this causes excessive voltage rise as well as heating that can damage the capacitors. Specially calibrated spark gaps and short-circuiting switches are deployed within a predetermined time interval to avoid damage to the capacitors.

The interaction between a series-capacitor-compensated ac transmission system in electrical **resonance** and a turbine-generator mechanical system in torsional mechanical resonance results in the phenomenon of **sub-synchronous resonance** (SSR). Energy is exchanged between the electrical and mechanical systems at one or more natural frequencies of the combined system below the synchronous frequency of the system. The resulting mechanical oscillations can increase until mechanical failure takes place.

Techniques to counteract SSR include the following:

- *Supplementary excitation control:* The subsynchronous current and/or voltage is detected and the excitation current is modulated using high-gain feedback to vary the generator output voltage, which counters the subsynchronous oscillations [see El-Serafi and Shaltout, 1979].
- *Static filters:* These are connected in series with each phase of each main generator. Step-up transformers are employed. The filters are tuned to frequencies that correspond to the power system frequency and the troublesome machine natural modes of oscillations [see Tice and Bowler, 1975].
- *Dynamic filters:* In a manner similar to that of excitation control, the subsynchronous oscillation is detected, and a counter emf is generated by a thyristor cycloconverter or a similar device and injected in the power line through a series transformer [see Kilgore et al., 1975].
- *Bypassing series capacitors:* To limit transient torque buildup, complete or partial bypass with the aid of low set gaps.
- Amortisseur windings on the pole faces of the generator rotors can be employed to improve damping.
- A more recent damping scheme [see Hingorani, 1981] is based on measuring the half-cycle period of the series-capacitor voltage, and if this period exceeds a preset value, the capacitor's charge is dissipated into a resistor shunting the capacitor through two antiparallel thyristors.
- A passive SSR countermeasure scheme [see Edris, 1990] involves using three different combinations of inductive and capacitive elements on the three phases. The combinations will exhibit the required equal degree of capacitive compensation in the three phases at power frequency. At any other frequency, the three combinations will appear as unequal reactances in the three phases. In this manner, asynchronous oscillations will drive unsymmetrical three-phase currents in the generator's armature windings. This creates an mmf with a circular component of a lower magnitude, compared with the corresponding component if the currents were symmetrical. The developed interacting electromagnetic torque will be lower.

Synchronous Compensators

A synchronous compensator is a synchronous motor running without a mechanical load. Depending on the value of excitation, it can absorb or generate reactive power. The losses are considerable compared with static capacitors. When used with a voltage regulator, the compensator can run automatically overexcited at high-load current and underexcited at low-load current. The cost of installation of synchronous compensators is high relative to capacitors.

Shunt Capacitors

Shunt capacitors are used to supply capacitive kVAR to the system at the point where they are connected, with the same effect as an overexcited synchronous condenser, generator, or motor. Shunt capacitors supply reactive power to counteract the out-of-phase component of current required by an inductive load. They are either energized continuously or switched on and off during load cycles.

Figure 61.24(a) displays a simple circuit with shunt capacitor compensation applied at the load side. The line current I_L is the sum of the motor load current I_M and the capacitor current I_c. From the current phasor diagram of Fig. 61.24(b), it is clear that the line current is decreased with the insertion of the shunt capacitor. Figure 61.24(c) displays the corresponding voltage phasors. The effect of the shunt capacitor is to reduce the source voltage to V_{s1} from V_{s0}.

From the above considerations, it is clear that shunt capacitors applied at a point in a circuit supplying a load of lagging power factor have the following effects:

- Increase voltage level at the load
- Improve voltage regulation if the capacitor units are properly switched
- Reduce I^2R power loss and I^2X kVAR loss in the system because of reduction in current
- Increase power factor of the source generators

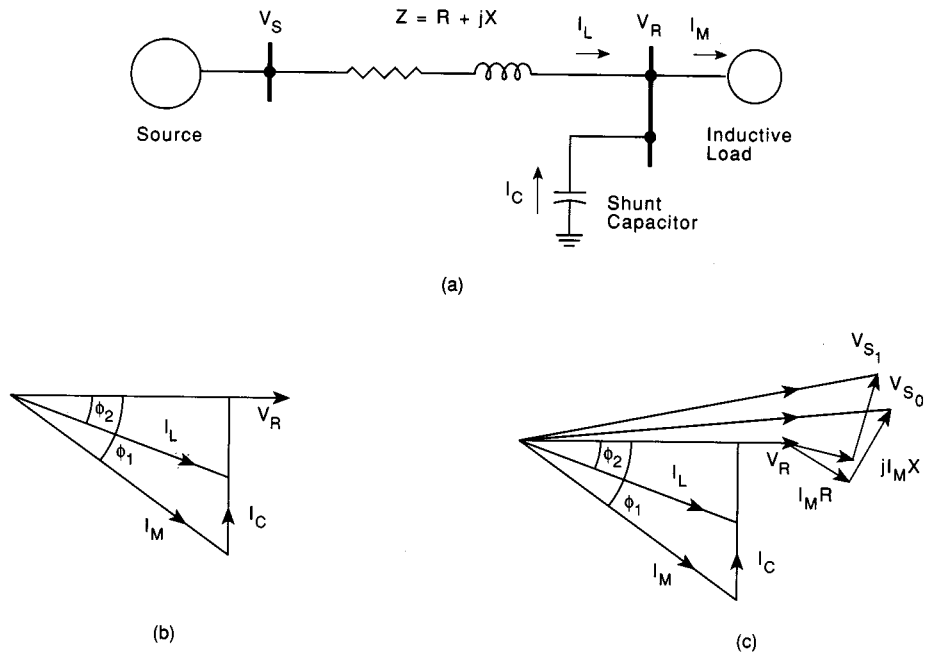


FIGURE 61.24 (a) Shunt-capacitor-compensated load; (b) current phasor diagram; (c) voltage phasor diagram.

- Decrease kVA loading on the source generators and circuits to relieve an overloaded condition or release capacity for additional load growth
- By reducing kVA load on the source generators, additional active power loading may be placed on the generators if turbine capacity is available
- Reduce demand kVA where power is purchased
- Reduce investment in system facilities per kW of load supplied

To reduce high inrush currents in starting large motors, a capacitor starting system is employed. This maintains acceptable voltage levels throughout the system. The high inductive component of normal reactive starting current is offset by the addition, during the starting period only, of capacitors to the motor bus. This differs from applying capacitors for motor power factor correction.

When used for voltage control, the action of shunt capacitors is different from that of synchronous condensers, since their reactive power varies as the square of the voltage, whereas the synchronous machine maintains approximately constant kVA for sudden voltage changes. The synchronous condenser has a greater stabilizing effect upon system voltages. The losses of the synchronous condenser are much greater than those of capacitors.

Note that in determining the amount of shunt capacitor kVAR required, since a voltage rise increases the lagging kVAR in the exciting currents of transformer and motors, some additional capacitor kVAR above that based on initial conditions without capacitors may be required to get the desired correction. If the load includes synchronous motors, it may be desirable, if possible, to increase the field currents to these motors.

The following are the relative merits of shunt and series capacitors:

- If the total line reactance is high, series capacitors are very effective.
- If the voltage drop is the limiting factor, series capacitors are effective; also, voltage fluctuations are evened out.
- If the reactive power requirements of the load are small, the series capacitor is of little value.
- If thermal considerations limit the current, then series capacitors are of little value since the reduction in line current associated with them is small.

Applying capacitors with harmonic-generating apparatus on a power system requires considering the potential of an excited harmonic resonance condition. Either a series or a shunt resonance condition may take place. In

actual electrical systems utilizing compensating capacitors, either type of resonance or a combination of both can occur if the resonant point happens to be close to one of the frequencies generated by harmonic sources in the system. The outcome can be the flow of excessive amounts of harmonic current or the appearance of excessive harmonic overvoltages, or both. Possible effects of this are excessive capacitor fuse operation, capacitor failure, overheating of other electrical equipment, or telephone interference.

Shunt Reactors

Shunt reactor compensation is usually required under conditions that are the opposite of those requiring shunt capacitor compensation (see Fig. 61.25). Shunt reactors are installed to remedy the following situations:

- Overvoltages that occur during low load periods at stations served by long lines as a result of the line's capacitance (Ferranti effect).
- Leading power factors at generating plants resulting in lower transient and steady-state stability limits, caused by reduced field current and the machine's internal voltage. In this case, shunt reactors are usually installed at either side of the generator's step-up transformers.
- Open-circuit line charging kVA requirements in extra-high-voltage systems that exceed the available generation capabilities.

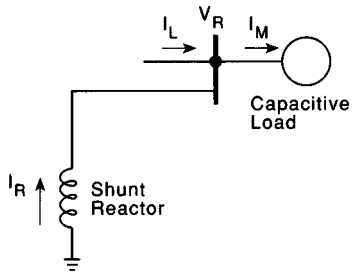


FIGURE 61.25 Shunt-reactor-compensated load.

Coupling from nearby energized lines can cause severe resonant overvoltages across the shunt reactors of unenergized compensated lines.

Static VAR Compensators (SVC)

Advances in **thyristor** technology for power systems applications have lead to the development of the static VAR compensators (SVC). These devices contain standard shunt elements (**reactors**, capacitors) but are controlled by thyristors [El-Hawary, 1995].

Static VAR compensators provide solutions to two types of compensation problems normally encountered in practical power systems [Gyugyi et al., 1978]. The first is load compensation, where the requirements are usually to reduce or cancel the reactive power demand of large and fluctuating industrial loads, such as electric arc furnaces and rolling mills, and to balance the real power drawn from the ac supply lines. These types of heavy industrial loads are normally concentrated in one plant and served from one network terminal, and thus can be handled by a local compensator connected to the same terminal. The second type of compensation is related to voltage support of transmission lines at a given terminal in response to disturbances of both load and generation. The voltage support is achieved by rapid control of the SVC reactance and thus its reactive power output. The main objectives of dynamic VAR compensation are to increase the stability limit of the ac power system, to decrease terminal voltage fluctuations during load variations, and to limit overvoltages subsequent to large disturbances. SVCs are essentially thyristor-controlled reactive power devices.

The two fundamental thyristor-controlled reactive power device configurations are [Olwegard et al., 1981]:

- Thyristor-switched shunt capacitors (TSC): The idea is to split a **capacitor bank** into sufficiently small capacitor steps and switch those steps on and off individually. Figure 61.26(a) shows the concept of the TSC. It offers stepwise control, virtually no transients, and no harmonic generation. The average delay for executing a command from the regulator is half a cycle.
- Thyristor-switched shunt reactors (TCR): In this scheme the fundamental frequency current component through the reactor is controlled by delaying the closing of the thyristor switch with respect to the natural zero crossings of the current. Figure 61.26(b) shows the concept of the TCR. Harmonic currents are generated from the phase-angle-controlled reactor.

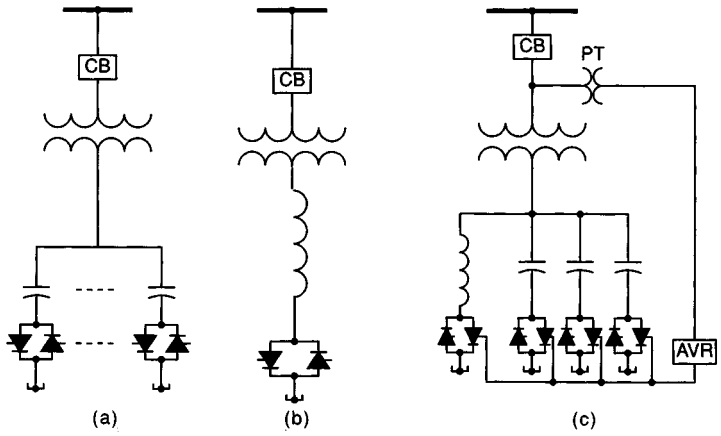


FIGURE 61.26 Basic static VAR compensator configurations. (a) Thyristor-switched shunt capacitors (TSC); (b) thyristor-switched shunt reactors (TCR); (c) combined TSC/TCR.

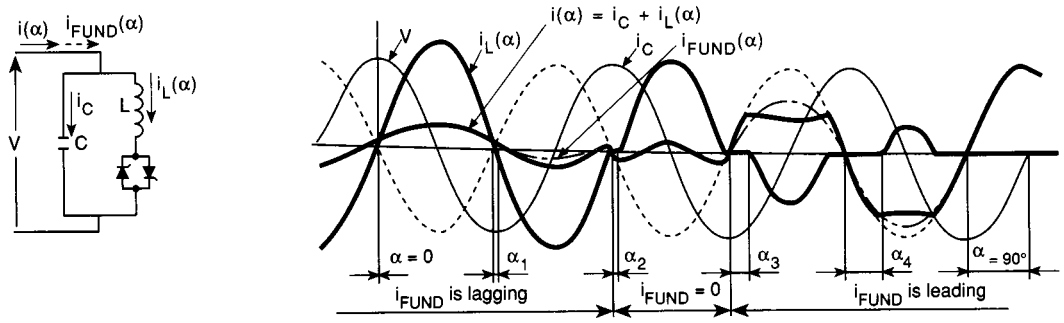


FIGURE 61.27 Basic fixed-capacitor, thyristor-controlled reactor-type compensator and associated waveforms.

The magnitude of the harmonics can be reduced using two methods. In the first, the reactor is split into smaller steps, while only one step is phase-angle controlled. The other reactor steps are either on or off. This decreases the magnitude of all harmonics. The second method involves the 12-pulse arrangement, where two identical connected thyristor-controlled reactors are used, one operated from a wye-connected secondary winding, the other from a delta-connected winding of a step-up transformer. TCR units are characterized by continuous control, and there is a maximum of one half-cycle delay for executing a command from the regulator.

In many applications, the arrangement of an SVC consists of a few large steps of thyristor-switched capacitor and one or two thyristor-controlled reactors, as shown in Fig. 61.26(c). The following are some practical schemes.

Fixed-Capacitor, Thyristor-Controlled Reactor (FC-TCR) Scheme

This scheme was originally developed for industrial applications, such as arc furnace "flicker" control [Gyugyi and Taylor, 1980]. It is essentially a TCR (controlled by a delay angle α) in parallel with a fixed capacitor. Figure 61.27 shows a basic fixed-capacitor, thyristor-controlled reactor-type compensator and associated waveforms. Figure 61.28 displays the steady-state reactive power versus terminal voltage characteristics of a static VAR compensator. In the figure, B_C is the imaginary part of the admittance of the capacitor C, and B_L is the imaginary part of the equivalent admittance of the reactor L at delay angle α. The relation between the output VARs and the applied voltage is linear over the voltage band of regulation. In practice, the fixed capacitor is usually replaced by a filter network that has the required capacitive reactance at the power system frequency but exhibits a low impedance at selected frequencies to absorb troublesome harmonics.

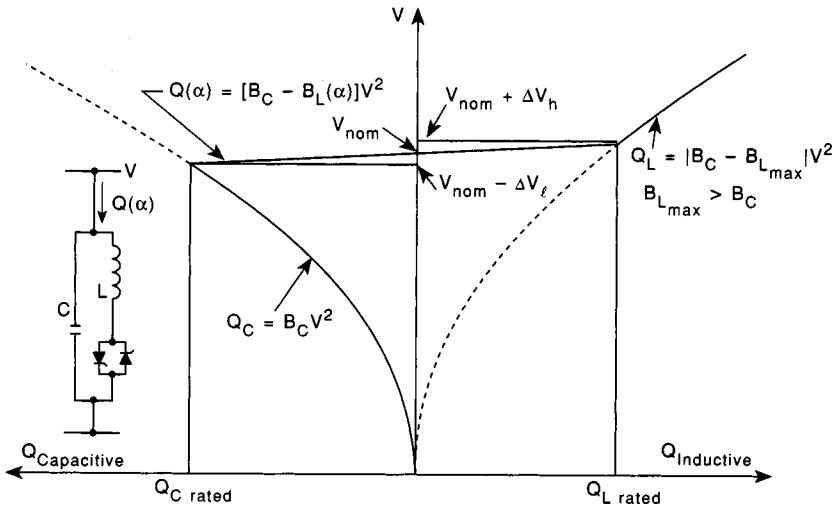


FIGURE 61.28 The steady-state reactive power versus terminal voltage characteristics of a static VAR compensator.

The behavior and response of the FC-TCR type of compensator under large disturbances is uncontrollable, at least during the first few cycles following the disturbance. The resulting voltage transients are essentially determined by the fixed capacitor and the power system impedance. This can lead to overvoltage and resonance problems.

At zero VAR demand, the capacitive and reactive VARs cancel out, but the capacitor bank's current is circulated through the reactor bank via the thyristor switch. As a result, this configuration suffers from no load (standby) losses. The losses decrease with increasing the capacitive VAR output and, conversely, increase with increasing the inductive VAR output.

Thyristor-Switched Capacitor, Thyristor-Controlled Reactor (TSC-TCR) Scheme

This hybrid compensator was developed specifically for utility applications to overcome the disadvantages of the FC-TCR compensators (behavior under large disturbances and loss characteristic). Figure 61.29 shows a basic circuit of this compensator. It consists in general of a thyristor-controlled reactor bank (or banks) and a number of capacitor banks, each in series with a solid-state switch, which is composed of either a reverse-parallel-connected thyristor pair or a thyristor in reverse parallel with a diode. The reactor's switch is composed of a reverse-parallel-connected thyristor pair that is capable of continuously controlling the current in the reactor from zero to maximum rated current.

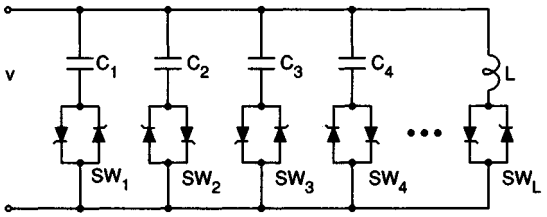


FIGURE 61.29 Basic thyristor-switched capacitor, thyristor-controlled reactor-type compensator.

The total capacitive range is divided into n operating intervals, where n is the number of capacitor banks in the compensator. In the first interval one capacitor bank is switched in, and at the same time the current in the TCR bank is adjusted so that the resultant VAR output from capacitor and reactor matches the VAR demand. In the ith interval the output is controllable in the range $[(i-1)\text{VAR}_{max}/n]$ to $(i\,\text{VAR}_{max}/n)$ by switching in the ith capacitor bank and using the TCR bank to absorb the surplus capacitive VARs. This scheme can be considered as a conventional FC-TCR, where the rating of the reactor bank is kept relatively small ($1/n$ times the maximum VAR output) and the value of the capacitor bank is changed in discrete steps so as to keep the operation of the reactor bank within its normal control range.

The losses of the TSC-TCR compensator at zero VARs output are inherently low, and they increase in proportion to the VAR output.

The mechanism by which SVCs introduce damping into the system can be explained as a result of the change in system voltage due to switching of a capacitor/reactor. The electrical power output of the generators is changed immediately due to the change in power transfer capability and the change in load power requirements.

Among the early applications of SVC for power system damping is the application to the Scandinavian system as discussed in Olwegard et al. [1981]. More recently, SVC control for damping of system oscillations based on local measurements has been proposed. The scheme uses phase-angle estimates based on voltage and power measurements at the SVC location as the control signal [see Lerch et al., 1991].

For a general mathematical model of an SVC and an analysis of its stabilizing effects, see Hammad [1986]. Representing the SVC in transient analysis programs is an important consideration [see Gole and Sood, 1990; Lefebvre and Gerin-Lajoie, 1992].

It is important to recognize that applying static VAR compensators to series-compensated ac transmission lines results in three distinct resonant modes [Larsen et al., 1990]:

- Shunt-capacitance resonance involves energy exchange between the shunt capacitance (line charging plus any power factor correction or SVCs) and the series inductance of the lines and the generator.
- Series-line resonance involves energy exchange between the series capacitor and the series inductance of the lines, transformers, and generators. The resonant frequency will depend on the level of series compensation.
- Shunt-reactor resonance involves energy exchange between shunt reactors at the intermediate substations of the line and the series capacitors.

The applications of SVCs are part of the broader area of flexible ac transmission systems (FACTS) [Hingorani, 1993]

Defining Terms

Capacitor bank: An assembly at one location of capacitors and all necessary accessories, such as switching equipment, protective equipment, and controls, required for a complete operating installation.

Reactor: A device whose primary purpose is to introduce reactance into a circuit. Inductive reactance is frequently abbreviated inductor.

Resonance: The enhancement of the response of a physical system to a periodic excitation when the excitation frequency is equal to a natural frequency of the system.

Shunt: A device having appreciable impedance connected in parallel across other devices or apparatus and diverting some of the current from it. Appreciable voltage exists across the shunted device or apparatus, and an appreciable current may exist in it.

Shunt reactor: A reactor intended for connection in shunt to an electric system to draw inductive current.

Subsynchronous resonance: An electric power system condition where the electric network exchanges energy with a turbine generator at one or more of the natural frequencies of the combined system below the synchronous frequency of the system.

Thyristor: A bistable semiconductor device comprising three or more junctions that can be switched from the off state to the on state, or vice versa, such switching occurring within at least one quadrant of the principal voltage-current characteristic.

Related Topic

1.2 Capacitors and Inductors

References

I.S. Benko, B. Bhargava, and W.N. Rothenbuhler, "Prototype NGH subsynchronous resonance damping scheme, part II—Switching and short circuit tests," *IEEE Trans. Power Syst.*, vol. 2, pp. 1040–1049, 1987.

L.E. Bock and G.R. Mitchell, "Higher line loadings with series capacitors," *Transmission Magazine*, March 1973.

E.W. Bogins and H.T. Trojan, "Application and design of EHV shunt reactors," *Transmission Magazine*, March 1973.

C.E. Bowler, D.N. Ewart, and C. Concordia, "Self excited torsional frequency oscillations with series capacitors," *IEEE Trans. Power Appar. Syst.,* vol. 93, pp. 1688–1695, 1973.

G.D. Brewer, H.M. Rustebakke, R.A. Gibley, and H.O. Simmons, "The use of series capacitors to obtain maximum EHV transmission capability," *IEEE Trans. Power Appar. Syst.,* vol. 83, pp. 1090–1102, 1964.

C. Concordia, "System compensation, an overview," *Transmission Magazine*, March 1973.

S.E.M. de Oliveira, I. Gardos, and E.P. Fonseca, "Representation of series capacitors in electric power system stability studies," *IEEE Trans. Power Syst.,* vol. 6, no. 3, pp. 1119–1125, 1991.

A.A. Edris, "Series compensation schemes reducing the potential of subsynchronous resonance," *IEEE Trans. Power Syst.,* vol. 5, no. 1, pp. 219–226, 1990.

M.E. El-Hawary, *Electrical Power Systems: Design and Analysis,* Piscataway, N.J.: IEEE Press, 1995.

A.M. El-Serafi and A. A. Shaltout, "Damping of SSR Oscillations by Excitation Control," *IEEE PES Summer Meeting*, Vancouver, 1979.

A.M. Gole and V.K. Sood, "A static compensator model for use with electromagnetic transients simulation programs," *IEEE Trans. Power Delivery*, vol. 5, pp. 1398–1407, 1990.

L. Gyugyi, R.A. Otto, and T.H. Putman, "Principles and applications of static thyristor-controlled shunt compensators," *IEEE Trans. Power Appar. Syst.*, vol. PAS-97, pp. 1935–1945, 1978.

L. Gyugyi and E.R. Taylor, Jr., "Characteristics of static thyristor-controlled shunt compensators for power transmission system applications," *IEEE Trans. Power Appar. Syst.*, vol. PAS-99, pp. 1795–1804, 1980.

A.E. Hammad, "Analysis of power system stability enhancement by static VAR compensators," *IEEE Trans. Power Syst.*, vol. 1, pp. 222–227, 1986.

J.F. Hauer, "Robust damping controls for large power systems," *IEEE Control Systems Magazine*, pp. 12–18, January 1989.

R.A. Hedin, K.B. Stump, and N.G. Hingorani, "A new scheme for subsynchronous resonance damping of torsional oscillations and transient torque—Part II," *IEEE Trans. Power Appar. Syst.,* vol. PAS-100, pp. 1856–1863, 1981.

N.G. Hingorani, "A new scheme for subsynchronous resonance damping of torsional oscillations and transient torque—Part I," *IEEE Trans. Power Appar. Syst.,* vol. PAS-100, pp. 1852–1855, 1981.

N.G. Hingorani, B. Bhargava, G.F. Garrigue, and G.D. Rodriguez, "Prototype NGH subsynchronous resonance damping scheme, part I—Field installation and operating experience," *IEEE Trans. Power Syst.,* vol. 2, pp. 1034–1039, 1987.

N.G. Hingorani, "Flexible AC transmission," *IEEE Spectrum.* vol. 30, no. 4, pp. 40–45, 1993.

IEEE Subsynchronous Resonance Working Group, "Proposed terms and definitions for subsynchronous oscillations," *IEEE Trans. Power Appar. Syst.,* vol. PAS-99, pp. 506–511, 1980.

IEEE Subsynchronous Resonance Working Group, "Countermeasures to subsynchronous resonance problems," *IEEE Trans. Power Appar. Syst.,* vol. PAS-99, pp. 1810–1818, 1980.

IEEE Subsynchronous Resonance Working Group, "Series capacitor controls and settings as countermeasures to subsynchronous resonance," *IEEE Trans. Power Appar. Syst.,* vol. PAS-101, pp. 1281–1287, June 1982.

G. Jancke, N. Fahlen, and O. Nerf, "Series capacitors in power systems," *IEEE Transactions on Power Appar. Syst.*, vol. PAS-94, pp. 915–925, May/June 1975.

L.A. Kilgore, D.G. Ramey, and W.H. South, "Dynamic filter and other solutions to the subsynchronous resonance problem," *Proceedings of the American Power Conference*, vol. 37, p. 923, 1975.

E.W. Kimbark, *Power System Stability,* vol. I, *Elements of Stability Calculations,* New York: Wiley, 1948.

E.W. Kimbark, "Improvement of system stability by switched series capacitors," *IEEE Trans. Power Appar. Syst.*, vol. 85, pp. 180–188, February 1966.

J.J. LaForest, K.W. Priest, Ramirez, and H. Nowak, "Resonant voltages on reactor compensated extra-high-voltage lines," *IEEE Trans. Power Appar. Syst.*, vol. PAS-91, pp. 2528–2536, November/December 1972.

E.V. Larsen, D.H. Baker, A.F. Imece, L. Gerin-Lajoie, and G. Scott, "Basic aspects of applying SVC's to series-compensated ac transmission lines," *IEEE Trans. Power Delivery*, vol. 5, pp. 1466–1472, July 1990.

S. Lefebvre and L. Gerin-Lajoie, "A static compensator model for the EMTP," *IEEE Trans. Power Systems*, vol. 7, no. 2, pp. 477–486, May 1992.

E. Lerch, D. Povh, and L. Xu, "Advanced SVC control for damping power system oscillations," *IEEE Trans. Power Syst.*, vol. 6, pp. 524–531, May 1991.

S.M. Merry and E.R. Taylor, "Overvoltages and harmonics on EHV systems," *IEEE Trans. Power Appar. Syst.,* vol. PAS-91, pp. 2537–2544, November/December 1972.
A. Olwegard, K. Walve, G. Waglund, H. Frank, and S. Torseng, "Improvement of transmission capacity by thyristor controlled reactive power," *IEEE Trans. Power Appar. Syst.*, vol. PAS-100, pp. 3930–3939, 1981.
J.B. Tice and C.E.J. Bowler, "Control of phenomenon of subsynchronous resonance," *Proceedings of the American Power Conference*, vol. 37, pp. 916–922, 1975.
B.M. Weedy, *Electric Power Systems*, London: Wiley, 1972.

Further Information

An excellent source of information on the application of capacitors on power systems is the Westinghouse *Transmission and Distribution* book, published in 1964. A most readable treatment of improving system stability by series capacitors is given by Kimbark's paper [1966]. Jancke et al. [1975] give a detailed discussion of experience with the 400-kV series-capacitor compensation installations on the Swedish system and aspects of the protection system. Hauer [1989] presents a discussion of practical stability controllers that manipulate series and/or shunt reactance.

An excellent summary of the state of the art in static VAR compensators is the record of the IEEE Working Group symposium conducted in 1987 on the subject (see IEEE Publication 87TH0187-5-PWR, Application of Static VAR Systems for System Dynamic Performance).

For state-of-the-art coverage of subsynchronous resonance and countermeasures, two symposia are available: IEEE Publication 79TH0059-6-PWR, State-of-the-Art Symposium—Turbine Generator Shaft Torsionals, and IEEE Publication 81TH0086-9-PWR, Symposium on Countermeasures for Subsynchronous Resonance.

61.5 Fault Analysis in Power Systems

Charles Gross

A **fault** in an electrical power system is the unintentional and undesirable creation of a conducting path (a *short circuit*) or a blockage of current (an *open circuit*). The short-circuit fault is typically the most common and is usually implied when most people use the term *fault*. We restrict our comments to the short-circuit fault.

The causes of faults include lightning, wind damage, trees falling across lines, vehicles colliding with towers or poles, birds shorting out lines, aircraft colliding with lines, vandalism, small animals entering switchgear, and line breaks due to excessive ice loading. Power system faults may be categorized as one of four types: single line-to-ground, line-to-line, double line-to-ground, and balanced three-phase. The first three types constitute severe unbalanced operating conditions.

It is important to determine the values of system voltages and currents during faulted conditions so that protective devices may be set to detect and minimize their harmful effects. The time constants of the associated transients are such that sinusoidal steady-state methods may still be used. The method of symmetrical components is particularly suited to fault analysis.

Our objective is to understand how symmetrical components may be applied specifically to the four general fault types mentioned and how the method can be extended to any unbalanced three-phase system problem.

Note that phase values are indicated by subscripts, *a, b, c;* sequence (symmetrical component) values are indicated by subscripts 0, 1, 2. The transformation is defined by

$$\begin{bmatrix} \overline{V}_a \\ \overline{V}_b \\ \overline{V}_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix} \begin{bmatrix} \overline{V}_0 \\ \overline{V}_1 \\ \overline{V}_2 \end{bmatrix} = [T] \begin{bmatrix} \overline{V}_a \\ \overline{V}_b \\ \overline{V}_c \end{bmatrix}$$

$$\begin{bmatrix} \overline{V}_0 \\ \overline{V}_1 \\ \overline{V}_2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix} \begin{bmatrix} \overline{V}_0 \\ \overline{V}_1 \\ \overline{V}_2 \end{bmatrix} = [T]^{-1} \begin{bmatrix} \overline{V}_a \\ \overline{V}_b \\ \overline{V}_c \end{bmatrix}$$

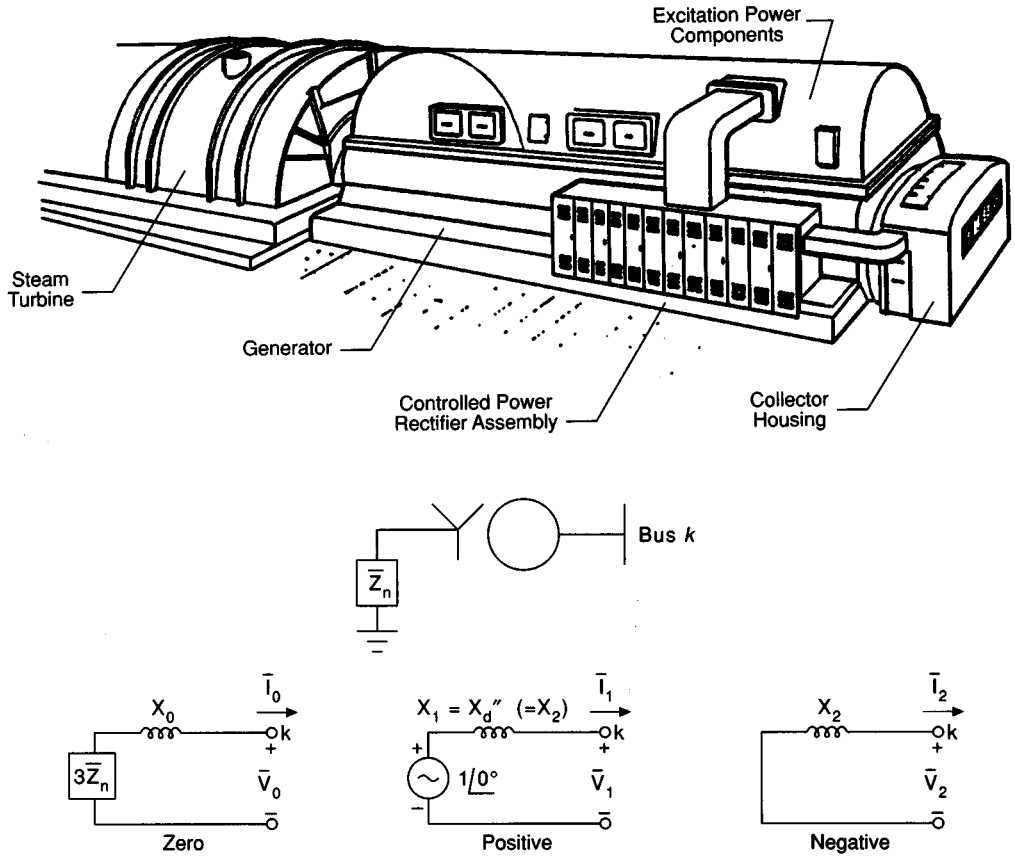


FIGURE 61.30 Generator sequence circuit models.

Simplifications in the System Model

Certain simplifications are possible and usually employed in fault analysis.

- Transformer magnetizing current and core loss will be neglected.
- Line shunt capacitance is neglected.
- Sinusoidal steady-state circuit analysis techniques are used. The so-called **dc offset** is accounted for by using correction factors.
- Prefault voltage is assumed to be $1\angle 0°$ per-unit. One per-unit voltage is at its nominal value prior to the application of a fault, which is reasonable. The selection of zero phase is arbitrary and convenient. Prefault load current is neglected.

For hand calculations, neglect series resistance is usually neglected (this approximation will not be necessary for a computer solution). Also, the only difference in the positive and negative sequence networks is introduced by the machine impedances. If we select the subtransient reactance X_d'' for the positive sequence reactance, the difference is slight (in fact, the two are identical for nonsalient machines). The simplification is important, since it reduces computer storage requirements by roughly one-third. Circuit models for generators, lines, and transformers are shown in Figs. 61.30, 61.31, and 61.32, respectively.

Our basic approach to the problem is to consider the general situation suggested in Fig. 61.33(a). The general terminals brought out are for purposes of external connections that will simulate faults. Note carefully the positive assignments of phase quantities. Particularly note that the currents flow *out of* the system. We can

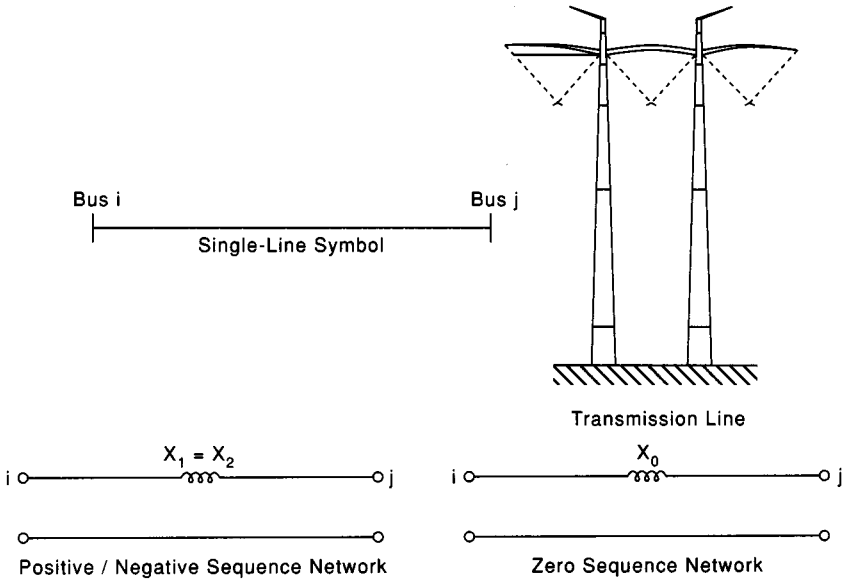


FIGURE 61.31 Line sequence circuit models.

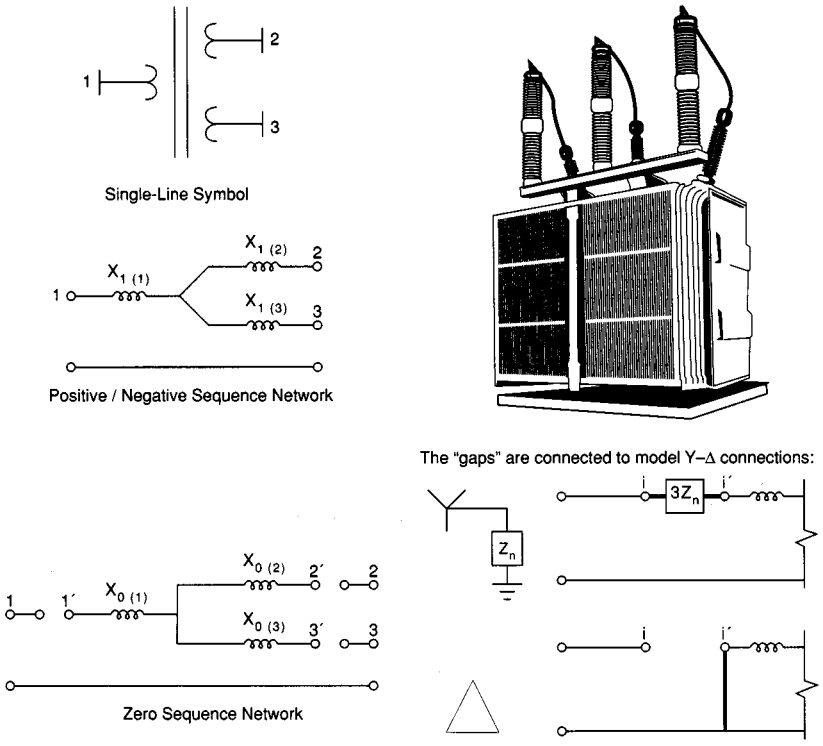


FIGURE 61.32 Transformer sequence circuit models.

construct general *sequence* equivalent circuits for the system, and such circuits are indicated in Fig. 61.33(b). The ports indicated correspond to the general three-phase entry port of Fig. 61.33(a). The positive sense of sequence values is compatible with that used for phase values.

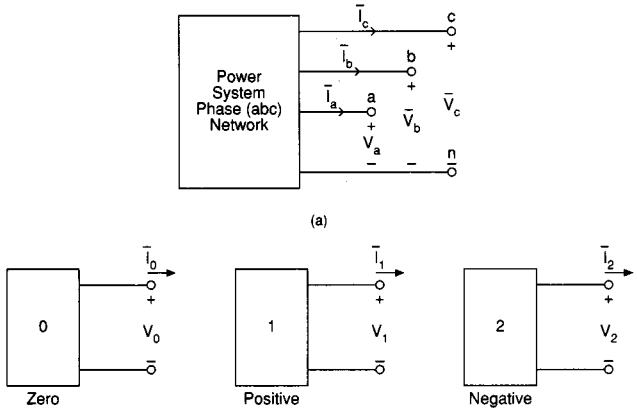


FIGURE 61.33 General fault port in an electric power system. (a) General fault port in phase (*abc*) coordinates; (b) corresponding fault ports in sequence (012) coordinates.

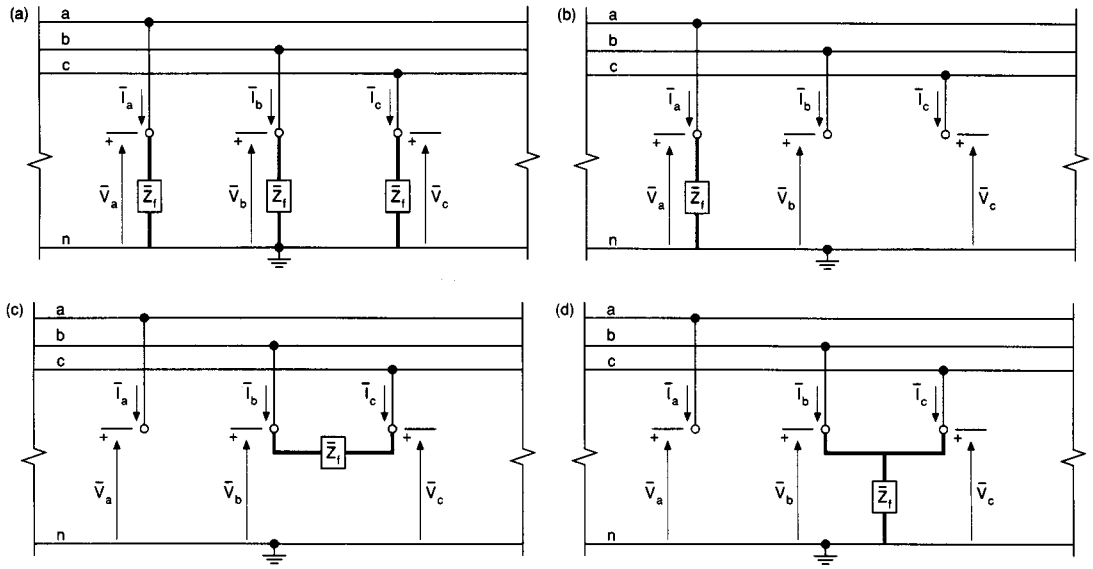


FIGURE 61.34 Fault types. (a) Three-phase fault; (b) single phase-to-ground fault; (c) phase-to-phase fault; (d) double phase-to-ground fault.

The Four Basic Fault Types

The Balanced Three-Phase Fault

Imagine the general three-phase access port terminated in a fault impedance ($\overline{Z}_f$) as shown in Fig. 61.34(a). The terminal conditions are

$$\begin{bmatrix} \overline{V}_a \\ \overline{V}_b \\ \overline{V}_c \end{bmatrix} = \begin{bmatrix} \overline{Z}_f & 0 & 0 \\ 0 & \overline{Z}_f & 0 \\ 0 & 0 & \overline{Z}_f \end{bmatrix} \begin{bmatrix} \overline{I}_a \\ \overline{I}_b \\ \overline{I}_c \end{bmatrix}$$

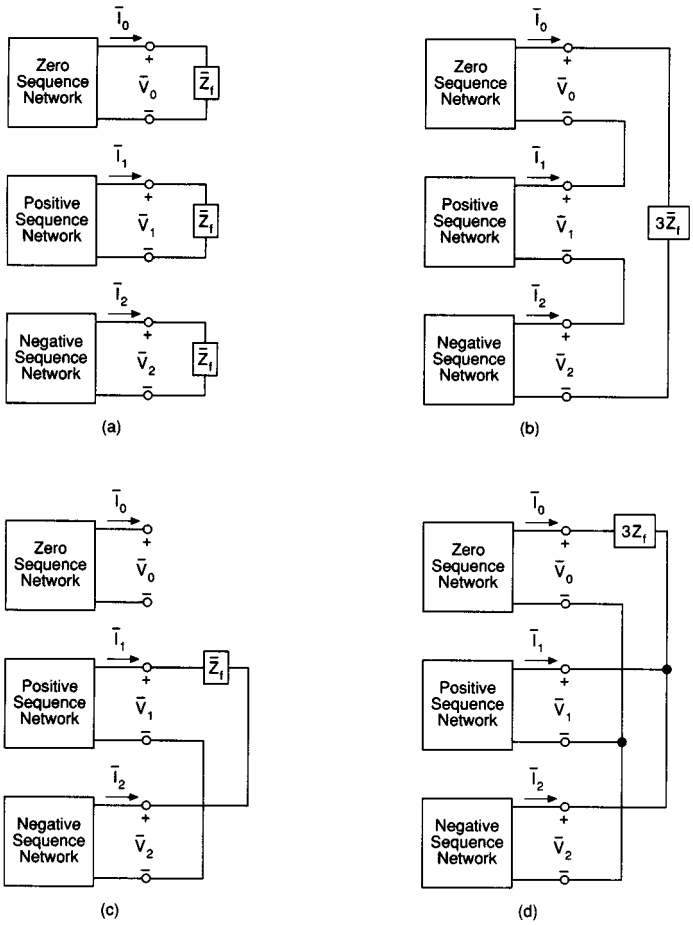


FIGURE 61.35 Sequence network terminations for fault types. (a) Balanced three-phase fault; (b) single phase-to-ground fault; (c) phase-to-phase fault; (d) double phase-to-ground fault.

Transforming to $[Z_{012}]$,

$$[Z_{012}] = [T]^{-1}\begin{bmatrix} \overline{Z}_f & 0 & 0 \\ 0 & \overline{Z}_f & 0 \\ 0 & 0 & \overline{Z}_f \end{bmatrix}[T] = \begin{bmatrix} \overline{Z}_f & 0 & 0 \\ 0 & \overline{Z}_f & 0 \\ 0 & 0 & \overline{Z}_f \end{bmatrix}$$

The corresponding network connections are given in Fig. 61.35(a). Since the zero and negative sequence networks are passive, only the positive sequence network is nontrivial.

$$\overline{V}_0 = \overline{V}_2 = 0 \tag{61.60}$$

$$\bar{I}_0 = \bar{I}_2 = 0 \tag{61.61}$$

$$\overline{V}_1 = \overline{Z}_f \bar{I}_1 \tag{61.62}$$

The Single Phase-to-Ground Fault

Imagine the general three-phase access port terminated as shown in Fig. 61.34(b). The terminal conditions are

$$\bar{I}_b = 0 \qquad \bar{I}_c = 0 \qquad \bar{V}_a = \bar{I}_a \bar{Z}_f$$

Therefore

$$\bar{I}_0 + a^2 \bar{I}_1 + a\bar{I}_2 = \bar{I}_0 + a\bar{I}_1 + a^2 \bar{I}_2 = 0$$

or

$$\bar{I}_1 = \bar{I}_2$$

Also

$$\bar{I}_b = \bar{I}_0 + a^2 \bar{I}_1 + a\bar{I}_2 = \bar{I}_0 + (a^2 + a)\bar{I}_1 = 0$$

or

$$\bar{I}_0 = \bar{I}_1 = \bar{I}_2 \tag{61.63}$$

Furthermore it is required that

$$\begin{aligned} \bar{V}_a &= \bar{Z}_f \bar{I}_a \\ \bar{V}_0 + \bar{V}_1 + \bar{V}_2 &= 3\bar{Z}_f \bar{I}_1 \end{aligned} \tag{61.64}$$

In general then, Eqs. (61.63) and (61.64) must be simultaneously satisfied. These conditions can be met by interconnecting the sequence networks as shown in Fig. 61.35(b).

The Phase-to-Phase Fault

Imagine the general three-phase access port terminated as shown in Fig. 61.34(c). The terminal conditions are such that we may write

$$\bar{I}_0 = 0 \qquad \bar{I}_b = -\bar{I}_c \qquad \bar{V}_b = \bar{Z}_f \bar{I}_b + \bar{V}_c$$

It follows that

$$\bar{I}_0 + \bar{I}_1 + \bar{I}_2 = 0 \tag{61.65}$$

$$\bar{I}_0 = 0 \tag{61.66}$$

$$\bar{I}_1 = -\bar{I}_2 \tag{61.67}$$

In general then, Eqs. (61.65), (61.66), and (61.67) must be simultaneously satisfied. The proper interconnection between sequence networks appears in Fig. 61.35(c).

The Double Phase-to-Ground Fault

Consider the general three-phase access port terminated as shown in Fig. 61.34(d). The terminal conditions indicate

$$\bar{I}_a = 0 \qquad \overline{V}_b = \overline{V}_c \qquad \overline{V}_b = (\bar{I}_b + \bar{I}_c)\overline{Z}_f$$

It follows that

$$\bar{I}_0 + \bar{I}_1 + \bar{I}_2 = \bar{0} \tag{61.68}$$

$$\overline{V}_1 = \overline{V}_2 \tag{61.69}$$

and

$$\overline{V}_0 - \overline{V}_1 = 3\overline{Z}_f \bar{I}_0 \tag{61.70}$$

For the general double phase-to-ground fault, Eqs. (61.68), (61.69), and (61.70) must be simultaneously satisfied. The sequence network interconnections appear in Fig. 61.35(d).

An Example Fault Study

Case: EXAMPLE SYSTEM
Run :
System has data for 2 Line(s); 2 Transformer(s);
4 Bus(es); and 2 Generator(s)

Transmission Line Data

Line	Bus	Bus	Seq	R	X	B	Srat
1	2	3	pos	0.00000	0.16000	0.00000	1.0000
			zero	0.00000	0.50000	0.00000	
2	2	3	pos	0.00000	0.16000	0.00000	1.0000
			zero	0.00000	0.50000	0.00000	

Transformer Data

Trans-former	HV Bus	LV Bus	Seq	R	X	C	Srat
1	2	1	pos	0.00000	0.05000	1.00000	1.0000
	Y	Y	zero	0.00000	0.05000		
2	3	4	pos	0.00000	0.05000	1.00000	1.0000
	Y	D	zero	0.00000	0.05000		

Generator Data

No.	Bus	Srated	Ra	Xd″	Xo	Rn	Xn	Con
1	1	1.0000	0.0000	0.200	0.0500	0.0000	0.0400	Y
2	4	1.0000	0.0000	0.200	0.0500	0.0000	0.0400	Y

Zero Sequence [Z] Matrix

0.0 + j(0.1144)	0.0 + j(0.0981)	0.0 + j(0.0163)	0.0 + j(0.0000)
0.0 + j(0.0981)	0.0 + j(0.1269)	0.0 + j(0.0212)	0.0 + j(0.0000)
0.0 + j(0.0163)	0.0 + j(0.0212)	0.0 + j(0.0452)	0.0 + j(0.0000)
0.0 + j(0.0000)	0.0 + j(0.0000)	0.0 + j(0.0000)	0.0 + j(0.1700)

Positive Sequence [Z] Matrix

0.0 + j(0.1310)	0.0 + j(0.1138)	0.0 + j(0.0862)	0.0 + j(0.0690)
0.0 + j(0.1138)	0.0 + j(0.1422)	0.0 + j(0.1078)	0.0 + j(0.0862)
0.0 + j(0.0862)	0.0 + j(0.1078)	0.0 + j(0.1422)	0.0 + j(0.1138)
0.0 + j(0.0690)	0.0 + j(0.0862)	0.0 + j(0.1138)	0.0 + j(0.1310)

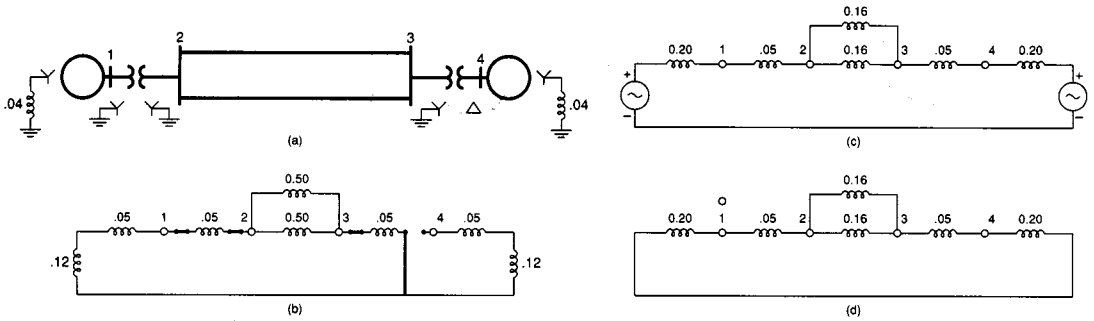


FIGURE 61.36 Example system. (a) Single-line diagram; (b) zero sequence network; (c) positive sequence network; (d) negative sequence network.

The single-line diagram and sequence networks are presented in Fig. 61.36.

Suppose bus 3 in the example system represents the fault location and $\bar{Z}_f = 0$. The positive sequence circuit can be reduced to its Thévenin equivalent at bus 3:

$$E_{T1} = 1.0\angle 0° \qquad \bar{Z}_{T1} = j0.1422$$

Similarly, the negative and zero sequence Thévenin elements are

$$\bar{E}_{T2} = 0 \qquad \bar{Z}_{T2} = j0.1422$$
$$\bar{E}_{T0} = 0 \qquad Z_{T0} = j0.0452$$

The network interconnections for the four fault types are shown in Fig. 61.37. For each of the fault types, compute the currents and voltages at the faulted bus.

Balanced Three-Phase Fault

The sequence networks are shown in Fig. 61.37(a). Obviously,

$$\bar{V}_0 = \bar{I}_0 = \bar{V}_2 = \bar{I}_2 = 0$$

$$\bar{I}_1 = \frac{1\angle 0°}{j0.1422} = -j7.032; \quad \text{also } \bar{V}_1 = 0$$

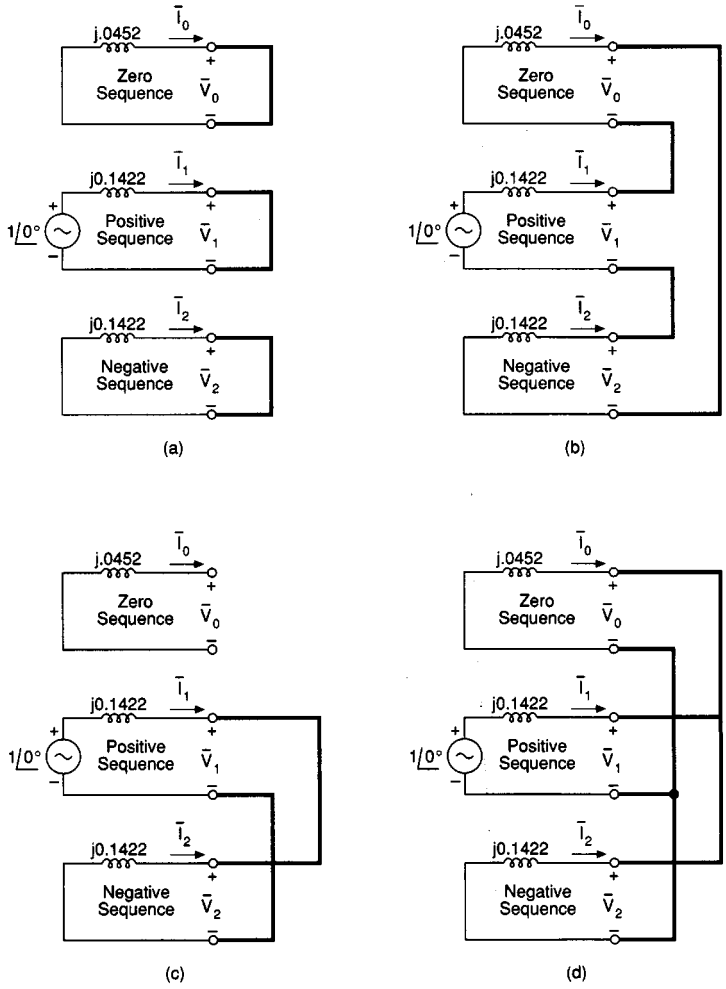


FIGURE 61.37 Example system faults at bus 3. (a) Balanced three-phase; (b) single phase-to-ground; (c) phase-to-phase; (d) double phase-to-ground.

To compute the phase values,

$$\begin{bmatrix} \bar{I}_a \\ \bar{I}_b \\ \bar{I}_c \end{bmatrix} = [T] \begin{bmatrix} \bar{I}_0 \\ \bar{I}_1 \\ \bar{I}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix} \begin{bmatrix} 0 \\ -j7.032 \\ 0 \end{bmatrix} = \begin{bmatrix} 7.032 \angle -90^\circ \\ 7.032 \angle 150^\circ \\ 7.032 \angle 30^\circ \end{bmatrix}$$

$$\begin{bmatrix} \bar{V}_a \\ \bar{V}_b \\ \bar{V}_c \end{bmatrix} = [T] \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Single Phase-to-Ground Fault

The sequence networks are interconnected as shown in Fig. 61.37(b).

$$\bar{I}_0 = \bar{I}_1 = \bar{I}_2 = \frac{1\angle 0^\circ}{j0.0452 + j0.1422 + j0.1422} = -j3.034$$

$$\begin{bmatrix} \bar{I}_a \\ \bar{I}_b \\ \bar{I}_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix} \begin{bmatrix} -j3.034 \\ -j3.034 \\ -j3.034 \end{bmatrix} = \begin{bmatrix} -j9.102 \\ 0 \\ 0 \end{bmatrix}$$

The sequence voltages are

$$\begin{aligned} \overline{V}_0 &= -j0.0452(-j3.034) = -1371 \\ \overline{V}_1 &= 1.0 - j0.1422(-j3.034) = 0.5685 \\ \overline{V}_2 &= -j0.1422(-j3.034) = -0.4314 \end{aligned}$$

The phase voltages are

$$\begin{bmatrix} \overline{V}_a \\ \overline{V}_b \\ \overline{V}_c \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{bmatrix} \begin{bmatrix} -0.1371 \\ 0.5685 \\ -0.4314 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.8901\angle -103.4^\circ \\ 0.8901\angle -103.4^\circ \end{bmatrix}$$

Phase-to-phase and double phase-to-ground fault values are calculated from the appropriate networks [Figs. 61.37(c) and (d)]. Complete results are provided.

Faulted Bus	Phase a	Phase b	Phase c
3	G	G	G

Sequence Voltages

Bus	V0		V1		V2	
1	0.0000/	0.0	0.3939/	0.0	0.0000/	0.0
2	0.0000/	0.0	0.2424/	0.0	0.0000/	0.0
3	0.0000/	0.0	0.0000/	0.0	0.0000/	0.0
4	0.0000/	0.0	0.2000/	–30.0	0.0000/	30.0

Phase Voltages

Bus	Va		Vb		Vc	
1	0.3939/	0.0	0.3939/	–120.0	0.3939/	120.0
2	0.2424/	0.0	0.2424/	–120.0	0.2424/	120.0
3	0.0000/	6.5	0.0000/	–151.2	0.0000/	133.8
4	0.2000/	–30.0	0.2000/	–150.0	0.2000/	90.0

Sequence Currents

Bus to Bus		I0		I1		I2	
1	2	0.0000/	167.8	3.0303/	–90.0	0.0000/	90.0
1	0	0.0000/	–12.2	3.0303/	90.0	0.0000/	–90.0
2	3	0.0000/	167.8	1.5152/	–90.0	0.0000/	90.0
2	3	0.0000/	167.8	1.5152/	–90.0	0.0000/	90.0
2	1	0.0000/	–12.2	3.0303/	90.0	0.0000/	–90.0
3	2	0.0000/	–12.2	1.5152/	90.0	0.0000/	–90.0
3	2	0.0000/	–12.2	1.5152/	90.0	0.0000/	–90.0
3	4	0.0000/	–12.2	4.0000/	90.0	0.0000/	–90.0
4	3	0.0000/	0.0	4.0000/	–120.0	0.0000/	120.0
4	0	0.0000/	0.0	4.0000/	60.0	0.0000/	–60.0

Faulted Bus	Phase a	Phase b	Phase c
3	G	G	G

Phase Currents

Bus to Bus		Ia		Ib		Ic	
1	2	3.0303/	–90.0	3.0303/	150.0	3.0303/	30.0
1	0	3.0303/	90.0	3.0303/	–30.0	3.0303/	–150.0
2	3	1.5151/	–90.0	1.5151/	150.0	1.5151/	30.0
2	3	1.5151/	–90.0	1.5151/	150.0	1.5151/	30.0
2	1	3.0303/	90.0	3.0303/	–30.0	3.0303/	–150.0
3	2	1.5151/	90.0	1.5151/	–30.0	1.5151/	–150.0
3	2	1.5151/	90.0	1.5151/	–30.0	1.5151/	–150.0
3	4	4.0000/	90.0	4.0000/	–30.0	4.0000/	–150.0
4	3	4.0000/	–120.0	4.0000/	120.0	4.0000/	–0.0
4	0	4.0000/	60.0	4.0000/	–60.0	4.0000/	–180.0

Faulted Bus	Phase a	Phase b	Phase c
3	G	0	0

Sequence Voltages

Bus	V0		V1		V2	
1	0.0496/	180.0	0.7385/	0.0	0.2615/	180.0
2	0.0642/	180.0	0.6731/	0.0	0.3269/	180.0
3	0.1371/	180.0	0.5685/	0.0	0.4315/	180.0
4	0.0000/	0.0	0.6548/	–30.0	0.3452/	210.0

Phase Voltages

Bus	Va		Vb		Vc	
1	0.4274/	0.0	0.9127/	–108.4	0.9127/	108.4
2	0.2821/	0.0	0.8979/	–105.3	0.8979/	105.3
3	0.0000/	89.2	0.8901/	–103.4	0.8901/	103.4
4	0.5674/	–61.8	0.5674/	–118.2	1.0000/	90.0

Sequence Currents

Bus	to Bus	I0		I1		I2	
1	2	0.2917/	–90.0	1.3075/	–90.0	1.3075/	–90.0
1	0	0.2917/	90.0	1.3075/	90.0	1.3075/	90.0
2	3	0.1458/	–90.0	0.6537/	–90.0	0.6537/	–90.0
2	3	0.1458/	–90.0	0.6537/	–90.0	0.6537/	–90.0
2	1	0.2917/	90.0	1.3075/	90.0	1.3075/	90.0
3	2	0.1458/	90.0	0.6537/	90.0	0.6537/	90.0
3	2	0.1458/	90.0	0.6537/	90.0	0.6537/	90.0
3	4	2.7416/	90.0	1.7258/	90.0	1.7258/	90.0
4	3	0.0000/	0.0	1.7258/	–120.0	1.7258/	–60.0
4	0	0.0000/	90.0	1.7258/	60.0	1.7258/	120.0

Faulted Bus	Phase a	Phase b	Phase c
3	G	0	0

Phase Currents

Bus	to Bus	Ia		Ib		Ic	
1	2	2.9066/	–90.0	1.0158/	90.0	1.0158/	90.0
1	0	2.9066/	90.0	1.0158/	–90.0	1.0158/	–90.0
2	3	1.4533/	–90.0	0.5079/	90.0	0.5079/	90.0
2	3	1.4533/	–90.0	0.5079/	90.0	0.5079/	90.0
2	1	2.9066/	90.0	1.0158/	–90.0	1.0158/	–90.0
3	2	1.4533/	90.0	0.5079/	–90.0	0.5079/	–90 0
3	2	1.4533/	90.0	0.5079/	–90.0	0.5079/	–90 0
3	4	6.1933/	90.0	1.0158/	90.0	1.0158/	90.0
4	3	2.9892/	–90.0	2.9892/	90.0	0.0000/	–90.0
4	0	2.9892/	90.0	2.9892/	–90.0	0.0000/	90.0

Faulted Bus	Phase a	Phase b	Phase c
3	0	C	B

Sequence Voltages

Bus	V0		V1		V2	
1	0.0000/	0.0	0.6970/	0.0	0.3030/	0.0
2	0.0000/	0.0	0.6212/	0.0	0.3788/	0.0
3	0.0000/	0.0	0.5000/	0.0	0.5000/	0.0
4	0.0000/	0.0	0.6000/	–30.0	0.4000/	30.0

Phase Voltages

Bus	Va		Vb		Vc	
1	1.0000/	0.0	0.6053/	–145.7	0.6053/	145.7
2	1.0000/	0.0	0.5423/	–157.2	0.5423/	157.2
3	1.0000/	0.0	0.5000/	–180.0	0.5000/	–180.0
4	0.8718/	–6.6	0.8718/	–173.4	0.2000/	90.0

Sequence Currents

Bus to Bus		I0		I1		I2	
1	2	0.0000/	–61.0	1.5152/	–90.0	1.5152/	90.0
1	0	0.0000/	119.0	1.5152/	90.0	1.5152/	–90.0
2	3	0.0000/	–61.0	0.7576/	–90.0	0.7576/	90.0
2	3	0.0000/	–61.0	0.7576/	–90.0	0.7576/	90.0
2	1	0.0000/	119.0	1.5152/	90.0	1.5152/	–90.0
3	2	0.0000/	119.0	0.7576/	90.0	0.7576/	–90.0
3	2	0.0000/	119.0	0.7576/	90.0	0.7576/	–90.0
3	4	0.0000/	119.0	2.0000/	90.0	2.0000/	–90.0
4	3	0.0000/	0.0	2.0000/	–120.0	2.0000/	120.0
4	0	0.0000/	90.0	2.0000/	60.0	2.0000/	–60.0

Faulted Bus	Phase a	Phase b	Phase c
3	0	C	B

Phase Currents

Bus to Bus		Ia		Ib		Ic	
1	2	0.0000/	180.0	2.6243/	180.0	2.6243/	0.0
1	0	0.0000/	180.0	2.6243/	0.0	2.6243/	180.0
2	3	0.0000/	–180.0	1.3122/	180.0	1.3122/	0.0
2	3	0.0000/	–180.0	1.3122/	180.0	1.3122/	0.0
2	1	0.0000/	180.0	2.6243/	0.0	2.6243/	180.0
3	2	0.0000/	–180.0	1.3122/	0.0	1.3122/	180.0
3	2	0.0000/	–180.0	1.3122/	0.0	1.3122/	180.0
3	4	0.0000/	–180.0	3.4641/	0.0	3.4641/	180.0
4	3	2.0000/	–180.0	2.0000/	180.0	4.0000/	0.0
4	0	2.0000/	0.0	2.0000/	0.0	4.0000/	– 180.0

Faulted Bus	Phase a	Phase b	Phase c
3	0	G	G

Sequence Voltages

Bus	V0		V1		V2	
1	0.0703/	0.0	0.5117/	0.0	0.1177/	0.0
2	0.0909/	0.0	0.3896/	0.0	0.1472/	0.0
3	0.1943/	–0.0	0.1943/	0.0	0.1943/	0.0
4	0.0000/	0.0	0.3554/	–30.0	0.1554/	30.0

Phase Voltages

Bus	Va		Vb		Vc	
1	0.6997/	0.0	0.4197/	–125.6	0.4197/	125.6
2	0.6277/	0.0	0.2749/	–130.2	0.2749/	130.2
3	0.5828/	0.0	0.0000/	–30.7	0.0000/	–139.6
4	0.4536/	–12.7	0.4536/	–167.3	0.2000/	90.0

Sequence Currents

Bus to Bus		I0		I1		I2	
1	2	0.4133/	90.0	2.4416/	– 90.0	0.5887/	90.0
1	0	0.4133/	–90.0	2.4416/	90.0	0.5887/	–90.0
2	3	0.2067/	90.0	1.2208/	– 90.0	0.2943/	90.0
2	3	0.2067/	90.0	1.2208/	– 90.0	0.2943/	90.0
2	1	0.4133/	–90.0	2.4416/	90.0	0.5887/	–90.0
3	2	0.2067/	– 90.0	1.2208/	90.0	0.2943/	– 90.0
3	2	0.2067/	– 90.0	1.2208/	90.0	0.2943/	– 90.0
3	4	3.8854/	– 90.0	3.2229/	90.0	0.7771/	– 90.0
4	3	0.0000/	0.0	3.2229/	– 120.0	0.7771/	120.0
4	0	0.0000/	–90.0	3.2229/	60.0	0.7771/	–60.0

Faulted Bus	Phase a	Phase b	Phase c
3	0	G	G

Phase Currents

Bus to Bus		Ia		Ib		Ic	
1	2	1.4396/	–90.0	2.9465/	153.0	2.9465/	27.0
1	0	1.4396/	90.0	2.9465/	–27.0	2.9465/	–153.0
2	3	0.7198/	–90.0	1.4733/	153.0	1.4733/	27.0
2	3	0.7198/	–90.0	1.4733/	153.0	1.4733/	27.0
2	1	1.4396/	90.0	2.9465/	–27.0	2.9465/	–153.0
3	2	0.7198/	90.0	1.4733/	–27.0	1.4733/	–153.0
3	2	0.7198/	90.0	1.4733/	–27.0	1.4733/	– 153.0
3	4	1.4396/	–90.0	6.1721/	–55.9	6.1721/	–124.1
4	3	2.9132/	–133.4	2.9132/	133.4	4.0000/	–0.0
4	0	2.9132/	46.6	2.9132/	–46.6	4.0000/	–180.0

Further Considerations

Generators are not the only sources in the system. All rotating machines are capable of contributing to fault current, at least momentarily. Synchronous and induction motors will continue to rotate due to inertia and function as sources of fault current. The impedance used for such machines is usually the transient reactance X'_d or the subtransient X''_d, depending on protective equipment and speed of response. Frequently motors smaller than 50 hp are neglected. Connecting systems are modeled with their Thévenin equivalents.

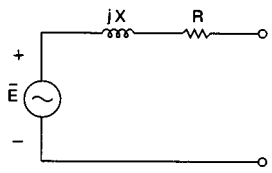


FIGURE 61.38 Positive sequence circuit looking back into faulted bus.

Although we have used ac circuit techniques to calculate faults, the problem is fundamentally transient since it involves sudden switching actions. Consider the so-called dc offset current. We model the system by determining its positive sequence Thévenin equivalent circuit, looking back into the positive sequence network at the fault, as shown in Fig. 61.38. The transient fault current is

$$i(t) = I_{ac}\sqrt{2}\cos(\omega t - \beta) + I_{dc}e^{-t/\tau}$$

This is a first-order approximation and strictly applies only to the three-phase or phase-to-phase fault. Ground faults would involve the zero sequence network also.

$$I_{ac} = \frac{E}{\sqrt{R^2 + X^2}} = \text{rms ac current}$$
$$I_{dc}(t) = I_{dc}e^{-t/\tau} = \text{dc offset current}$$

The maximum initial dc offset possible would be

$$\text{Max } I_{dc} = I_{max} = \sqrt{2}I_{ac}$$

The dc offset will exponentially decay with time constant τ, where

$$\tau = \frac{L}{R} = \frac{X}{\omega R}$$

The maximum dc offset current would be $I_{dc}(t)$

$$I_{dc}(t) = I_{dc}e^{-t/\tau} = \sqrt{2}I_{ac}e^{-t/\tau}$$

The *transient rms* current $I(t)$, accounting for both the ac and dc terms, would be

$$I(t) = \sqrt{I_{ac}^2 + I_{dc}^2(t)} = I_{ac}\sqrt{1 + 2e^{-2t/\tau}}$$

Define a multiplying factor k_i such that I_{ac} is to be multiplied by k_i to estimate the interrupting capacity of a breaker which operates in time T_{op}. Therefore,

$$k_i = \frac{I(T_{op})}{I_{ac}} = \sqrt{1 + 2e^{-2T_{op}/\tau}}$$

Observe that the maximum possible value for k_i is $\sqrt{3}$.

Example

In the circuit of Fig. 61.38, $E = 2400$ V, $X = 2\ \Omega$, $R = 0.1\ \Omega$, and $f = 60$ Hz. Compute k_i and determine the interrupting capacity for the circuit breaker if it is designed to operate in two cycles. The fault is applied at $t = 0$.

Solution

$$I_{ac} \cong \frac{2400}{2} = 1200 \text{ A}$$
$$T_{op} = \frac{2}{60} = 0.0333 \text{ s}$$
$$\tau = \frac{X}{\omega R} = \frac{2}{37.7} = 0.053$$
$$k_i = \sqrt{1 + 2e^{-2T_{op}/\tau}} = \sqrt{1 + 2e^{-0.0067/0.053}} = 1.252$$

Therefore

$$I = k_i I_{ac} = 1.252(1200) = 1503 \text{ A}$$

The Thévenin equivalent at the fault point is determined by normal sinusoidal steady-state methods, resulting in a first-order circuit as shown in Fig. 61.38. While this provides satisfactory results for the steady-state component I_{ac}, the *X/R* value so obtained can be in serious error when compared with the rate of decay of $I(t)$ as measured by oscillographs on an actual faulted system. The major reasons for the discrepancy are, first of all, that the system, for transient analysis purposes, is actually high-order, and second, the generators do not hold constant impedance as the transient decays.

Summary

Computation of fault currents in power systems is best done by computer. The major steps are summarized below:

- Collect, read in, and store machine, transformer, and line data in per-unit on common bases.
- Formulate the sequence impedance matrices.
- Define the faulted bus and Z_f. Specify type of fault to be analyzed.
- Compute the sequence voltages.
- Compute the sequence currents.
- Correct for wye-delta connections.
- Transform to phase currents and voltages.

For large systems, computer formulation of the sequence impedance matrices is required. Refer to Further Information for more detail. Zero sequence networks for lines in close proximity to each other (on a common right-of-way) will be mutually coupled. If we are willing to use the same values for positive and negative sequence machine impedances,

$$[Z_1] = [Z_2]$$

Therefore, it is unnecessary to store these values in separate arrays, simplifying the program and reducing the computer storage requirements significantly. The error introduced by this approximation is usually not important. The methods previously discussed neglect the prefault, or load, component of current; that is, the usual assumption is that currents throughout the system were zero prior to the fault. This is almost never strictly true; however, the error produced is small since the fault currents are generally much larger than the load currents. Also, the load currents and fault currents are out of phase with each other, making their sum more nearly equal to the larger components than would have been the case if the currents were in phase. In addition, selection of precise values for prefault currents is somewhat speculative, since there is no way of predicting what the loaded state of the system is when a fault occurs. When it is important to consider load currents, a power flow study is made to calculate currents throughout the system, and these values are superimposed on (added to) results from the fault study.

A term which has wide industrial use and acceptance is the *fault level* or **fault MVA** at a bus. It relates to the amount of current that can be expected to flow out of a bus into a three-phase fault. As such, it is an alternate way of providing positive sequence impedance information. Define

$$\begin{aligned}\text{Fault level in MVA at bus } i &= V_{i_{\text{pu nominal}}} I_{i_{\text{pu fault}}} S_{3\phi_{\text{base}}} \\ &= (1)\frac{1}{Z^1_{ii}} S_{3\phi_{\text{base}}} = \frac{S_{3\phi_{\text{base}}}}{Z^1_{ii}}\end{aligned}$$

Fault study results may be further refined by approximating the effect of dc offset.

The basic reason for making fault studies is to provide data that can be used to size and set protective devices. The role of such protective devices is to detect and remove faults to prevent or minimize damage to the power system.

Defining Terms

DC offset: The natural response component of the transient fault current, usually approximated with a first-order exponential expression.

Fault: An unintentional and undesirable conducting path in an electrical power system.

Fault MVA: At a specific location in a system, the initial symmetrical fault current multiplied by the prefault nominal line-to-neutral voltage (× 3 for a three-phase system).

Sequence (012) quantities: Symmetrical components computed from phase (*abc*) quantities. Can be voltages, currents, and/or impedances.

References

P. M. Anderson, *Analysis of Faulted Power Systems,* Ames: Iowa State Press, 1973.

M. E. El-Hawary, *Electric Power Systems: Design and Analysis,* Reston, Va.: Reston Publishing, 1983.

M. E. El-Hawary, *Electric Power Systems,* New York: IEEE Press, 1995.

O. I. Elgerd, *Electric Energy Systems Theory: An Introduction,* 2nd ed., New York: McGraw-Hill, 1982.

General Electric, *Short-Circuit Current Calculations for Industrial and Commercial Power Systems,* Publication GET-3550.

C. A. Gross, *Power System Analysis,* 2nd ed., New York: Wiley, 1986.

S. H. Horowitz, *Power System Relaying,* 2nd ed, New York: Wiley, 1995.

I. Lazar, *Electrical Systems Analysis and Design for Industrial Plants,* New York: McGraw-Hill, 1980.

C. R. Mason, *The Art and Science of Protective Relaying,* New York: Wiley, 1956.

J. R. Neuenswander, *Modern Power Systems,* Scranton, Pa.: International Textbook, 1971.

G. Stagg and A. H. El-Abiad, *Computer Methods in Power System Analysis,* New York: McGraw-Hill, 1968.

Westinghouse Electric Corporation, *Applied Protective Relaying,* Relay-Instrument Division, Newark, N.J., 1976.

A. J. Wood, *Power Generation, Operation, and Control,* New York: Wiley, 1996.

Further Information

For a comprehensive coverage of general fault analysis, see Paul M. Anderson, *Analysis of Faulted Power Systems,* New York, IEEE Press, 1995. Also see Chapters 9 and 10 of *Power System Analysis* by C.A. Gross, New York: Wiley, 1986.

61.6 Protection

Arun G. Phadke

Fundamental Principles of Protection

Protective equipment—**relays**—is designed to respond to system abnormalities (faults) such as short circuits. When faults occur, the relays must signal the appropriate circuit breakers to trip and isolate the faulted equipment. The protection systems not only protect the faulty equipment from more serious damage, they also protect the power system from the consequences of having faults remain on the system for too long. In modern high-voltage systems, the potential for damage to the power system—rather than to the individual equipment—is often far more serious, and power system security considerations dictate the design of the protective system. The protective system consists of four major subsystems as shown in Fig. 61.39. The **transducers** (*T*)

are current and voltage transformers, which transform high voltages and currents to a more manageable level. In the United States, the most common standard for current transformers is a secondary current of 5 A (or less) for steady-state conditions. In Europe, and in some other foreign countries, a 1-A standard is also common. The voltage transformer standard is 69.3 V line-to-neutral or 120 V line-to-line on the transformer secondary side. Standardization of the secondary current and voltage ratings of the transducers has permitted independent development of the transducers and relays. The power handling capability of the transducers is expressed in terms of the volt-ampere burden, which they can supply without significant waveform distortion. In general, the transient response of the transducers is much more critical in relaying applications.

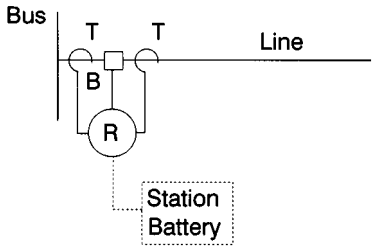


FIGURE 61.39 Elements of a protection system.

The second element of the protection system is the relay (*R*). This is the device that, using the current, voltage, and other inputs, can determine if a fault exists on the system, for which action on the part of the relay is needed. We will discuss relays in greater detail in the following. The third element of the protection chain is the circuit breaker (*B*), which does the actual job of interrupting the flow of current to the fault. Modern high-voltage circuit breakers are capable of interrupting currents of up to 100,000 A, against system voltages of up to 800,000 V, in about 15 to 30 ms. Lower-voltage circuit breakers are generally slower in operating speed. The last element of the protection chain is the station battery, which powers the relays and circuit breakers. The battery voltage has also been standardized at 125 V, although some other voltage levels may prevail in generating stations and in older substations.

The relays and circuit breakers must remove the faulted equipment from the system as quickly as possible. Also, if there are many alternative ways of deenergizing the faulty equipment, the protection system must choose a strategy that will remove from service the minimum amount of equipment. These ideas are embodied in the concepts of zones of protection, relay speed, and reliability of protection.

Zones of Protection

To make sure that a protection system removes the minimum amount of equipment from the power system during its operation, the power system is divided into zones of protection. Each zone has its associated protection system. A fault inside the zone causes the associated protection system to operate. A fault in any other zone must not cause an operation. A zone of protection usually covers one piece of equipment, such as a transmission line. The zone boundary is defined by the location of transducers (usually current transformers) and also by circuit breakers that will operate to isolate the zone. A set of zones of protection is shown in Fig. 61.40. Note that all zones are shown to overlap with their neighbors. This is to ensure that no point on the system is left unprotected. Occasionally, a circuit breaker may not exist at a zone boundary. In such cases, the tripping must be done at some other remote circuit breakers. For example, consider protection zone *A* in Fig. 61.40. A fault in that zone must be isolated by tripping circuit breakers *X* and *Y*. While the breaker *X* is near the transformer and can be tripped locally, *Y* is remote from the station, and some form of communication channel must be used to transfer the trip command to *Y*. Although most zones of protection have a precise extent, there are some zones that have a loosely defined reach. These are known as *open* zones and are most often encountered in transmission line protection.

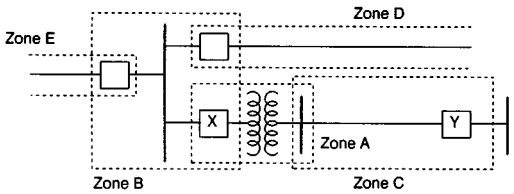


FIGURE 61.40 Zones of protection for a power system. Zones overlap; most zones are bounded by breakers.

Speed of Protection

The faster the operation of a protection function, the quicker is the prospect of removing a fault from the system. Thus, all protection systems are made as fast as possible. However, there are considerations that dictate

against making the protection faster than a minimum limit. Also, occasionally, it may be necessary to slow down a protection system in order to satisfy some specific system need. In general, the fastest protection available operates in about 5 to 10 ms after the inception of a fault [Thorp et al., 1979]. If the protection is made faster than this, it is likely to become "trigger happy" and operate falsely when it should not. When a protection system is intended as a backup system for some other protection, it is necessary to deliberately slow it down so that the primary protection may operate in its own time before the backup system will operate. This calls for a deliberate slowing of the backup protection. Depending upon the type of backup system being considered, the protection may sometimes be slowed down to operate in up to several seconds.

Reliability of Protection

In the field of relaying, **reliability** implies certain very specific concepts [Mason, 1956]. A reliable protection system has two attributes: *dependability* and *security*. A dependable relay is one that always operates for conditions for which it is designed to operate. A secure relay is one that will not operate for conditions for which it is not intended to operate. In modern power systems, the failure to operate when a fault occurs—lack of dependability—has very serious consequences for the power system. Therefore, most protective systems are made secure by duplicating relaying equipment, duplicating relaying functions, and providing several levels of backup protection. Thus modern systems tend to be very dependable, i.e., every fault is cleared, perhaps by more than one relay. As a consequence, security is somewhat degraded: modern protection systems will, occasionally, act and trip equipment falsely. Such occurrences are rare, but not uncommon. As power systems become leaner, i.e., they have insufficient margins of reserve generation and transmission, lack of security can be quite damaging. This has led to recent reevaluation of the proper balance between security and dependability of the protection systems.

Overcurrent Protection

The simplest fault detector is a sensor that measures the increase in current caused by the fault. The fuse is the simplest overcurrent protection; in fact, it is the complete protection chain—sensor, relay, and circuit breaker—in one package. Fuses are used in lower-voltage (distribution) circuits. They are difficult to set in high-voltage circuits, where load and fault currents may be of the same order of magnitude. Furthermore, they must be replaced when blown, which implies a long duration outage. They may also lead to system unbalances. However, when applicable, they are simple and inexpensive.

Inverse-Time Characteristic

Overcurrent relays sense the magnitude of the current in the circuit, and when it exceeds a preset value (known as the *pickup setting* of the relay), the relay closes its output contact, energizing the trip coil of the appropriate circuit breakers. The pickup setting must be set above the largest load current that the circuit may carry and must be smaller than the smallest fault current for which the relay must operate. A margin factor of 2 to 3 between the maximum load on the one hand and the minimum fault current on the other and the pickup setting of the relay is considered to be desirable. The overcurrent relays usually have an *inverse-time* characteristic as shown in Fig. 61.41. When the current exceeds the pickup setting, the relay operating time decreases in inverse proportion to the current magnitude. Besides this built-in feature in the relay mechanism, the relay also has a *time-dial* setting, which shifts the inverse-time curve vertically, allowing for more flexibility in setting the relays. The time dial has 11 discrete settings, usually labeled 1/2, 1, 2, . . ., 10, the lowest setting providing the fastest operation. The inverse-time characteristic offers an ideal relay for providing primary and backup protection in one package.

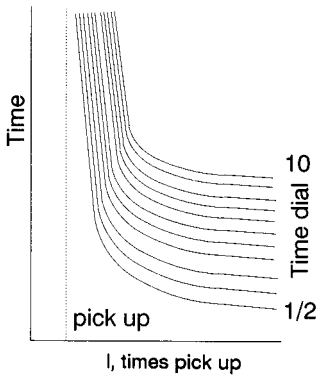


FIGURE 61.41 Inverse-time relay characteristic.

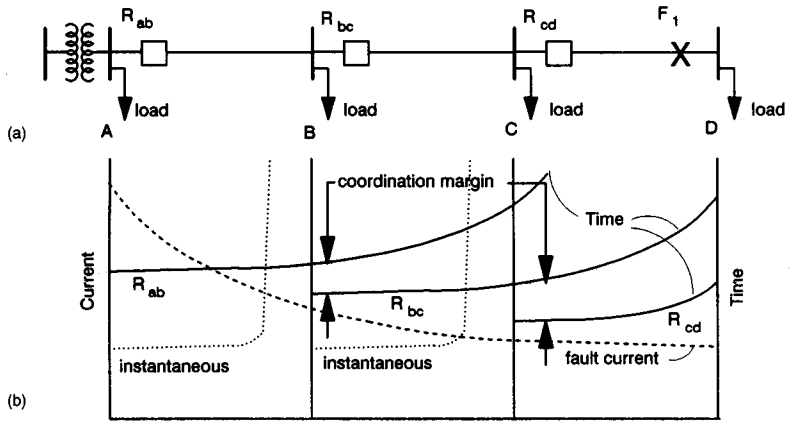


FIGURE 61.42 Coordination of inverse-time overcurrent and instantaneous relays for a radial system.

Coordination Principles

Consider the radial transmission system shown in Fig. 61.42. The transformer supplies power to the feeder, which has four loads at buses *A*, *B*, *C*, and *D*. For a fault at F_1, the relay R_{cd} must operate to open the circuit breaker B_{cd}. The relay R_{bc} is responsible for a zone of protection, which includes the entire zone of R_{cd}. This constitutes a remote backup for the protection at bus *C*. The backup relay (R_{bc}) must be slower than the primary relay (R_{cd}), its associated circuit breaker, with a safety margin. This delay in operating of the backup relay is known as the *coordination delay* and is usually about 0.3 s. In a similar fashion, R_{ab} backs up R_{bc}. The magnitude of the fault current varies as shown in Fig. 61.42(b), as the location of the fault is moved along the length of the feeder. We may plot the inverse time characteristic of the relay with the fault location as the abscissa, recalling that a smaller current magnitude gives rise to a longer operating time for the relay. The coordinating time delay between the primary and backup relays is also shown. It can be seen that, as we move from the far end of the feeder toward the source, the fault clearing time becomes progressively longer. The coordination is achieved by selecting relays with a time dial setting that will provide the proper separation in operating times.

The effect of cumulative coordination-time delays is slowest clearing of faults with the largest fault currents. This is not entirely satisfactory from the system point of view, and wherever possible, the inverse-time relays are supplemented by *instantaneous* overcurrent relays. These relays, as the name implies, have no intentional time delays and operate in less than one cycle. However, they cannot coordinate with the downstream relays and therefore must not operate ("see") for faults into the protection zone of the downstream relay. This criterion is not always possible to meet. However, whenever it can be met, instantaneous relays are used and provide a preferable compromise between fast fault clearing and coordinated backup protection.

Directional Overcurrent Relays

When power systems become meshed, as for most subtransmission and high-voltage transmission networks, inverse time overcurrent relays do not provide adequate protection under all conditions. The problem arises because the fault current can now be supplied from either end of the transmission line, and discrimination between faults inside and outside the zone of protection is not always possible. Consider the loop system shown in Fig. 61.43. Notice that in this system there must be a circuit breaker at each end of the line, as a fault on the line cannot be interrupted by opening one end alone. Zone *A* is the zone of protection for the line *A–D*. A fault at F_1 must be detected by the relays R_{ad} and R_{da}. The current through the circuit breaker B_{da} for the fault F_1 must be the determining quantity for the operation of the relay R_{da}. However, the impedances of the lines may be such that the current through the breaker B_{da} for the fault F_2 may be higher than the current for the fault F_1. Thus, if current magnitude alone is the criterion, the relay R_{da} would operate for fault F_2, as well as for the fault F_1. Of course, operation of R_{da} for F_2 is inappropriate, as it is outside its zone of protection, zone *A*. This

problem is solved by making the overcurrent relays directional. By this is meant that the relays will respond as overcurrent relays only if the fault is in the forward direction from the relays, i.e., in the direction in which their zone of protection extends. The directionality is provided by making the relay sensitive to the phase angle between the fault current and a reference quantity, such as the line voltage at the relay location. Other reference sources are also possible, including currents in the neutral of a transformer bank at the substation.

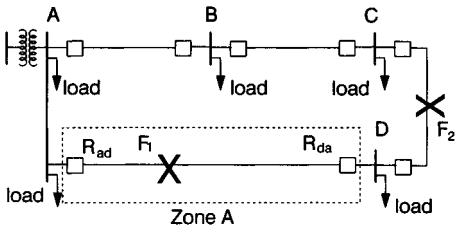


FIGURE 61.43 Protection of a loop (network) system with directional overcurrent relays.

Distance Protection

As the power networks become more complex, protection with directional overcurrent relays becomes even more difficult, if not impossible. Recall that the pickup setting of the relays must be set above the maximum load which the line is expected to carry. However, a network system has so many probable configurations due to various circuit breaker operations that the maximum load becomes difficult to define. For the same reason, the minimum fault current—the other defining parameter for the pickup setting—also becomes uncertain. Under these circumstances, the setting of the pickup of the overcurrent relays, and their reach, which will satisfy all the constraints, becomes impossible. Distance relays solve this problem.

Distance relays respond to a ratio of the voltage and current at the relay location. The ratio has the dimensions of an impedance, and the impedance between the relay location and fault point is proportional to the distance of the fault. As the zone boundary is related to the distance between the sending end and the receiving end of the transmission line, the distance to the fault forms an ideal relaying parameter. The distance is also a unique parameter in that it is independent of the current magnitude. It is thus free from most of the difficulties associated with the directional overcurrent relays mentioned above.

In a three-phase power system, 10 types of faults are possible: three single phase-to-ground faults, three phase-to-phase faults, three double phase-to-ground faults, and one three-phase fault. It turns out that relays responsive to the ratio of delta voltages and delta currents measure the correct distance to all multiphase faults. The delta quantities are defined as the difference between any two phase quantities; for example, $E_a - E_b$ is the delta voltage between a and b phases. Thus for a multiphase fault between phases x and y,

$$\frac{E_x - E_y}{I_x - I_y} = Z_1$$

where x and y can be a, b, or c and Z_1 is the positive sequence impedance between the relay location and the fault. For ground distance relays, the faulted phase voltage, and a compensated faulted phase current must be used

$$\frac{E_x}{I_x + mI_0} = Z_1$$

where m is a constant depending upon the line impedances and I_0 is the zero sequence current in the transmission line. A full complement of relays consists of three phase distance relays and three ground distance relays. As explained before, the phase relays are energized by the delta quantities, while the ground distance relays are energized by each of the phase voltages and the corresponding compensated phase currents. In many instances, ground distance protection is not preferred, and the time overcurrent relays may be used for ground fault protection.

Step-Distance Protection

The principle of distance measurement for faults is explained above. A relaying system utilizing that principle must take into account several features of the measurement principle and develop a complete protection scheme. Consider the system shown in Fig. 61.44. The distance relay R_{ab} must protect line AB, with its zone of protection as indicated by the dashed line. However, the distance calculation made by the relay is not precise enough for

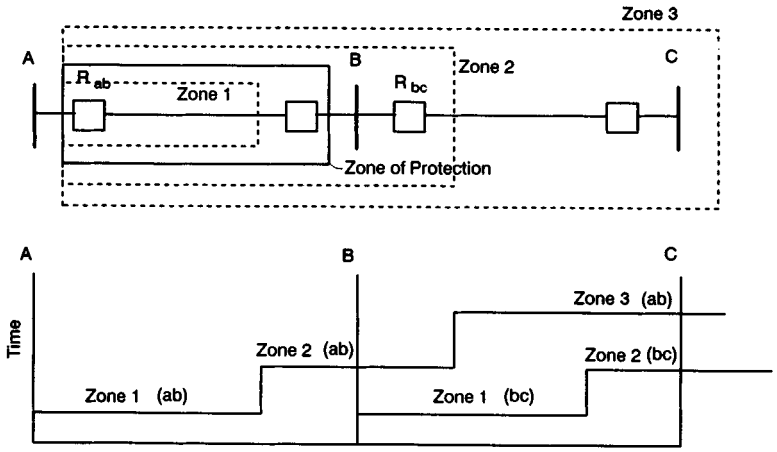


FIGURE 61.44 Zones of protection in a step-distance protection scheme. Zone 3 provides backup for the downstream line relays.

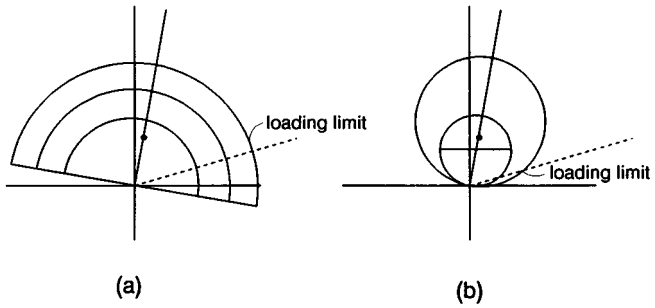


FIGURE 61.45 (a) Directional impedance characteristic. (b) Mho characteristic. Loadability limits as shown.

it to be able to distinguish between a fault just inside the zone and a fault just outside the zone, near bus *B*. This problem is solved by providing a two-zone scheme, such that if a fault is detected to be in zone 1, the relay trips instantaneously, and if the fault is detected to be inside zone 2, the relay trips with a time delay of about 0.3 s. Thus for faults near the zone boundary, the fault is cleared with this time delay, while for near faults, the clearing is instantaneous. This arrangement is referred to as a *step-distance* protection scheme, consisting of an underreaching zone (zone 1), and an overreaching zone (zone 2). The relays of the neighboring line (*BC*) can also be backed up by a third zone of the relay, which reaches beyond the zone of protection of relay R_{bc}. Zone 3 operation is delayed further to allow the zone 1 or zone 2 of R_{bc} to operate and clear the fault on line *BC*.

The distance relays may be further subdivided into categories depending upon the shape of their protection characteristics. The most commonly used relays have a directional distance, or a mho characteristic. The two characteristics are shown in Fig. 61.45. The directional impedance relay consists of two functions, a directional detection function and a distance measurement function. The mho characteristic is inherently directional, as the mho circle, by relay design, passes through the origin of the *RX* plane. Figure 61.45 also shows the multiple zones of the step distance protection.

Loadability of Distance Relays

The load carried by a transmission line translates into an apparent impedance as seen by the relay, given by

$$Z_{\text{app}} = \frac{|E|^2}{P - jQ}$$

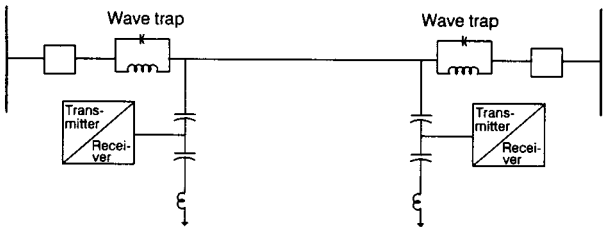


FIGURE 61.46 Carrier system for pilot protection of lines. Transmitter and receiver are connected to relays.

where $P–jQ$ is the load complex power and E is the voltage at the bus where a distance relay is connected. This impedance maps into the RX plane, as do all other apparent impedances, and hence the question arises whether this apparent load impedance could be mistaken for a fault by the distance relay. Clearly, this depends upon the shape of the distance relay characteristic employed. The loadability of a distance relay refers to the maximum load power (minimum apparent impedance) that the line can carry before a protective zone of a distance relay is penetrated by the apparent impedance. A typical load line is shown in Fig. 61.45. It is clear from this figure that the mho characteristic has a higher loadability than the directional impedance relay. In fact, other relay characteristics can be designed so that the loadability of a relay is increased even further.

Other Uses of Distance Relays

Although the primary use of distance relays is in protecting transmission lines, some other protection tasks can also be served by distance relays. For example, loss-of-field protection of generators is often based upon distance relays. Out-of-step relays and relays for protecting reactors may also be distance relays. Distance relays are also used in pilot protection schemes described next, and as backup relays for power apparatus.

Pilot Protection

Pilot protection of transmission lines uses communication channels (pilot channels) between the line terminals as an integral element of the protection system. In general, pilot schemes may be subdivided into categories according to the medium of communication used. For example, the pilot channels may be wire pilots, leased telephone circuits, dedicated telephone circuits, microwave channels, power line carriers, or fiber optic channels. Pilot protection schemes may also be categorized according to their function, such as a tripping pilot or a blocking pilot. In the former, the communication medium is used to send a tripping signal to a remote line terminal, while in the latter, the pilot channel is used to send a signal that prevents tripping at the remote terminal for faults outside the zone of protection of the relays. The power line carrier system is the most common system used in the United States. It uses a communication channel with a carrier signal frequency ranging between 30 and 300 kHz, the most common bands being around 100 kHz. The modulated carrier signal is coupled into one or more phases of the power line through coupling capacitors. In almost all cases, the capacitors of the capacitive-coupled voltage transformers are used for this function (see Fig. 61.46). The carrier signal is received at both the sending and the receiving ends of the transmission line by tuned receivers. The carrier signal is blocked from flowing into the rest of the power system by blocking filters, which are parallel resonant circuits, known as *wave traps*.

Coverage of 100% of Transmission Line

The step-distance scheme utilizes the zone 1 and zone 2 combination to protect 100% of the transmission line. The middle portion of the transmission line, which lies in zone 1 of relays at the two ends of the line, is protected at high speed from both ends. However, for faults in the remaining portion of the line, the near end clears the fault at high speed, i.e., in zone 1 time, while the remote end clears the fault in zone 2 time. In effect, such faults remain on the system for zone 2 time, which may be of the order 0.3 to 0.5 s. This becomes undesirable in modern power systems where the margin of stability may be quite limited. In any case, it is good protection practice to protect the entire line with high-speed clearing of all internal faults from both ends of the transmission line. Pilot protection accomplishes this task.

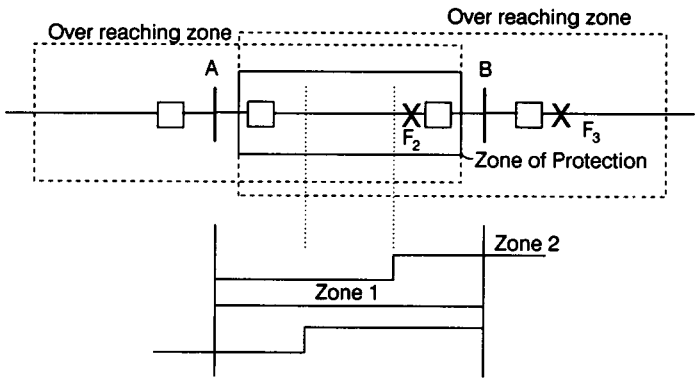


FIGURE 61.47 Pilot protection with overreaching zones of protection. This is most commonly used in a directional comparison blocking scheme.

Directional Comparison Blocking Scheme

Consider the fault at F_2 shown in Fig. 61.47. As discussed above, this fault will be cleared in zone 1 time by the step-distance relay at bus *B*, while the relay at bus *A* will clear the fault in zone 2 time. Since the relays at bus *B* can determine, with a high degree of certainty, that a fault such as F_2 is indeed inside the zone of protection of the relays, one could communicate this knowledge to terminal *A*, which can then cause the local circuit breaker to trip for the fault F_2. If the entire relaying and communication task can be accomplished quickly, 100% of the line can be protected at high speed. One of the most commonly used methods of achieving this function is to use overreaching zones of protection at both terminals, *and* if a fault is detected to be inside this zone, and if the remote terminal confirms that the fault is inside the zone of protection, then the local relay may be allowed to trip. In actual practice, the complement of this information is used to block the trip at the remote end. Thus, the remote end, terminal *B* in this case, detects faults that are outside the zone of protection and, for those faults, sends a signal which asks the relay at terminal *A* to block the tripping command. Thus, for a fault such as F_3, the relay at *A* will trip, unless the communication is received from terminal *B* that this particular fault is outside the zone of protection—as indeed fault F_3 happens to be. This mode, known as a blocking carrier, is preferred, since a loss of the carrier signal created by an internal fault, or due to causes that are unrelated to the fault, will not prevent the trip at the remote end. This is a highly dependable protection system, and precisely because of that it is somewhat less secure. Nevertheless, as discussed previously, most power systems require that a fault be removed as quickly as possible, even if in doing so for a few faults an unwarranted trip may result.

Other Pilot Protection Schemes

Several other types of pilot protection schemes are available. The choice of a specific scheme depends upon many factors. Some of these factors are importance of the line to the power system, the available communication medium, dependability of the communication medium, loading level of the transmission line, susceptibility of the system to transient stability oscillations, presence of series or shunt compensating devices, multiterminal lines, etc. A more complete discussion of all these issues will be found in the references [Westinghouse, 1982; Blackburn, 1987; Horowitz and Phadke, 1992].

Computer Relaying

Relaying with computers began to be discussed in technical literature in the mid-1960s. Initially, this was an academic exercise, as neither the computer speeds nor the computer costs could justify the use of computers for relaying. However, with the advent of high-performance microprocessors, computer relaying has become a very practical and attractive field of research and development. All major manufacturers of electric power equipment have computer relays to meet all the needs of power system engineers. Computer relaying is also being taught in several universities and has provided a very fertile field of research for graduate students.

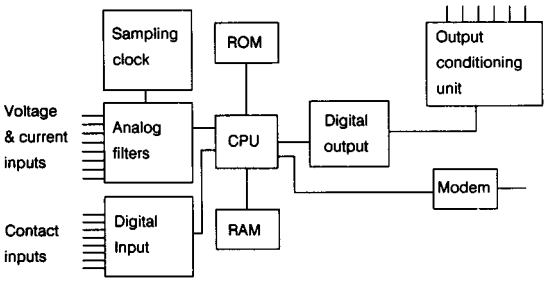


FIGURE 61.48 Block diagram of a computer relay architecture.

Computer relaying has also uncovered new ways of measuring power system parameters and may influence future development of power system monitoring and control functions.

Incentives for Computer Relaying

The acceptance of **computer relays** has been due to economic factors which have made microcomputers relatively inexpensive and computationally powerful. In addition to this economic advantage, the computer relays are also far more versatile. Through their self-diagnostic capability, they provide an assurance of availability. Thus, even if they should suffer the same (or even greater) number of failures in the field as traditional relays, their failures could be communicated to control centers and a maintenance crew called to repair the failures immediately. This type of diagnostic capability was lacking in traditional protection systems and often led to failures of relays, which went undetected for extended periods. Such hidden failures have been identified as one of the main sources of power system blackouts.

The computing power available with computer relays has also given rise to newer and better protection functions in several instances. Improved protection of transformers, multiterminal lines, fault location, and reclosing are a few of the protection functions where computer relaying is likely to have a significant impact. Very significant developments in the computer relaying field are likely to occur in the coming years.

Architecture for a Computer Relay

There are many ways of implementing computer-based relays. Figure 61.48 is a fairly typical block diagram of a computer relay architecture. The input signals consisting of voltage and currents and contact status are filtered to remove undesired frequency components and potentially damaging surges. These signals are sampled by the CPU under the control of a sampling clock. Typical sampling frequency used in a modern digital relay varies between 4 and 32 times the nominal power system frequency. The sampled data is processed by the CPU with a digital filtering algorithm, which estimates the appropriate relaying quantity. A typical relaying quantity may be the rms value of a current, the voltage or current phasor, or the apparent impedance. The estimated parameters are then compared with prestored relay characteristics, and the appropriate control action is initiated. The decision of the relay is communicated to the substation equipment, such as the circuit breaker, through the output ports. These outputs must also be filtered to block any surges from entering the relay through the output lines. In most cases, the relay can also communicate with the outside world through a modem. The data created by a fault is usually saved by the relaying computer and can be used for fault analysis or for sequence-of-event analysis following a power system disturbance. The user may interface with the relay through a keyboard, a control panel, or a communication port. In any case, provision must be made to enter relay settings in the relay and to save these settings in case the station power supply fails. Although the block diagram in Fig. 61.48 shows different individual subsystems, the actual hardware composition of the subsystems is dependent on the computer manufacturer. Thus, we may find several microprocessors in a given implementation, each controlling one or more subsystems. Also, the hardware technology is in a state of flux, and in a few years, we may see an entirely different realization of the computer relays.

Experience and Future Trends

Field experience with the computer relays has been excellent so far. The manufacturers of traditional relays have adopted this technology in a big way. As more experience is gained with the special requirements of computer relays, it is likely that other—nontraditional—relay manufacturers will enter the field.

It seems clear that in computer relaying, power system engineers have obtained a tool with exciting new possibilities. Computers, with the communication networks now being developed, can lead to improved monitoring, protection, and control of power systems. An entirely new field, adaptive relaying, has been introduced recently [Phadke and Horowitz, 1990]. The idea is that protection systems should adapt to changing conditions of the power networks. In doing so, protection systems become more sensitive and reliable. Another development, which can be traced to computer relaying, is that of synchronized phasor measurements in power systems [Phadke and Thorp, 1991]. The development of the Global Positioning System (GPS) satellites has made possible the synchronization of sampling clocks used by relays and other measuring devices across the power system. This technology is expected to have a major impact on static and dynamic state estimation and on control of the electric power networks.

Defining Terms

Computer relays: Relays that use digital computers as their logic elements.

Distance protection: Relaying principle based upon estimating fault location (distance) and providing a response based upon the distance to the fault.

Electromechanical relays: Relays that use electromechanical logic elements.

Pilot: A communication medium used by relays to help reach a reliable diagnosis of certain faults.

Relays: Devices that detect faults on power equipment and systems and take appropriate control actions to deenergize the faulty equipment.

Reliability: For relays, reliability implies *dependability*, i.e., certainty of operating when it is supposed to, and *security*, certainty of not operating when it is not supposed to.

Solid state relays: Relays that use solid state analog components in their logic elements.

Transducers: Current and voltage transformers that reduce high-magnitude signals to standardized low-magnitude signals which relays can use.

Related Topic

1.3 Transformers

References

J.L. Blackburn, "Protective relaying," Marcel Dekker, 1987.

S.H. Horowitz and A.G. Phadke, *Power System Relaying*, Research Studies Press, New York: Wiley & Sons, 1992.

C.R. Mason, *The Art and Science of Protective Relaying*, New York: Wiley & Sons, 1956.

A.G. Phadke and S.H. Horowitz, "Adaptive relaying," *IEEE Computer Applications in Power*, vol. 3, no. 3, pp. 47–51, July 1990.

A.G. Phadke and J.S. Thorp, "Improved control and protection of power systems through synchronized phasor measurements," in *Analysis and Control System Techniques for Electric Power Systems*, part 3, C.T. Leondes, Ed., San Diego: Academic Press, pp. 335–376, 1991.

J.S. Thorp, A.G. Phadke, S.H. Horowitz, and J.E. Beehler, "Limits to impedance relaying," *IEEE Trans. PAS*, vol. 98, no. 1, pp. 246–260, January/February 1979.

Westinghouse Electric Corporation, "Applied Protective Relaying," 1982.

Further Information

In addition to the references provided, papers sponsored by the Power System Relaying Committee of the IEEE and published in the *IEEE Transactions on Power Delivery* contain a wealth of information about protective relaying practices and systems. Publications of CIGRÉ also contain papers on relaying, through their Study Committee 34 on protection. Relays and relaying systems usually follow standards, issued by IEEE in this country, and by such international bodies as the IEC in Europe. The field of computer relaying has been covered in *Computer Relaying for Power Systems*, by A.G. Phadke and J.S. Thorp (New York: Wiley, 1988).

61.7 Transient Operation of Power Systems

R. B. Gungor

Stable operations of power transmission systems have been a great concern of utilities since the beginning of early power distribution networks. The transient operation and the stability under transient operation are studied for existing systems, as well as the systems designed for future operations.

Power systems must be stable while operating normally at steady state for slow system changes under switching operations, as well as under emergency conditions, such as lightning strikes, loss of some generation, or loss of some transmission lines due to faults.

The tendency of a power system (or a part of it) to develop torques to maintain its stable operation is known as **stability**. The determination of the stability of a system then is based on the static and dynamic characteristics of its synchronous generators. Although large induction machines may contribute energy to the system during the *subtransient* period that lasts one or two cycles at the start of the **disturbance**, in general, induction machine loads are treated as static loads for **transient stability** calculations. This is one of the simplification considerations, among others.

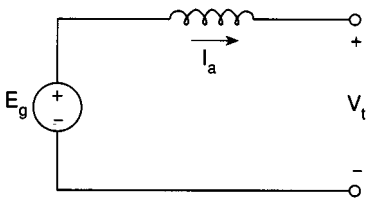


FIGURE 61.49 Per-phase model of an ideal synchronous generator.

The per-phase model of an ideal synchronous generator with nonlinearities and the stator resistance neglected is shown in Fig. 61.49, where E_g is the generated (excitation) voltage and X_s is the steady-state direct axis *synchronous reactance*. In the calculation of transient and subtransient currents, X_s is replaced by *transient reactance* X'_s and *subtransient reactance* X''_s, respectively.

Per-phase electrical power output of the generator for this model is given by Eq. (61.71).

$$P_e = \frac{E_g V_t}{X_s} \sin \delta = P_{\max} \sin \delta \tag{61.71}$$

where δ is the **power angle**, the angle between the generated voltage and the terminal voltage.

The simple power-angle relation of Eq. (61.71) can be used for real power flow between any two voltages separated by a reactance. For the synchronous machine, the total real power is three times the value calculated by Eq. (61.71), when voltages in volts and the reactance in ohms are used. On the other hand, Eq. (61.71) gives *per-unit* power when per-unit voltages and reactance are used.

Figure 61.50 shows a sketch of the power-angle relation of Eq. (61.71). Here the power P_1 is carried by the machine under δ_1, and P_2 under δ_2. For gradual changes in the output power up to $P_{\max}$ for $\delta = 90^\circ$, the machine will be stable. So we can define the **steady-state stability** limit as

$$\delta \leq 90^\circ \quad \frac{\partial P}{\partial \delta} > 0 \tag{61.72}$$

A sudden change in the load of the generator, e.g., from P_1 to P_2, will cause the rotor to slow down so that the power angle δ is increased to supply the additional power to the load. However, the deceleration of the rotor cannot stop instantaneously. Hence, although at δ_2 the developed power is sufficient to supply the load, the rotor will overshoot δ_2 until a large enough opposite torque is built up to stop deceleration. Now the excess energy will start accelerating the rotor to decrease δ. Depending on the inertia and damping, these oscillations will die out or the machine will become unstable and lose its synchronism to drop out of the system. This is the basic **transient operation** of a synchronous generator. Note that during this operation it may be possible for δ to become larger than 90° and the machine still stay stable. Thus $\delta = 90^\circ$ is not the transient stability limit.

Figure 61.51 shows typical power-angle versus time relations.

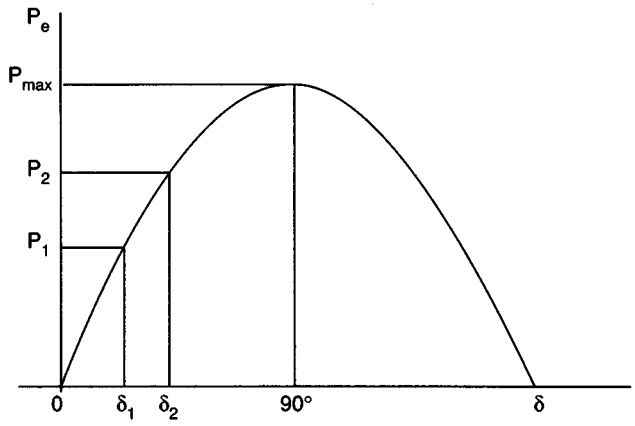


FIGURE 61.50 Power-angle characteristics of ideal synchronous generator.

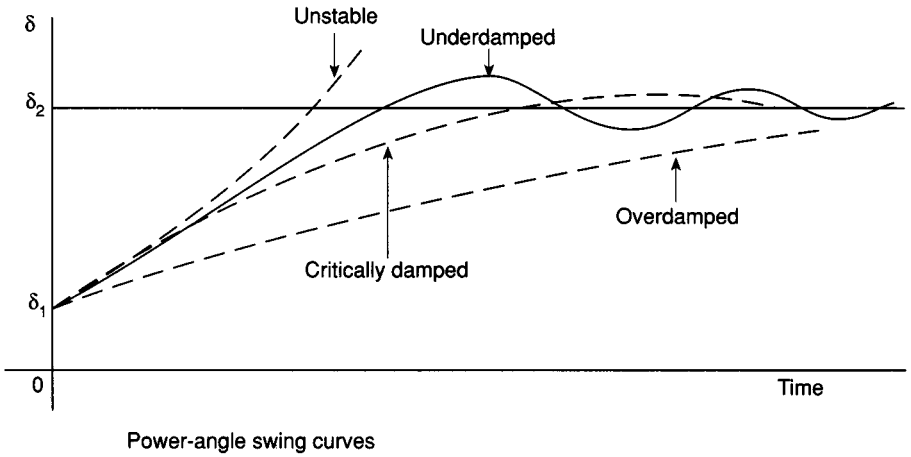


FIGURE 61.51 Typical power angle–time relations.

In the discussions to follow, the damping (stabilizing) effects of (1) the excitation systems; (2) the speed governors; and (3) the damper windings (copper squirrel-cage embedded into the poles of the synchronous generators) are omitted.

Stable Operation of Power Systems

Figure 61.52 shows an *N*-bus power system with *G* generators.

To study the stability of multimachine transmission systems, the resistances of the transmission lines and transformers are neglected and the reactive networks are reduced down to the generator internal voltages by dropping the loads and eliminating the load buses. One such reduced network is sketched in Fig. 61.53.

The power flow through the reactances of a reduced network are

$$P_{ij} = \frac{E_i E_j}{X_{ij}} \sin \delta_{ij} \qquad i, j = 1, 2, \ldots, G \tag{61.73}$$

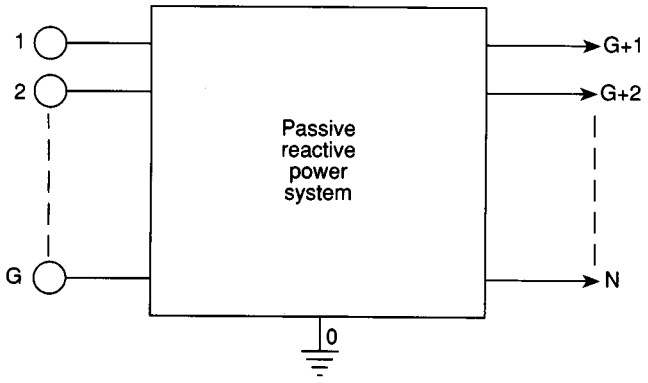


FIGURE 61.52 A multimachine reactive power system.

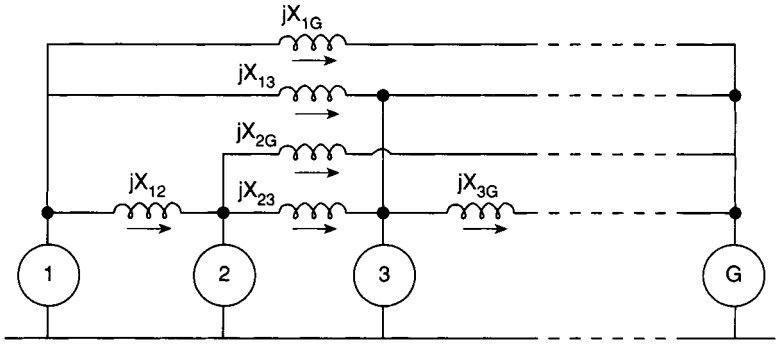


FIGURE 61.53 Multiport reduced reactive network.

The generator powers are

$$P_i = \sum_{k=1}^{G} P_{ik} \tag{61.74}$$

The system will stay stable for

$$\frac{\partial P_i}{\partial \delta_{ij}} > 0 \qquad i = 1, 2, \ldots, G \tag{61.75}$$

Equation (61.75) is observed for two machines at a time by considering all but two (say k and n) of the powers in Eq. (61.74) as constants. Since the variations of all powers but k and n are zero, we have

$$dP_i = \frac{\partial P_i}{\partial \delta_{i1}} d\delta_{i1} + \frac{\partial P_i}{\partial \delta_{i2}} d\delta_{i2} + \cdots + \frac{\partial P_i}{\partial \delta_{iG}} d\delta_{iG} = 0 \tag{61.76}$$

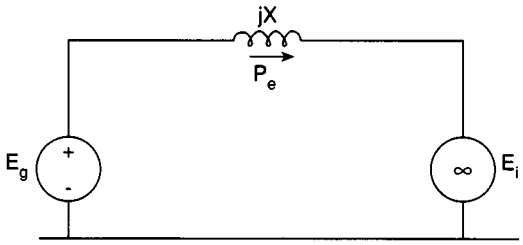


FIGURE 61.54 An ideal generator connected to an infinite bus.

These G-2 equations are simultaneously solved for G-2 $d\delta_{ij}$s, then these are substituted in dP_k and dP_n equations to calculate the partial derivatives of P_k and P_n with respect to δ_{kn} to see if Eqs. (61.75) for $i = k$ and $i = n$ are satisfied. Then the process is repeated for the remaining pairs.

Although the procedure outlined seems complicated, it is not too difficult to produce a computer algorithm for a given system.

To study the transient stability, dynamic operations of synchronous machines must be considered. An ideal generator connected to an infinite bus (an ideal source) through a reactance is sketched in Fig. 61.54.

The so-called *swing equation* relating the accelerating (or decelerating) power (difference between shaft power and electrical power as a function of δ) to the second derivative of the power angle is given in Eq. (61.77).

$$\begin{aligned} P_a &= P_s - P_e \\ M\frac{d^2\delta}{dt^2} &= P_s - \frac{E_g E_i}{X}\sin\delta \end{aligned} \tag{61.77}$$

where $M = HS/180f$ (MJ/electrical degree); H is the inertia constant (MJ/MVA); S is the machine rating (MVA); f is the frequency (Hz); P_s is the shaft power (MW).

For a system of G machines, a set of G swing equations as given in Eq. (61.78) must be solved simultaneously.

$$M_i\frac{d^2\delta_i}{dt^2} = P_{s_i} - P_{\max_i}\sin\delta_i \qquad i = 1, 2, \ldots, G \tag{61.78}$$

The swing equation of the single-machine system of Fig. 61.54 can be solved either graphically or analytically. For graphical integration, which is called *equal-area criterion,* we represent the machine by its subtransient reactance, assuming that electrical power can be calculated by Eq. (61.71), and during the transients the shaft power P_s remains constant. Then, using the power-angle curve(s), we sketch the locus of operating point on the curve(s) and equate the areas for stability. Figure 61.55 shows an example for which the shaft power of the machine is suddenly increased from the initial value of P_o to P_s.

The excess energy (area A_1) will start to accelerate the rotor to increase δ from δ_o to δ_m for which the area (A_2) above P_s equals the area below. These areas are

$$\begin{aligned} A_1 &= P_s(\delta_s - \delta_o) - \int_{\delta_o}^{\delta_s} P_{\max}\sin\delta\, d\delta \\ A_2 &= \int_{\delta_s}^{\delta_m} P_{\max}\sin\delta\, d\delta - P_s(\delta_m - \delta_s) \end{aligned} \tag{61.79}$$

Substituting, the values of P_o, P_s, δ_o, and δ_s, δ_m can be calculated.

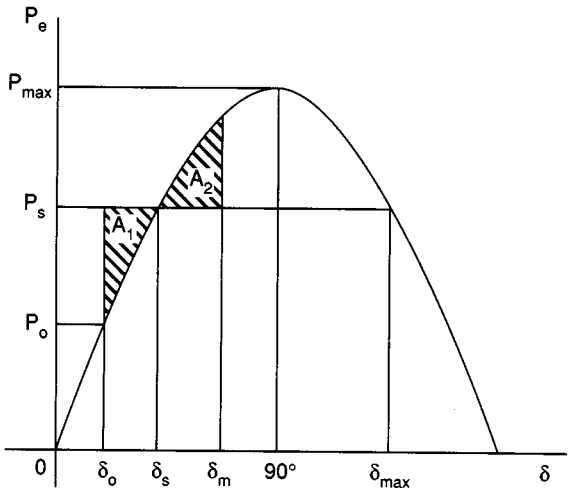


FIGURE 61.55 A sudden loading of a synchronous generator.

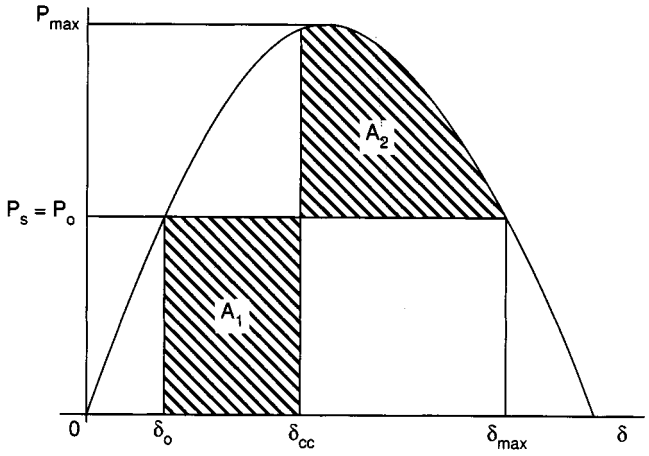


FIGURE 61.56 Critical clearing angle for stability.

Figure 61.56 illustrates another example, where a three-phase fault reduces the power transfer to infinite bus to zero. δ_{cc} is the **critical clearing angle** beyond which the machine will not stay stable.

The third example, shown in Fig. 61.57, indicates that the power transfers before, during, and after the fault are different. Here the system is stable as long as $\delta_m \leq \delta_{max}$.

For the analytical solution of the swing equation a *numerical integration* technique is used (Euler's method, modified Euler's method, Runge-Kutta method, etc.). The latter is most commonly used for computer algorithms.

The solution methods developed are based on various assumptions. As before, machines are represented by subtransient reactances, electrical powers can be calculated by Eq. (61.71), and the shaft power does not change during transients. In addition, the velocity increments are assumed to start at the beginning of time increments, and acceleration increments start at the middle of time increments; finally, an average acceleration can be used where acceleration is discontinuous (e.g., where circuit breakers open or close).

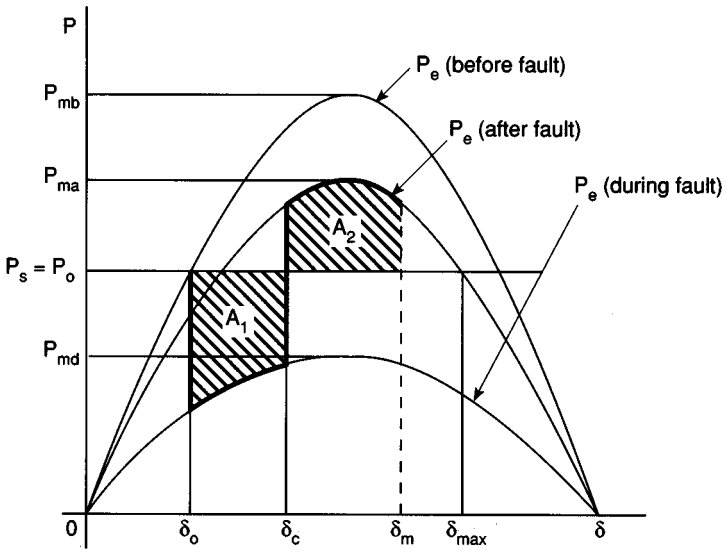


FIGURE 61.57 Power-angle relation for power transfer during fault.

Figure 61.58 shows a sketch of angle, velocity, and acceleration changes related to time as outlined above. Under these assumptions the next value of the angle δ can be obtained from the previous value as

$$\delta_{k+1} = \delta_k + \Delta_{k+1}\delta = \delta_k + \Delta_k\delta + \frac{(\Delta t)^2}{M} P_{ak} \tag{61.80}$$

where the accelerating power is

$$P_{ak} = P_s - P_{ek}$$

and

$$P_{ek} = P_{\max k} \sin \delta_k$$

For hand calculations a table, as shown in Table 61.5, can be set up for fast processing.

TABLE 61.5 Numerical Calculations of Swing Equations

n	t	P_{max}	P_e	P_{ak}	$\frac{(\Delta t)^2 P_a}{M}$	$\Delta_{k+1}\delta$	δ_k
0	0_-						
0	0_+						
0	0_{av}						
1	Δt						
2	$2\Delta t$						
3	$3\Delta t$						
4	$4\Delta t$						
5	$5\Delta t$						
6	$6\Delta t$						

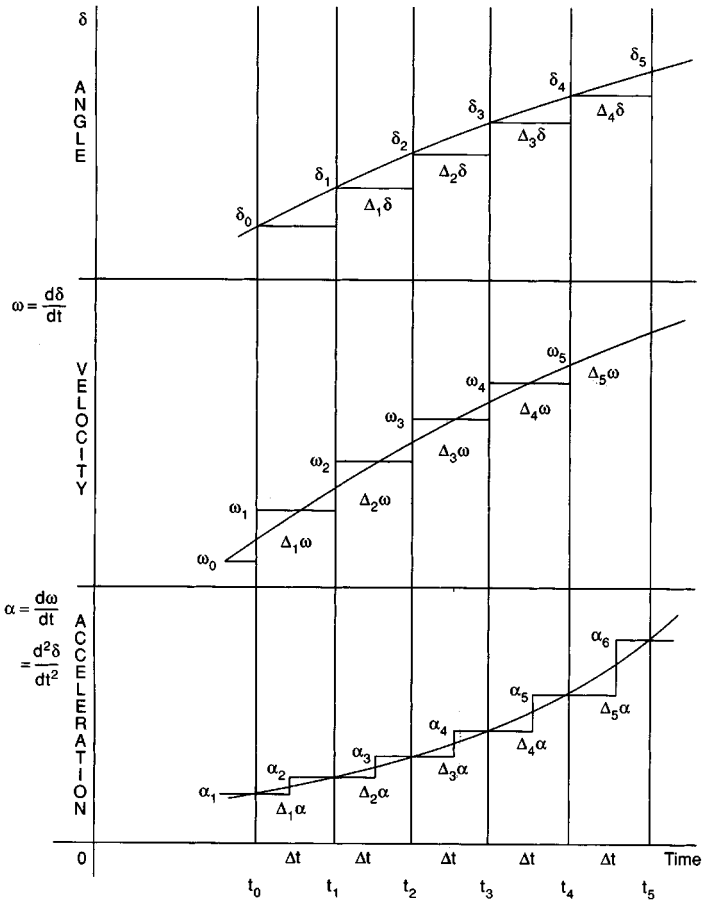


FIGURE 61.58 Incremental angle, velocity, and acceleration changes versus time.

Computer algorithms are developed by using the before-fault, during-fault, and after-fault Z_{BUS} matrix of the reactive network reduced to generator internal voltages with generators represented by their subtransient reactances. Each generator's swing curve is obtained by numerical integration of its power angle for a specified condition, then a set of swing curves is tabulated or graphed for observation of the transient stability. An example with partial calculated data and a line plot for such a study are included on the next page.

Defining Terms

Critical clearing angle: Power angle corresponding to the critical clearing time.

Critical clearing time: The maximum time at which a fault must be cleared for the system to stay transiently stable.

Disturbance (fault): A sudden change or a sequence of changes in the components or the formation of a power system.

Large disturbance: A disturbance for which the equations for dynamic operation cannot be linearized for analysis.

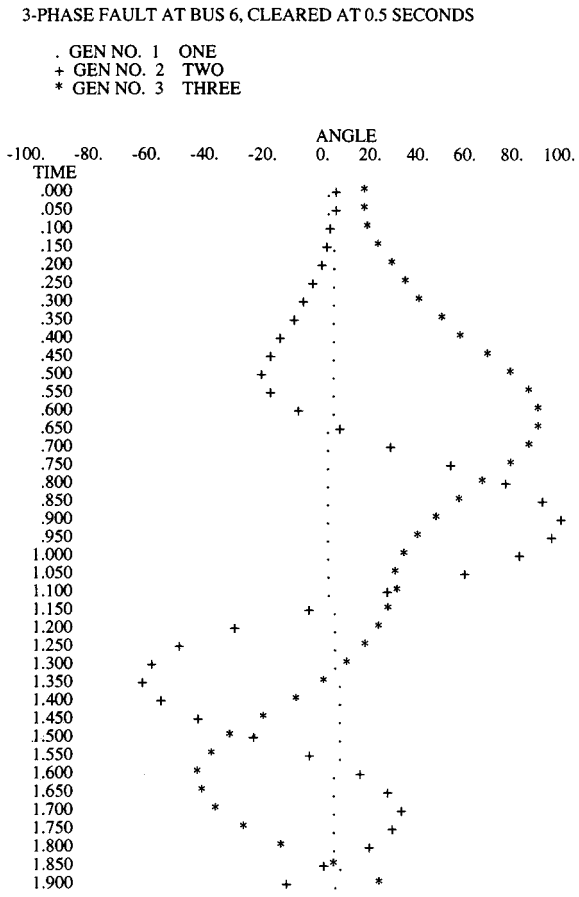
Power angle: The electrical angle between the generated and terminal voltages of a synchronous generator.

Small disturbance: A disturbance for which the equations for dynamic operation can be linearized for analysis.

Stability: The tendency of a power system (or a part of it) to develop torques to maintain its stable operation for a disturbance.

Transient stability program
7-Bus system with 3 generators
3-Phase fault at bus 6, cleared at
0.5 seconds by removing the line 1-6

Time	Gen	Angle	Power
.000	1	3.46	118.38
.000	2	5.80	111.90
.000	3	15.16	95.65
.050	1	3.46	118.59
.050	2	5.31	110.95
.050	3	15.86	96.24
.100	1	3.46	119.21
.100	2	3.84	108.12
.100	3	17.97	97.99
.150	1	3.46	120.15
.150	2	1.47	103.59
.150	3	21.45	100.83
.200	1	3.46	121.31
.200	2	–1.66	97.62
.200	3	26.27	104.66
.500	1	3.46	55.48
.500	2	–26.55	–215.72
.500	3	79.92	481.86
.900	1	3.46	–198.56
.900	2	100.99	458.41
.900	3	49.43	72.78
1.950	1	3.46	125.86
1.950	2	–30.18	–216.29
1.950	3	41.40	425.31
2.000	1	3.46	125.86
2.000	2	–34.60	–216.29
2.000	3	57.78	425.31



Steady-state stability: A power system is steady-state stable if it reaches another steady-state operating point after a small disturbance.

Transient operation: A power system operating under abnormal conditions because of a disturbance.

Transient stability: A power system is transiently stable if it reaches a steady-state operating point after a large disturbance.

Related Topic

12.1 Introduction

References

J. Arrillaga, C.P. Arnold, and B.J. Harker, *Computer Modeling of Electrical Power Systems*, New York: Wiley, 1983.

A.R. Bergen, *Power System Analysis*, Englewood Cliffs, N.J.: Prentice-Hall, 1986.

H.E. Brown, *Solution of Large Networks by Matrix Methods,* New York: Wiley, 1985.

A.A. Fouad and V. Vittal, *Power System Transient Stability Analysis*, Englewood Cliffs, N.J.: Prentice-Hall, 1992.

J.D. Glover and M. Sarma, *Power System Analysis and Design*, Boston: PWS Publishers, 1987.

C.A. Gross, *Power System Analysis*, 2nd ed., New York: Wiley, 1986.

R.B. Gungor, *Power Systems,* San Diego: Harcourt Brace Jovanovich, 1988.

G.T. Heydt, *Computer Analysis Methods for Power Systems*, New York: Macmillan, 1986.

W.D. Stevenson, *Elements of Power System Analysis*, 4th ed., New York: McGraw-Hill, 1982.

Y. Wallach, *Calculations & Programs for Power System Networks*, Englewood Cliffs, N.J.: Prentice-Hall, 1986.

Further Information

In addition to the references listed above, further and more recent information can be found in IEEE publications, such as *IEEE Transactions on Power Systems*, *IEEE Transactions on Power Delivery*, *IEEE Transactions on Energy Conversion*, and *IEEE Transactions on Automatic Control*.

Power Engineering Review and *Computer Applications in Power* of the IEEE are good sources for paper summaries.

Finally, *IEEE Transactions on Power Apparatus and Systems* dating back to the 1950s can be consulted.

61.8 Planning

J. Duncan Glover

An electric utility transmission system performs three basic functions: delivers outputs from generators to the system, supplies power to the distribution system, and provides for power interchange with other utilities. The electric utility industry has developed planning principles and criteria to ensure that the transmission system reliably performs these basic functions.

The North American Electric Reliability Council (NERC) has provided definitions of the terms **reliability**, **adequacy**, and **security** (see Defining Terms at the end of this section).

System reliability may be viewed from two perspectives: short-term reliability and long-term reliability. The system operator is primarily concerned with real-time security aspects in the short term, that is, supplying steady, uninterrupted service under existing operating conditions and as they occur over the next few minutes, hours, days, or months. The transmission planning engineer, however, is concerned not only with security aspects in the short term but also adequacy and security aspects in the long term, as many as 25 or more years into the future.

The actual construction of a major transmission facility requires three to five years or more, depending largely on the siting and certification process. As such, the planning process requires up to ten years prior to operation of these facilities to ensure that they are available when required. The long lead times, environmental impacts, and high costs required for new transmission facilities require careful, near-optimal planning. Future changes in system operating conditions, such as changes in spatial load and generation patterns, create uncertainties that challenge the transmission planning engineer to select the best technical solution among several alternatives with due consideration of nontechnical factors. Transmission planning strives to maintain an optimal balance between system reliability, environmental impacts, and cost under future uncertainties.

Before transmission planning is started, long-term load forecasting and generation planning are completed. In long-term load forecasting, peak and off-peak loads in each area of the system under study are projected, year by year, from the present up to 25 years into the future. Such forecasts are based on present and past load trends, population growth patterns, and economic indicators. In generation planning, generation resources are selected with sufficient generation reserve margins to meet projected customer loads with adequate quality and reliability in an economic manner. New generating units both at new plant sites and at existing plants are selected, and construction schedules are established to ensure that new generation goes on-line in time to meet projected loads.

The results of long-term load forecasting and generation planning are used by transmission planning engineers to design the future transmission system so that it performs its basic functions. The following are selected during the transmission planning process.

- Routes for new lines
- Number of circuits for each route or right-of-way
- EHV versus HVDC lines
- Overhead versus underground line construction
- Types of towers for overhead lines
- Voltage levels

- Line ratings
- Shunt reactive and series capacitive line compensation
- Number and locations of substations
- Bus and circuit breaker configurations at substations
- Circuit breaker ratings
- Number, location, and ratings of bulk-power-system transformers
- Number, location, and ratings of voltage-regulating transformers and phase-shifting transformers
- Number, location, and ratings of static VAR systems, synchronous condensers, and shunt capacitor banks for voltage control
- Basic insulation levels (BILs)
- Surge arrester locations and ratings
- Protective relaying schemes
- Communications facilities
- Upgrades of existing circuits
- Reinforcements of system interconnections

Planning Tools

As electric utilities have grown in size and the number of interconnections has increased, making the above selections during the planning process has become increasingly complex. The increasing cost of additions and modifications has made it imperative that planning engineers consider a wide range of design options and perform detailed studies on the effects on the system of each option based on a number of assumptions: normal and emergency operating conditions, peak and off-peak loadings, and present and future years of operation. A large volume of network data must be collected and accurately handled. To assist the planning engineer, the following digital computer programs are used [Glover and Sarma, 1994]:

1. *Power-flow programs.* Power-flow (also called load-flow) programs compute voltage magnitudes, phase angles, and transmission line power flows for a power system network under steady-state operating conditions. Other output results, including transformer tap settings, equipment losses, and reactive power outputs of generators and other devices, are also computed. To do this, the locations, sizes, and operating characteristics of all loads and generation resources of the system are specified as inputs. Other inputs include the network configuration as well as ratings and other characteristics of transmission lines, transformers, and other equipment. Today's computers have sufficient storage and speed to compute in less than 1 min power-flow solutions for networks with more than 2000 buses and 2500 transmission lines. High-speed printers then print out the complete solution in tabular form for analysis by the planning engineer. Also available are interactive power-flow programs, whereby power-flow results are displayed on computer screens in the form of single-line diagrams; the engineer uses these to modify the network from a keyboard or with a mouse and can readily visualize the results. Spreadsheet analyses are also used. The computer's large storage and high-speed capabilities allow the engineer to run the many different cases necessary for planning.
2. *Transient stability programs.* Transient stability programs are used to study power systems under disturbance conditions to predict whether synchronous generators remain in synchronism and system stability is maintained. System disturbances can be caused by the sudden loss of a generator or a transmission line, by sudden load increases or decreases, and by short circuits and switching operations. The stability program combines power-flow equations and generator dynamic equations to compute the angular swings of machines during disturbances. The program also computes critical clearing times for network faults and allows the planning engineer to investigate the effects of various network modifications, machine parameters, disturbance types, and control schemes.
3. *Short-circuits programs.* Short-circuits programs compute three-phase and line-to-ground fault currents in power system networks in order to evaluate circuit breakers and relays that detect faults and

control circuit breakers. Minimum and maximum short-circuit currents are computed for each circuit breaker and relay location under various system operating conditions, such as lines or generating units out of service, in order to specify circuit breaker ratings and protective relay schemes.

4. *Transients programs.* Transients programs compute the magnitudes and shapes of transient overvoltages and currents that result from switching operations and lightning strikes. Planning engineers use the results of transients programs to specify BILs for transmission lines, transformers, and other equipment and to select surge arresters that protect equipment against transient overvoltages.

Research efforts aimed at developing computerized, automated transmission planning tools are ongoing. Examples and references are given in Back et al. [1989] and Smolleck et al. [1989]. Other programs for transmission planning include production-cost, investment-cost, relay-coordination, power-system database management, transformer thermal analysis, and transmission line design programs. Some of the vendors that offer software packages for transmission planning are given as follows:

- ABB Network Control Ltd., Switzerland
- CYME International, Burlington, Mass.
- EDSDA Micro Corporation, Bloomfield, Mich.
- Electric Power Consultants, Inc., Scotia, N.Y.
- Electrocon International, Inc., Ann Arbor, Mich.
- Power Technologies, Inc., Schenectady, N.Y.
- Operation Technology, Inc., Irvine, Calif.

Basic Planning Principles

The electric utility industry has established basic planning principles intended to provide a balance among all power system components so as not to place too much dependence on any one component or group of components. Transmission planning criteria are developed from these principles along with actual system operating history and reasonable contingencies. These planning principles are given as follows:

1. Maintain a balance among power system components based on size of load, size of generating units and power plants, the amount of power transfer on any transmission line or group of lines, and the strength of interconnections with other utilities. In particular:
   a. Avoid excessive generating capacity at one unit, at one plant, or in one area.
   b. Avoid excessive power transfer through any single transformer, through any transmission line, circuit, tower, or right-of-way, or though any substation.
   c. Provide interconnection capacity to neighboring utilities that is commensurate with the size of generating units, power plants, and system load.
2. Provide transmission capability with ample margin above that required for normal power transfer from generators to loads in order to maintain a high degree of flexibility in operation and to meet a wide range of contingencies.
3. Provide for power system operation such that all equipment loadings remain within design capabilities.
4. Utilize switching arrangements, associated relay schemes, and controls that permit:
   a. Effective operation and maintenance of equipment without excessive risk of uncontrolled power interruptions.
   b. Prompt removal and isolation of faulted components.
   c. Prompt restoration in the event of loss of any part of the system.

Equipment Ratings

Transmission system loading criteria used by planning engineers are based on equipment ratings. Both normal and various emergency ratings are specified. Emergency ratings are typically based on the time required for either emergency operator actions or equipment repair times. For example, up to 2 h may be required following a major event such as loss of a large generating unit or a critical transmission facility in order to bring other

generating resources on-line and to perform appropriate line-switching operations. The time to repair a failed transmission line typically varies from 2 to 10 days, depending on the type of line (overhead, underground cable in conduit, or pipe-type cable). The time required to replace a failed bulk-power-system transformer is typically 30 days. As such, ratings of each transmission line or transformer may include normal, 2-h emergency, 2- to 10-day emergency, and in some cases 30-day emergency ratings.

The rating of an overhead transmission line is based on the maximum temperature of the conductors. Conductor temperature affects the conductor sag between towers and the loss of conductor tensile strength due to annealing. If the temperature is too high, proscribed conductor-to-ground clearances [ANSI, 1993] may not be met, or the elastic limit of the conductor may be exceeded such that it cannot shrink to its original length when cooled. Conductor temperature depends on the current magnitude and its time duration, as well as on ambient temperature, wind velocity, solar radiation, and conductor surface conditions. Standard assumptions on ambient temperature, wind velocity, etc., are selected, often conservatively, to calculate overhead transmission line ratings [ANSI/IEEE Std. 738–85, 1985]. It is common practice to have summer and winter normal line ratings, based on seasonal ambient temperature differences. Also, in locations with higher prevailing winds, such as coastal areas, larger normal line ratings may be selected. Emergency line ratings typically vary from 110 to 120% of normal ratings. Recently, real-time monitoring of actual conductor temperatures along a transmission line has been used for on-line dynamic transmission line ratings [Henke and Sciacca, 1989].

Normal ratings of bulk-power-system transformers are determined by manufacturers' nameplate ratings. Nameplate ratings are based on the following ANSI/IEEE standard conditions: (1) continuous loading at nameplate output; (2) 30°C average ambient temperature (never exceeding 40°C); and (3) 110°C average hot-spot conductor temperature (never exceeding 120°C) for 65°C-average-winding-rise transformers [ANSI/IEEE C57.92-1981, 1990]. For 55°C-average-winding-rise transformers, the hot-spot temperature limit is 95°C average (never exceeding 105°C). The actual output that a bulk-power-system transformer can deliver at any time with normal life expectancy may be more or less than the nameplate rating, depending on the ambient temperature and actual temperature rise of the windings. Emergency transformer ratings typically vary from 130 to 150% of nameplate ratings.

Planning Criteria

Transmission system planning criteria have been developed from the above planning principles and equipment ratings as well as from actual system operating data, probable operating modes, and equipment failure rates. These criteria are used to plan and build the transmission network with adequate margins to ensure a reliable supply of power to customers under reasonable equipment-outage contingencies. The transmission system should perform its basic functions under a wide range of operating conditions. Transmission planning criteria include equipment loading criteria, transmission voltage criteria, stability criteria, and regional planning criteria.

Equipment Loading Criteria

Typical equipment loading criteria are given in Table 61.6. With no equipment outages, transmission equipment loadings should not exceed normal ratings for all realistic combinations of generation and interchange. Operation of all generating units including base-loaded and peaking units during peak load periods as well as operation of various combinations of generation and interchange during off-peak periods should be considered. Also, normal ratings should not be exceeded with all transmission lines and transformers in service and with any generating unit out of service.

With any single-contingency outage, emergency ratings should not be exceeded. One loading criterion is not to exceed 2-h emergency ratings when any transmission line or transformer is out of service. This gives time to perform switching operations and change generation levels, including use of peaking units, to return to normal loadings.

With some of the likely double-contingency outages, the transmission system should supply all system load without exceeding emergency ratings. One criterion is not to exceed 2- to 10-day emergency ratings when any line and any transformer are out of service or when any line and any generator are out of service. This gives time to repair the line. With the outage of any transformer and any generator, 30-day emergency ratings should not be exceeded, which gives time to install a spare transformer.

TABLE 61.6 Typical Transmission Equipment Loading Criteria

Equipment Out of Service	Rating Not to Be Exceeded	Comment
None	Normal	
Any generator	Normal	
Any line or any transformer	2-h emergency	Before switching.
Any line and any transformer*	2- to 10-day emergency	After switching required for both outages. Line repair time.
Any line and any generator*	2- to 10-day emergency	After switching required for both outages. Line repair time.
Any transformer and any generator*	30-day emergency	After switching required for both outages. Install spare transformer.

*Some utilities do not include double-contingency outages in transmission system loading criteria.

TABLE 61.7 Typical Minimum Transmission Voltage Criteria

	Planned Minimum Transmission Voltage at Substations, % of Nominal		
System Condition	Generator Station	EHV Station	HV Station
Normal	102	98	95–97.5
Single-contingency outage	100	96	92.5–95
Double-contingency outage*	98	94	92.5

*Some utilities do not include double-contingency outages in planned minimum transmission voltage criteria.

The loading criteria in Table 61.6 do not include all types of double-contingency outages. For example, the outage of a double-circuit transmission line or two transmission lines in the same right-of-way is not included. Also, the loss of two transformers in the same load area is not included. Under these double-contingency outages, it may be necessary to shed load at some locations during heavy load periods. Although experience has shown that these outages are relatively unlikely, their consequences should be evaluated in specific situations. Factors to be evaluated include the size of load served, the degree of risk, and the cost of reinforcement.

Specific loading criteria may also be required for equipment serving critical loads and critical load areas. One criterion is to maintain service to critical loads under a double-contingency outage with the prior outage of any generator.

Transmission Voltage Criteria

Transmission voltages should be maintained within suitable ranges for both normal and reasonable emergency conditions. Abnormal transmission voltages can cause damage or malfunction of transmission equipment such as circuit breakers or transformers and adversely affect many customers. Low transmission voltages tend to cause low distribution voltages, which in turn cause increased distribution losses as well as higher motor currents at customer loads and at power plant auxiliaries. Transmission voltage planning criteria are intended to be conservative.

Maximum planned transmission voltage is typically 105% of rated nominal voltage for both normal and reasonable emergency conditions. Typical minimum planned transmission voltages are given in Table 61.7. System conditions in Table 61.7 correspond to equipment out of service in Table 61.6. Single-contingency outages correspond to the loss of any line, any transformer, or any generator. Double-contingency outages correspond to the loss of any transmission line and transformer, any transmission line and generator, any transformer and generator, or any two generators.

Typical planned minimum voltage criteria shown in Table 61.7 for EHV (345 kV and higher) substations and for generator substations are selected to maintain adequate voltage levels at interconnections, at power plant auxiliary buses, and on the lower-voltage transmission systems. Typical planned minimum voltage criteria for lower HV (such as 138 kV, 230 kV) transmission substations vary from 95 to 97.5% of nominal voltage under normal system conditions to as low as 92.5% of nominal under double-contingency outages.

Equipment used to control transmission voltages includes voltage regulators at generating units (excitation control), tap-changing transformers, regulating transformers, synchronous condensers, shunt reactors, shunt capacitor banks, and static VAR devices. When upgrades are selected during the planning process to meet planned transmission voltage criteria, some of this equipment should be assumed out of service.

Stability Criteria

System stability is the ability of all synchronous generators in operation to stay in synchronism with each other while moving from one operating condition to another. Steady-state stability refers to small changes in operating conditions, such as normal load changes. Transient stability refers to larger, abrupt changes, such as the loss of the largest generator or a short circuit followed by circuit breakers opening, where synchronism or loss of synchronism occurs within a few seconds. Dynamic stability refers to longer time periods, from minutes up to a half hour following a large, abrupt change, where steam generators (boilers), automatic generation control, and system operator actions affect stability.

In the planning process, steady-state stability is evaluated via power-flow programs by the system's ability to meet equipment loading criteria and transmission voltage criteria under steady-state conditions. Transient stability is evaluated via stability programs by simulating system transient response for various types of disturbances, including short circuits and other abrupt network changes. The planning engineer designs the system to remain stable for the following typical disturbances:

1. With all transmission lines in service, a permanent three-phase fault (short circuit) occurs on any transmission line, on both transmission lines on any double-circuit tower, or at any bus; the fault is successfully cleared by primary relaying.
2. With any one transmission line out of service, a permanent three-phase fault occurs on any other transmission line; the fault is successfully cleared by primary relaying.
3. With all transmission lines in service, a permanent three-phase fault occurs on any transmission line; backup relaying clears the fault after a time delay, due to a circuit breaker failure.

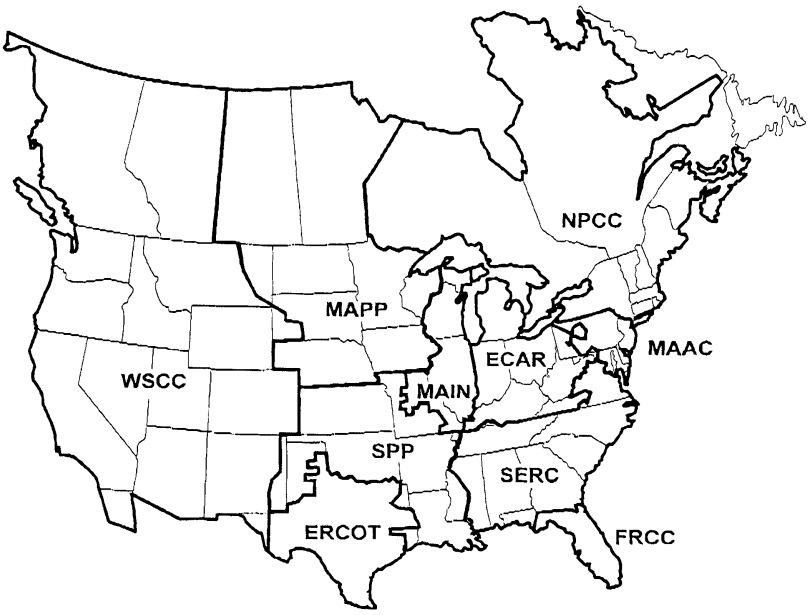
Regional Planning Criteria

The North American Electric Reliability Council (NERC) defines nine geographical regions in North America, as shown in Fig. 61.59 [NERC, 1988]. Transmission planning studies are performed at two levels: (1) individual electric utility companies separately perform planning studies of their internal systems and (2) companies jointly participate in NERC committees or working groups to perform regional and interregional planning studies. The purpose of regional planning studies is to evaluate the transfer capabilities between interconnected utilities and the impact of severe disturbances.

One typical regional criterion is that the incremental power transfer capability, in addition to scheduled interchange, should provide a reasonable generation reserve margin under the following conditions: peak load, the most critical transmission line out of service, no component overloaded.

Another criterion is that severe disturbances to the interconnected transmission network should not result in system instability, widespread cascading outages, voltage collapse, or system blackouts. [NERC, 1988, 1989, and 1991]. Severe disturbances include the following:

1. With any three generating units or any combination of units up to 30% of system load out of service in an area, a sudden outage of any transmission line or any transformer occurs.
2. With any two generating units or any combination of units up to 20% of system load out of service in an area, a sudden outage of any generator or any double-circuit transmission line occurs.
3. With any transmission line or transformer out of service in an area, a sudden outage of any other transmission line or transformer occurs.
4. With any transmission line or transformer out of service in an area as well as any two generating units or any combination of units up to 20% of system load, a sudden outage of a transmission line occurs.
5. A sudden outage of all generating units at a power plant occurs.
6. A sudden outage of either a transmission substation or all transmission lines on a common right-of-way occurs.
7. A sudden outage of a large load or a major load center occurs.



ECAR
East Central Area Reliability Coordination Agreement

ERCOT
Electric Reliability Council of Texas

FRCC
Florida Reliability Coordinating Council

MAAC
Mid-Atlantic Area Council

MAIN
Mid-America Interconnected Network, Inc.

MAPP
Mid-Continent Area Power Pool

NPCC
Northeast Power Coordinating Council

SERC
Southeastern Electric Reliability Council

SPP
Southwest Power Pool

WSCC
Western Systems Coordinating Council

Affiliate

ASCC
Alaska Systems Coordinating Council

FIGURE 61.59 Nine regional reliability councils established by NERC. (*Source: 1996 Annual Report,* Princeton, N.J.: North American Electric Reliability Council, 1997. With permission.)

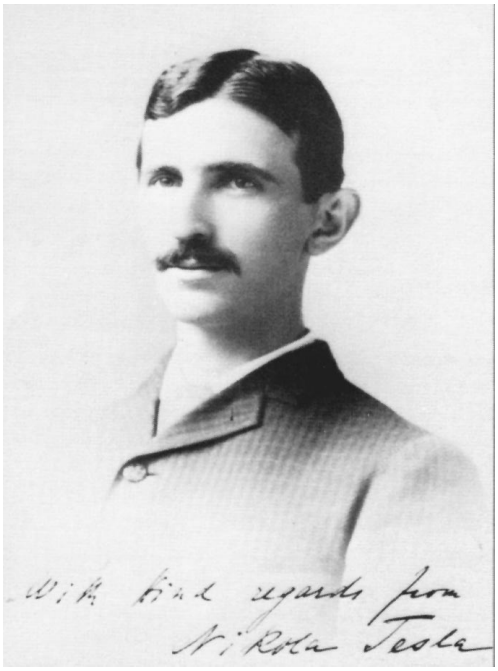
When evaluating the impacts of the above severe disturbances, regional planning studies should consider steady-state stability, transient stability, and dynamic stability. These studies should also consider the effects of three-phase faults and slow fault clearing due to improper relaying or failure of a circuit breaker to open, as well as the anticipated load range and various operating conditions.

Value-Based Transmission Planning

Recently some utilities have begun to use a value-of-service concept in transmission planning [EPRI, 1986]. This concept establishes a method of assigning a dollar value to various levels of reliability in order to balance reliability and cost. For each particular outage, the amount and dollar value of unserved energy are determined. Dollar value of unserved energy is based on rate surveys of various types of customers. If the cost of the transmission project required to eliminate the outage exceeds the value of service, then that project is given a lower priority. As such, reliability is quantified, and benefit-to-cost ratios are used to compare and prioritize planning options.

Nikola Tesla (1856–1943)

Nikola Tesla was born of Serbian parents in the village of Smiljan, in what is now Yugoslavia. He showed his technical brilliance early, but felt that his native country offered him only limited opportunities. In 1884 he emigrated to the United States and began working for Thomas Edison. He soon struck out on his own, however, for Edison had little use for Tesla's bold new ideas — in particular, his brilliant solution to the problems of applying alternating current in light and power systems. Tesla's polyphase ac system was brought to market by George Westinghouse, and after an acrimonious struggle with the Edison interests, which were wedded to the use of direct current (dc), the Tesla system became the standard in the twentieth century. Tesla's other inventions included the synchronous ac motor, devices for generating high voltage and high frequency currents, and contributions to radio technology. Tesla received the Edison Medal of the American Institute of Electrical Engineers in 1916. (Courtesy of the IEEE Center for the History of Electrical Engineering.)



Defining Terms

The North American Electric Reliability Council (NERC) defines *reliability* and the related terms *adequacy* and *security* as follows [NERC, 1988]:

Adequacy: The ability of the bulk-power electric system to supply the aggregate electric power and energy requirements of the consumers at all times, taking into account scheduled and unscheduled outages of system components.

Reliability: In a bulk-power electric system, reliability is the degree to which the performance of the elements of that system results in power being delivered to consumers within accepted standards and in the amount desired. The degree of reliability may be measured by the frequency, duration, and magnitude of adverse effects on consumer service.

Security: The ability of the bulk-power electric system to withstand sudden disturbances such as electric short circuits or unanticipated loss of system components.

References

ANSI C2-1993, National Electrical Safety Code, 1993 Edition, Piscataway, N.J.: IEEE, 1993.

ANSI/IEEE C57.92-1981, IEEE Guide for Loading Mineral-Oil Immersed Power Transformers Up to and Including 100 MVA with 55°C or 65°C Average Winding Rise, Piscataway, N.J.: IEEE, 1990.

ANSI/IEEE Std. 738-1985, Calculation of Bare Overhead Conductor Temperature and Ampacity under Steady-State Conditions, Piscataway, N.J.: IEEE, 1985.

H. Back et al., "PLATINE—A new computerized system to help in planning the power transmission networks," *IEEE Trans. Power Systems*, vol. 4, no. 1, pp.242–247, 1989.

Electric Power Research Institute (EPRI), Value of Service Reliability to Consumers, Report EA-4494, Palo Alto, Calif.: EPRI, March 1986.
J.D. Glover and M.S. Sarma, *Power System Analysis and Design with Personal Computer Applications*, 2nd ed., Boston: PWS Publishing Co., 1994.
R.K. Henke and S.C. Sciacca, "Dynamic thermal rating of critical lines—A study of real-time interface requirements," *IEEE Computer Applications in Power*, pp. 46–51, July 1989.
NERC, *Reliability Concepts*, Princeton, N.J.: North American Electric Reliability Council, February 1985.
NERC, *Overview of Planning Reliability Criteria*, Princeton, N.J.: North American Electric Reliability Council, April 1988.
NERC, *Electricity Transfers and Reliability*, Princeton, N.J.: North American Electric Reliability Council, October 1989.
NERC, *A Survey of the Voltage Collapse Phenomenon*, Princeton, N.J.: North American Electric Reliability Council, 1991.
H.A. Smolleck et al., "Translation of large data-bases for microcomputer-based application software: Methodology and a case study," *IEEE Comput. Appl. Power,* pp. 40–45, July 1989.

Further Information

The North American Electric Reliability Council (NERC) was formed in 1968, in the aftermath of the November 9, 1965, northeast blackout, to promote the reliability of bulk-electric-power systems of North America. Transmission planning criteria presented here are partially based on NERC criteria as well as on specific criteria used by transmission planning departments from three electric utility companies: American Electric Power Service Corporation, Commonwealth Edison Company, and Pacific Gas & Electric Company. NERC's publications, developed by utility experts, have become standards for the industry. In most cases, these publications are available at no charge from NERC, Princeton, N.J.

Arrillaga, J. “Power Quality”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

62

Power Quality

62.1 Power Quality Disturbances
Periodic Waveform Distortion • Voltage Fluctuations and Flicker • Brief Interruptions, Sags, and Swells • Unbalances • Transients
62.2 Power Quality Monitoring
62.3 Power Quality Conditioning

Jos Arrillaga
University of Canterbury (New Zealand)

Ideally, power should be supplied without interruptions at constant frequency, constant voltage and with perfectly sinusoidal and, in the case of three-phase, symmetrical waveforms. Supply reliability constitutes a recognized independent topic, and is not usually discussed under power quality. The specific object of power quality is the "*pureness*" of the supply including voltage variations and waveform **distortion**.

Power system **disturbances** and the continually changing demand of consumers give rise to voltage variations. Deviation from the sinusoidal voltage supply can be due to transient phenomena or to the presence of non-linear components.

The power network is not only the main source of energy supply but also the conducting vehicle for possible interferences between consumers. This is a subject that comes under the general heading of electromagnetic compatibility (EMC).

EMC refers to the ability of electrical and electronic components, equipment, and systems to operate satisfactorily without causing interference to other equipment or systems, or without being affected by other operating systems in that electromagnetic environment.

EMC is often perceived as interference by electromagnetic radiation between the various elements of a system. The scope of EMC, however, is more general and it also includes conductive propagation and coupling by capacitance, inductance (self and mutual) encompassing the whole frequency spectrum.

A power quality problem is any occurrence manifested in voltage, current, or frequency deviation that results in failure or misoperation of equipment. The newness of the term reflects the newness of the concern. Decades ago, power quality was not a worry because it had no effect on most loads connected to electric distribution systems.

Therefore, power quality can also be defined as the ability of the electrical power system to transmit and deliver electrical energy to the consumers within the limits specified by EMC standards.

62.1 Power Quality Disturbances

Following standard criteria [IEC, 1993], the main deviations from a perfect supply are

- periodic waveform distortion (harmonics, interharmonics)
- voltage fluctuations, flicker
- short voltage interruptions, dips (sags), and increases (swells)
- three-phase unbalance
- transient overvoltages

The main causes, effects and possible control of these disturbances are considered in the following sections.

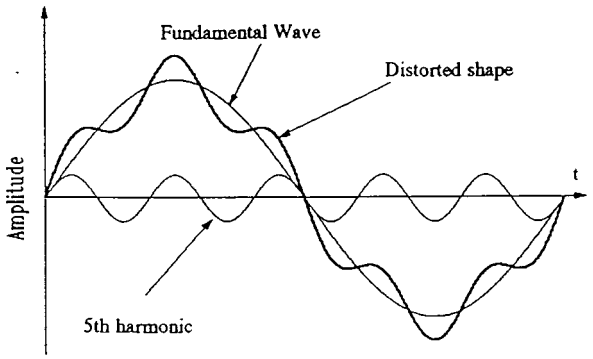


FIGURE 62.1 Example of a distorted sine wave.

Periodic Waveform Distortion

Harmonics are sinusoidal voltages or currents having frequencies that are whole multiples of the frequency at which the supply system is designed to operate (e.g., 50 Hz or 60 Hz). An illustration of fifth harmonic distortion is shown in Fig. 62.1. When the frequencies of these voltages and currents are not an integer of the fundamental they are termed interharmonics.

Both harmonic and interharmonic distortion is generally caused by equipment with non-linear voltage/current characteristics.

In general, distorting equipment produces harmonic currents that, in turn, cause harmonic voltage drops across the impedances of the network. Harmonic currents of the same frequency from different sources add vectorially.

The main detrimental effects of harmonics are [Arrillaga et al., 1985]

- maloperation of control devices, main signalling systems, and protective relays
- extra losses in capacitors, transformers, and rotating machines
- additional noise from motors and other apparatus
- telephone interference
- The presence of power factor correction capacitors and cable capacitance can cause shunt and series resonances in the network producing voltage amplification even at a remote point from the distorting load.

As well as the above, interharmonics can perturb **ripple control signals** and at sub-harmonic levels can cause flicker.

To keep the harmonic voltage content within the recommended levels, the main solutions in current use are

- the use of high pulse rectification (e.g., smelters and **HVdc** converters)
- passive filters, either tuned to individual frequencies or of the band-pass type
- active filters and conditioners

The harmonic sources can be grouped in three categories according to their origin, size, and predictability, i.e., small and predictable (domestic and residential), large and random (arc furnaces), and large and predictable (static converters).

Small Sources

The residential and commercial power system contains large numbers of single-phase converter-fed power supplies with capacitor output smoothing, such as TVs and PCs, as shown in Fig. 62.2. Although their individual rating is insignificant, there is little diversity in their operation and their combined effect produces considerable odd-harmonic distortion. The gas discharge lamps add to that effect as they produce the same harmonic components.

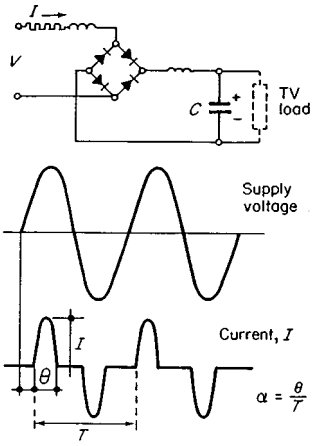


FIGURE 62.2 Single-phase bridge supply for a TV set.

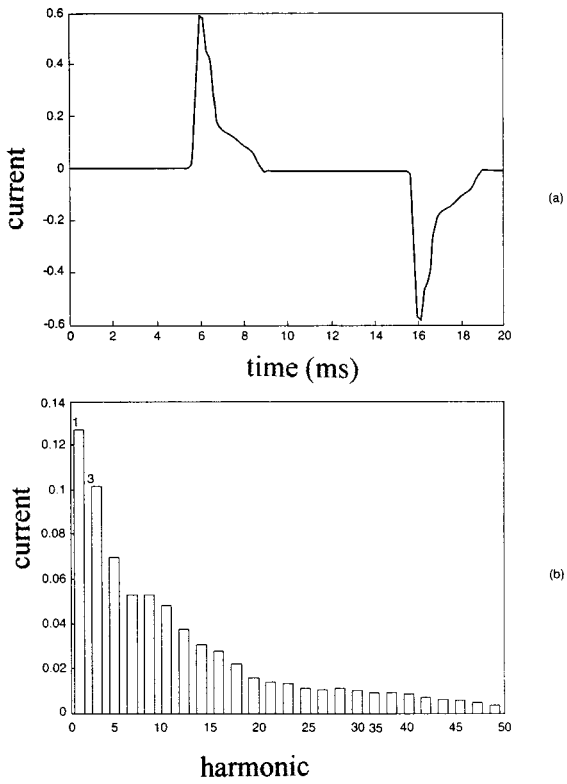


FIGURE 62.3 Current waveform (a) and harmonic spectrum (b) of a high efficiency lamp.

Figure 62.3 illustrates the current waveform and harmonic spectrum of a typical high efficiency lamp. The **total harmonic distortion (THD)** of such lamps can be between 50 and 150%.

Large and Random Sources

The most common and damaging load of this type is the arc furnace. Arc furnaces produce random variations of harmonic and interharmonic content which is uneconomical to eliminate by conventional filters.

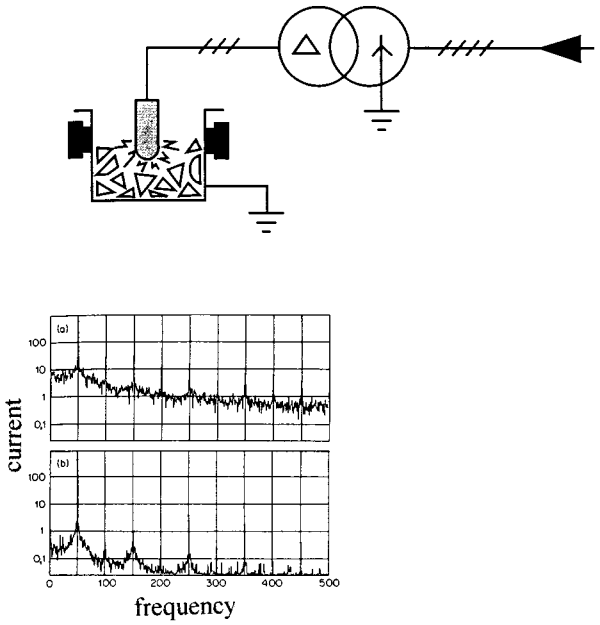


FIGURE 62.4 Typical frequency spectra of arc furnace operation. (a) During fusion; (b) during refining.

Figure 62.4 shows a snap-shot of the frequency spectra produced by an arc furnace during the melting and refining processes, respectively. These are greatly in excess of the recommended levels.

These loads also produce voltage fluctuations and flicker. Connection to the highest possible voltage level and the use of series reactances are among the measures currently taken to reduce their impact on power quality.

Static Converters

Large power converters, such as those found in smelters and HVdc transmission, are the main producers of harmonic current and considerable thought is given to their local elimination in their design.

The standard configuration for industrial and HVdc applications is the twelve-pulse converter, shown in Fig. 62.5. The "*characteristic*" harmonic currents for the configuration are of orders $12\,K \pm 1$ and their amplitudes are inversely proportional to the harmonic order, as shown by the spectrum of Fig. 62.6(b) which correspond to the time waveform of Fig. 62.6(a). These are, of course, maximum levels for ideal system conditions, i.e., with an infinite (zero impedance) ac system and a perfectly flat direct current (i.e., infinite smoothing reactance).

When the ac system is weak and the operation not perfectly symmetrical, **uncharacteristic harmonics** appear [Arrillaga, 1983].

While the characteristic harmonics of the large power converter are reduced by filters, it is not economical to reduce in that way the uncharacteristic harmonics and, therefore, even small injection of these harmonic currents can, via parallel resonant conditions, produce very large voltage distortion levels.

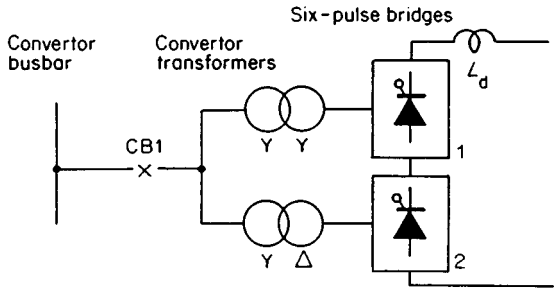


FIGURE 62.5 Twelve-pulse converter.

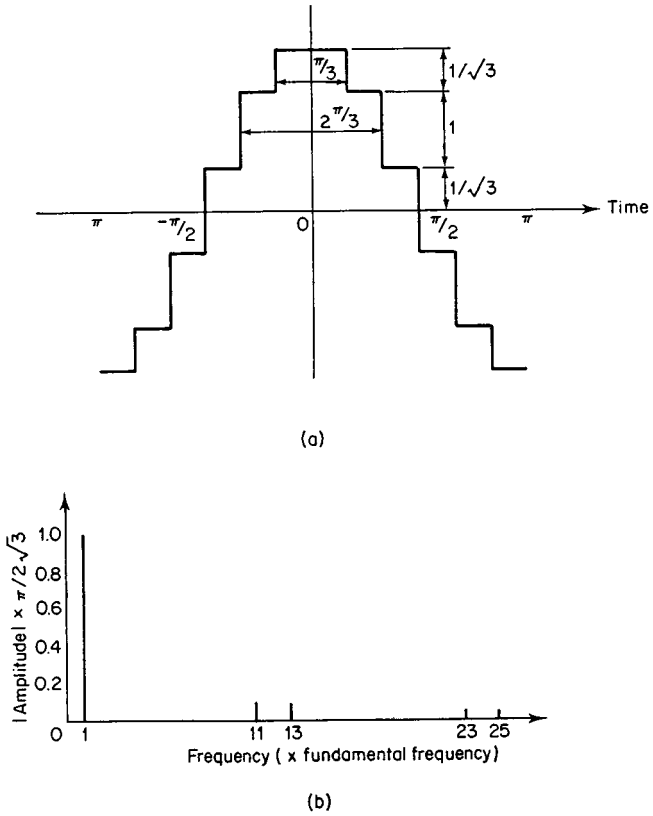


FIGURE 62.6 Twelve-pulse converter current. (a) Waveform; (b) spectrum.

An example of uncharacteristic converter behavior is the presence of fundamental frequency on the dc side of the converter, often induced from ac transmission lines in the proximity of the dc line, which produces second harmonic and direct current on the ac side.

Even harmonics, particularly the second, are very disruptive to power electronic devices and are, therefore, heavily penalized in the regulations.

The flow of dc current in the ac system is even more distorting, the most immediate effect being asymmetrical saturation of the converters or other transformers with a considerable increase in even harmonics which, under certain conditions, can lead to **harmonic instabilities** [Chen et al., 1996].

Another common example is the appearance of triplen harmonics. Asymmetrical voltages, when using a common firing angle control for all the valves, result in current pulse width differences between the three phases which produce triplen harmonics. To prevent this effect, modern large power converters use the equidistant firing concept instead [Ainsworth, 1968]. However, this controller cannot eliminate second harmonic amplitude modulation of the dc current which, via the converter modulation process, returns third harmonic current of positive sequence. This current can flow through the converter transformer regardless of its connection and penetrate far into the ac system. Again, the presence of triplen harmonics is discouraged by stricter limits in the regulations.

Voltage Fluctuations and Flicker

This group includes two broad categories, i.e.,

- step voltage changes, regular or irregular in time, such as those produced by welding machines, rolling mills, mine winders, etc. [Figs. 62.7(a) and (b)].
- cyclic or random voltage changes produced by corresponding variations in the load impedance, the most typical case being the arc furnace load (Fig. 62.7(c)).

Generally, since voltage fluctuations have an amplitude not exceeding ±10%, most equipment is not affected by this type of disturbance. Their main disadvantage is flicker, or fluctuation of luminosity of an incandescent lamp. The important point is that it is impossible, in practice, to change the characteristics of the filament. The physiological discomfort associated with this phenomenon depends on the amplitude of the fluctuations, the rate of repetition for voltage changes, and the duration of the disturbance. There is, however, a perceptibility threshold below which flicker is not visible.

Flicker is mainly associated with the arc furnaces because they draw different amounts of current each power cycle. The upshot is a modulation of the system voltage magnitude in the vicinity of the furnace. The modulation frequency is in the band 0 to 30 Hz, which is in the range that can cause noticeable flicker of light bulbs.

The flicker effect is usually evaluated by means of a flickermeter (IEC Publication 868). Moreover, the amplitude of modulation basically depends on the ratio between the impedance of the disturbing installation and that of the supply network.

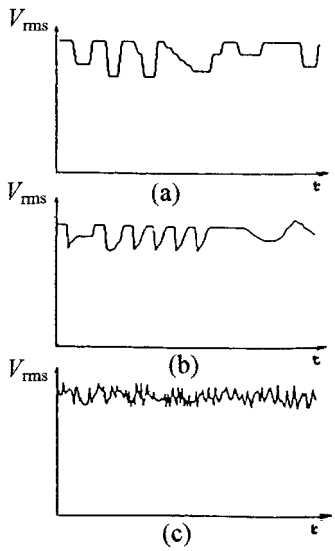


FIGURE 62.7 Voltage fluctuations.

Brief Interruptions, Sags, and Swells

Voltage Dips (SAGS)

A voltage dip is a sudden reduction (between 10 and 90%) of the voltage, at a point in the electrical system, such as that shown in Fig. 62.8, and lasting for 0.5 cycle to several seconds.

Dips with durations of less than half a cycle are regarded as transients.

A voltage dip may be caused by switching operations associated with temporary disconnection of supply, the flow of heavy current associated with the start of large motor loads or the flow of fault currents. These events may emanate from customers' systems or from the public supply network.

The main cause of momentary voltage dips is probably the lightning strike. In the majority of cases, the voltage drops to about 80% of its nominal value. In terms of duration, dips tend to cluster around three values: 4 cycles (the typical clearing time for faults), 30 cycles (the instantaneous reclosing time for breakers), and 120 cycles (the delayed reclosing time of breakers). The effect of a voltage dip on equipment depends on both its magnitude and its duration; in about 42% of the cases observed to date they are severe enough to exceed the tolerance standard adopted by computer manufacturers.

Possible effects are:

- extinction of discharge lamps
- incorrect operation of control devices
- speed variation or stopping of motors
- tripping of contactors
- computer system crash or measuring errors in instruments equipped with electronic devices
- commutation failure in HVdc converters [Arrillaga, 1983]

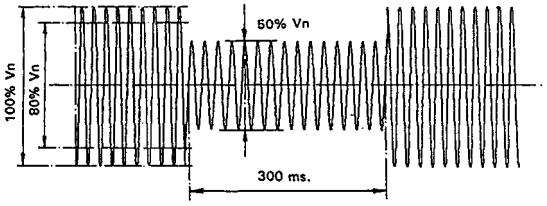


FIGURE 62.8 Voltage sag.

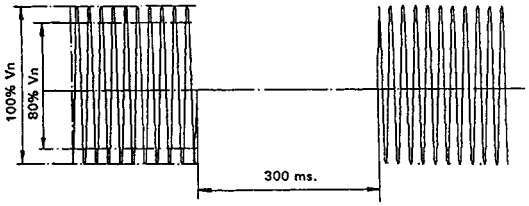


FIGURE 62.9 Voltage interruption.

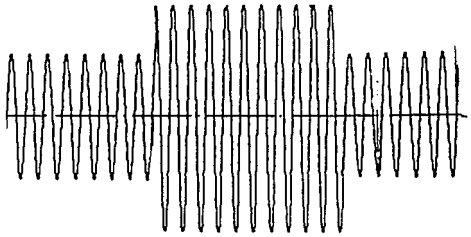


FIGURE 62.10 Voltage swell.

Brief Interruptions

Brief interruptions can be considered as voltage sags with 100% amplitude (see Fig. 62.9). The cause may be a blown fuse or breaker opening and the effect an expensive shutdown. For instance, a five-cycle interruption at a glass factory has been estimated as $200,000, and a major computer center reports that a 2-second outage can cost approximately $600,000. The main protection of the customer against such events is the installation of uninterruptible power supplies or power quality conditioners (discussed later).

Brief Voltage Increases (SWELLS)

Voltage swells, shown in Fig. 62.10, are brief increases in rms voltage that sometimes accompany voltage sags. They appear on the unfaulted phases of a three-phase circuit that has developed a single-phase short circuit. They also occur following load rejection.

Swells can upset electric controls and electric motor drives, particularly common adjustable-speed drives, which can trip because of their built-in protective circuitry. Swells may also stress delicate computer components and shorten their life.

Possible solutions to limit this problem are, as in the case of sags, the use of uninterruptible power supplies and conditioners.

Unbalances

Unbalance describes a situation, as shown in Fig. 62.11, in which the voltages of a three-phase voltage source are not identical in magnitude, or the phase differences between them are not 120 electrical degrees, or both. It affects motors and other devices that depend on a well-balanced three-phase voltage source.

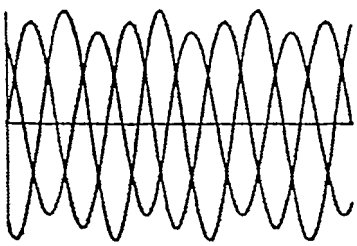


FIGURE 62.11 Voltage unbalance.

The degree of unbalances is usually defined by the proportion of negative and zero **sequence components**.

The main causes of unbalance are single-phase loads (such as electric railways) and untransposed overhead transmission lines.

A machine operating on an unbalanced supply will draw a current with a degree of unbalance several times that of the supply voltage. As a result, the three-phase currents may differ considerably and temperature rise in the machine will take place.

Motors and generators, particularly the large and more expensive ones, may be fitted with protection to detect extreme unbalance. If the supply unbalance is sufficient, the "*single-phasing*" protection may respond to the unbalanced currents and trip the machine.

Polyphase converters, in which the individual input phase voltages contribute in turn to the dc output, are also affected by an unbalanced supply, which causes an undesirable ripple component on the dc side, and non-characteristic harmonics on the ac side.

Transients

Voltage disturbances shorter than sags or swells are classified as transients and are caused by sudden changes in the power system [Greenwood, 1971]. They can be impulsive, generally caused by lightning and load switching, and oscillatory, usually due to capacitor-bank switching.

Capacitor switching can cause resonant oscillations leading to an overvoltage some three to four times the nominal rating, causing tripping or even damaging protective devices and equipment. Electronically based controls for industrial motors are particularly susceptible to these transients.

According to their duration, transient overvoltages can be divided into:

- switching surge (duration in the range of *ms*)
- impulse, spike (duration in the range of μs)

Surges are high-energy pulses arising from power system switching disturbances, either directly or as a result of resonating circuits associated with switching devices. They also occur during step load changes.

Impulses in microseconds, as shown in Fig. 62.12, result from direct or indirect lightning strokes, arcing, insulation breakdown, etc.

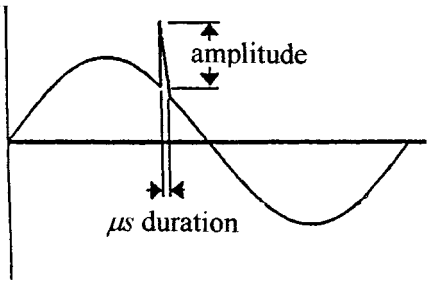


FIGURE 62.12 Impulse.

Protection against surges and impulses is normally achieved by surge-diverters and arc-gaps at high voltages and avalanche diodes at low voltages.

Faster transients in nanoseconds due to electrostatic discharges, an important category of EMC, are not normally discussed under Power Quality.

62.2 Power Quality Monitoring

Figure 62.13 illustrates the various components of a power quality detection system, i.e., voltage and current transducers, information transmission, instrumentation, and displays.

The most relevant information on power quality monitoring requirements can be found in the document IEC 1000-4.7. This document provides specific recommendations on monitoring accuracy in relation to the operating condition of the power system.

With reference to monitoring of individual frequencies, the maximum recommended relative errors for the magnitude and phase are 5% and 5°, respectively, under normal operating conditions and with constant voltage or current levels. However, such precision must be maintained for voltage variations of up to 20% (of nominal value) and 100% (peak value). For current measurements, the precision levels apply for overcurrents of up to 20% and peaks of 3 times rms value (on steady state) and 10 times the nominal current for a 1-sec duration.

Errors in the frequency response of current transformers occur due to capacitive effects, which are not significant in the harmonic region (say, up to the 50th harmonic), and also due to magnetizing currents. The

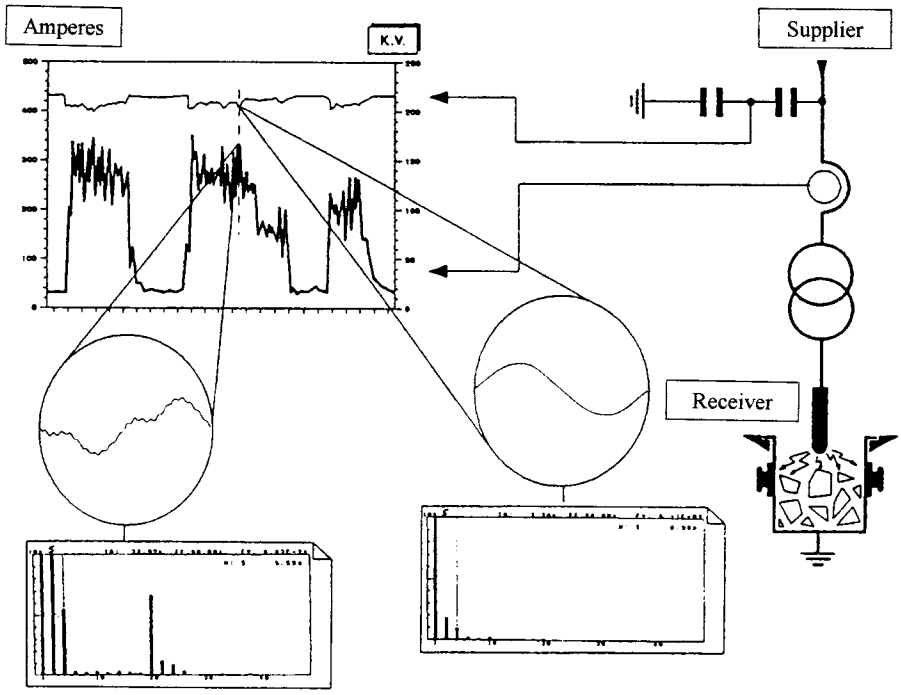


FIGURE 62.13 Power quality monitoring components.

latter can be minimized by reducing the current transformer load and improving the power factor; the ideal load being a short-circuited secondary with a *clamp* to monitor the current. Alternative transducers are being proposed for high frequency measurements using optical, magneto-optical, and Hall effect principles.

The iron-core voltage transformers respond well to harmonic frequencies for voltages up to 11 kV. Due to insulation capacitance, these transformers are not recommended for much higher voltages. The conventional capacitive voltage transformers (CVTs) are totally inadequate due to low frequency resonances between the capacitive divider and the output magnetic transformer; special portable capacitive dividers, without the output transformers, are normally used for such measurements. Again, alternative transducer principles, as for the current transformer, are being proposed for future schemes.

The signal transmission from the transducers to the control room passes through very noisy electromagnetic environments and the tendency is to use fiber optic cables, designed to carry either analog or digital samples of information in the time domain.

The time domain information is converted by signal or harmonic analyzers into the frequency domain; the instrumentation is also programmed to derive any required power quality indexes, such as THD (total harmonic distortion), EDV (equivalent distortion voltage), EDI (equivalent distortion current), etc.

The signal processing is performed by either analog or digital instrumentation, though the latter is gradually displacing the former. Most digital instruments in existence use the **FFT** (Fast Fourier Transform). The processing of information can be continuous or discontinuous depending on the characteristic of the signals under measurement with reference to waveform distortion. Document IEC 1000-4.7 lists the following types:

- quasi stationary harmonics
- fluctuating harmonics
- intermittent harmonics
- interharmonics

Only in the case of quasi stationary waveforms can the use of discontinuous monitoring be justified; examples of this type are the well-defined loads such as TV and PC sets.

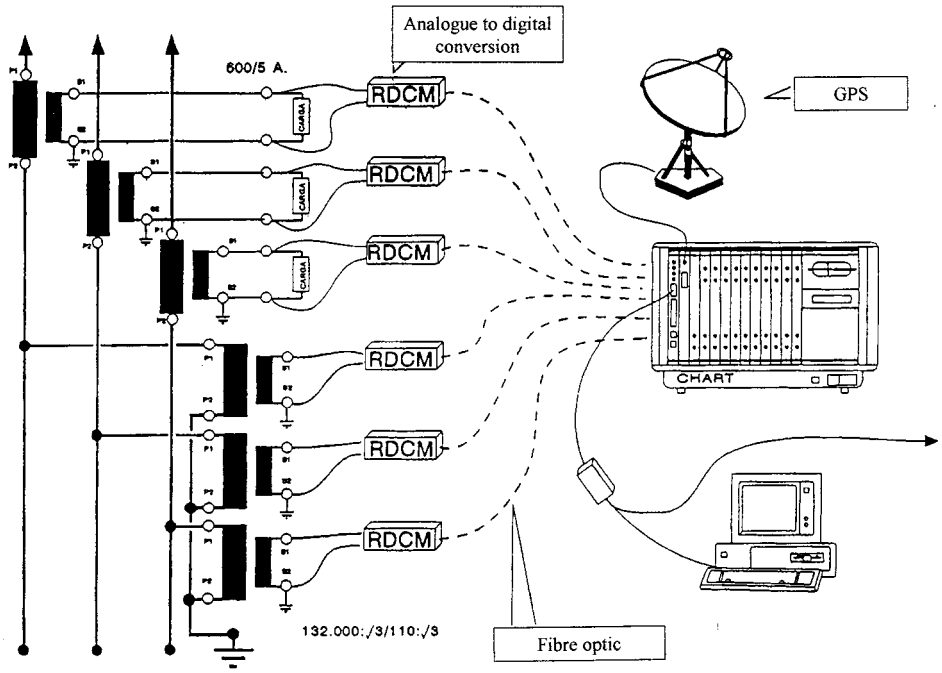


FIGURE 62.14 Simultaneous measurement of voltages and currents in a three-phase line.

In the remaining categories, it is necessary to perform real time continuous monitoring; examples of loads producing non-stationary distortion are arc furnaces and rolling mills.

Most of the instruments commercially available are not designed specifically for power system application, i.e., they are not multi-phase and cannot process continuous information. At the time of writing, the only system capable of multi-channel three-phase real time continuous monitoring is CHART [Miller and Dewe, 1992] which, although originally designed for harmonic monitoring, is capable of deriving continuous information of other power quality indexes such as flicker. It is based on the Intel Multi-bus II architecture and the RMX 386 operating system. An illustration of the system, shown in Fig. 62.14, includes remote data conversion modules, digital fiber optic transmission, **GPS** synchronization, central parallel processing, and ethernet-connected PCs for distant control and display.

62.3 Power Quality Conditioning

A common device in current use to ensure supply continuity for critical loads is the UPS or uninterruptible power supply. For brief interruptions, the UPS are of the static type, using batteries as the energy source and involving a rectifier/inverter system. A block diagram of a typical UPS is shown in Fig. 62.15 [Heydt, 1991].

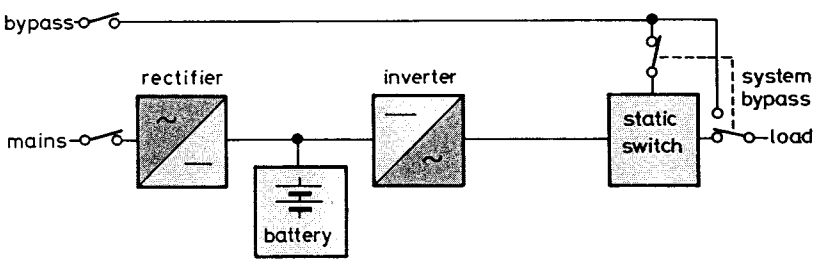


FIGURE 62.15 Uninterruptible power supply.

In the next few years power quality enhancements, in terms of reduced interruptions and voltage variations, can be expected by the application of power electronic controllers to utility distribution systems and/or at the supply end of many industrial and commercial customers.

Among the solutions already available are the solid state circuit breaker, the static condensers (or statcon), and the dynamic voltage restorer [Hingorani, 1995].

In a solid state circuit breaker, thyristors connected back-to-back form an ac switch module, several of which are, in turn, connected in series to acquire the required voltage rating. The breaker will interrupt the circuit at the first zero of the ac current. This means a delay of a few milliseconds, which should be acceptable for most applications.

Figure 62.16 shows a simplified illustration of a statcon which is made up of GTOs (Gate Turn Off) or similar devices such as insulated-gate bipolar transistors (IGBTs) or MOS-controlled thyristors (MCTs). The converter is driven by a dc storage device such as a dc capacitor, battery, or superconducting magnetic storage, and an ac transformer.. The dynamic voltage restorer, shown schematically in Fig. 62.17, turns a distorted waveform, including voltage dips, into the required waveform. The device injects the right amount of voltage by way of a series-connected transformer into the distribution feeder between the power supply side and load side.

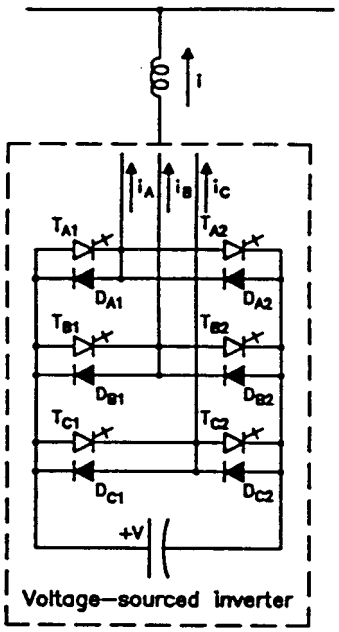


FIGURE 62.16 Static condenser.

The dynamic voltage restorer is similar to the statcon, with a transformer, converter, and storage, except that the transformer is connected in series with the busbar feeding the sensitive load. Compensation occurs in both directions, making up for the voltage dips and reducing the overvoltage. The response is very fast, occurring within a few milliseconds.

The capacity of the dc storage capacitor, in both the statcon and the dynamic voltage restorer, determines the duration of the correction provided for individual voltage dips. It can be a few cycles or seconds long. To enhance the load support capability, a storage battery with a booster electronic circuit can be connected in parallel with the capacitor.

Superconducting magnetic energy storage can be very effective to provide power for short periods. When the storage is not supporting the load, the converter will automatically charge the storage from the utility system, to be ready for the next event.

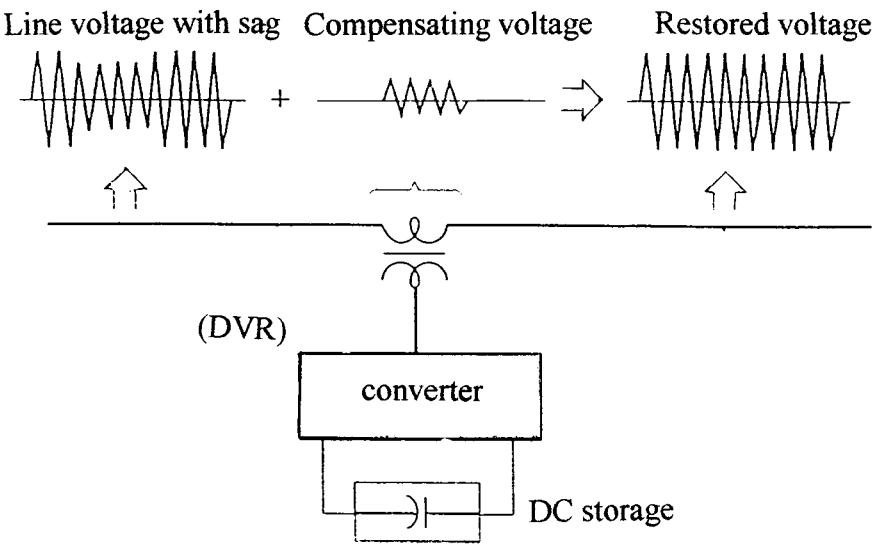


FIGURE 62.17 Dynamic voltage restorer.

Defining Terms

Distortion: Any deviation from a perfectly sinusoidal wave.
Disturbance: Any sudden change in the intended power, voltage, or current supply.
FFT (fast Fourier transform): Efficient computation of the discrete Fourier transform.
GPS (global positioning satellite): Used for time stamping and synchronization of multi-measurements at different geographical locations.
Harmonic instability: Extreme distortion of the voltage waveform at a particular frequency that causes inverter maloperation.
HVdc: High voltage direct current transmission.
Ripple control signal: A burst of pulses at a fixed non-harmonic frequency injected into the power system for the purpose of load management control.
Sequence components: Three symmetrical sets of voltages or currents equivalent to an asymmetrical three-phase unbalanced set.
THD (total harmonic distortions): The ratio of rms value of the harmonic content to the rms value of the generated frequency (in %).
Uncharacteristic harmonics: Static converter harmonics of orders different from $Pk \pm 1$ where P is the pulse number.

Related Topic

5.3 Distortion

References

J. Ainsworth, "The phase-locked oscillator. A new control system for controlled static convertors", *Trans. IEEE,* PAS-87, pp. 859-865, 1968.
J. Arrillaga, *High Voltage Direct Current Transmission*, London: IEE-Peter Peregrinus, 1983.
J. Arrillaga, D.A. Bradley, and P.S. Bolger, *Power System Harmonics*, London: John Wiley & Sons, 1985.
S. Chen, A.R. Wood, and J. Arrillaga, "HVdc converter transformer core saturation instability: a frequency domain analysis", *IEE Proc.—Gener. Transm. Distrib,* 143(1), 75-81, 1996.
A. Greenwood, *Electrical Transients in Power Systems*, New York: Wiley Interscience, 1971.
J. Heydt, *Electric Power Quality*, Stars in a Circle Publications, 1991.
N.G. Hingorani, "Introducing custom power", *IEEE Spectrum*, 41-48, June 1995.
International Electrotechical Commission Group, IEC TL 77, 1993.
A.J. Miller and M.B. Dewe, "Multichannel continuous harmonic analysis in real time", *Trans. IEEE Power Delivery,* 7(4), 1913-1919, 1992.

Further Information

Electric Power Quality by J. Heydt and *Power System Harmonics* by J. Arrillaga et al. are the only texts discussing the topic, though the latter is currently out of print. Two international conferences take place biennially specifically related to Power Quality; these are PQA (Power Quality: end use applications and perspectives) and the IEEE sponsored ICHQP (International Conference on Harmonics and Quality of Power).

Important information can also be found in the regular updates of the IEC and CENELEC standards, CIGRE, CIRED, UIC, and UNIPEDE documents and national guides such as the IEEE 519-1992.

Finally, the IEE and IEEE Journals on Power Transmission and Delivery, respectively, publish regularly important contributions in this area.

theory assumptions are both violated for the case when a loadflow does solve after discontinuous parameter change because the parameter variation is not continuous and smooth and the power system model may not be continuously differentiable at the point ($\mathbf{x}_0$, $\mathbf{p}_0$). The P-V curve, or Q-V curve, or eigenvalues and eigenvectors could be computed and used to assess proximity to voltage instability after each equipment outage or discontinuous parameter change when a loadflow solution exists to establish whether the solutions is stable or unstable at values of $\mathbf{p}$ above $\mathbf{p}_0$. The computation of the P-V curve, Q-V curve, or eigenvalues and eigenvectors requires significant computation and is not practical for screening thousands of contingencies for voltage instability or for assessing proximity to instability although they are used to assess stability and proximity to instability after a few selected contingencies. These methods also do not explicitly take into account the many discontinuities in the model and eigenvalues that occur for continuous parameter and discontinuous parameter changes. In many cases, the eigenvalue changes due to discontinuities is virtually all the change that occurs in an eigenvalue that approaches instability [IEEE, 1993] and the above methods have particular difficulty in such cases. The above methods cannot assess the agents that lose voltage instability for a particular event and cannot diagnose a cure when the loadflow has no solution for an equipment outage, wheeling or transaction combination, or both. These methods can provide a cure when a loadflow solution exists but its capabilities have not been compared to the f Security Assessment and Diagnosis proposed cure. The Voltage Stability
Security Assessment and Diagnosis (VSSAD) [Schlueter, 1998d] overcomes the above difficulties because:

1. It determines the number of discontinuities in any eigenvalue that have already occurred due to generator PV to load PQ bus type changes that are associated with an eigenvalue compared to the total number that are needed to produce voltage instability when the eigenvalue becomes negative. The eigenvalue is associated with a coherent bus group (voltage control area) [Schlueter, 1998a; f]. The subset of generators that experience PV-PQ bus type changes (reactive reserve basin) for computing a Q-V curve at any bus in that bus group are proven to capture the number of discontinuities in that eigenvalue [Schlueter, 1998a; f]. An eigenvalue approximation for the agent, composed of the test voltage control area where the Q-V current is computed and its reactive reserve basin, is used to theoretically justify the definitions of a voltage control area and the reactive reserve basin of an agent. The VSSAD agents are thus proven to capture eigenvalue structure of the loadflow jacobian evaluated at any operating point ($\mathbf{x}_0$, $\mathbf{p}_0$). The reactive reserve on generators in each voltage control area of a reactive reserve basin is proven to measure proximity to each of the remaining discontinuities in the eigenvalue required for bifurcation.
2. It can handle strictly discontinuous (equipment outage or large transfer or wheeling transaction changes) or continuous model or parameter change (load increase, transfer increases, and wheeling increases) whereas the above methods are restricted to continuous changes to assess stability or instability at a point $\mathbf{p}_0$.
3. It can simultaneously and quickly assess proximity to voltage instability for all agents where each has a bifurcating eigenvalue. Proximity to instability of any agent is measured by assessing (1) the percentage of voltage control areas containing generators in a reactive reserve basin with non-zero reserves, and (2) the percentage of base case reactive reserves remaining on reactive reserve basin voltage control areas that have not yet exhausted reserves [Schlueter, 1998b; f].
4. It can assess the cure for instability for contingencies that do not have a solution. The cure can be either (1) adding needed reactive reserve on specific generators to obtain a solution that is voltage stable, (2) adding reactive supply resources needed in one or more agents, or (3) the reduction in generation and load in one or more agents or between one or more agents to obtain a solution and assure that it is a stable solution. These cures can be obtained in an automated fashion [Schlueter, 1998b; f]. The diagnosis can also indicate if the lack of a solution is due to convergence difficulties or round-off error if the diagnosis indicates the contingency combination does not produce sufficient network reactive losses to cause instability or any agent.
5. It can provide operating constraints or security constraints on each agent's reactive reserve basin reserves that prevent voltage instability in an agent in a manner identical to how thermal constraints prevent thermal overload on a branch and voltage constraints prevent bus voltage limit violation at a bus [Schlueter, 1998c; f].

6. The reactive reserve basin operating constraints allow optimization that assures that correcting one voltage instability problem due to instability in one or more agents will not produce other voltage stability problems in the rest of the system [Schlueter, 1998c; f].
7. The reactive reserve basin constraints after an equipment outage and operating change combination allows optimization of transmission capacity that specifically corrects that particular equipment outage and transaction change induced voltage instability with minimum control change [Schlueter, 1998c; f].
8. It requires very little computation per contingency and can find multiple contingencies that cause voltage instability by simulating only a small percentage of the possible multiple contingencies [Schlueter, 1998d].

Kinds of Loadflow Instability

Two kinds of voltage instability have been associated with a loadflow model: loss of control voltage instability and clogging voltage instability [Schlueter, 1998d]. Loss of control voltage instability is caused by exhaustion of reactive power supply that produces loss of voltage control on some of the generators or synchronous condensers. Loss of voltage control on these reactive supply devices implies both lack of any further reactive supply from these devices and loss of control of voltage that will increase network reactive losses that absorb a portion of the flow of reactive power supply and prevent it from reaching the subregion needing that reactive supply. Loss of voltage control develops because of equipment outages (generator, transmission line, and transformer), operating condition changes (wheeling, interchange, and transfer transactions), and load/generation pattern changes. Loss of control voltage instability occurs in the subtransmission and transmission system [Schlueter, 1998d]. It produces either saddle node or singularity-induced bifurcation in a differential algebraic model. On the other hand, clogging develops because of increasing reactive power losses, and switching shunt capacitors and tap changers reaching their limits. These network reactive losses, due to increasing magnetic field and shunt capacitive supply withdrawal, can completely block reactive power supply from reaching the subregion with need [Schlueter, 1998d]. Clogging voltage instability can produce algebraic bifurcation in a differential algebraic model. The VSSAD method can diagnose whether the voltage instability occurs due to clogging or loss of control voltage instability for each equipment outage, transaction combination, or both that have no solution.

Theoretical Justification of the Diagnosis in VSSAD

A bifurcation subsystem analysis has been developed that theoretically justifies the diagnosis performed by [Schlueter, 1997; 1998a; b; d; f]. This bifurcation subsystem analysis for a loadflow model attempts to break the loadflow model into a subsystem model and external model

$$f\left(x_s, x_e, p\right) = \begin{bmatrix} f_s\left(x_s, x_e, p\right) \\ f_e\left(x_s, x_e, p\right) \end{bmatrix} = \begin{bmatrix} O_{n_1} \\ O_{n_2} \end{bmatrix} \tag{63.47}$$

and to break the state x into two components $x = \begin{bmatrix} x_s \\ x_e \end{bmatrix}$ where x_s is the dimension of $f_s(x_s, x_e, p) = O_{n_i}$.

The bifurcation occurs at $p^\star = p_o + \mu_o^\star \underline{n}$ when

$$\begin{bmatrix} \frac{\partial f_s}{\partial x_s}\left(x^\star, p^\star\right) & \frac{\partial f_s}{\partial x_e}\left(x^\star, p^\star\right) \\ \frac{\partial f_e}{\partial x_s}\left(x^\star, p^\star\right) & \frac{\partial f_e}{\partial x_e}\left(x^\star, p^\star\right) \end{bmatrix} \begin{bmatrix} u_i\left(p^\star\right) \\ w_i\left(p^\star\right) \end{bmatrix} = \lambda_i\left(p^\star\right) \begin{bmatrix} u_i(\\ w_i(\end{bmatrix} \tag{63.48}$$

The vector $\begin{bmatrix} u_i(p^*) \\ w_i(p^*) \end{bmatrix}$ is the right eigenvector of eigenvalue $\lambda_i(p^*) = 0$ at bifurcation point p^*. A bifurcation subsystem exists if two conditions hold:

$$\frac{\partial f_s}{\partial x_s}(x^*, p^*) \quad u_i(p^*) = 0 \tag{63.49}$$

$$\frac{\partial f_s}{\partial x_e}\frac{\partial f_e^{-1}}{\partial x_e}\frac{\partial f_e}{\partial x_s}u_i(p^*) = 0 \tag{63.50}$$

The first condition is called the bifurcation subsystem condition and the second is called the geometric decoupling condition. Finding a bifurcation subsystem for any bifurcation of the full system model requires finding the combination of correct dimension, correct subset of equations, and correct subset of variables such that the subsystem experiences the bifurcation (Eq. (63.49)) of the full system model (Eq. (63.48)) but also produces that bifurcation since the external model is completely uncoupled from the bifurcation subsystem in the direction of the right eigenvector (Eq. (63.50)). The right eigenvector is an approximation of the center manifold at bifurcation, and the center manifold is the subsystem that actually experiences the bifurcation and is obtained via a nonlinear transformation of the model. The expectation of finding a bifurcation subsystem for any loadflow bifurcation, noting the above requirements for identifying such a bifurcation subsystem, is that the difficulty in finding a bifurcation subsystem would be great even though one may exist for some bifurcations. The results in [Schlueter, 1998b; f] prove that one cannot only describe the bifurcation subsystem (where) for every clogging voltage instability and for every loss of control voltage instability, but also can theoretically establish diagnostic information on when, proximity, and cure for a specific bifurcation in a specific bifurcation subsystem for clogging or for loss of control voltage instability [Schlueter, 1998b; f].

The analysis establishes that:

1. The real power balance equations are a bifurcation subsystem for angle instability when the loadflow model is decoupled ($\frac{dP}{dV}$ and $\frac{dQ}{d\theta}$ are assumed null) [Schlueter, 1998b; f].
2. The reactive power balance equations are a bifurcation subsystem for voltage instability when the loadflow model is assumed decoupled [Schlueter, 1998b; f].
3. A voltage control area is the bifurcation subsystem (agent) for clogging voltage instability. The agent is vulnerable to voltage instability for loss of generation in the agent, line outage in the agent boundary, or increased real and reactive flow across the agent boundary based on analysis of the lower bound approximation of the eigenvalue associated with that agent. The cure for clogging voltage instability in this agent is to reduce the real and reactive flow across the boundary of the agent [Schlueter, 1998b; f].
4. A voltage control area and its associated reactive reserve basin are the bifurcation subsystem (agent) for loss of control voltage instability. The agent is vulnerable to voltage instability for loss of generation in the agent, line outages, transfer or wheeling transactions that reduce reactive reserve basin reserves based on analysis of the lower bound approximation of the eigenvalue associated with that agent. The cure for voltage instability in the agent is to add reactive reserves on the reactive reserve basin via capacitor insertion, generator voltage setpoint changes on reactive reserve basin generators, or reverse tap position changes on underload tap changers [Schlueter, 1998b; f].
5. The percentage of reserves unexhausted in the reactive reserve basin is theoretically justified as a proximity measure for clogging instability in any clogging voltage instability agent. The percentage of voltage control areas in a reactive reserve basin with unexhausted reactive reserve is theoretically justified as a proximity measure for each loss of control voltage instability agent [Schlueter, 1998b; f].
6. Exhaustion of reactive reserves in a particular locally most vulnerable agent's reactive reserve basin causes cascading exhaustion of reactive reserves and loss of control voltage instability in agents with successively

larger reactive reserve basins. This partially explains why voltage collapse occurs [Schlueter, 1998a; d; f] which is a cascading loss of stability in several agents.

The automated diagnostic procedures in VSSAD are thus theoretically justified via this bifurcation subsystem analysis.

Future Research

Research is needed to:

1. Develop improved nonlinear dynamic load models that are valid at any particular instant and that are valid when voltage decline is severe. The lack of accurate load models makes it difficult to accurately simulate the time behavior and/or assess the cause of the voltage instability. The lack of knowledge of what constitutes an accurate load model makes accurate postmortem simulation of a particular blackout a process of making trial and error assumptions on the load model structure to obtain as accurate a simulation as possible that conforms with time records of the event. Accurate predictive simulation of events that have not occurred is very difficult [Taylor, et al. 1998].
2. Explain (a) why each specific cascading sequence of bifurcations inevitably occurs in a differential algebraic model, and (b) the dynamic signature associated with each bifurcation sequence. Work is underway to explain why instability in generator and load dynamics can inevitably cause a singularity-induced bifurcation to occur. The time signature for singularity-induced bifurcation changes dependence on why it occurs is discussed in [Schlueter, 1998e; Liu, 1998].
3. Extend bifurcation subsystem analysis to the differential algebraic model and link the bifurcation subsystem in a differential algebraic model, to those obtained in the loadflow model. The bifurcation subsystems for different Hopf and saddle node bifurcations can explain why the subsystem experiences instability, as well as how to prevent instability as has been possible for bifurcation subsystems in the algebraic model. Knowledge of bifurcation subsystems in the algebraic model may assist in identifying bifurcation subsystems in the differential algebraic model.
4. Develop a protective or corrective control for voltage instability. A protective control would use constraints on the current operating condition for contingencies predicted to cause voltage instability if they occurred. These constraints on the current operation would prevent voltage instability if and when the contingency occurred. A corrective control would develop a control that correct the instability in the bifurcation subsystems experiencing instability only after the equipment outages or operating changes predicted to produce voltage instability have occurred. The implementation of the corrective control requires a regional 5-s updated data acquisition system and control implementation similar to that used in Electricité de France and elsewhere in Europe.

Defining Terms

Power system stability: The property of a power system that enables it to remain in a state of operating equilibrium under normal operating conditions and to converge to another acceptable state of equilibrium after being subjected to a disturbance. Instability occurs when the above is not true or when the system loses synchronism between generators and between generators and loads.

Small signal stability: The ability of the power system to maintain synchronism under small disturbances [Kundur, 1994].

Transient stability: The ability of a power system to maintain synchronism for a severe transient disturbance [Kundur, 1994].

Rotor angle stability: The ability of the generators in a power system to remain in synchronism after a severe transient disturbance [Kundur, 1994].

Voltage viability: The ability of a power system to maintain acceptable voltages at all buses in the system after being subjected to a disturbance. Loss of viability can occur if voltage at some bus or buses are below acceptable levels [Kundur, 1994]. Loss of viability is not voltage instability.

Voltage stability: The ability of the combined generation and transmission system to supply load after a disturbance, increased load, or change in system conditions without an uncontrollable and progressive decrease in voltage [Kundur, 1994]. Loss of voltage instability may stem from the attempt of load dynamics to restore power consumption beyond the capability of the combined transmission and generation system. Both small signal and transient voltage instability can occur.

Voltage collapse: An instability that produces a cascading (1) loss of stability in subsystems, and/or (2) outage of equipment due to relaying actions.

Bifurcation: A sudden change in system response from a smooth, continuous, slow change in parameters p.

References

T.M. Apostol, *Mathematical Analysis,* Second Edition, Addison-Wesley Publishing, 1974.

K. Ben-Kilani, Bifurcation Subsystem Method and its Application to Diagnosis of Power System Bifurcations Produced by Discontinuities, Ph.D. Dissertation, Michigan State University, August 1997.

C.A. Canizares, F.L. Alvarado, C.L. DeMarco, I. Dobson, and W.F. Long, Point of collapse methods applied to AC/DC power system, *IEEE Trans. on Power System,* 7, 673–683, 1992.

I. Dobson and Liming Lu, Using an iterative method to compute a closest saddle node bifurcation in the load power parameter space of an electric power system, in *Proceedings of the Bulk Power System Voltage Phenomena. II. Voltage Stability and Security,* Deep Creek Lake, MD, 1991.

T.Y. Guo and R.A. Schlueter, Identification of generic bifurcation and stability problems in a power system differential algebraic model, *IEEE Trans. on Power Systems,* 9, 1032–1044, 1994.

IEEE Working Group on Voltage Stability, Suggested Techniques for Voltage Stability Analysis, IEEE Power Engineering Society Report, 93TH0620-5PWR, 1993.

P. Kundur, *Power System Stability and Control,* Power System Engineering Series, McGraw-Hill, 1994.

S. Liu, Bifureation Dynamics as a Cause of Recent Voltage Collapse Problems on the WSCC System, Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1998.

N.D. Reppen and R.R. Austria, Application of the optimal power flow to analysis of voltage collapse limited power transfer, in *Bulb Power System Voltage Phenomena. II. Voltage Stability and Security,* August 1991, Deep Creek Lake, MD.

Survey of Voltage Collapse Phenomena: Summary of Interconnection Dynamics Task Force's Survey on Voltage Collapse Phenomena, Section III Incidents, North American Reliability Council Report, August, 1991.

P.W. Sauer, C. Rajagopalan, B. Lesieutre, and M.A. Pai, Dynamic Aspects of voltage/power characteristics, *IEEE Trans. on Power Systems,* 7, 990–1000, 1992.

R.A. Schlueter, K. Ben-Kilani, and U. Ahn, Impact of modeling accuracy on type, kind, and class of stability problems in a power system model, *Proceedings of the ECC & NSF International Workshop on Bulk Power System Voltage Stability, Security and Control Phenomena-III,* pp. 117–156. August 1994.

R.A. Schlueter, A structure based hierarchy for intelligent voltage stability control in planning, scheduling, and stabilizing power systems, *Proceedings of the EPRI Conference on Future of Power Delivery in the 21st Century,* La Jolla, CA, November 1997.

R.A. Schlueter and S. Liu, Justification of the voltage stability security assessment as an improved modal analysis procedure, *Proceedings of the Large Engineering System Conference on Power System Engineering,* pp. 273–279, June 1998.

R.A. Schlueter, K. Ben-Kilani, and S. Liu, Justification of the voltage security assessment method using the bifurcation subsystem method, *Proceedings of the Large Engineering System Conference on Power Systems,* pp. 266-272, June 1998.

R.A. Schlueter and S. Liu, A structure based hierarchy for intelligent voltage stability control in operation planning, scheduling, and dispatching power systems, *Proceedings of the Large Engineering System Conference on Power System Engineering,* pp. 280–285, June 1998.

R.A. Schlueter, A voltage stability security assessment method, *IEEE Trans. on Power Systems,* 13, 1423-1438, 1998.

R.A. Schlueter, S. Liu, K. Ben-Kilani, and I.-P. Hu, Static voltage instability in generator flux decay dynamics as a cause of voltage collapse, accepted for publication in the *Journal on Electric Power System Research, July 1998.*

R. Schlueter, S. Liu, and N. Alemadi, Intelligent Voltage Stability Assessment Diagnosis, and Control of Power Systems Using a Modal Structure, Division of Engineering Research Technical Report, December 1998 and distributed to attendees of *Bulk Power System Dynamics and Control IV; Restructuring*, August 24–28, 1998, Santorini, Greece.

C. Taylor, *Power System Voltage Stability,* Power System Engineering Series, McGraw-Hill, New York, 1994.

C. Taylor, D. Kostorev, and W. Mittlestadt, Model validation for the August 10, 1996 WSCC outage, *IEEE Winter Meeting,* paper PE-226-PWRS-0-12-1997.

T. Van Cutsem, A method to compute reactive power margins with respect to voltage collapse, in *IEEE Trans. on Power Systems,* 6, 145–156, 1991.

T. Van Cutsem and C. Vournas, Voltage stability of electric power systems, *Power Electronic and Power System Series,* Kluwer Academic Publisher, Boston, MA, 1998.

V. Venkatasubramanian, X. Jiang, H. Schattler, and J. Zaborszky, Current status of the taxonomy theory of large power system dynamics, DAE systems with hard limits, *Proceedings of the Bulk Power System Voltage; Phenomena-III Stability, Security and Control,* pp. 15–103, August 1994.

Further Reading

There are several good books that discuss voltage stability. Kundur [1994] is the most complete in describing the modeling required to perform voltage stability as well as some of the algebraic model-based methods for assessing proximity to voltage instability. Van Cutsem and Vournas' book [1998] provides the only dynamical systems discussion of voltage instability and provides a picture of the various dynamics that play a role in producing voltage instability. Methods for analysis and simulation of the voltage instability dynamics are presented. This analysis and simulation is motivated by a thorough discussion of the network, generator, and load dynamics models and their impacts on voltage instability. Taylor [1994] provides a tutorial review of voltage stability, the modeling needed, and simulation tools required and how they can be used to perform a planning study on a particular utility or system.

The *IEEE Transactions on Power Systems* is a reference for the most recent papers on voltage viability and voltage instability problems. The *Journal of Electric Power Systems Research* and *Journal on Electric Machines and Power Systems* also contain excellent papers on voltage instability.

Gross, C.A. “Power Transformers”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

64

Power Transformers

64.1 Transformer Construction
The Transformer Core • Core and Shell Types • Transformer Windings • Taps

64.2 Power Transformer Modeling
The Three-Winding Ideal Transformer Equivalent Circuit • A Practical Three-Winding Transformer Equivalent Circuit • The Two-Winding Transformer

64.3 Transformer Performance

64.4 Transformers in Three-Phase Connections
Phase Shift in Y–Δ Connections • The Three-Phase Transformer • Determining Per-Phase Equivalent Circuit Values for Power Transformers: An Example

64.5 Autotransformers

Charles A. Gross
Auburn University

64.1 Transformer Construction

The Transformer Core

The core of the power TRANSFORMER is usually made of laminated cold-rolled magnetic steel that is grain oriented such that the rolling direction is the same as that of the flux lines. This type of core construction tends to reduce the eddy current and hysteresis losses. The eddy current loss P_e is proportional to the square of the product of the maximum flux density B_M (T), the frequency f (Hz), and thickness t (m) of the individual steel lamination.

$$P_e = K_e(B_M t f)^2 \quad \text{(W)} \tag{64.1}$$

K_e is dependent upon the core dimensions, the specific resistance of a lamination sheet, and the mass of the core. Also,

$$P_h = K_h f B_M^n \quad \text{(W)} \tag{64.2}$$

In Eq. (64.2), P_h is the hysteresis power loss, n is the Steinmetz constant $(1.5 < n < 2.5)$ and K_h is a constant dependent upon the nature of core material and varies from $3 \times 10^{-3} m$ to $20 \times 10^{-3} m$, where m = core mass in kilograms.

The core loss therefore is

$$P_e = P_e + P_h \tag{64.3}$$

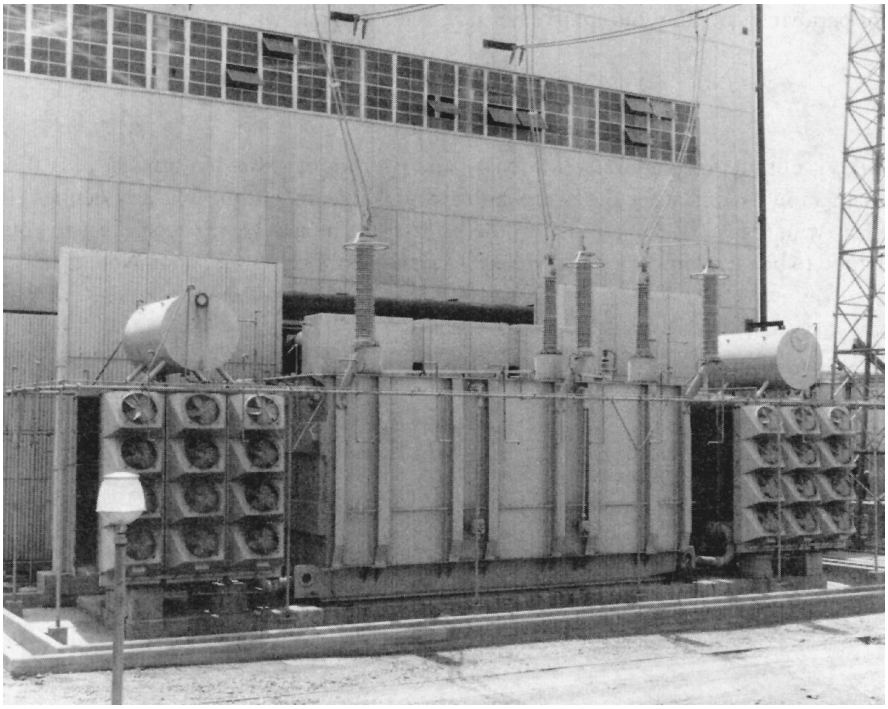
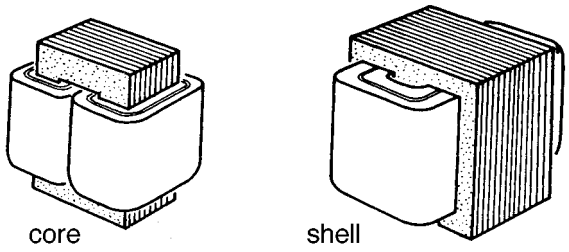


FIGURE 64.1 230kVY:17.1kVΔ 1153-MVA 3ϕ power transformer. (Photo courtesy of General Electric Company.)

Core and Shell Types

Transformers are constructed in either a shell or a core structure. The shell-type transformer is one where the windings are completely surrounded by transformer steel in the plane of the coil. Core- type transformers are those that are not shell type. A power transformer is shown in Fig. 64.1.

Multiwinding transformers, as well as polyphase transformers, can be made in either shell- or core-type designs.



Transformer Windings

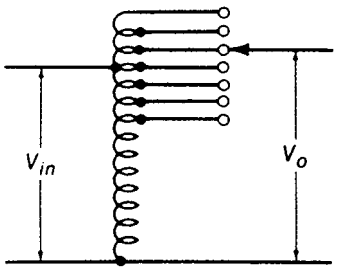
The windings of the power transformer may be either copper or aluminum. These conductors are usually made of conductors having a circular cross section; however, larger cross-sectional area conductors may require a rectangular cross section for efficient use of winding space.

The life of a transformer insulation system depends, to a large extent, upon its temperature. The total temperature is the sum of the ambient and the temperature rise. The temperature rise in a transformer is intrinsic to that transformer at a fixed load. The ambient temperature is controlled by the environment the transformer is subjected to. The better the cooling system that is provided for the transformer, the higher the "kVA" rating for the same ambient. For example, the kVA rating for a transformer can be increased with forced air (fan) cooling. Forced oil and water cooling systems are also used. Also, the duration of operating time at high temperature directly affects insulation life.

Other factors that affect transformer insulation life are vibration or mechanical stress, repetitive expansion and contraction, exposure to moisture and other contaminants, and electrical and mechanical stress due to overvoltage and short-circuit currents.

Paper insulation is laid between adjacent winding layers. The thickness of this insulation is dependent on the expected electric field stress. In large transformers oil ducts are provided using paper insulation to allow a path for cooling oil to flow between coil elements.

The short-circuit current in a transformer creates enormous forces on the turns of the windings. The short-circuit currents in a large transformer are typically 8 to 10 times larger than rated and in a small transformer are 20 to 25 times rated. The forces on the windings due to the short-circuit current vary as the square of the current, so whereas the forces at rated current may be only a few newtons, under short-circuit conditions these forces can be tens of thousands of newtons. These mechanical and thermal stresses on the windings must be taken into consideration during the design of the transformer. The current-carrying components must be clamped firmly to limit movement. The solid insulation material should be precompressed and formed to avoid its collapse due to the thermal expansion of the windings.



Taps

Power transformer windings typically have taps, as shown. The effect on transformer models is to change the turns ratio.

64.2 Power Transformer Modeling

The electric power **transformer** is a major power system component which provides the capability of reliably and efficiently changing (transforming) ac voltage and current at high power levels. Because electrical power is proportional to the product of voltage and current, for a specified power level, low current levels can exist only at high voltage, and vice versa.

The Three-Winding Ideal Transformer Equivalent Circuit

Consider the three coils wrapped on a common core as shown in Fig. 64.2(a). For an infinite core permeability (μ) and windings made of material of infinite conductivity (σ):

$$v_1 = N_1 \frac{d\phi}{dt} \quad v_2 = N_2 \frac{d\phi}{dt} \quad v_3 = N_3 \frac{d\phi}{dt} \tag{64.4}$$

where ϕ is the core flux. This produces:

$$\frac{v_1}{v_2} = \frac{N_1}{N_2} \quad \frac{v_2}{v_3} = \frac{N_2}{N_3} \quad \frac{v_3}{v_1} = \frac{N_3}{N_1} \tag{64.5}$$

For sinusoidal steady state performance:

$$\overline{V}_1 = \frac{N_1}{N_2}\overline{V}_2 \quad \overline{V}_2 = \frac{N_2}{N_3}\overline{V}_3 \quad \overline{V}_3 = \frac{N_3}{N_1}\overline{V}_1 \tag{64.6}$$

where $\overline{V}$, etc. are complex phasors.

The circuit symbol is shown in Fig. 64.2(b). Ampere's law requires that

$$\oint \hat{H} \cdot \hat{dl} = i_{\text{enclosed}} = 0 \tag{64.7}$$

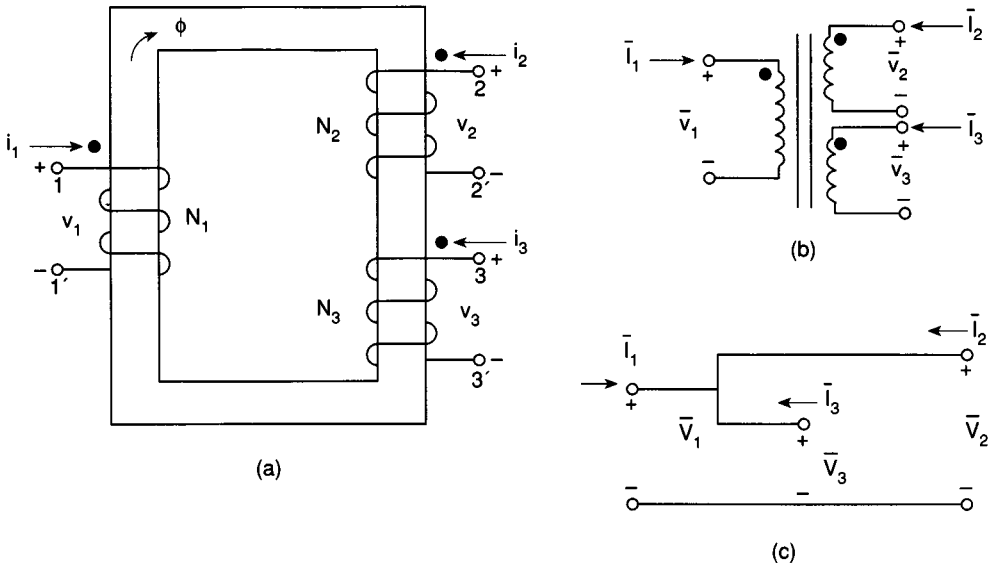


FIGURE 64.2 Ideal three-winding transformer. (a) Ideal three-winding transformer; (b) schematic symbol; (c) per-unit equivalent circuit.

$$0 = N_1 i_1 + N_2 i_2 + N_3 i_3 \tag{64.8}$$

Transform Eq. (64.8) into phasor notation:

$$N_1 \bar{I}_1 + N_2 \bar{I}_2 + N_3 \bar{I}_3 = 0 \tag{64.9}$$

Equations (64.6) and (64.9) are basic to understanding transformer operation. Consider Eq. (64.6). Also note that $-V_1$, $-V_2$, and $-V_3$ must be in phase, with dotted terminals defined positive. Now consider the total input complex power $-S$.

$$\bar{S} = \bar{V}_1 \bar{I}_1^{\star} + \bar{V}_2 \bar{I}_2^{\star} + \bar{V}_3 \bar{I}_3^{\star} = 0 \tag{64.10}$$

Hence, ideal transformers can absorb neither real nor reactive power.

It is customary to scale system quantities (*V, I, S, Z*) into dimensionless quantities called per-unit values. The basic per-unit scaling equation is

$$\text{Per-unit value} = \frac{\text{actual value}}{\text{base value}}$$

The base value always carries the same units as the actual value, forcing the per-unit value to be dimensionless. Base values normally selected arbitrarily are V_{base} and S_{base}. It follows that:

$$I_{\text{base}} = \frac{S_{\text{base}}}{V_{\text{base}}}$$

$$Z_{\text{base}} = \frac{V_{\text{base}}}{I_{\text{base}}} = \frac{V_{\text{base}}^2}{S_{\text{base}}}$$

When per-unit scaling is applied to transformers V_{base} is usually taken as V_{rated} as in each winding. S_{base} is common to all windings; for the two- winding case S_{base} is S_{rated}, since S_{rated} is common to both windings.

Per-unit scaling simplifies transformer circuit models. Select two primary base values, V_{1base} and S_{1base}. Base values for windings 2 and 3 are:

$$V_{2_{base}} = \frac{N_2}{N_1} V_{1_{base}} \qquad V_{3_{base}} = \frac{N_3}{N_1} V_{1_{base}} \tag{64.11}$$

and

$$S_{1_{base}} = S_{2_{base}} = S_{3_{base}} = S_{base} \tag{64.12}$$

By definition:

$$I_{1_{base}} = \frac{S_{base}}{V_{1_{base}}} \qquad I_{2_{base}} = \frac{S_{base}}{V_{2_{base}}} \qquad I_{3_{base}} = \frac{S_{base}}{V_{3_{base}}} \tag{64.13}$$

It follows that

$$I_{2_{base}} = \frac{N_1}{N_2} I_{1_{base}} \qquad I_{3_{base}} = \frac{N_1}{N_3} I_{1_{base}} \tag{64.14}$$

Thus, Eqs. (64.3) and (64.6) scaled into per-unit become:

$$\overline{V}_{1_{pu}} = \overline{V}_{2_{pu}} = \overline{V}_{3_{pu}} \tag{64.15}$$

$$\bar{I}_{1_{pu}} + \bar{I}_{2_{pu}} + \bar{I}_{3_{pu}} = 0 \tag{64.16}$$

The basic per-unit equivalent circuit is shown in Fig. 64.2(c). The extension to the *n*-winding case is clear.

A Practical Three-Winding Transformer Equivalent Circuit

The circuit of Fig. 64.2(c) is reasonable for some power system applications, since the core and windings of actual transformers are constructed of materials of high μ and σ, respectively, though of course not infinite. However, for other studies, discrepancies between the performance of actual and ideal transformers are too great to be overlooked. The circuit of Fig. 64.2(c) may be modified into that of Fig. 64.3 to account for the most important discrepancies. Note:

R_1,R_2,R_3 Since the winding conductors cannot be made of material of infinite conductivity, the windings must have some resistance.

X_1,X_2,X_3 Since the core permeability is not infinite, not all of the flux created by a given winding current will be confined to the core. The part that escapes the core and seeks out parallel paths in surrounding structures and air is referred to as *leakage* flux.

R_c,X_m Also, since the core permeability is not infinite, the magnetic field intensity inside the core is not zero. Therefore, some current flow is necessary to provide this small H. The path provided in the circuit for this "magnetizing" current is through X_m. The core has internal power losses, referred to as *core loss*, due to hystereses and eddy current phenomena. The effect is accounted for in the resistance R_c. Sometimes R_c and X_m are neglected.

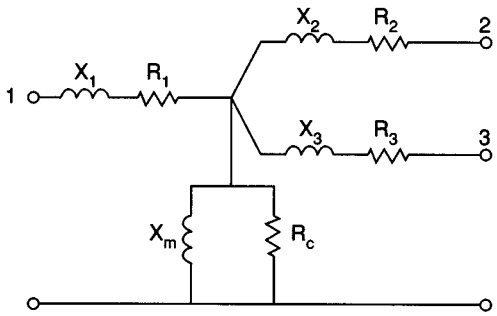


FIGURE 64.3 A practical equivalent circuit.

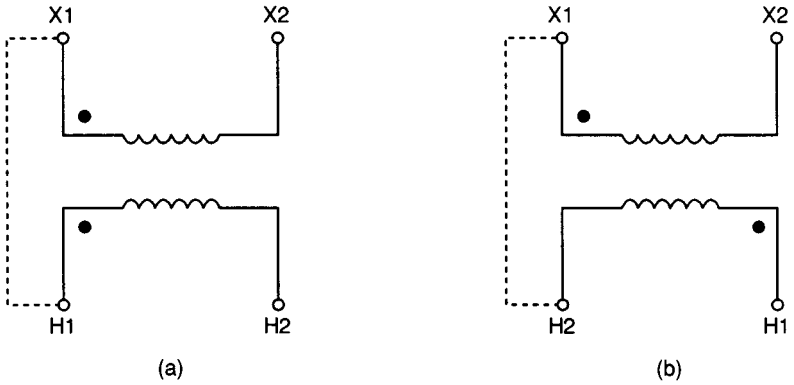


FIGURE 64.4 Transformer polarity terminology: (a) subtractive; (b) additive.

The circuit of Fig. 64.3 is a refinement on that of Fig. 64.2(c). The values R_1, R_2, R_3, X_1, X_2, X_3 are all small (less than 0.05 per-unit) and R_c, X_m, large (greater than 10 per-unit). The circuit of Fig. 64.3 requires that all values be in per-unit. Circuit data are available from the manufacturer or obtained from conventional tests. It must be noted that although the circuit of Fig. 64.3 is commonly used, it is not rigorously correct because it does not properly account for the mutual couplings between windings.

The terms **primary** and **secondary** refer to source and load sides, respectively (i.e., energy flows from primary to secondary). However, in many applications energy can flow either way, in which case the distinction is meaningless. Also, the presence of a third winding (tertiary) confuses the issue. The terms *step up* and *step down* refer to what the transformer does to the voltage from source to load. ANSI standards require that for a two-winding transformer the high-voltage and low-voltage terminals be marked as H1-H2 and X1-X2, respectively, with H1 and X1 markings having the same significance as *dots* for **polarity** markings. [Refer to ANSI C57 for comprehensive information.] *Additive* and *subtractive transformer polarity* refer to the physical positioning of high-voltage, low-voltage *dotted* terminals as shown in Fig. 64.4. If the dotted terminals are adjacent, then the transformer is said to be *subtractive*, because if these adjacent terminals (H1-X1) are connected together, the voltage between H2 and X2 is the *difference* between primary and secondary. Similarly, if adjacent terminals X1 and H2 are connected, the voltage (H1-X2) is the *sum* of primary and secondary values.

The Two-Winding Transformer

The device can be simplified to two windings. Common two-winding transformer circuit models are shown in Fig. 64.5.

$$\overline{Z}_e = \overline{Z}_1 + \overline{Z}_2 \tag{64.17}$$

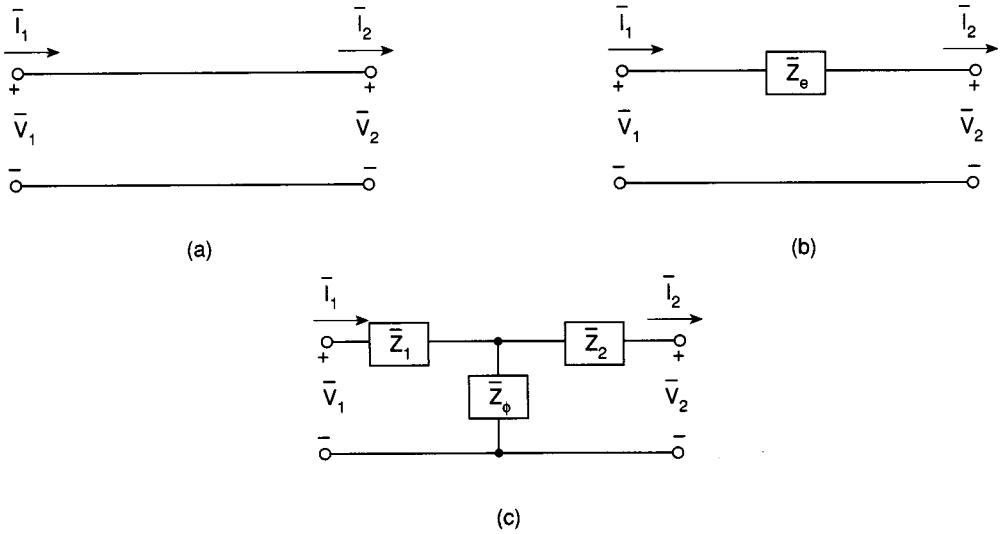


FIGURE 64.5 Two-winding transformer-equivalent circuits. All values in per-unit. (a) Ideal case; (b) no load current negligible; (c) precise model.

$$\bar{Z}_m = \frac{R_c(jX_m)}{R_c + jX_m} \tag{64.18}$$

Circuits (a) and (b) are appropriate when $-Z_m$ is large enough that magnetizing current and core loss is negligible.

64.3 Transformer Performance

There is a need to assess the quality of a particular transformer design. The most important measure for performance is the concept of efficiency, defined as follows:

$$\eta = \frac{P_{\text{out}}}{P_{\text{in}}} \tag{64.19}$$

where P_{out} is output power in watts (kW, MW) and P_{in} is input power in watts (kW, MW).

The situation is clearest for the two-winding case where the output is clearly defined (i.e., the secondary winding), as is the input (i.e., the primary). Unless otherwise specified, the output is understood to be rated power at rated voltage at a user-specified power factor. Note that

$$\Sigma L = P_{\text{in}} - P_{\text{out}} = \text{sum of losses}$$

The transformer is frequently modeled with the circuit shown in Fig. 64.6. Transformer losses are made up of the following components:

Electrical losses:

$$I'^2_1 R_{\text{eq}} = I_1^2 R_1 + I_2^2 R_2 \tag{64.20a}$$

$$\text{Primary winding loss} = I_1^2 R_1 \tag{64.20b}$$

$$\text{Secondary winding loss} = I_2^2 R_2 \tag{64.20c}$$

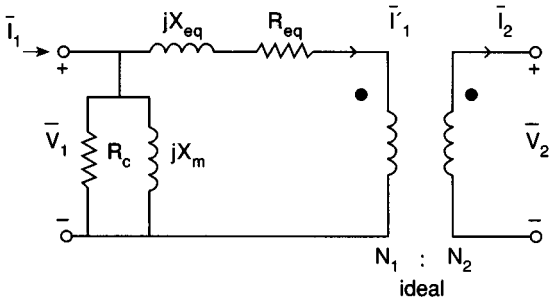


FIGURE 64.6 Transformer circuit model.

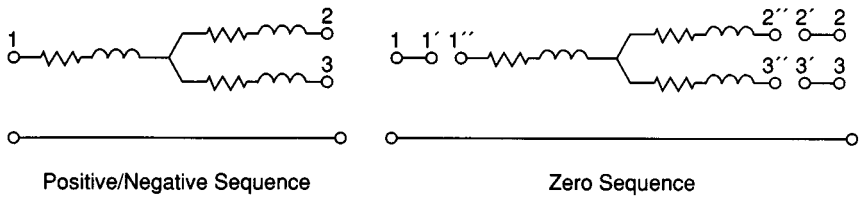


FIGURE 64.7 Sequence equivalent transformer circuits.

Magnetic (core) loss:

$$P_c = P_e + P_h = V_1^2/R_c \tag{64.21}$$

Core eddy current loss = P_e

Core hysterisis loss = P_h

Hence:

$$\Sigma L = I'^2_1 R_{eq} + V_1^2/R_c \tag{64.22}$$

A second concern is fluctuation of secondary voltage with load. A measure of this situation is called *voltage regulation*, which is defined as follows:

$$\text{Voltage Regulation (VR)} = \frac{V_{2NL} - V_{2FL}}{V_{2FL}} \tag{64.23}$$

where V_{2FL} = rated secondary voltage, with the transformer supplying rated load at a user-specified power factor, and V_{2NL} = secondary voltage with the load removed (set to zero), holding the primary voltage at the full load value.

A complete performance analysis of a 100 kVA 2400/240 V single-phase transformer is shown in Table 64.1.

64.4 Transformers in Three-Phase Connections

Transformers are frequently used in three-phase connections. For three identical three-winding transformers, nine windings must be accounted for. The three sets of windings may be individually connected in wye or delta in any combination. The symmetrical component transformation can be used to produce the sequence equivalent circuits shown in Fig. 64.7 which are essentially the circuits of Fig. 64.3 with R_c and X_m neglected.

The positive and negative sequence circuits are valid for both wye and delta connections. However, Y–Δ connections will produce a phase shift which is not accounted for in these circuits.

TABLE 64.1 Analysis of a Single-Phase 2400:240V 100-kVA Transformer

Voltage and Power Ratings

HV (Line-V)	LV (Line-V)	S (Total-kVA)
2400	240	100

Test Data

Short Circuit (HV) Values	Open Circuit (LV) Values
Voltage = 211.01	240.0 volts
Current = 41.67	22.120 amperes
Power = 1400.0	787.5 watts

Equivalent Circuit Values (in ohms)

Values referred to		HV Side	LV Side	Per-Unit
Series Resistance	=	0.8064	0.008064	0.01400
Series Reactance	=	4.9997	0.049997	0.08680
Shunt Magnetizing Reactance	=	1097.10	10.9714	19.05
Shunt Core Loss Resistance	=	7314.30	73.1429	126.98

Power Factor (—)	Efficiency (%)	Voltage Regulation (%)	Power Factor (—)	Efficiency (%)	Voltage Regulation (%)
0.0000 lead	0.00	–8.67	0.9000 lag	97.54	5.29
0.1000 lead	82.92	–8.47	0.8000 lag	97.21	6.50
0.2000 lead	90.65	–8.17	0.7000 lag	96.81	7.30
0.3000 lead	93.55	–7.78	0.6000 lag	96.28	7.86
0.4000 lead	95.06	–7.27	0.5000 lag	95.56	8.26
0.5000 lead	95.99	–6.65	0.4000 lag	94.50	8.54
0.6000 lead	96.62	–5.89	0.3000 lag	92.79	8.71
0.7000 lead	97.07	–4.96	0.2000 lag	89.56	8.79
0.8000 lead	97.41	–3.77	0.1000 lag	81.09	8.78
0.9000 lead	97.66	–2.16	0.0000 lag	0.00	8.69
1.0000 —	97.83	1.77			

Rated load performance at power factor = 0.866 lagging.

Secondary Quantities; LOW Voltage Side

	SI Units	Per-Unit
Voltage	240 volts	1.0000
Current	416.7 amperes	1.0000
Apparent power	100.0 kVA	1.0000
Real power	86.6 kW	0.8660
Reactive power	50.0 kvar	0.5000
Power factor	0.8660 lag	0.8660

Primary Quantities; HIGH Voltage Side

	SI Units	Per-Unit
Voltage	2539 volts	1.0577
Current	43.3 amperes	1.0386
Apparent power	109.9 kVA	1.0985
Real power	88.9 kW	0.8888
Reactive power	64.6 kvar	0.6456
Power factor	0.8091 lag	0.8091

Efficiency = 97.43%; voltage regulation = 5.77%.

The zero sequence circuit requires special modification to account for wye, delta connections. Consider winding 1:

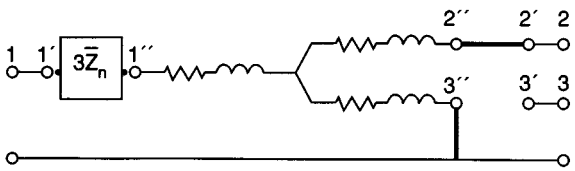
1. Solid grounded wye — short 1′ to 1″.
2. Ground wye through $-Z_n$ — connect 1′ to 1″ through $3-Z_n$.
3. Ungrounded wye — leave 1′ to 1″ open.
4. Delta — short 1″ to reference.

Winding sets 2 and 3 interconnections produce similar connection constraints at terminals 2′–2″ and 3′–3″, respectively.

Example. Three identical transformers are to be used in a three-phase system. They are connected at their terminals as follows:

Winding set 1 wye, grounded through $-Z_n$
Winding set 2 wye, solid ground
Winding set 3 delta

The zero sequence network is as shown.



Phase Shift in Y–Δ Connections

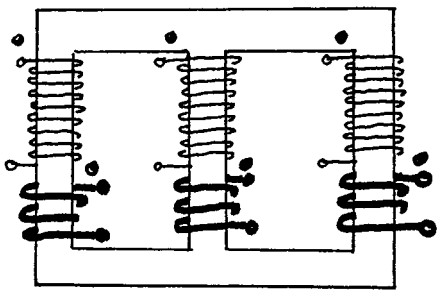
The positive and negative sequence networks presented in Fig. 64.7 are misleading in one important detail. For Y–Y or Δ–Δ connections, it is always possible to label the phases in such a way that there is no phase shift between corresponding primary and secondary quantities. However, for Y–Δ or Δ–Y connections, it is impossible to label the phases in such a way that no phase shift between corresponding quantities is introduced. ANSI standard C57.12.10.17.3.2 is as follows:

> *For either wye-delta or delta-wye connections, phases shall be labeled in such a way that positive sequence quantities on the high voltage side lead their corresponding positive sequence quantities on the low voltage side by 30°. The effect on negative sequence quantities may be the reverse, i.e., HV values* lag *LV values by 30°.*

This 30° phase shift is *not* accounted for in the sequence networks of Fig. 64.7. The effect only appears in the positive and negative sequence networks; the zero sequence network quantities are unaffected.

The Three-Phase Transformer

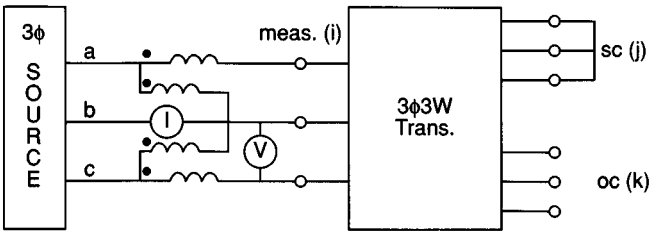
It is possible to construct a device (called a three-phase transformer) which allows the phase fluxes to share common magnetic return paths. Such designs allow considerable savings in core material, and corresponding economies in cost, size, and weight. Positive and negative sequence impedances are equal; however, the zero sequence impedance may be different. Otherwise the circuits of Fig. 64.7 apply as discussed previously.



Determining Per-Phase Equivalent Circuit Values for Power Transformers

One method of obtaining such data is through testing. Consider the problem of obtaining transformer equivalent circuit data from short-circuit tests. A numerical example will clarify per-unit scaling considerations.

(a)



(b)

Line Voltage		3ph Power Ratings	S3ph Base = 100 MVA
Primary	15 kV	100 MVA	Zbase = 2.25
Secondary	115 kV	100 MVA	Zbase = 132.25
Tertiary	4.157 kV	20 MVA	Zbase = 0.1728

(c)

Voltage (line)	Current (line)	Power (3ph)	1	2	3
1200.0 V	3849.0 A	889.0 kW	meas	sc	oc
1840.0 V	100.0 A	100.0 kW	oc	meas	sc
33.0 V	2776.0 A	120.0 kW	sc	oc	meas

(d)

	R	X	Z		R	X	Z
Z12:	0.00889	0.07950	0.08000	Z1:	0.00686	0.20145	0.20157
Z23:	0.02520	0.07627	0.08033	Z2:	0.00203	–0.12195	0.12196
Z31:	0.03004	0.39967	0.40080	Z3:	0.02318	0.19822	0.19957

FIGURE 64.8 Transformer circuit data from short-circuit tests. (a) Setup for transformer short-circuit tests; (b) transformer data; (c) short-circuit test data; (d) short-circuit impedance values in per-unit.

The short-circuit test circuit arrangement is shown in Fig. 64.8. The objective is to derive equivalent circuit data from the test data provided in Fig. 64.8. Note that measurements are made in winding "*i*", with winding "*j*" shorted, and winding "*k*" left open. The short circuit impedance, looking into winding "*i*" with the transformer so terminated is designated as Z_{ij}. The indices *i*, *j*, and *k*, can be 1, 2, or 3.

The impedance calculations are done in per-unit; base values are provided in Fig. 64.8(c). The transformer ratings of the transformer of Fig. 64.2(a) would conventionally be provided as follows:

3ϕ 3W Transformer
15kVY/115kVY/4.157kVΔ
100/100/20 MVA

where 3ϕ means that the transformer is a three-phase piece of equipment (as opposed to an interconnection of three single-phase devices). 3W means three three-phase windings (actually nine windings). Usually the schematic is supplied also. The 15 kV rating is the *line* (phase-to-phase) value; three-phase apparatus is always rated in *line* values. "Y" means winding No. 1 is internally wye connected. 115kVY means that 115 kV is the line voltage rating, and winding No. 2 is wye connected. In 4.157kVΔ, again, "4.157kV" is the line voltage rating, and winding No. 3 is delta connected. 100/100/20 MVA are the *total* (3ϕ) power ratings for the primary, secondary, and tertiary winding, respectively; three-phase apparatus is always rated in three-phase terms.

The per-unit bases for $S_{3\phi base}$ = 100 MVA are presented in Fig. 64.8(b). Calculating the short-circuit impedances from the test data in Fig. 64.8(c):

$$Z_{ij} = \frac{V_{i_{line}}/\sqrt{3}}{I_{i_{line}}}$$

$$R_{ij} = \frac{R_{3\phi}/3}{I^2_{1_{line}}}$$

$$X_{ij} = \sqrt{Z^2_{ij} - R^2_{ij}}$$

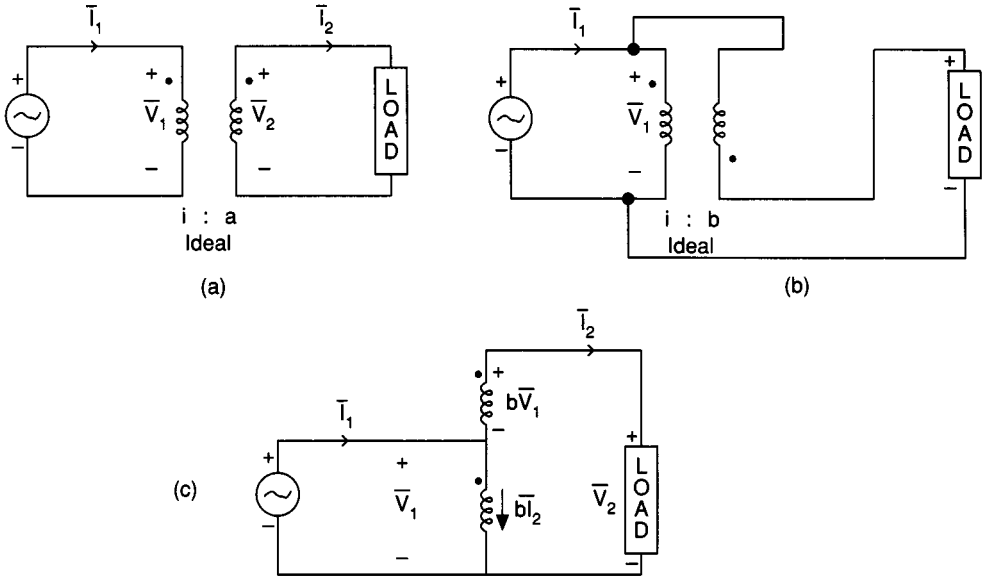


FIGURE 64.9 Autotransformer connection. (a) Conventional step-up connection; (b) autotransformer connection; (c) part (b) redrawn.

Now calculate the transformer impedances from the short-circuit impedances:

$$\bar{Z}_1 = \frac{1}{2}\left(\bar{Z}_{12} - \bar{Z}_{23} + \bar{Z}_{31}\right)$$

$$\bar{Z}_2 = \frac{1}{2}\left(\bar{Z}_{23} - \bar{Z}_{13} + \bar{Z}_{12}\right)$$

$$\bar{Z}_3 = \frac{1}{2}\left(\bar{Z}_{31} - \bar{Z}_{12} + \bar{Z}_{23}\right)$$

Results are shown in Fig. 64.8(d). Observe that the Y–Δ winding connections had no impact on the calculations.

Another detail deserves mention. Although the real and reactive parts of the short-circuit impedances ($-Z_{12}$, $-Z_{23}$, $--Z_{31}$) will always be positive, this is not true for the transformer impedances ($—Z_1$, $--Z_2$, $--Z_3$). One or more of these can be, and frequently is, negative for actual short-circuit data. Negative values underscore that the circuit of Fig. 64.7 is a *port equivalent* circuit, producing correct values at the winding terminals.

64.5 Autotransformers

Transformer windings, though magnetically coupled, are electrically isolated from each other. It is possible to enhance certain performance characteristics for transformers by electrically interconnecting primary and secondary windings. Such devices are called **autotransformers**. The benefits to be realized are lower cost, smaller size and weight, higher efficiency, and better voltage regulation. The basic connection is illustrated in Fig. 64.9. The issues will be demonstrated with an example.

Consider the conventional connection, shown in Fig. 64.9(a).

$$\bar{V}_2 = a\bar{V}_1$$

$$\bar{I}_2 = \frac{1}{a}\bar{I}_1$$

$$S_{\text{rating}} = V_1 I_1 = V_2 I_2 = S_{\text{load}}$$

Now for the autotransformer:

$$\overline{V}_2 = \overline{V}_1 + b\overline{V}_1 = (1 + b)\overline{V}_1$$
$$\bar{I}_1 = \bar{I}_2 + b\bar{I}_2 = (1 + b)\bar{I}_2$$

For the same effective ratio

$$1 + b = a$$

Therefore each winding rating is:

$$S_{\text{rated}} = S_{\text{load}}\left(\frac{b}{1+b}\right)$$

For example if $b = 1$ ($a = 2$)

$$S_{\text{rating}} = 1/2\ S_{\text{load}}$$

meaning that the transformer rating is only 50% of the load.

The principal advantage of the autotransformer is the increased power rating. Also, since the losses remain the same, expressed as a percentage of the new rating, they go down, and correspondingly, the efficiency goes up. The machine impedances in per unit drop for similar reasons. A disadvantage is the loss of electrical isolation between primary and secondary. Also, low impedance is not necessarily good, as we shall see when we study faults on power systems. Autotransformers are used in three-phase connections and in voltage control applications.

Defining Terms

Autotransformer: A transformer whose primary and secondary windings are electrically interconnected.
Polarity: Consideration of in-phase or out-of-phase relations of primary and secondary ac currents and voltages.
Primary: The source-side winding.
Secondary: The load-side winding.
Tap: An electrical terminal that permits access to a winding at a particular physical location.
Transformer: A device which converts ac voltage and current to different levels at essentially constant power and frequency.

Related Topics

1.3 Transformers • 3.4 Power and Energy • 3.5 Three-Phase Circuits

References

ANSI Standard C57, New York: American National Standards Institute.
S. J. Chapman, *Electric Machinery Fundamentals,* 2nd ed, New York: McGraw-Hill, 1991.
V. Del Toro, *Basic Electric Machines*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.
M.E. El-Hawary, *Electric Power Systems: Design and Analysis*, Reston, Va.: Reston Publishing, 1983.
O.I. Elgerd, *Electric Energy Systems Theory: An Introduction,* 2nd ed., New York: McGraw-Hill, 1982.
R. Feinburg, *Modern Power Transformer Practice,* New York: Wiley, 1979.
A.E. Fitzgerald, C. Kingsley, and S. Umans, *Electric Machinery*, 5th ed., New York: McGraw-Hill, 1990.

C. A. Gross, *Power Systems Analysis*, 2nd ed., New York: Wiley, 1986.
N.N. Hancock, *Matrix Analysis of Electrical Machinery*, 2nd ed., Oxford: Pergamon, 1974.
E. Lowden, *Practical Transformer Design Handbook,* 2nd ed, Blue Ridge Summit, Pa.: TAB, 1989.
G. McPherson, *An Introduction to Electrical Machines and Transformers*, New York: Wiley, 1981.
A. J. Pansini, *Electrical Transformers,* Englewood Cliffs, NJ: Prentice-Hall, 1988.
G.R. Slemon, *Magnetoelectric Devices*, New York: Wiley, 1966.
R. Stein and W. T. Hunts, Jr., *Electric Power System Components: Transformers and Rotating Machines*, New York: Van Nostrand Reinhold, 1979.

Further Information

For a comprehensive coverage of general transformer theory, see Chapter 2 of *Electric Machines* by G.R. Slemon and A. Straughen (Addison-Wesley, 1980). For transformer standards, see ANSI Standard C57. For a detailed explanation of transformer per-unit scaling, see Chapter 5 of *Power Systems Analysis* by C.A. Gross (John Wiley, 1986). For design information see *Practical Transformer Design Handbook* by E. Lowden (TAB, 1989).

Karady, G.G. “Energy Distribution”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

65

Energy Distribution

65.1 Introduction
65.2 Primary Distribution System
65.3 Secondary Distribution System
65.4 Radial Distribution System
65.5 Secondary Networks
65.6 Load Characteristics
65.7 Voltage Regulation
65.8 Capacitors and Voltage Regulators

George G. Karady
Arizona State University

65.1 Introduction

Distribution is the last section of the electrical power system. Figure 65.1 shows the major components of the electric power system. The power plants convert the energy stored in the fuel (coal, oil, gas, nuclear) or hydro into electric energy. The energy is supplied through step-up transformers to the electric network. To reduce energy transportation losses, step-up transformers increase the voltage and reduce the current. The high-voltage network, consisting of transmission lines, connects the power plants and high-voltage **substations** in parallel. The typical voltage of the high-voltage transmission network is between 240 and 765 kV. The high-voltage substations are located near the load centers, for example, outside a large town. This network permits load sharing among power plants and assures a high level of reliability. The failure of a line or power plant will not interrupt the energy supply.

The subtransmission system connects the high-voltage substations to the distribution substations. These stations are directly in the load centers. For example, in urban areas, the distance between the distribution stations is around 5 to 10 miles. The typical voltage of the subtransmission system is between 138 and 69 kV. In high load density areas, the subtransmission system uses a network configuration that is similar to the high-voltage network. In medium and low load density areas, the loop or radial connection is used. Figure 65.1 shows a typical radial connection.

The distribution system has two parts, primary and secondary. The primary distribution system consists of overhead lines or underground cables, which are called **feeders**. The feeders run along the streets and supply the distribution transformers that step the voltage down to the secondary level (120–480 V). The secondary distribution system contains overhead lines or underground cables supplying the consumers directly (houses, light industry, shops, etc.) by single- or three-phase power. Separate, dedicated primary feeders supply industrial customers requiring several megawatts of power. The subtransmission system directly supplies large factories consuming over 50 MW.

65.2 Primary Distribution System

The most frequently used voltages and wiring in the primary distribution system are listed in Table 65.1.

Primary distribution, in low load density areas, is a radial system. This is economical but yields low reliability. In large cities, where the load density is very high, a primary cable network is used. The distribution substations

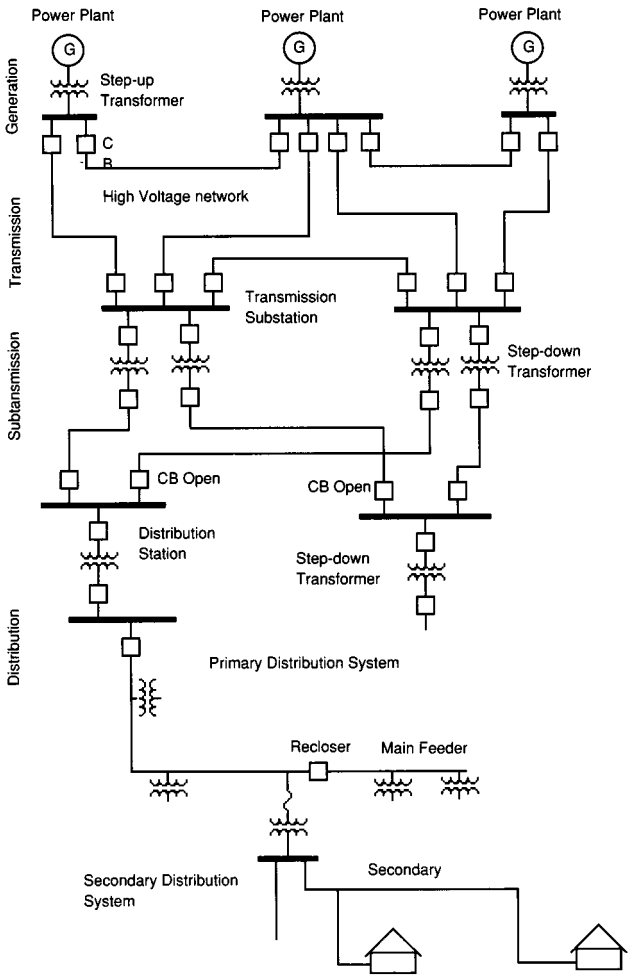


FIGURE 65.1 Electric energy system.

TABLE 65.1 Typical Primary Feeder Voltages (line-to-line)

Class, kV	Voltage, kV	Wiring
2.5	2.4	3-wire delta
5	4.16	4-wire Y
8.66	7.2	4-wire Y
15	12.47	3-wire delta/4-wire Y
25	22.9	4-wire Y
35	34.5	4-wire Y

are interconnected by the feeders (lines or cables). Circuit breakers (CBs) are installed at both ends of the feeder for short-circuit protection. The loads are connected directly to the feeders through fuses. The connection is similar to the one-line diagram of the high-voltage network shown in Fig. 65.1. The high cost of the network limits its application. A more economical and fairly reliable arrangement is the loop connection, when the main feeder is supplied from two independent distribution substations. These stations share the load. The problem with this connection is the circulating current that occurs when the two supply station voltages are different. The loop arrangement significantly improves system reliability.

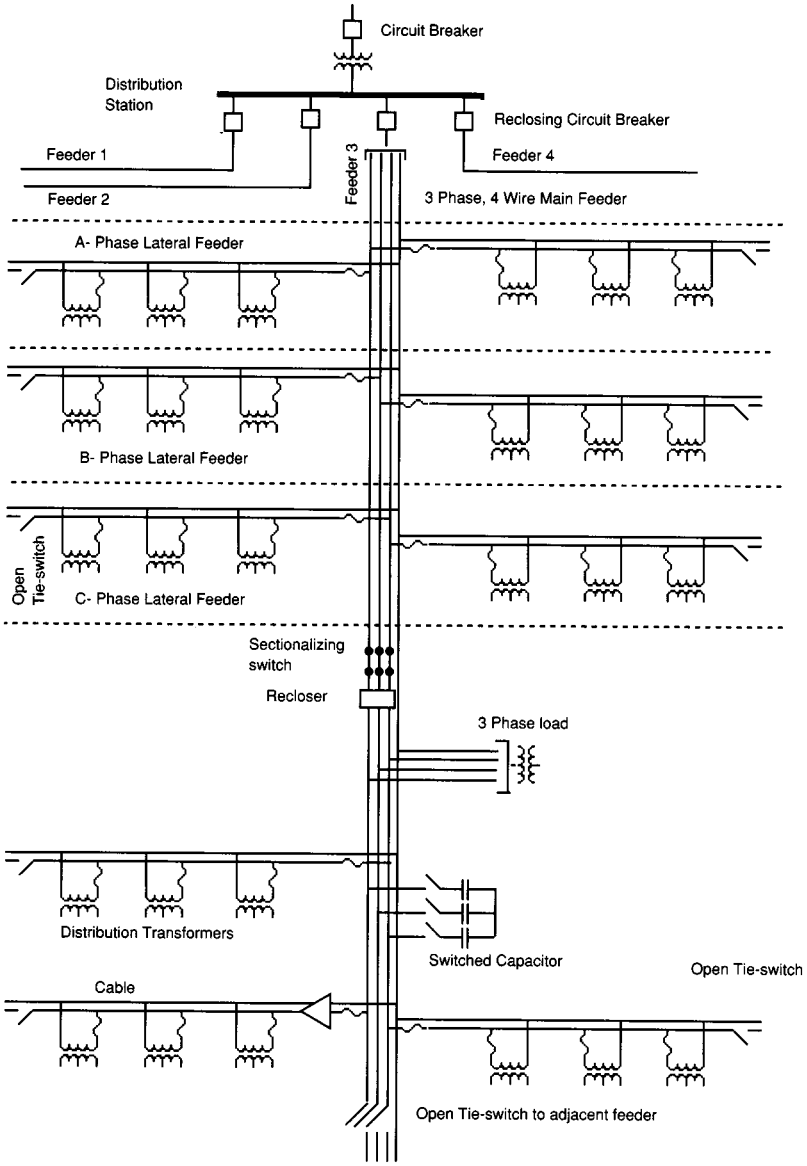


FIGURE 65.2 Radial primary distribution system.

The circulating current can be avoided by using the open-loop connection. This is a popular, frequently used circuit. Figure 65.2 shows a typical open-loop primary feeder. The distribution substation has four outgoing main feeders. Each feeder supplies a different load area and is protected by a reclosing CB.

The three-phase four-wire main feeders supply single-phase lateral feeders. A **recloser** and a sectionalizing switch divide the main feeder into two parts. The normally open tie-switch connects the feeder to the adjacent distribution substation. The fault between the CB and recloser opens the reclosing CB. The CB recloses after a few cycles. If the fault is not cleared, the opening and reclosing process is repeated two times. If the fault has not been cleared before the third reclosing, the CB remains open. Then the sectionalizing switch opens and the tie-switch closes. This energizes the feeder between the recloser and the tie-switch from the neighboring feeder. Similarly, the fault between the recloser and tie-switch activates the recloser. The recloser opens and recloses three times. If the fault is not cleared, the recloser remains open and separates the faulty part of the

feeder. This method is particularly effective in overhead lines where temporary faults are often caused by lightning, wind, and metal balloons.

A three-phase switched **capacitor bank** is rated two-thirds of the total average reactive load and installed two-thirds of the distance out on the feeder from the source. The capacitor bank improves the power factor and reduces voltage drop at heavy loads. However, at light loads, the capacitor is switched off to avoid overvoltages.

Some utilities use voltage regulators at the primary feeders. The voltage regulator is an autotransformer. The secondary coil of the transformer has 32 taps, and a switch connects the selected tap to the line to regulate the voltage. The problem with the **tap changer** is that the lifetime of the switch is limited. This permits only a few operations per day.

The lateral single-phase feeders are supplied from different phases to assure equal phase loading. Fuse cutouts protect the lateral feeders. These fuses are coordinated with the fuses protecting the distribution transformers. The fault in the distribution transformer melts the transformer fuse first. The lateral feeder fault operates the cutout fuse before the recloser or CB opens permanently.

A three-phase line supplies the larger loads. These loads are protected by CBs or high-power fuses.

Most primary feeders in rural areas are overhead lines using pole-mounted distribution transformers. The capacitor banks and the reclosing and sectionalizing switches are also pole-mounted. Overhead lines reduce the installation costs but reduce aesthetics.

In urban areas, an underground cable system is used. The switchgear and transformers are placed in underground vaults or ground-level cabinets. The underground system is not affected by weather and is highly reliable. Unfortunately, the initial cost of an underground cable is significantly higher than an overhead line with the same capacity. The high cost limits the underground system to high-density urban areas and housing developments. Flooding can be a problem.

65.3 Secondary Distribution System

The secondary distribution system provides electric energy to the customers through the distribution transformers and secondary cables. Table 65.2 shows the typical voltages and wiring arrangements.

In residential areas, the most commonly used is the single-phase three-wire 120/240-V radial system, where the lighting loads are supplied by the 120 V and the larger household appliances (air conditioner, range, oven, and heating) are connected to the 240-V lines. Depending on the location, either underground cables or overhead lines are used for this system.

In urban areas, with high-density mixed commercial and residential loads, the three-phase 208/120-V four-wire network system is used. This network assures higher reliability but has significantly higher costs. Underground cables are used by most secondary networks.

High-rise buildings are supplied by a three-phase four-wire 480/277-V spot network. The fluorescent lighting is connected to a 277-V and the motor loads are supplied by a 480-V source. A separate local 120-V system supplies the outlets in the various rooms. This 120-V radial system is supplied by small transformers from the 480-V network.

TABLE 65.2 Secondary Voltages and Connections

Class	Voltage	Connection	Application
1-phase	120/240	Three-wire	Residential
3-phase	208/120	Four-wire	Commercial/residential
3-phase	480/277	Four-wire	High-rise buildings
3-phase	380/220	Four-wire	General system, Europe
3-phase	120/240	Four-wire	Commercial
3-phase	240	Three-wire	Commercial/industrial
3-phase	480	Three-wire	Industrial
3-phase	240/480	Four-wire	Industrial

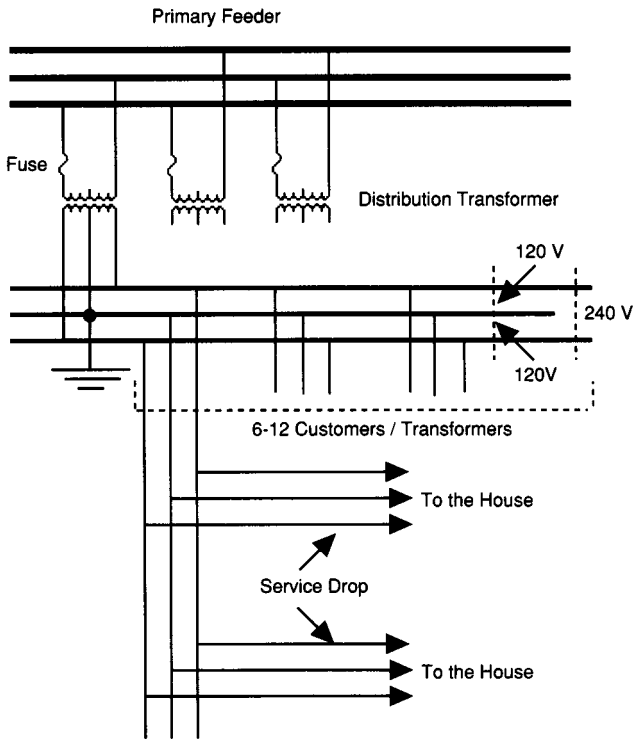


FIGURE 65.3 Typical 120/240-V radial secondary system.

65.4 Radial Distribution System

A typical overhead single-phase three-wire 120/240-V secondary system is shown in Fig. 65.3. The three distribution transformers are mounted on separate primary feeder poles and supplied from different phases. Each transformer supplies 6 to 12 houses. The transformers are protected by fuses. The secondary feeders and the service drops are not protected individually. The secondary feeder uses insulated No. 1/0 or 4/0 aluminum conductors. The average secondary length is from 200 to 600 ft. The typical load is from 15 to 30 W/ft.

The underground distribution system is used in modern suburban areas. The transformers are pad-mounted or placed in an underground vault. A typical 50-kVA transformer serves 5 to 6 houses, with each house supplied by an individual cable.

The connection of a typical house is shown in Fig. 65.4. The incoming secondary service drop supplies the kW and kWh meter. The modern, mostly electronic meters measure 15-min kW demand and the kWh energy consumption. It records the maximum power demand and energy consumption. The electrical utility maintains the distribution system up to the secondary terminals of the meter. The homeowner is responsible for the service panel and house wiring. The typical service panel is equipped with a main switch and circuit breaker. The main switch permits the deenergization of the house and protects against short circuits. The smaller loads are supplied by 120 V and the larger loads by 240 V. Each outgoing line is protected by a circuit breaker. The neutral has to be grounded at the service panel, just past the meter. The water pipe was used for grounding in older houses. In new houses a metal rod, driven in the earth, provides proper grounding. In addition, a separate bare wire is connected to the ground. The ground wire connects the metal parts of the appliances and service panel box together to protect against ground-fault-produced electric shocks.

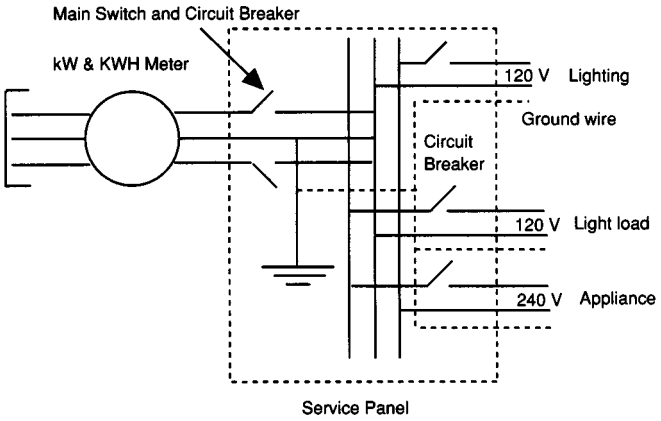


FIGURE 65.4 Residential electrical connection.

65.5 Secondary Networks

The secondary network is used in urban areas with high load density. Figure 65.5 shows a segment of a typical secondary network.

The secondary feeders form a mesh or grid that is supplied by transformers at the node points. The multiple supply assures higher reliability and better load sharing. The loads are connected directly to the low-voltage grid, without any protection equipment. The network is protected by fuses and network protector circuit breakers installed at the secondary transformers. A short circuit blows the fuses and limits the current. The network protectors automatically open on reverse current and reclose when the voltage on the primary feeder is restored after a fault.

65.6 Load Characteristics

The distribution system load varies during the day. The maximum load occurs in the early evening or late afternoon, and the minimum load occurs at night. The design of the distribution system requires both values, because the voltage drop is at the maximum during the peak load, and overvoltage may occur during the minimum load. The power companies continuously study the statistical variation of the load and can predict the expected loads on the primary feeders with high accuracy. The feeder design considers the expected peak load or maximum demand and the future load growth.

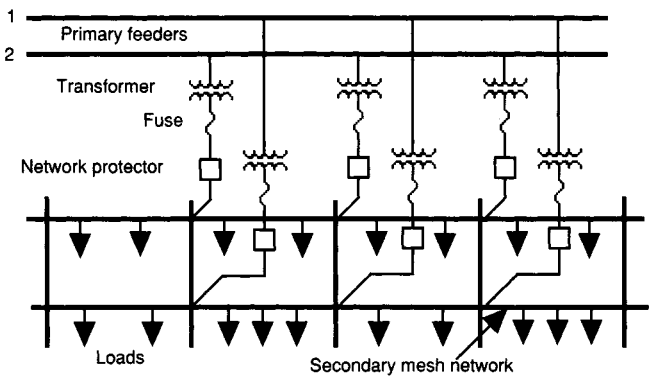


FIGURE 65.5 Typical segment of a secondary distribution network.

The economic conductor cross-section calculation requires the determination of average losses. The average loss is calculated by the loss factor (LSF), which is determined by statistical analyses of load variation.

$$\text{LSF} = \frac{\text{average loss}}{\text{loss at peak load}}$$

The average load is determined by the load factor (LF), which is the ratio of average load to peak load. The load factor for an area is determined by statistical analyses of the load variation in past years. The approximate relation between the loss factor and load factor is

$$\text{LSF} = 0.3\text{LF} + 0.7\text{LF}^2$$

This equation is useful because the load factor is measured continuously by utilities, and more accurate values are available for the load factor than for the loss factor. Typical values are given in Table 65.3.

TABLE 65.3 Typical Annual Load Factor Values

Type of Load	Load Factor
Residential	0.48
Commercial	0.66
Industrial	0.72

TABLE 65.4 Typical Demand Factors for Multifamily Dwellings

Number of Dwellings	Demand Factor, %
3 to 5	45
18 to 20	38
39 to 42	28
62 & over	23

The connected load or demand can be estimated accurately in residential and industrial areas. The connected load or demand is the sum of continuous ratings of apparatus connected to the system. However, not all equipment is used simultaneously. The actual load in a system is significantly lower than the connected load. The demand factor is used to estimate the actual or maximum demand. The demand factor (DF) is defined by

$$\text{DF} = \frac{\text{maximum demand}}{\text{total connected demand}}$$

The demand factor depends on the number of customers and the type of load. Typical demand factor values are given in Table 65.4.

65.7 Voltage Regulation

The voltage supplied to each customer should be within the ±5% limit, which, at 120 V, corresponds to 114 and 126-V. Figure 65.6 shows a typical voltage profile for a feeder at light and heavy load conditions. The figure shows that at heavy load, the voltage at the end of the line will be less than the allowable minimum voltage. However, at the light load condition the voltage supplied to each customer will be within the allowable limit. Calculation of the voltage profile, voltage drop, and feeder loss is one of the major tasks in distribution system design. The concept of voltage drop and loss calculation is demonstrated using the feeder shown in Fig. 65.6.

To calculate the voltage drop, the feeder is divided into sections. The sections are determined by the loads. Assuming a single-phase system, the load current is calculated by Eq. (65.1):

$$|I_i| = \frac{P_i}{V \cos \varphi_i}, \qquad I_i = |I_i| (\cos \varphi_i + \sin \varphi_i) \tag{65.1}$$

where P is the power of the load, V is the rated voltage, and φ is the power factor.

The section current is the sum of the load currents. Equation (65.2) gives the section current between load i and $i - 1$:

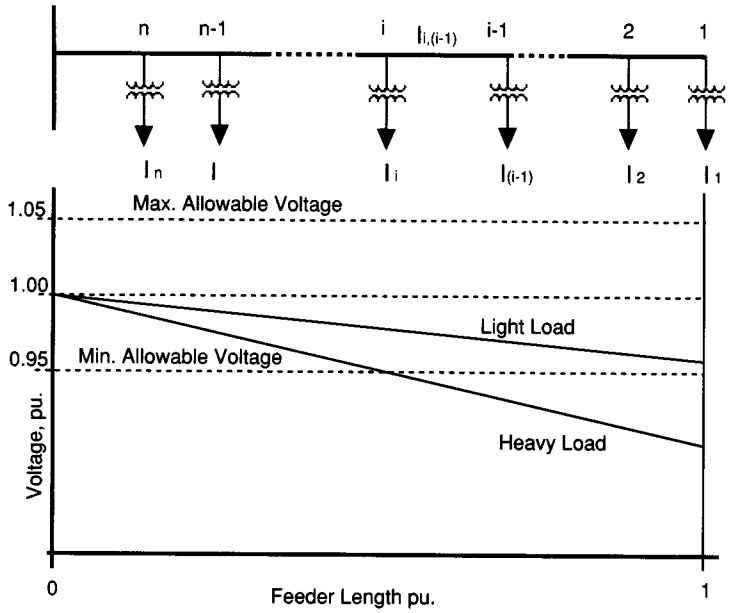


FIGURE 65.6 Feeder voltage profile.

$$I_{(i,\,i-1)} = \sum_{1}^{i-1} I_i \tag{65.2}$$

The electrical parameters of the overhead feeders are the resistance and reactance, which are given in Ω/mi. The underground feeders have significant capacitance in addition to the reactance and resistance. The capacitance is given in μF/mi. The actual values for overhead lines can be calculated using the conductor diameter and phase-to-phase and phase-to-ground distances [Fink and Beaty, 1978]. The residential underground system generally uses single-conductor cables with polyethylene insulation. The older systems use rubber insulation with neoprene jacket. Circuit parameters should be obtained from manufacturers. The distribution feeders are short transmission lines. Even the primary feeders are only a few miles long. This permits the calculation of the section resistance and reactance by multiplying the Ω/mi values by the length of the section. The length of the section in a single-phase two-wire system is two times the actual length. In a balanced three-phase system, it is the simple length. In a single-phase three-wire system the voltage drop on the neutral conductor must be calculated. Further information may be obtained from Pansini [1991].

Equation (65.3) gives the voltage drop, with a good approximation, for section i, $(i-1)$. The total voltage drop is the sum of the sections voltage drops.

$$e_{i,(i-1)} = \left|I_{i,(i-1)}\right|(R_{i,(i-1)} \cos \varphi_{i,(i-1)} + X_{i,(i-1)} \sin \varphi_{i,(i-1)}) \tag{65.3}$$

Equation (65.4) gives the losses on the line:

$$\text{Loss}_i = \sum_{1}^{i-1} (I_{i,(i-1)})^2 R_{i,(i-1)} \tag{65.4}$$

The presented calculation method describes the basic concept of feeder design; more details can be found in the literature.

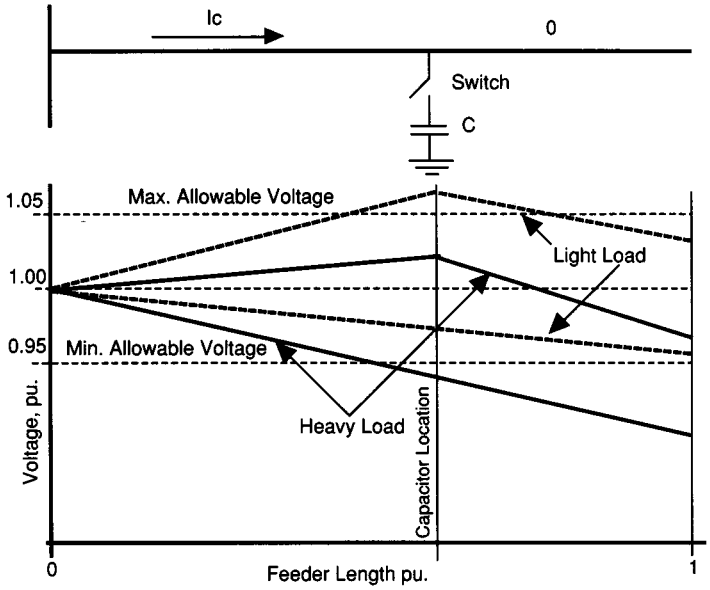


FIGURE 65.7 Capacitor effect on voltage profile.

65.8 Capacitors and Voltage Regulators

The voltage drop can be reduced by the application of a shunt capacitor. As shown in Fig. 65.7, a properly selected and located shunt capacitor assures that the voltage supplied to each of the customers will be within the allowable limit at the heavy load condition. However, at light load, the same capacitor will increase the voltage above the allowable limit. Most capacitors in the distribution system use switches. The capacitor is switched off during the night when the load is light and switched on when the load is heavy. The most frequent use of capacitors is on the primary feeders. In an overhead system, three-phase capacitor banks with vacuum switches are installed on the poles. Residential underground systems require less shunt capacitance for voltage control due to the reduced reactance. Even so, shunt capacitors are used for power factor correction and loss reduction.

The optimum number, size, and location of capacitor banks on a feeder is determined by detailed computer analyses. The concept of optimization includes the minimization of the operation, installation, and investment costs. The most important factor that affects the selection is the distribution and power factor of loads. In residential areas, the load is uniformly distributed. In this case the optimum location of the capacitor bank is around two-thirds of the length of the feeder.

The effect of capacitor bank can be studied by adding the capacitor current to the load current. The capacitor current flows between the supply and the capacitor as shown in Fig. 65.7. Its value can be calculated from Eq. (65.5) for a single-phase system:

$$I_c = j\omega CV, \qquad \omega = 2\pi f \tag{65.5}$$

where *C is the capacitance, f is the frequency (60 Hz), and V is the voltage to ground.*

The capacitive current is added to the inductive load current, reducing the total current, the voltage drop, and losses. The voltage drop and loss can be calculated from Eqs. (65.2) to (65.5).

The voltage regulator is a tap-changing transformer, which is located, in most cases, at the supply end of the feeder. The tap changer increases the supply voltage, which in turn increases the voltage above the allowable minimum at the last load. The tap changer transformer has two windings. The excitation winding is connected in parallel. The regulating winding is connected in series with the feeder. The latter has taps and a tap changer

switch. The switch changes the tap position according to the required voltage. The tap changing requires the short interruption of load current. The frequent current interruptions reduce the lifetime of the tap changer switch. This problem limits the number of tap changer operations to between one to three per day.

Defining Terms

Capacitor bank: Consists of capacitors connected in parallel. Each capacitor is placed in a metal can and equipped with bushings.

Feeder: Overhead lines or cables which are used to distribute the load to the customers. They interconnect the distribution substations with the loads.

Recloser: A circuit breaker which is designed to interrupt short-circuit current and reclose the circuit after interruption.

Substation: A junction point in the electric network. The incoming and outgoing lines are connected to a busbar through circuit breakers.

Tap changer: A transformer. One of the windings is equipped with taps. The usual number of taps is 32. Each tap provides a 1% voltage regulation. A special circuit breaker is used to change the tap position.

Related Topics

1.2 Capacitors and Inductors • 3.1 Voltage and Current Laws • 3.2 Node and Mesh Analysis • 3.4 Power and Energy • 67.4 Load Management

References

D.F.S. Brass et al., in *Electric Power Distribution, 415 V–33 kV,* E.O. Taylor and G.A. Boal (eds.), London: Edward Arnold, 1966, p. 272.

D.G. Fink and H.W. Beaty, *Standard Handbook for Electrical Engineers,* 11th ed., New York: McGraw-Hill, 1978, sec. 18.

T. Gönen, *Electric Power Distribution System Engineering,* New York: Wiley, 1986.

T. Gönen, *Electric Power Transmission System Engineering,* New York: Wiley, 1988, p. 723.

A.J. Pansini, *Power Transmission and Distribution,* Liburn, Ga.: The Fairmont Press, 1991.

E.P. Parker, *McGraw-Hill Encyclopedia of Energy,* New York: McGraw-Hill, 1981, p. 838.

Various, *Electrical Transmission and Distribution Reference Book,* W. Central Station Engineers, East Pittsburgh: Westinghouse Electric Corporation, 1950, p. 824.

Various, *Distribution Systems. Electric Utility Engineering Reference Books,* J. Billard (ed.), East Pittsburgh: Westinghouse Electric Corporation, 1965, p. 567.

Various, *EHV Transmission Line Reference Book,* G.E.C. Project EHV (ed.), New York: Edison Electric Institute, 1968, p. 309.

B.M. Weedy, *Underground Transmission of Electric Power,* New York: Wiley, 1980, p. 294.

W.L. Weeks, *Transmission and Distribution of Electrical Energy,* New York: Harper & Row, 1981, p. 302.

Further Information

Other recommended publications include J. M. Dukert, *A Short Energy History of the United States,* Edison Electric Institute, 1980. Also, the *IEEE Transactions on Power Delivery* publishes distribution papers sponsored by the Transmission and Distribution Committee. These papers deal with the latest development in the distribution area. Every-day problems are presented in two magazines: *Transmission & Distribution* and *Electrical World.*

Liu, C.C., Vu, K.T., Yu, Y., Galler, D., Strange, E.G., Ong, Chee-Mun “Electrical Machines”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

66

Electrical Machines

Chen-Ching Liu
University of Washington

Khoi Tien Vu
ABB Transmission Technical Institute

Yixin Yu
Tianjing University

Donald Galler
Massachusetts Institute of Technology

Elias G. Strangas
Michigan State University

Chee-Mun Ong
Purdue University

66.1 Generators
AC Generators • DC Generators
66.2 Motors
Motor Applications • Motor Analysis
66.3 Small Electric Motors
Single Phase Induction Motors • Universal Motors • Permanent Magnet AC Motors • Stepping Motors
66.4 Simulation of Electric Machinery
Basics in Modeling • Modular Approach • Mathematical Transformations • Base Quantities • Simulation of Synchronous Machines • Three-Phase Induction Machines

66.1 Generators

Chen-Ching Liu, Khoi Tien Vu, and Yixin Yu

Electric generators are devices that convert energy from a mechanical form to an electrical form. This process, known as electromechanical energy conversion, involves magnetic fields that act as an intermediate *medium.* There are two types of generators: alternating current (ac) and direct current (dc). This section explains how these devices work and how they are modeled in analytical or numerical studies.

The input to the machine can be derived from a number of energy sources. For example, in the generation of large-scale electric power, coal can produce steam that drives the shaft of the machine. Typically, for such a thermal process, only about 1/3 of the raw energy (i.e., from coal) is converted into mechanical energy. The final step of the energy conversion is quite efficient, with an efficiency close to 100%.

The generator's operation is based on Faraday's law of electromagnetic induction. In brief, if a coil (or winding) is linked to a varying magnetic field, then an electromotive force, or voltage, emf, is induced across the coil. Thus, generators have two essential parts: one creates a magnetic field, and the other where the emf's are induced. The magnetic field is typically generated by electromagnets (thus, the field intensity can be adjusted for control purposes), whose windings are referred to as field windings or **field circuits**. The coils where the emf's are induced are called *armature* windings or **armature circuits**. One of these two components is stationary (stator), and the other is a rotational part (rotor) driven by an external torque. Conceptually, it is immaterial which of the two components is to rotate because, in either case, the armature circuits always "see" a varying magnetic field. However, practical considerations lead to the common design that for ac generators, the field windings are mounted on the rotor and the armature windings on the stator. In contrast, for dc generators, the field windings are on the stator and armature on the rotor.

AC Generators

Today, most electric power is produced by synchronous generators. Synchronous generators rotate at a constant speed, called **synchronous speed**. This speed is dictated by the operating frequency of the system and the machine structure. There are also ac generators that do not necessarily rotate at a fixed speed such as those

found in windmills (induction generators); these generators, however, account for only a very small percentage of today's generated power.

Synchronous Generators

Principle of Operation. For an illustration of the steady-state operation, refer to Fig. 66.1 which shows a cross section of an ac machine. The rotor consists of a winding wrapped around a steel body. A dc current is made to flow in the rotor winding (or field winding), and this results in a magnetic field (rotor field). When the rotor is made to rotate at a constant speed, the three stationary windings aa', bb', and cc' experience a periodically varying magnetic field. Thus, emf's are induced across these windings in accordance with Faraday's law. These emf's are ac and periodic; each period corresponds to one revolution of the rotor. Thus, for 60-Hz electricity, the rotor of Fig. 66.1 has to rotate at 3600 revolutions per minute (rpm); this is the synchronous speed of the given machine. Because the windings aa', bb', and cc' are displaced equally in space from each other (by 120 degrees), their emf waveforms are displaced in time by 1/3 of a period. In other words, the machine of Fig. 66.1 is capable of generating three-phase electricity. This machine has two poles since its rotor field resembles that of a bar magnet with a north pole and a south pole.

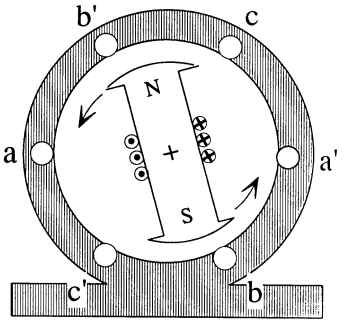


FIGURE 66.1 Cross section of a simple two-pole synchronous machine. The rotor body is salient. Current in rotor winding: ⊗ into the page, ⊙ out of the page.

When the stator windings are connected to an external (electrical) system to form a closed circuit, the steady-state currents in these windings are also periodic. These currents create magnetic fields of their own. Each of these fields is pulsating with time because the associated current is ac; however, the combination of the three fields is a **revolving field**. This revolving field arises from the space displacements of the windings and the phase differences of their currents. This combined magnetic field has two poles and rotates at the same speed and direction as the rotor. In summary, for a loaded synchronous (ac) generator operating in a steady state, there are two fields rotating at the same speed: one is due to the rotor winding and the other due to the stator windings. It is important to observe that the armature circuits are in fact exposed to two rotating fields, one of which, the armature field, is caused by and in fact tends to counter the effect of the other, the rotor field. The result is that the induced emf in the armature can be reduced when compared with an unloaded machine (i.e., open-circuited stator windings). This phenomenon is referred to as **armature reaction**.

It is possible to build a machine with p poles, where $p = 4, 6, 8, \ldots$ (even numbers). For example, the cross-sectional view of a four-pole machine is given in Fig. 66.2. For the specified direction of the (dc) current in the rotor windings, the rotor field has two pairs of north and south poles arranged as shown. The emf induced in a stator winding completes one period for every pair of north and south poles sweeping by; thus, each revolution of the rotor corresponds to two periods of the stator emf's. If the machine is to operate at 60 Hz then the rotor needs to rotate at 1800 rpm. In general, a p-pole machine operating at 60 Hz has a rotor speed of $3600/(p/2)$ rpm. That is, the lower the number of poles is, the higher the rotor speed has to be. In practice, the number of poles is dictated by the mechanical system (prime mover) that drives the rotor. Steam turbines operate best at a high speed; thus, two- or four-pole machines are suitable. Machines driven by hydro turbines usually have more poles.

Usually, the stator windings are arranged so that the resulting armature field has the same number of poles as the rotor field. In practice, there are many possible ways to arrange these windings; the essential idea, however, can be understood via the simple arrangement shown in Fig. 66.2. Each phase consists of a pair of windings (thus occupies four slots on the stator structure), e.g., those for phase a are labeled a_1a_1' and a_2a_2'. Geometry suggests that, at any time instant, equal emf's are induced across the windings of the same phase. If the individual windings are connected in series as shown in Fig. 66.2, their emf's add up to form the phase voltage.

Mathematical/Circuit Models. There are various models for synchronous machines, depending on how much detail one needs in an analysis. In the simplest model, the machine is equivalent to a constant voltage source in series with an impedance. In more complex models, numerous nonlinear differential equations are involved.

Steady-state model. When a machine is in a steady state, the model requires no differential equations. The representation, however, depends on the rotor structure: whether the rotor is cylindrical (round) or salient.

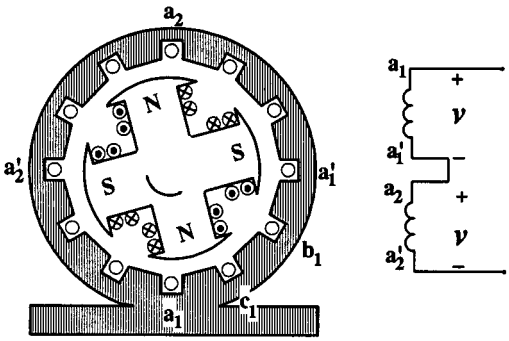


FIGURE 66.2 *Left,* cross section of a four-pole synchronous machine. Rotor has a salient pole structure. *Right,* schematic diagram for phase *a* windings.

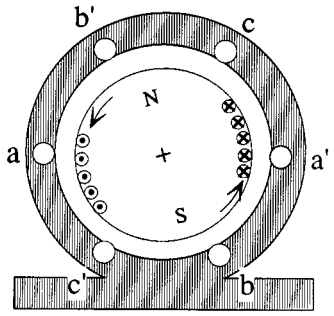


FIGURE 66.3 Cross section of a two-pole round-rotor synchronous machine.

The rotors depicted in Figs. 66.1 and 66.2 are salient since the poles are protruding from the shaft. Such structures are mechanically weak, since at a high speed (3600 rpm and 1800 rpm, respectively) the centrifugal force becomes a serious problem. Practically, for high-speed turbines, round-rotor (or cylindrical-rotor) structures are preferred. The cross section of a two-pole, round-rotor machine is depicted in Fig. 66.3. From a practical viewpoint, salient rotors are easier to build because each pole and its winding can be manufactured separately and then mounted on the rotor shaft. For round rotors, slots need to be reserved in the rotor where the windings can be placed.

The mathematical model for round-rotor machines is much simpler than that for salient-rotor ones. This stems from the fact that the rotor body has a permeability much higher than that of the air. In a steady state, the stator field and the rotor body are at a standstill relative to each other. (They rotate at the same speed as discussed earlier.) If the rotor is salient, it is easier to establish the magnetic flux lines along the direction of the rotor body (when viewed from the cross section). Therefore, for the same set of stator currents, different positions of the rotor alter the stator field in different ways; this implies that the induced emf's are different. If the rotor is round, then the relative position of the rotor structure does not affect the stator field. Hence, the associated mathematical model is simplified.

In the following, the steady-state models of the round-rotor and salient-rotor generators are explained.

Refer to Fig. 66.3 which shows a two-pole round-rotor machine. Without loss of generality, one can select phase *a* (i.e., winding aa') for the development of a mathematical model of the machine. As mentioned previously, the (armature or stator) winding of phase *a* is exposed to two magnetic fields: rotor field and stator field.

1. Rotor field. Its flux as seen by winding aa' varies with the rotor position; the flux linkage is largest when the N–S axis is perpendicular to the winding surface and minimum (zero) when this axis aligns with the surface. Thus, one can express the flux due to the rotor field as seen by winding aa' as $\lambda_1 = L(\theta)I_F$ where θ is to denote the angular position of the N–S axis (of the rotor field) relative to the surface of aa', I_F is the rotor current (a dc current), and L is a periodic function of θ.
2. Stator field. Its flux as seen by winding aa' is a combination of three individual fields which are due to currents in the stator windings, i_a, i_b, and i_c. This flux can be expressed as $\lambda_2 = L_s i_a + L_m i_b + L_m i_c$, where L_s (L_m) is the self (mutual) inductance. Because the rotor is round, L_s and L_m are not dependent on θ, the relative position of the rotor and the winding. Typically, the sum of the stator currents $i_a + i_b + i_c$ is near zero; thus, one can write $\lambda_2 = (L_s - L_m)i_a$.

The total flux seen by winding aa' is $\lambda = \lambda_1 - \lambda_2 = L(\theta)I_F - (L_s - L_m)i_a$, where the minus sign in $\lambda_1 - \lambda_2$ is due to the fact that the stator field opposes the rotor field. The induced emf across the winding aa' is $d\lambda/dt$, the time derivative of λ:

$$e_a = \frac{d\lambda}{dt} = \frac{dL}{dt} I_F - (L_s - L_m)\frac{di_a}{dt} \triangleq e_F - (L_s - L_m)\frac{di_a}{dt}$$

The time-varying quantities are normally sinusoidal, and for practical purposes, can be represented by phasors. Thus the above expression becomes:

$$\bar{E}_a = \bar{E}_F - (L_s - L_m)j\omega_0 \bar{I}_a \triangleq \bar{E}_F - jX_s\bar{I}_a$$

where ω_0 is the angular speed (rad/s) of the rotor in a steady state. This equation can be modeled as a voltage source $-E_F$ behind a reactance jX_s, as shown in Fig. 66.4; this reactance is usually referred to as *synchronous reactance*. The resistor R_a in the diagram represents the winding resistance, and V_t is the voltage measured across the winding.

As mentioned, the theory for salient-rotor machines is more complicated. In the equation $\lambda_2 = L_s i_a + L_m i_b + L_m i_c$, the terms L_s and L_m are now dependent on the (relative) position of the rotor. For example (refer to Fig. 66.1), L_s is maximum when the rotor is in a vertical position and minimum when the rotor is 90° away.

In the derivation of the mathematical/circuit model for salient-rotor machines, the stator field B_2 can be resolved into two components; when the rotor is viewed from a cross section, one component aligns along the rotor and the other is perpendicular to the rotor (Fig. 66.5). The component B_d, which directly opposes the rotor field, is said to belong to the *direct axis*; the other component, B_q, is weaker and belongs to the *quadrature axis*. The model for a salient-rotor machine consists of two circuits, direct-axis circuit and quadrature-axis circuit, each similar to Fig. 66.4. Any quantity of interest, such as I_a, the current in winding aa', is made up of two components, one from each circuit. The round-rotor machine can be viewed as a special case of the salient-pole theory where the corresponding parameters of the d-axis and q-axis circuits are equal.

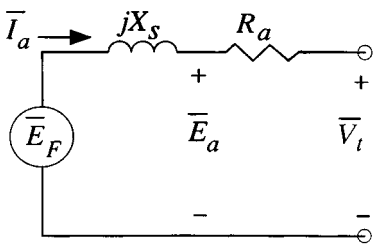


FIGURE 66.4 Per-phase equivalent circuit of round-rotor synchronous machines. $-E_F$ is the internal voltage (phasor form) and V_t is the terminal volt-

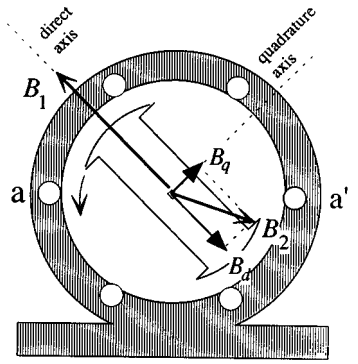


FIGURE 66.5 In the salient-pole theory, the stator field (represented by a single vector B_2) is decomposed into B_d and B_q. Note that $|B_d| > |B_q|$.

Dynamic models. When a power system is in a steady state (i.e., operated at an equilibrium), the electrical output of each generator is equal to the power applied to the rotor shaft. (Various losses have been neglected without affecting the essential ideas provided in this discussion.) Disturbances occur frequently in power systems, however. Examples of disturbances are load changes, short circuits, and equipment outages. A disturbance results in a mismatch between the power input and output of generators, and therefore the rotors depart from their synchronous-speed operation. Intuitively, the impact is more severe for machines closer to the disturbance. When a system is perturbed, there are several possibilities for its subsequent behavior. If the disturbance is small, the machines may soon reach a new steady speed, which is close to or identical to their synchronous speed, in which case the system is said to be stable. It may also happen that some machines speed up while others slow down. In a more complicated situation, a rotor may oscillate about its synchronous speed. This results in an unstable case. An unstable situation can result in abnormal changes in system frequency and voltage and, unless properly controlled, may lead to damage to machines (e.g., broken shafts). To study these phenomena, dynamic models are required. Details of a dynamic model depend on a number of factors such as location of disturbance and time duration of interest. An overview of dynamic generator models is given here. In essence, there are two aspects that need be modeled: electromechanical and electromagnetic.

1. *Electromechanical equations.* Electromechanical equations are to model the effect of input–output imbalance on the rotor speed (and therefore on the operating frequency). The rotor of each machine can be described by the so-called **swing equation**,

$$M\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} = P_{\text{in}} - P_{\text{out}}$$

where θ denotes the rotor position relative to a certain rotating frame, M the inertia of rotor, and D damping. The term $d\theta/dt$ represents the angular velocity and $d^2\theta/dt^2$ is the angular acceleration of the rotor. The preceding differential equation is derived from Newton's law for rotational motions and, in some respects, resembles the dynamical equation of a swinging pendulum (with $P_{\text{in}} \sim$ driving torque, and $P_{\text{out}} \sim$ restoring torque). The term P_{in}, which drives the rotor shaft, can be considered constant in many cases. The term P_{out}, the power sent out to the system, may behave in a very complicated way. Qualitatively, P_{out} tends to increase (respectively, decrease) as the rotor position moves forward (respectively, backward) relative to the synchronous rotating frame. However, such a stable operation can take place only when the system is capable of absorbing (respectively, providing) the extra power. In a multimachine system, conflict might arise when various machines compete with each other in sending out more (or sending out less) electrical power; as a result, the stabilizing effect might be reduced or even lost.

2. *Electromagnetic equations.* The (nonlinear) electromagnetic equations are derived from Faraday's law of electromagnetic induction—induced emf's are proportional to the rate of change of the magnetic fluxes. A general form is as follows:

$$\begin{cases} e_d = \dfrac{d}{dt}\lambda_d + \lambda_q\dfrac{d}{dt}\theta - ri_d \\ e_q = \dfrac{d}{dt}\lambda_q + \lambda_d\dfrac{d}{dt}\theta - ri_q \end{cases} \tag{66.1}$$

where

$$\begin{cases} \lambda_d = G(s)i_F - X_d(s)i_d \\ \lambda_q = -X_q(s)i_q \end{cases} \tag{66.2}$$

The true terminal voltage, e.g., e_a for phase a, can be obtained by combining the direct-axis and quadrature-axis components e_d and e_q, respectively, which are given in Eq. (66.1). On each line of Eq. (66.1), the induced emf is the combination of two sources: the first is the rate of change of the flux on the same axis [$(d/dt)\lambda_d$ on the first line, $(d/dt)\lambda_q$ on the second]; the second comes into effect only when a disturbance makes the rotor and stator fields depart from each other [given by $(d/dt)\theta$]. The third term in the voltage equation represents the ohmic loss associated with the stator winding.

Equation (66.2) expresses the fluxes in terms of relevant currents: flux is equal to inductance times current, with inductances $G(s)$, $X_d(s)$, $X_q(s)$ given in an operational form (s denotes the derivative operator).

Figure 66.6 gives a general view of the input–output state description of machine's dynamic model, the state variables of which appear in Eqs. (66.1) and (66.2).

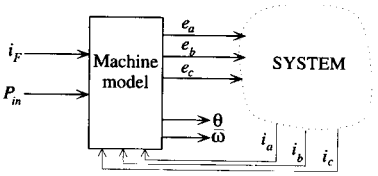


FIGURE 66.6 A block diagram depicting a qualitative relationship among various electrical and mechanical quantities of a synchronous machine. e_a, e_b, e_c are phase voltages; i_a, i_b, i_c phase currents; i_F rotor field current; θ relative position of rotor; ω deviation of rotor speed from synchronous speed; P_{in} mechanical power input. The state variables appear in Eqs. (66.1) and (66.2).

3. *Miscellaneous.* In addition to the basic components of a synchronous generator (rotor, stator, and their windings), there are auxiliary devices which help maintain the machine's operation within acceptable limits. Three such devices are mentioned here: governor, damper windings, and excitation control system.

- Governor. This is to control the mechanical power input P_{in}. The control is via a feedback loop where the speed of the rotor is constantly monitored. For instance, if this speed falls behind the synchronous speed, the input is insufficient and has to be increased. This is done by opening up the valve to increase the steam for turbogenerators or the flow of water through the penstock for hydrogenerators. Governors are mechanical systems and therefore have some significant time lags (many seconds) compared to other electromagnetic phenomena associated with the machine. If the time duration of interest is short, the effect of governor can be ignored in the study; that is, P_{in} is treated as a constant.
- Damper windings (armortisseur windings). These are special conducting bars buried in notches on the rotor surface, and the rotor resembles that of a squirrel-cage-rotor induction machine (see Section 66.2). The damper windings provide an additional stabilizing force for the machine when it is perturbed from an equilibrium. As long as the machine is in a steady state, the stator field rotates at the same speed as the rotor, and no currents are induced in the damper windings. That is, these windings exhibit no effect on a steady-state machine. However, when the speeds of the stator field and the rotor become different (because of a disturbance), currents are induced in the damper windings in such a way as to keep, according to Lenz's law, the two speeds from separating.
- Excitation control system. Modern excitation systems are very fast and quite efficient. An excitation control system is a feedback loop that aims at keeping the voltage at machine terminals at a set level. To explain the main feature of the excitation system, it is sufficient to consider Fig. 66.4. Assume that a disturbance occurs in the system, and as a result, the machine's terminal voltage V_t drops. The excitation system boosts the internal voltage E_F; this action can increase the voltage V_t and also tends to increase the reactive power output.

From a system viewpoint, the two controllers of excitation and governor rely on local information (machine's terminal voltage and rotor speed). In other words, they are decentralized controls. For large-scale systems, such designs do not always guarantee a desired stable behavior since the effect of interconnection is not taken into account in detail.

Synchronous Machine Parameters. When a disturbance, such as a short circuit at the machine terminals, takes place, the dynamics of a synchronous machine will be observed before a new steady state is reached. Such a process typically takes a few seconds and can be divided into subprocesses. The damper windings (armortisseur) exhibit their effect only during the first few cycles when the difference in speed between the rotor and the perturbed stator field is significant. This period is referred to as *subtransient.* The next and longer period, which is between the subtransient and the new steady state, is called *transient.*

Various parameters associated with the subprocesses can be visualized from an equivalent circuit. The *d*-axis and *q*-axis (dynamic) equivalent circuits of a synchronous generator consist of resistors, inductors, and voltage sources. In the subtransient period, the equivalent of the damper windings needs to be considered. In the transient period, this equivalent can be ignored. When the new steady state is reached, the current in the rotor winding becomes a constant (dc); thus, one can further ignore the equivalent inductance of this winding. This approximate method results in three equivalent circuits, listed in order of complexity: subtransient, transient, and steady state. For each circuit, one can define parameters such as (effective) reactance and time constant. For example, the *d*-axis circuit for the transient period has an effective reactance X'_d and a time constant T'_{do} (computed from the *R-L* circuit) when open circuited. The parameters of a synchronous machine can be computed from experimental data and are used in numerical studies. Typical values for these parameters are given in Table 66.1.

References on synchronous generators are numerous because of the historical importance of these machines in large-scale electric energy production. [Sarma, 1979] includes a derivation of the steady-state and dynamic models, dynamic performance, excitation, and trends in development of large generators. [Chapman, 1991]

TABLE 66.1 Typical Synchronous Generator Parameters[a]

Parameter	Symbol	Round Rotor	Salient-Pole Rotor with Damper Windings
Synchronous reactance			
d-axis	X_d	1.0–2.5	1.0–2.0
q-axis	X_q	1.0–2.5	0.6–1.2
Transient reactance			
d-axis	X'_d	0.2–0.35	0.2–0.45
q-axis	X'_q	0.5–1.0	0.25–0.8
Subtransient reactance			
d-axis	X''_d	0.1–0.25	0.15–0.25
q-axis	X''_q	0.1–0.25	0.2–0.8
Time constants			
Transient			
Stator winding open-circuited	T'_{do}	4.5–13	3.0–8.0
Stator winding short-circuited	T'_d	1.0–1.5	1.5–2.0
Subtransient			
Stator winding short-circuited	T''_d	0.03–0.1	0.03–0.1

[a] Reactances are per unit, i.e., normalized quantities. Time constants are in seconds.

Source: M.A. Laughton and M.G. Say, eds., *Electrical Engineer's Reference Book,* Stoneham, Mass.: Butterworth, 1985.

and [McPherson, 1981] are among the basic sources of reference in electric machinery, where many practical aspects are given. An introductory discussion of power system stability as related to synchronous generators can be found in [Bergen, 1986]. A number of handbooks that include subjects on ac as well as dc generators are also available in [Laughton and Say, 1985; Fink and Beaty, 1987; and Chang, 1982].

Superconducting Generators

The demand for electricity has increased steadily over the years. To satisfy the increasing demand, there has been a trend in the development of generators with very high power rating. This has been achieved, to a great extent, by improvement in materials and cooling techniques. Cooling is necessary because the loss dissipated as heat poses a serious problem for winding insulation. The progress in machine design based on conventional methods appears to reach a point where further increases in power ratings are becoming difficult. An alternative method involves the use of superconductivity.

In a superconducting generator, the field winding is kept at a very low temperature so that it stays superconductive. An obvious advantage to this is that no resistive loss can take place in this winding, and therefore a very large current can flow. A large field current yields a very strong magnetic field, and this means that many issues considered important in the conventional design may no longer be critical. For example, the conventional design makes use of iron core for armature windings to achieve an appropriate level of magnetic flux for these windings; iron cores, however, contribute to heat loss—because of the effects of hysteresis and eddy currents—and therefore require appropriate designs for winding insulation. With the new design, there is no need for iron cores since the magnetic field can be made very strong; the absence of iron allows a simpler winding insulation, thereby accommodating additional armature windings.

There is, however, a limit to the field current increase. It is known that superconductivity and diamagnetism are closely related; that is, if a material is in the superconducting state, no magnetic lines of force can enter its interior. Increasing the current produces more and more magnetic lines of force, and this can continue until the dense magnetic field can penetrate the material. When this happens, the material fails to stay superconductive, and therefore resistive loss can take place. In other words, a material can stay superconductive until a certain *critical field strength* is reached. The critical field strength is dependent on the material and its temperature.

A typical superconducting design of an ac generator, as in the conventional design, has the field winding mounted on the rotor and armature winding on the stator. The main differences between the two designs lie in the way cooling is done. The rotor has an inner body which is to support a winding cooled to a very low temperature by means of liquid helium. The liquid helium is fed to the winding along the rotor axis. To maintain the low temperature, thermal insulation is needed, and this can be achieved by means of a vacuum space and a radiation shield. The outer body of the rotor shields the rotor's winding from being penetrated by the armature fields so that the superconducting state will not be destroyed. The stator structure is made of nonmagnetic material, which must be mechanically strong. The stator windings (armature) are not superconducting and are typically cooled by water. The immediate surroundings of the machine must be shielded from the strong magnetic fields; this requirement, though not necessary for the machine's operation, can be satisfied by the use of a copper or laminated iron screen.

From a circuit viewpoint, superconducting machines have smaller internal impedance relative to the conventional ones (refer to equivalent circuit shown in Fig. 66.4). Recall that the reactance jX_s stems from the fact that the armature circuits give rise to a magnetic field that tends to counter the effect of the rotor winding. In the conventional design, such a magnetic field is enhanced because iron core is used for the rotor and stator structures; thus jX_s is large. In the superconducting design, the core is basically air; thus, jX_s is smaller. The difference is generally a ratio of 5:1 in magnitude. An implication is that, at the same level of output current I_a and terminal voltage V_t, it requires of the superconducting generator a smaller induced emf E_F or, equivalently, a smaller field current.

It is expected that the use of superconductivity adds another 0.4% to the efficiency of generators. This improvement might seem insignificant (compared to an already achieved figure of 98% by the conventional design) but proves considerable in the long run. It is estimated that given a frame size and weight, a superconducting generator's capacity is three times that of a conventional one. However, the new concept has to deal with such practical issues as reliability, availability, and costs before it can be put into large-scale operation.

[Bumby, 1983] provides more details on superconducting electric machines with issues such as design, performance, and application of such machines.

Induction Generators

Conceptually, a three-phase induction machine is similar to a synchronous machine, but the former has a much simpler rotor circuit. A typical design of the rotor is the squirrel-cage structure, where conducting bars are embedded in the rotor body and shorted out at the ends. When a set of three-phase currents (waveforms of equal amplitude, displaced in time by one-third of a period) is applied to the stator winding, a rotating magnetic field is produced. (See the discussion of a revolving magnetic field for synchronous generators in the section "Principle of Operation".) Currents are therefore induced in the bars, and their resulting magnetic field interacts with the stator field to make the rotor rotate in the same direction. In this case, the machine acts as a motor since, in order for the rotor to rotate, energy is drawn from the electric power source. When the machine acts as a motor, its rotor can never achieve the same speed as the rotating field (this is the synchronous speed) for that would imply no induced currents in the rotor bars. If an external mechanical torque is applied to the rotor to drive it beyond the synchronous speed, however, then electric energy is pumped to the power grid, and the machine will act as a generator.

An advantage of induction generators is their simplicity (no separate field circuit) and flexibility in speed. These features make induction machines attractive for applications such as windmills.

A disadvantage of induction generators is that they are highly inductive. Because the current and voltage have very large phase shifts, delivering a moderate amount of power requires an unnecessarily high current on the power line. This current can be reduced by connecting capacitors at the terminals of the machine. Capacitors have negative reactance; thus, the machine's inductive reactance can be compensated. Such a scheme is known as capacitive compensation. It is ideal to have a compensation in which the capacitor and equivalent inductor completely cancel the effect of each other. In windmill applications, for example, this faces a great challenge because the varying speed of the rotor (as a result of wind speed) implies a varying equivalent inductor. Fortunately, strategies for ideal compensation have been designed and put to commercial use.

In [Chapman, 1991], an analysis of induction generators and the effect of capacitive compensation on machine's performance are given.

DC Generators

To obtain dc electricity, one may prefer an available ac source with an electronic rectifier circuit. Another possibility is to generate dc electricity directly. Although the latter method is becoming obsolete, it is still important to understand how a dc generator works. This section provides a brief discussion of the basic issues associated with dc generators.

Principle of Operation

As in the case of ac generators, a basic design will be used to explain the essential ideas behind the operation of dc generators. Figure 66.7 is a schematic diagram showing an end of a simple dc machine.

The stator of the simple machine is a permanent magnet with two poles labeled N and S. The rotor is a cylindrical body and has two (insulated) conductors embedded in its surface. At one end of the rotor, as illustrated in Fig. 66.7, the two conductors are connected to a pair of copper segments; these semicircular segments, shown in the diagram, are mounted on the shaft of the rotor. Hence, they rotate together with the rotor. At the other end of the rotor, the two conductors are joined to form a coil.

Assume that an external torque is applied to the shaft so that the rotor rotates at a certain speed. The rotor winding formed by the two conductors experiences a periodically varying magnetic field, and hence an emf is induced across the winding. Note that this voltage periodically alternates in sign, and thus, the situation is conceptually the same as the one encountered in ac generators. To make the machine act as a dc source, viewed from the terminals, some form of rectification needs be introduced. This function is made possible with the use of copper segments and brushes.

According to Fig. 66.7, each copper segment comes into contact with one brush half of the time during each rotor revolution. The placement of the (stationary) brushes guarantees that one brush always has positive potential relative to the other. For the chosen direction of rotation, the brush with higher potential is the one directly beneath the N-pole. (Should the rotor rotate in the reverse direction, the opposite is true.) Thus, the brushes can serve as the terminals of the dc source. In electric machinery, the rectifying action of the copper segments and brushes is referred to as **commutation**, and the machine is called a commutating machine.

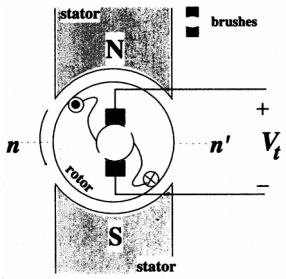


FIGURE 66.7 A basic two-pole dc generator. V_t is the voltage across the machine terminals. ⊗# and ⊙# indicate the direction of currents (into or out of the page) that would flow if a closed circuit is made.

A qualitative sketch of V_t, the voltage across terminals of an unloaded simple dc generator, as a function of time is given in Fig. 66.8. Note that this voltage is not a constant. A unidirectional current can flow when a resistor is connected across the terminals of the machine.

The pulsating voltage waveform generated by the simple dc machine usually cannot meet the requirement of practical applications. An improvement can be made with more pairs of conductors. These conductors are placed in slots that are made equidistant on the rotor surface. Each pair of conductors can generate a voltage waveform similar to the one in Fig. 66.8, but there are time shifts among these waveforms due to the spatial displacement among the conductor pairs. For instance, when an individual voltage is minimum (zero), other voltages are not. If these voltage waveforms are added, the result is a near constant voltage waveform. This improvement of the dc waveform requires many pairs of the copper segments and a pair of brushes.

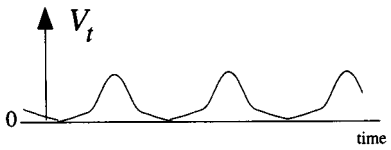


FIGURE 66.8 Open-circuited terminal voltage of the simple dc generator.

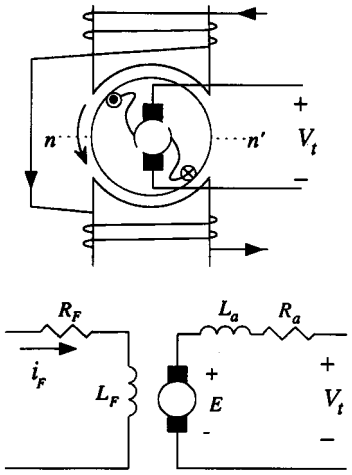


FIGURE 66.9 A simple two-pole dc generator with a stator winding to produce a magnetic field. *Top*, main components of the machine; *bottom*, coupled-circuit representation; the circuit on the left represents the field winding; the induced emf E is controlled by i_F.

When the generator is connected to an electrical load, load currents flow through the rotor conductors. Therefore, a magnetic field is set up in addition to that of the permanent magnet. This additional field generally weakens the magnetic flux seen by the rotor conductors. A direct consequence is that the induced emf's are less than those in an unloaded machine. Similar to the case of ac generators, this phenomenon is referred to as armature reaction, or flux-weakening effect.

The use of brushes in the design of dc generators can cause a serious problem in practice. Each time a brush comes into contact with two adjacent copper segments, the corresponding conductors are short-circuited. For a loaded generator, such an event occurs when the currents in these conductors are not zero, resulting in flashover at the brushes. This means that the life span of the brushes can be drastically reduced and that frequent maintenance is needed. A number of design techniques have been developed to mitigate this problem.

Mathematical/Circuit Model

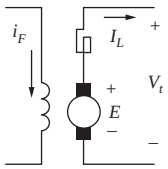
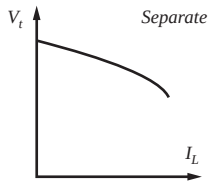
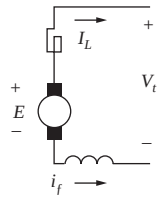
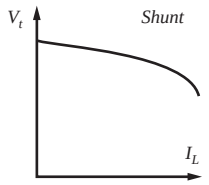
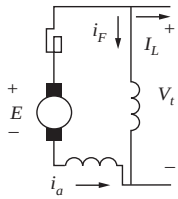
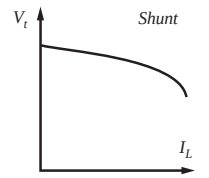
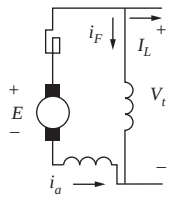
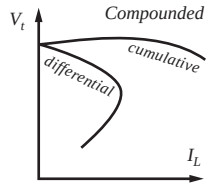
The (no-load) terminal voltage V_t of a dc generator depends on several factors. First, it depends on the construction of the machine (e.g., the number of conductors). Second, the voltage magnitude depends on the magnetic field of the stator: the stronger the field is, the higher the voltage becomes. Third, since the induced emf is proportional to the rate of change of the magnetic flux (Faraday's law), the terminals have higher voltage with a higher machine speed. One can write

$$V_{t(\text{no load})} = K\lambda n$$

where K is a constant representing the first factor, λ is magnetic flux, and n is rotor speed. The foregoing equation provides some insights into the voltage control of dc generators. Among the three terms, it is impractical to modify K, which is determined by the machine design. Changing n over a wide range may not be feasible since this is limited by what drives the rotor. Changing the magnetic flux λ can be done if the permanent magnet is replaced by an electromagnet, and this is how the voltage control is done in practice. The control of λ is made possible by adjusting the current fed to this electromagnet. Figure 66.9 shows the modified design of the simple dc generator. The stator winding is called the *field winding*, which produces excitation for the machine. The current in the field winding is adjusted by means of a variable resistor connected in series with this winding. It is also possible to use two field windings in order to have more flexibility in control.

The use of field winding(s) on the stator of the dc machine leads to a number of methods to produce the magnetic field. Depending on how the field winding(s) and the rotor winding are connected, one may have

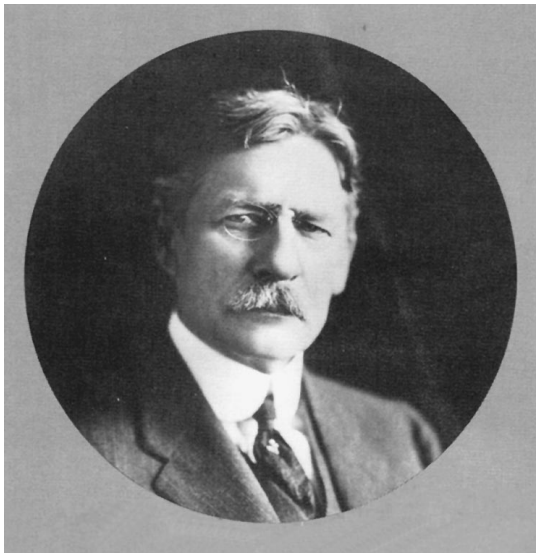
TABLE 66.2 Excitation Methods and Voltage Current Characteristics for DC Generators

Excitation Methods	Characteristics
Separate	
	
For low currents, the curve is nearly a straight line. As load current increases, the armature reaction becomes more severe and contributes to the nonlinear drop.	
Series	
	
At no load, there is no field current, and voltage is due to the residual flux of the stator core. The voltage rises rapidly over the range of low currents, but the resistive drop soon becomes dominant.	
Shunt	
	
Voltage buildup depends on the residual flux. The shunt field resistance must be less than a critical value.	
Compounded	
	
There are two field windings. Depending on how they are set up, one may have *cumulative* if the two fields are additive, *differential* if the two fields are subtractive.	*Cumulative:* An increase in load current increases the resistive drop, yet creates more flux. At high currents, however, resistive drop becomes dominant. *Differential:* An increase in load current not only increases the resistive drop, but also reduces the net flux. Voltage drops drastically.

shunt excitation, series excitation, etc. Each connection yields a different terminal characteristic. The possible connections and the resulting current–voltage characteristics are given in Table 66.2.

[Chapman, 1991] and [Fink and Beaty, 1987] provide more detailed discussions of dc generators. Specifically, [Chapman, 1991] shows how the characteristics are derived for various excitation methods.

Frank Julian Sprague (1857–1934)



Frank Sprague was a true entrepreneur in the new field of electrical technology. After a brief stint on Thomas Edison's staff, Sprague went out on his own, founding Sprague Electric Railway and Motor Company in 1884. In 1887, Sprague equipped the first modern trolley railway in the United States. Sprague's successful construction of a streetcar system for Richmond, Virginia, in 1888 was the beginning of the great electric railway boom. Sprague followed this system with 100 other such systems, both in America and Europe, during the next two years. In less than 15 years, more than 20,000 miles (32,000 km) of electric street railway were built.

In addition to his work in railroads, Sprague's diverse talents led to his development of electric elevators, an ac induction smelting furnace, miniature electric power units for use in small appliances, and as a member of the U.S. Naval Consulting Board during World War I, he developed fuses and air and depth bombs. Sprague was awarded the AIEE's Edison Medal in 1910. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

Defining Terms

Armature circuit: A winding where the load current is carried.

Armature reaction: The phenomenon in which the magnetic field due to currents in the armature circuit counters the effect of the field circuit.

Commutation: A mechanical technique in which rectification can be achieved in dc machines.

Field circuit: A set of windings that produces a magnetic field so that the electromagnetic induction can take place in electric machines.

Revolving fields: A magnetic field created by multiphase currents on spatially displaced windings in rotating machines; the field revolves in the air gap.

Swing equation: A nonlinear differential equation describing the rotor dynamics of an ac synchronous machine.

Synchronous speed: A characteristic speed of synchronous and induction machines with a revolving field; it is determined by the rotor structure and the line frequency.

Related Topics

2.2 Ideal and Practical Sources • 3.4 Power and Energy • 104.1 Welding and Bonding

References

M. S. Sarma, *Synchronous Machines (Their Theory, Stability, and Excitation Systems),* New York: Gordon and Breach, 1979.
J. R. Bumby, *Superconducting Rotating Electrical Machines,* New York: Oxford University Press, 1983.
S. J. Chapman, *Electric Machinery Fundamentals,* New York: McGraw-Hill, 1991.
G. McPherson, *An Introduction to Electrical Machines and Transformers,* New York: Wiley, 1981.
A. R. Bergen, *Power Systems Analysis,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
M. A. Laughton and M. G. Say, Eds., *Electrical Engineer's Reference Book,* Stoneham, Mass.: Butterworth, 1985.
D. G. Fink and H. W. Beaty, Eds., *Standard Handbook for Electrical Engineers,* New York: McGraw-Hill, 1987.
S. S. L. Chang, ed., *Fundamentals Handbook of Electrical and Computer Engineering,* New York: Wiley, 1982.

Further Information

Several handbooks, e.g., *Electrical Engineer's Reference Book* and *Standard Handbook for Electrical Engineers,* give more details on the machine design. [Bumby, 1983] covers the subject of superconducting generators. Some textbooks in the area of rotating machines are listed as [Sarma, 1979; Chapman, 1991; McPherson, 1981].

The quarterly journal *IEEE Transactions on Energy Conversion* covers the field of rotating machinery and power generation. Another IEEE quarterly journal, *IEEE Transactions on Power Systems,* is devoted to the general aspects of power system engineering and power engineering education.

The bimonthly journal *Electric Machines and Power Systems,* published by Hemisphere Publishing Corporation, covers the broad field of electromechanics, electric machines, and power systems.

66.2 Motors

Donald Galler

Electric motors are the most commonly used prime mover in industry. The classification of the types of ac and dc motors commonly used in industrial applications is shown in Fig. 66.10.

Motor Applications

DC Motors

Permanent magnet (PM) field motors occupy the low end of the horsepower (hp) range and are commercially available up to about 10 hp. Below 1 hp they are used for servo applications, such as in machine tools, for robotics, and in high-performance computer peripherals.

Wound field motors are used above about 10 hp and represent the highest horsepower range of **dc motor** application. They are commercially available up to several hundred horsepower and are commonly used in traction, hoisting, and other applications where a wide range of speed control is needed. The shunt wound dc motor is commonly found in industrial applications such as grinding and machine tools and in elevator and hoist applications. Compound wound motors have both a series and shunt field component to provide specific torque-speed characteristics. Propulsion motors for transit vehicles are usually compound wound dc motors.

AC Motors

Single-phase ac motors occupy the low end of the horsepower spectrum and are offered commercially up to about 5 hp. Single-phase **synchronous motors** are only used below about 1/10 of a horsepower. Typical applications are timing and motion control, where low torque is required at fixed speeds. Single-phase **induction motors** are used for operating household appliances and machinery from about 1/3 to 5 hp.

Polyphase ac motors are primarily three-phase and are by far the largest electric prime mover in all of industry. They are offered in ranges from 5 up to 50,000 hp and account for a large percentage of the total motor industry in the world. In number of units, the three-phase **squirrel cage induction motor** is the most common. It is commercially available from 1 hp up to several thousand horsepower and can be used on

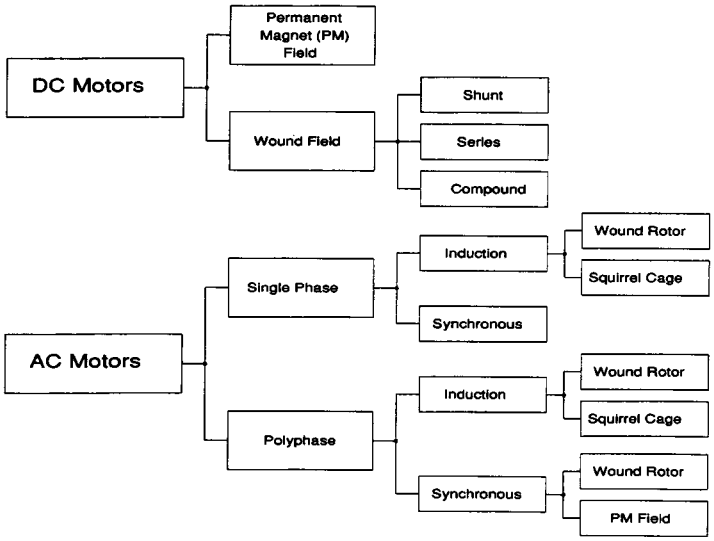


FIGURE 66.10 Classification of ac and dc motors for industrial applications.

conventional ac power or in conjunction with adjustable speed ac drives. Fans, pumps, and material handling are the most common applications.

When the torque-speed characteristics of a conventional ac induction motor need to be modified, the **wound rotor induction motor** is used. These motors replace the squirrel cage rotor with a wound rotor and slip rings. External resistors are used to adjust the torque-speed characteristics for speed control in such applications as ac cranes, hoists, and elevators.

Three-phase synchronous motors can be purchased with PM fields up to about 5 hp and are used for applications such as processing lines and transporting film and sheet materials at precise speeds.

In the horsepower range above about 10,000 hp, three-phase synchronous motors with wound fields are used rather than large squirrel cage induction motors. Starting current and other characteristics can be controlled by the external field exciter. Three-phase synchronous motors with wound fields are available up to about 50,000 hp.

Motor Analysis

DC Motor Analysis

The **separately excited dc motor** is the simplest of all dc motors and is the one most commonly found in industrial applications. The equivalent circuit is shown in Fig. 66.11. An adjustable dc voltage V is applied to the motor terminals. This voltage is impressed across the series combination of the armature resistance R_a and the back emf V_a generated by the armature. The field is energized with a separate dc power supply, usually at 300 or 500 V dc.

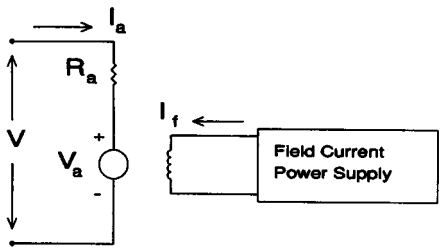


FIGURE 66.11 Equivalent circuit of separately excited dc motor.

The terminal voltage is given as

$$V = I_a R_a + V_a \tag{66.3}$$

The torque in steady state is

$$T = K_t I_a \Phi \tag{66.4}$$

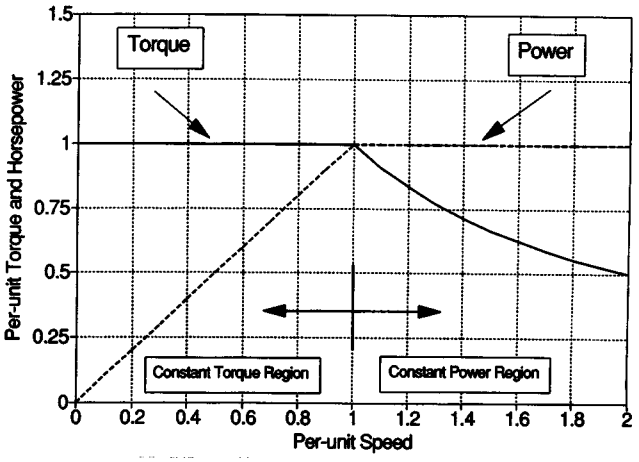


FIGURE 66.12 Torque-speed capability for the separately excited dc motor.

and the generated armature voltage is

$$V = K_a \omega \Phi \tag{66.5}$$

where Φ is the magnitude of the flux produced by the field winding and is proportional to the field current I_f. The torque constant K_t and the armature constant K_a are numerically equal in a consistent set of units. ω is the shaft speed in radians/second.

Solving the three equations gives the steady-state speed as

$$\omega = \frac{V - T(R_a / K_t \Phi)}{K_a \Phi} \tag{66.6}$$

The input power and output power are

$$P_{\text{in}} = I_a V \tag{66.7}$$

$$P_{\text{out}} = \omega T = I_a V - I_a^2 R_a \tag{66.8}$$

The efficiency (neglecting power loss in the field) is

$$\eta = \frac{P_{\text{out}}}{P_{\text{in}}} = \frac{\omega T}{I_a V} \tag{66.9}$$

A simplified torque-speed curve is shown in Fig. 66.12. The torque capability is constant up to the base speed of the motor while the armature and field currents are held constant. The speed is controlled by armature voltage in this range. Operation above base speed is accomplished by reducing the field current. This is called *field weakening*. The motor operates at constant power in this range, and the torque reduces with increasing speed.

Synchronous Motor Analysis

Synchronous motor analysis may be conducted using either a round rotor or salient pole model for the motor. The round rotor model is used in the following discussion. The equivalent circuit is shown in Fig. 66.13. The model consists of two ac voltages V_1 and V_2 connected by an impedance $Z = R + jX$. Analysis is facilitated by use of the phasor diagram shown in Fig. 66.14. The power delivered through the impedance to the load is

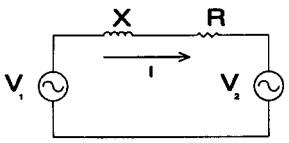


FIGURE 66.13 Per-phase equivalent circuit model for the synchonous motor (round rotor model).

$$P_2 = V_2 I \cos \phi_2 \tag{66.10}$$

where ϕ_2 is the phase angle of I with respect to V_2. The phasor current

$$I = \frac{V_1 - V_2}{Z} \tag{66.11}$$

is expressed in polar form as

$$\begin{aligned} I &= \frac{V_1 \angle \delta - V_2 \angle 0^\circ}{Z \angle \phi_2} \\ &= \frac{V_1}{Z} \delta - \angle \phi - \frac{V_2}{Z} \angle -\phi_z \end{aligned} \tag{66.12}$$

The equations make use of the fact that the three-phase operation is symmetrical and uses a "per-phase" equivalent circuit. This will also be true for the induction motor, which is analyzed in the following section.

The real part of I is

$$I \cos \phi_2 = \frac{V_1}{Z} \cos(\delta - \phi_z) - \frac{V_2}{Z} \cos(-\phi_z) \tag{66.13}$$

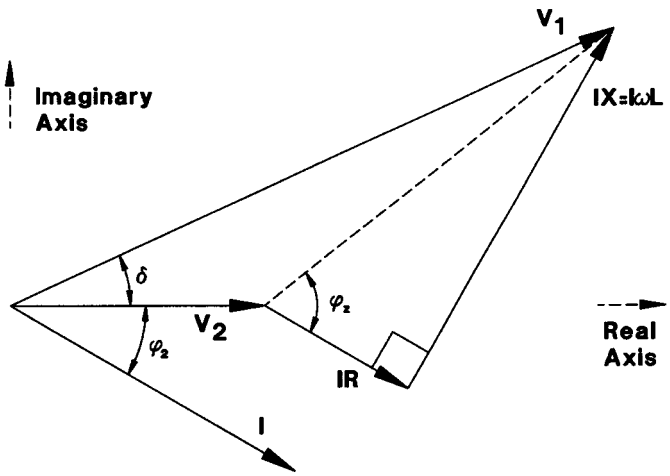


FIGURE 66.14 Phasor diagram for the ac synchronous motor (round rotor model).

Using Eq. (66.13) in Eq. (66.10) gives

$$P_2 = \frac{V_1 V_2}{Z} \cos(\delta - \phi_z) - \frac{V_2^2 R}{Z^2} \tag{66.14}$$

Letting $\alpha = 90° - \phi_z = \arctan R/X$ gives the output power as

$$P_2 = \frac{V_1 V_2}{Z} \sin(\delta + \alpha) - \frac{V_2^2 R}{Z^2} \tag{66.15}$$

and the input power as

$$P_1 = \frac{V_1 V_2}{Z} \sin(\delta - \alpha) + \frac{V_1^2 R}{Z^2} \tag{66.16}$$

Usually R is neglected and

$$P_1 = P_2 = \frac{V_1 V_2}{X} \sin \delta \tag{66.17}$$

which shows that the power is maximum when $\delta = 90°$ and is

$$P_{\text{MAX}} = \frac{V_1 V_2}{X} \tag{66.18}$$

The current can be found from Eqs. (66.15) and (66.16) since the only loss occurs in R. Setting

$$I^2 R = P_2 - P_1 \tag{66.19}$$

and solving for I gives

$$I = \sqrt{(P_2 - P_1)/R} \tag{66.20}$$

which is the input line current.

The power factor is

$$\cos \theta = \frac{P_1}{V_1 I} \tag{66.21}$$

and $\theta = \delta + \phi_2$ as shown in Fig. 66.14.

All the foregoing values are per-phase values. The total input power is

$$P_{\text{in}} = 3P_1 \tag{66.22}$$

The mechanical output power is

$$\begin{aligned} P_{\text{out}} &= T\omega \\ &= 3 \cdot P_2 \end{aligned} \tag{66.23}$$

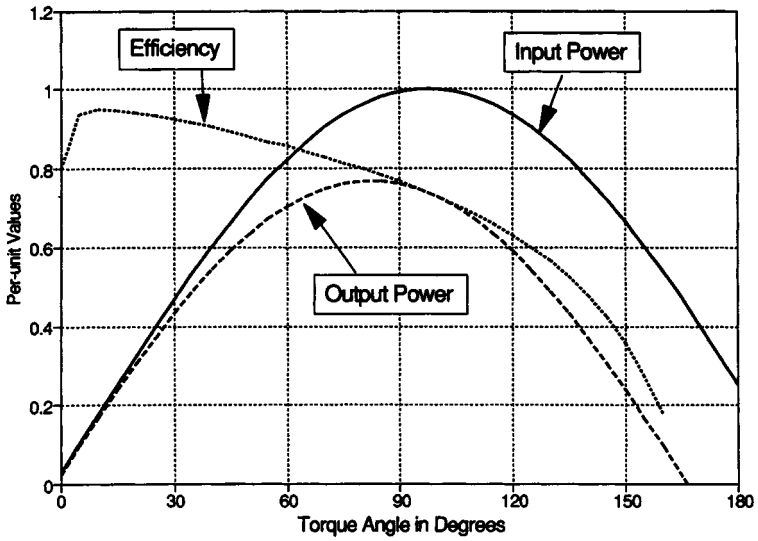


FIGURE 66.15 Synchronous motor performance.

and the torque is

$$T = 3 \cdot P_{\text{out}} / \omega \tag{66.24}$$

where ω is the rotational speed of the motor expressed in radians per second.

Synchronous motor operation is determined by the torque angle δ and is illustrated in Fig. 66.15 for a typical motor. Input power, output power, and current are shown on a per-unit basis. Torque is not shown but is related to output power only by a constant.

Induction Motor Analysis

The characteristic algebraic equations for the steady-state power, torque, and efficiency of the ac induction motor are derived from the per-phase equivalent circuit of Fig. 66.16. All voltages and currents are in sinusoidal steady state. The derivation of the equations can be simplified by defining the complex motor impedance as

$$Z_m = \frac{\alpha}{\zeta} + j\frac{\beta}{\zeta} \tag{66.25}$$

By defining the following constants as

$$\begin{aligned} M_1 &= R_1 R_2^2 \\ M_2 &= R_2 L_m{}^2 \\ M_3 &= L_2 + L_m \\ M_4 &= L_1 + L_m \\ M_5 &= R_1 M_3^2 + M_2 \end{aligned} \tag{66.26}$$

the terms of Eq. (66.25) become

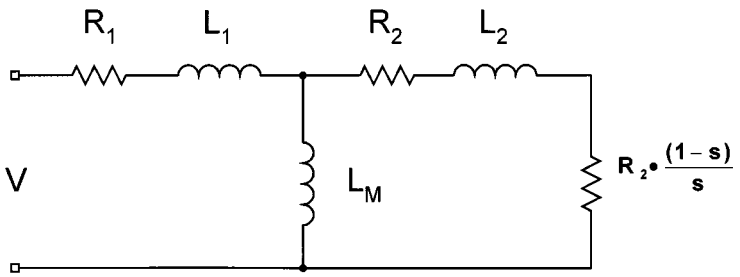


FIGURE 66.16 Equivalent circuit of ac induction motor.

$$\zeta = R_2^2 + \omega_s^2 M_3^2 \tag{66.27}$$

$$\alpha = \zeta R_1 + (\omega_m + \omega_s)\omega_s M_2 \tag{66.28}$$

$$\beta = (\omega_m + \omega_s)\,[\zeta L_1 + L_M R_2^2 + \omega_s^2 M_3 L_2 L_m] \tag{66.29}$$

The angular velocity ω_s is the slip frequency and is defined as follows:

$$\omega_s = \omega_f - \omega_m \tag{66.30}$$

where ω_f is the frequency applied to the **stator** and

$$\omega_m = \omega / N_p \tag{66.31}$$

is the **rotor** angular velocity in terms of an equivalent stator frequency. N_p is the number of stator pole pairs. The average mechanical output power of the motor is the power in the resistance $R_2\,\omega_m/\omega_s$ and is given as

$$P_{\text{out}} = \frac{3V^2\zeta\omega_m\omega_s M_2}{\alpha^2 + \beta^2} \tag{66.32}$$

where V is the rms line-neutral voltage. Since

$$\begin{aligned} T &= \frac{P_{\text{out}}}{\omega} \\ &= \frac{P_{\text{out}} N_p}{\omega_m} \end{aligned} \tag{66.33}$$

the torque becomes

$$T = \frac{3V^2\zeta N_p\omega_s M_2}{\alpha^2 + \beta^2} \tag{66.34}$$

The motor efficiency is defined as

$$\eta = \frac{P_{\text{out}}}{P_{\text{in}}} \tag{66.35}$$

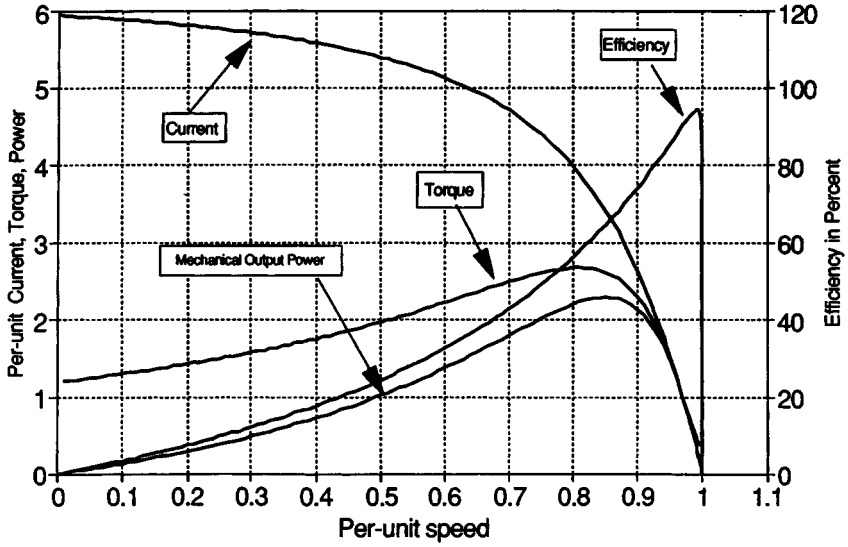


FIGURE 66.17 Induction motor operating characteristics, fixed voltage, and frequency.

where the input power is

$$P_{\text{in}} = \frac{3V_f^2 \cos\theta}{|Z_m|} = \frac{3V_f^2\,\alpha}{|Z_m|^2\,\zeta} \tag{66.36}$$

Using Eqs. (66.32) and (66.36), the efficiency becomes

$$\eta = \frac{\omega_m \omega_s M_2}{\alpha} \tag{66.37}$$

Typical performance characteristics of the induction motor are shown in Fig. 66.17.

Classical analysis represents all the motor expressions in terms of the slip, *s*, which is defined as

$$s = \frac{\omega_f - \omega_m}{\omega_f} \tag{66.38}$$

where ω_m is the equivalent mechanical frequency of the rotor, $\omega_m = \omega / N_p$, and ω_f is the angular velocity of the stator field in radians/second.

In this format, the output power is

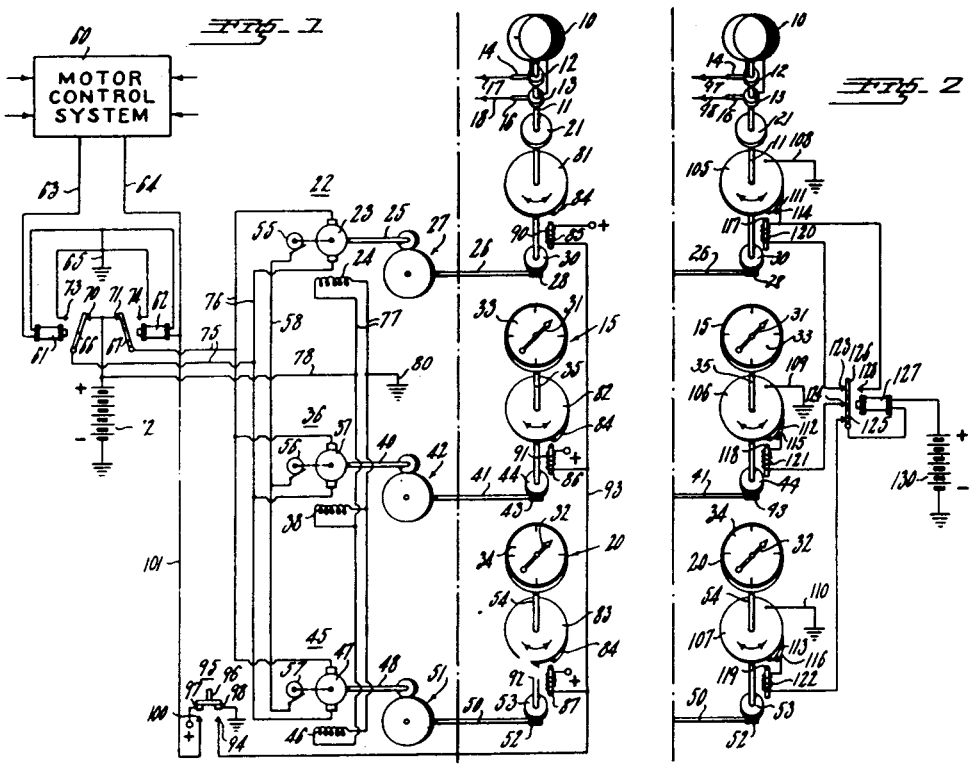
$$P = I_2^2 R_2 \cdot \frac{(1 - s)}{s} \tag{66.39}$$

Automatic Motor Synchronization Control

William P. Lear
Patented July 2, 1946
#2,403,098

Lear described a system for synchronizing instrumentation throughout an aircraft using DC servo motors instead of mechanical linkages that loaded down the master instrument. Its greater application came in using it to control and maintain altitude and heading by synchronizing the aircraft's control surfaces and using the servos to adjust them. This "autopilot" helped reduce pilot fatigue on long flights and was one of the developments that made commercial air practical during the 1950s and beyond. The servo control principles described are still used in automated air and sea navigation today.

Lear is perhaps best known for his development of small corporate jet aircraft known as Learjets in the 1960s. He patented the first practical car radio in the 1930s that launched today's giant Motorola Company. He also developed the eight-track tape system for autos in the 1960s and before his death in 1978, he designed the Lear fan, a high speed propeller aircraft made entirely from composites. (Copyright © 1995, DewRay Products, Inc. Used with permission.)



The maximum torque, T_m, occurs at a slip of

$$S_m = \frac{R_2}{\sqrt{R_1^2 + \left(X_1 + X_2\right)^2}} \tag{66.40}$$

where X_1 and X_2 are the stator and rotor reactances corresponding to L_1 and L_2. If R_1 and X are neglected, the torque can be expressed as

$$T = 2T_m \cdot \frac{ss_m}{s^2 + s_m^2} \tag{66.41}$$

but this expression loses accuracy if $s < 0.1$ where most practical operation takes place. Another expression,

$$T = 2T_m \cdot \frac{ss_m}{s^2 + s_m\left(1 + sR_1/R_2\right)^2} \cdot \left(\left(1 + \left(1 + s_m R_1/R_2\right)^2\right)\right)$$

may be used and is useful over the whole slip range. The full equation set of the previous discussion should be used where variable frequency and variable voltage operation is used, such as in adjustable speed drives. These equations are accurate for all regions of motor and generator operation.

ACand DC Motor Terms

General Terms

ω: Shaft angular velocity in radians/second
P_{out}: Electrical output power
P_{in}: Electrical input power
η: Efficiency
T: Shaft torque

DC Motor Terms

I_a: Armature current
I_f: Field current
V_a: Back emf generated by armature
V: Motor terminal voltage
R_a: Armature resistance
K_t: Torque constant
K_a: Armature constant
Φ: Field flux

AC Induction Motor Terms

L_1: Stator winding inductance
R_1: Stator winding resistance
L_2: Rotor winding inductance
R_2: Rotor winding resistance
L_M: Magnetizing inductance
N_p: Number of pole pairs in stator winding
ω_f: Frequency of voltage applied to stator
ω_m: Rotor equivalent mechanical frequency
ω_s: Slip frequency, $\omega_s = \omega_f - \omega_m$
s: Slip $s = (\omega_f - \omega_m)/\omega_f$
T_M: Maximum torque
s_M: Slip at maximum torque

AC Synchronous Motor Terms

V_1: Terminal voltage
V_2: Back emf generated by rotor
R: Rotor circuit resistance
X: Rotor circuit reactance
Z: Rotor circuit impedance $Z = R + jX$
δ: Torque angle (between V_1 and V_2)
ϕ_2: Angle between I and V_2
ϕ_Z: Rotor circuit reactance angle $\phi_z = \tan^{-1} X/R$
α: $90° - \phi_z$
θ: Power factor angle $\theta = \delta + \phi_2$

Defining Terms

DC motor: A dc motor consists of a stationary active part, usually called the field structure, and a moving active part, usually called the armature. Both the field and armature carry dc.

Induction motor: An ac motor in which a primary winding on the stator is connected to the power source and polyphase secondary winding on the rotor carries induced current.

Permanent magnet dc motor: A dc motor in which the field flux is supplied by permanent magnets instead of a wound field.

Rotor: The rotating member of a motor including the shaft. It is commonly called the armature on most dc motors.

Separately excited dc motor: A dc motor in which the field current is derived from a circuit which is independent of the armature.

Squirrel cage induction motor: An induction motor in which the secondary circuit (on the rotor) consists of bars, short-circuited by end rings. This forms a squirrel cage conductor structure which is disposed in slots in the rotor core.

Stator: The portion of a motor that includes and supports the stationary active parts. The stator includes the stationary portions of the magnetic circuit and the associated windings and leads.

Synchronous motor: An ac motor in which the average speed of normal operation is exactly proportional to the frequency to which it is connected. A synchronous motor generally has rotating field poles which are excited by dc.

Wound rotor induction motor: An induction motor in which the secondary circuit consists of a polyphase winding or coils connected through a suitable circuit. When provided with slip rings, the term *slip-ring induction motor* is used.

Related Topics

2.2 Ideal and Practical Sources • 104.2 Large Drives

References

P. C. Sen, *Thyristor DC Drives,* New York: John Wiley, 1981.

P. C. Sen, *Principles of Electric Machines and Power Electronics,* 2nd ed., New York: John Wiley, 1997.

G. R. Slemon, *Electric Machines and Drives,* Reading, Mass.: Addison-Wesley, 1992.

I. Boldea and S. A. Nasar, *Vector Control of AC Drives,* Boca Raton, Fla.: CRC Press, 1992.

M. G. Say and E. O. Taylor, *Direct Current Machines,* 2nd ed., London: Pitman Publishing, 1986.

R. H. Engelmann and W. H. Middendorf, *Handbook of Electric Motors,* New York: Marcel Dekker, 1995.

D. W. Novotny and T. A. Lipo, *Vector Control and Dynamics of AC Drives,* Oxford: Clarendon Press, 1996.

Further Information

The theory of ac motor drive operation is covered in the collection of papers edited by Bimal K. Bose, *Adjustable Speed AC Drive Systems* (IEEE, 1981). A good general text is *Electric Machinery,* by Fitzgerald, Kingsley, and Umans. The analysis of synchronous machines is covered in the book *Alternating Current Machines,* by M.G. Say (Wiley, 1984). *Three-Phase Electrical Machines — Computer Simulation* by J. R. Smith (Wiley, 1993) covers computer modeling and simulation techniques.

66.3 Small Electric Motors

Elias G. Strangas

Introduction

Small electrical machines carry a substantial load in residential environments, but also in industrial environments, where they are mostly used to control processes.

In order to adapt to the limitations of the power available, the cost requirements, and the widely varying operating requirements, small motors are available in a great variety of designs. Some of the small motors require electronics in order to start and operate, while others can start and run directly connected to the supply line.

AC motors that can start directly from the line are mostly of the induction type. Universal motors are also used extensively for small AC powered, handheld tools. They can either run directly from the line or have their speed adjusted through electronics.

Stepping motors of many varying designs require electronics to operate. They are used primarily to position a tool or a component and are seldom used to provide steady rotating motion.

Besides these motors, permanent magnet AC motors are replacing rapidly both DC and induction motors for accurate speed and position control, but also to decrease size and increase efficiency. They require power and control electronics to start and run.

Single Phase Induction Motors

To produce rotation, a multi-phase stator winding is often used in an AC motor, supplied from a symmetric and balanced system of currents. The magnetomotive force of these windings interacts with the magnetic field of the rotor (induced or applied) to produce a torque. In three-phase induction motors, the rotor field is created by currents that are induced due to the relative speed of the rotor and the synchronously rotating stator field.

In an induction motor that is supplied by a single-phase stator current, it is not as clear how a rotating magnetomotive force can be created and a torque be produced. Two different concepts will be used to generate torque.

The first, conceptually simpler design concept, involves the generation of a second current which flows in a second winding of the stator. This auxiliary winding is spatially displaced on the stator. This brings the motor design close to the multi-phase principle. The current in the auxiliary winding has to be out of phase with the current in the main winding, and this is accomplished through the use of increased resistance in it or a capacitor in series with it. A motor can operate in this fashion over its entire speed range.

Once the motor is rotating, the second design concept allows that one of the phases, the auxiliary one, be disconnected. The current in the remaining main winding alone produces only a pulsating flux, which can be analyzed as the sum of two rotating fields of equal amplitude but opposite direction. These fields, as seen from the moving rotor, rotate at different speeds, hence inducing in it currents of different frequency and amplitude. If the speed of the rotor is w_r, the applied frequency to the stator is f and the number of pole pairs in the motor is p, the frequencies of the currents induced in the rotor are $pw_r - f$ and $pw_r + f$. These unequal currents in turn produce unequal torques in the two directions, with a nonzero net torque.

The various designs of single-phase induction motors result from the variety of ways that the two phases are generated and by whether the auxiliary phase remains energized after starting.

Shaded Pole Motors

These motors are simple, reliable, and inefficient. The stator winding is not distributed on the rotor surface, but rather it is concentrated on salient poles. The auxiliary winding, which has to produce flux out of phase with the main winding, is nothing but a hardwired shorted turn around a portion of the main pole as Fig. 66.18.

Because of the shorted turn, the flux out of the shaded part of the pole lags behind the flux out of the main pole. The motor always rotates from the main to the shaded pole, and it is not possible to change directions.

Shaded pole motors are inefficient and have high starting and running current and low starting torque. They are used where reliability and cost are important, while their small size makes unimportant the overall effect of their disadvantages, e.g., small fans. Their size ranges from 0.002 to 0.1 hp.

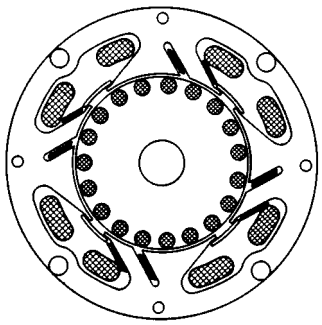


FIGURE 66.18 A shaded pole motor with tapered poles and magnetic wedges. (*Source:* C. G. Veinott and J. E. Martin, *Fractional and Subfractional Horsepower Electric Motors,* New York: McGraw-Hill, 1986. With permission.)

Resistance Split-Phase Motors

These motors have an auxiliary winding which simply has higher resistance than the main winding and is displaced spatially on the stator by about 90°. Both windings are distributed on the stator surface and are connected to the line voltage, but the different time constants between them makes the

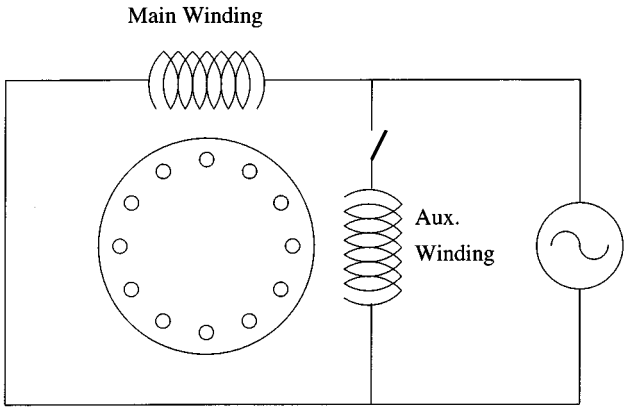


FIGURE 66.19 Connections of a resistive, split-phase motor.

current in the auxiliary winding lead that of the main. This arrangement results in a nonzero, but relatively low starting torque and high starting current.

The use of the auxiliary winding is limited only to starting—the motor runs more efficiently without it, as a single phase motor described earlier. A switch, activated by speed (centrifugal) or by stator temperature, disconnects the auxiliary winding shortly after starting. Figure 66.19 represents schematically the connections of this type of motor.

These motors represent an improvement in efficiency and starting torque over shaded pole motors, at the expense of increased cost and lower reliability. They are built to larger sizes, but their application is limited by the high starting current.

Capacitor Motors

Another way to generate a phase angle of current in the auxiliary winding is to include a capacitor in series with it. The capacitor can be disconnected after starting in a capacitor start motor. Their operation is similar to that of the resistance split-phase motor, but they have better starting characteristics and are made as large as 5 hp. Figure 66.20 shows schematically the wiring diagram of the capacitor start motor.

To optimize both starting and running, different values of the capacitor are used. One value of the capacitor is calculated to minimize starting current and maximize starting torque, while the other is designed to maximize efficiency at the operating point. A centrifugal switch handles the changeover. Such motors are built for up to 10 hp, and their cost is relatively high because of the switch and two capacitors. Figure 66.21 shows schematically the wiring diagram of the capacitor start and run motor.

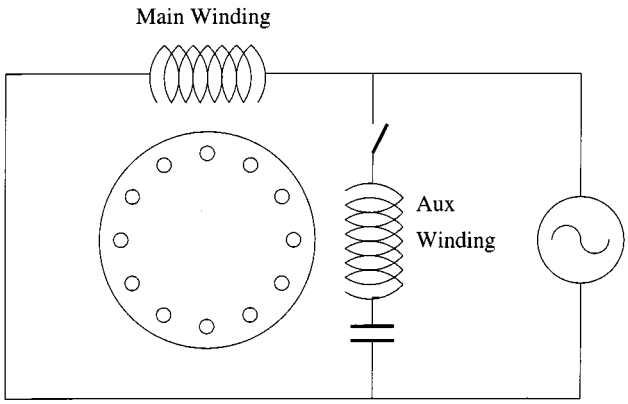


FIGURE 66.20 Conenctions of a capacitor start motor.

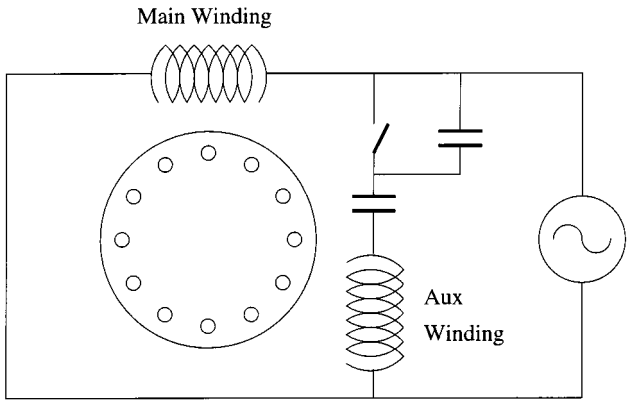


FIGURE 66.21 Connections of a capacitor-start, capacitor-run motor.

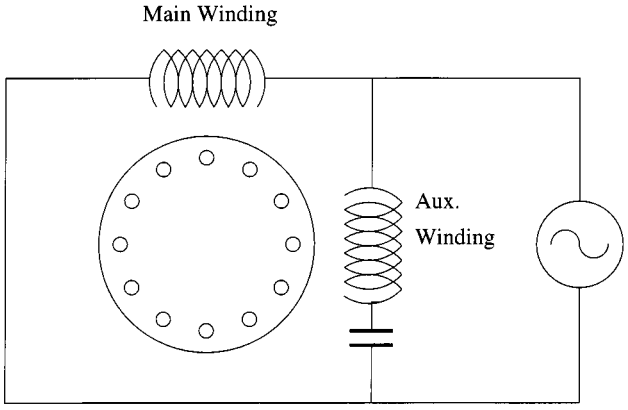


FIGURE 66.22 Connections of a permanent split capacitor motor.

A permanent split capacitor motor uses the same capacitor throughout the speed range of the motor. Its value requires a compromise between the values of the two-capacitor motors. The result is a motor design optimized for a particular application, e.g., a compressor or a fan. Figure 66.22 shows schematically the wiring diagram of the permanent split capacitor motor.

Universal Motors

These motors can be supplied from either DC or AC. Their design is essentially similar to a DC motor with series windings. When operated as AC motors, supplied say by a 60 Hz source, the current in the armature and the field windings reverses 120 times per second. As the torque is roughly proportional to both armature and field currents, connecting these windings in series guarantees that the current reverses in both at the same time, retaining the unidirectional torque. Figure 66.23 shows a schematic diagram of the connections of universal motors.

They can run at speeds up to 20,000 rpm, thus being very compact for a given horsepower. Their most popular applications include portable drills, food mixers, and fans.

Universal motors supplied from AC lend themselves easily to variable speed applications. A potentiometer, placed across the line voltage, controls the firing of a TRIAC thus varying the effective value of the voltage at the motor.

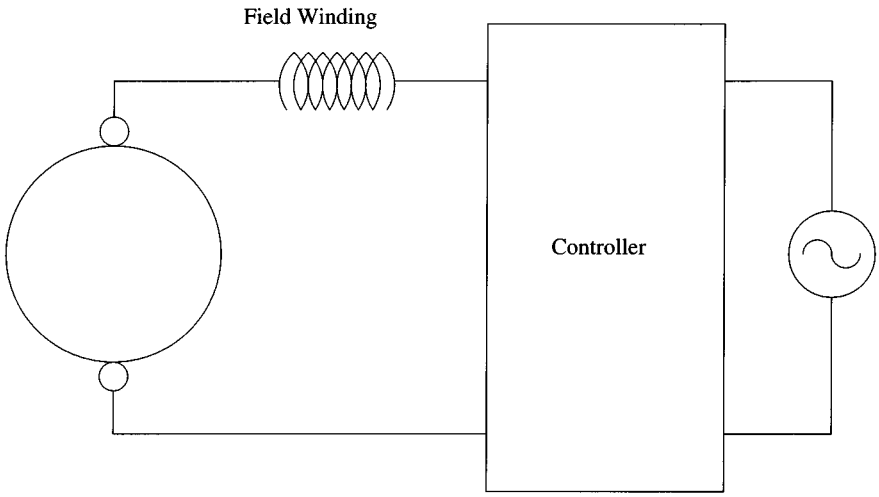


FIGURE 66.23 Connections of a universal motor.

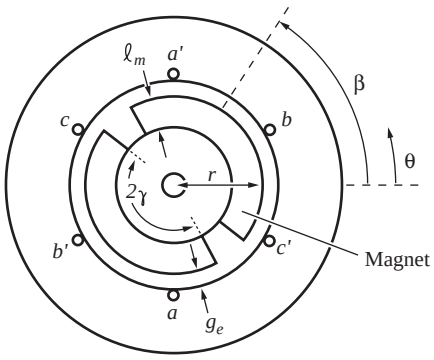


FIGURE 66.24 Surface mounted magnets on a Permanent Magnet AC motor.

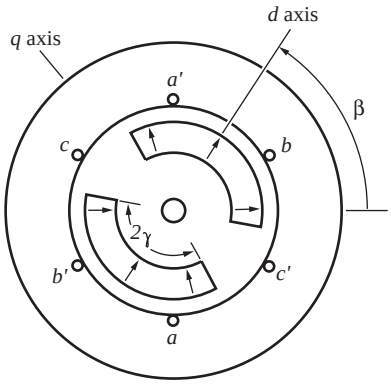


FIGURE 66.25 Inset (interior) magnets on a permanent magnet AC motor.

Permanent Magnet AC Motors

When compared to induction motors, permanent magnet motors have higher steady state torque for the same size and better efficiency. They carry a polyphase winding in the stator, which can be either rectangular or sinusoidally distributed. The rotor has a steel core, with permanent magnets mounted on it or inset. These magnets can be made from a variety of materials, such as rare earth, ceramic, etc.

Figure 66.24 shows a schematic of the cross-section of a motor with surface mounted magnets, and Fig. 66.25 shows a schematic of a motor with inset magnets.

The stator windings are supplied by a DC source through power electronic switches that constitute an inverter. Which switches are to be conducting at any time is determined by a controller, which in turn uses as inputs a speed or torque command and a measurement or an estimate of the rotor position. Figure 66.26 shows a schematic of the motor cross-section and of the inverter.

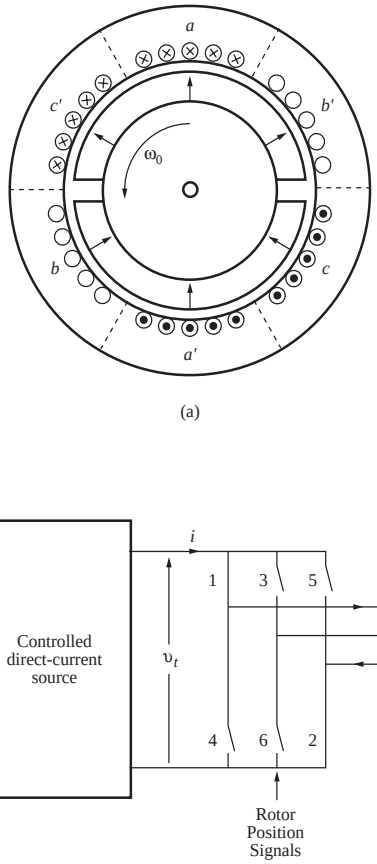


FIGURE 66.26 Permanent magnet AC motor and inverter.

When the stator windings are rectangular and are energized based only on the rotor position, the resulting set of PM motor, inverter, and controller is called a brushless DC motor. The developed torque is proportional to the airgap flux, B_g, and the stator current, I_s.

$$T = kB_g I_s$$

Due to the rotor speed, w_0 a voltage, e, (back emf) is induced to the stator windings.

$$e = kB_g w_0$$

Stepping Motors

These motors convert a series of power pulses to a corresponding series of equal angular movements. These pulses can be delivered at a variable rate, allowing the accurate positioning of the rotor without feedback. They can develop torque up to 15 Nm and can handle 1500 to 2500 pulses per second. They have zero steady state error in positioning and high torque density. An important characteristic of stepping motors is that when one phase is activated they do not develop a rotating but rather a holding torque, which makes them retain accurately their position, even under load.

Stepping motors are conceptually derived either from a variable reluctance motor or from a permanent magnet synchronous motor.

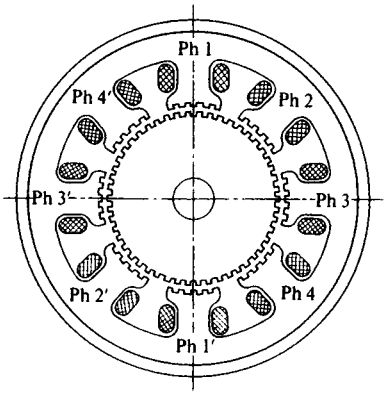


FIGURE 66.27 Cross-sectional view of a four-phase variable reluctance motor. Number of rotor teeth 50, step number 200, step angle 1.8°. (*Source:* Oxford University Press, 1989. With permission.)

One design of stepping motors, based on the doubly salient switched reluctance motor, uses a large number of teeth in the rotor (typically 45) to create saliency, as shown in Fig. 66.27. In this design, when the rotor teeth are aligned in say Phase 1, they are misaligned in Phases 2 and 3. A pulse of current in Phase 2 will cause a rotation so that the alignment will occur at Phase 2. If, instead, a pulse to Phase 3 is given, the rotor will move the same distance in the opposite rotation.

The angle corresponding to a pulse is small, typically 3° to 5°, resulting from alternatively exciting one stator phase at a time.

A permanent magnet stepping motor uses permanent magnets in the rotor. Figure 66.28 shows the steps in the motion of a four-phase PM stepping motor.

Hybrid stepping motors come in a variety of designs. One, shown in Fig. 66.29, consists of two rotors mounted on the same shaft, displaced by one half tooth. The permanent magnet is placed axially between the rotors, and the magnetic flux flows radially at the air gaps, closing through the stator circuit. Torque is created by the interaction of two magnetic fields, that due to the magnets and that due to the stator currents. This design allows a finer step angle control and higher torque, as well as smoother torque during a step.

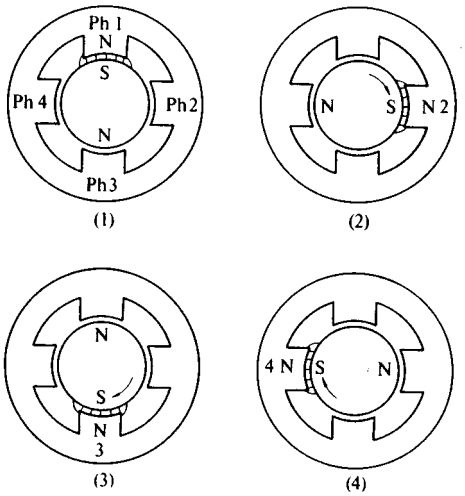


FIGURE 66.28 Steps in the operation of a permanent magnet stepping motor. (*Source:* T. Kenjo, *Stepping Motors and Their Microprocessor Controls,* Oxford University Press, 1989. With permission.)

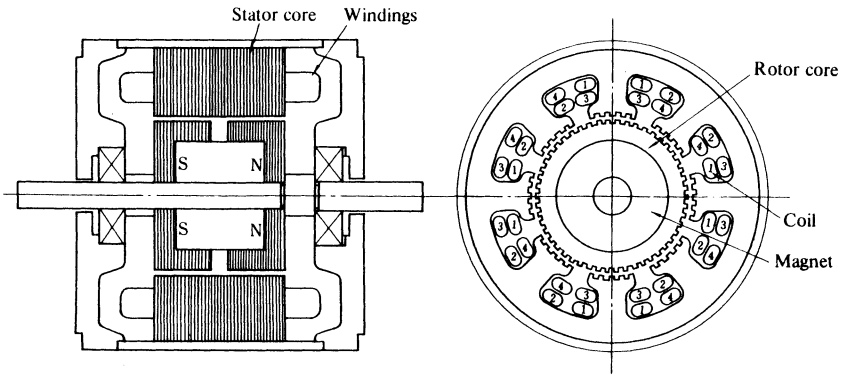


FIGURE 66.29 Construction of a hybrid stepping motor. (*Source:* Oxford University Press, 1989. With permission.)

Fundamental to the operation of stepping motors is the utilization of power electronic switches, and of a circuit providing the timing and duration of the pulses. A characteristic of a specific stepping motor is the maximum frequency it can operate at starting or running without load.

As the frequency of the pulses to a running motor is increased, eventually the motor loses synchronism. The relation between the frictional load torque and maximum pulse frequency is called the pull-out characteristic.

References

G. R. Slemon, *Electrical Machines and Drives,* Addison-Wesley, 1992.

T. Kenjo, *Stepping Motors and Their Microprocessor Controls,* Oxford University Press, 1984.

R. H. Engelman and W. H. Middendorf, Eds., *Handbook of Electric Motors,* New York: Marcel Dekker, 1995.

R. Miller and M. R. Miller, *Fractional Horsepower Electric Motors,* Bobs Merrill Co., 1984.

G. G. Veinott and J. E. Martin, *Fractional and Subfractional Horsepower Electric Motors,* New York: McGraw-Hill, 1986.

T. J. E. Miller, *Brushless Permanent-Magnet and Reluctance Motor Drives,* Oxford University Press, 1989.

S. A. Nasar, I. Boldea, and L. E. Unnewehr, *Permanent Magnet, Reluctance and Self-Synchronous Motors,* Boca Raton, Fla.: CRC Press, 1993.

Further Information

There is an abundance of books and literature on small electrical motors. *IEEE Transactions on Industry Applications, Power Electronics, Power Delivery and Industrial Electronics* all have articles on the subject. In addition, IEE and other publications and conference records can provide the reader with specific and useful information.

Electrical Machines and Drives [Slemon, 1992] is one of the many excellent textbooks on the subject. *Stepping Motors and their Microprocessor Controls* [Kenjo, 1984] has a thorough discussion of stepping motors, while *Fractional and Subfractional Horsepower Electric Motors* [Veinott and Martin, 1986] covers small AC and DC motors. *Brushless Permanent-Magnet and Reluctance Motor Drives* [Miller, 1989] and *Permanent Magnet, Reluctance and Self-Synchronous Motors* [Nasar et al., 1993] reflect the increased interest in reluctance and brushless DC motors, and provide information on their theory of operation, design and control. Finally, *Fractional Horsepower Electric Motors* [Miller and Miller, 1984] gives a lot of practical information about the application of small motors.

66.4 Simulation of Electric Machinery

Chee-Mun Ong

Simulation has been an option when the physical system is too large or expensive to experiment with or simply not available. Today, with powerful simulation packages, simulation is becoming a popular option for conducting studies and for learning, especially when well-established models are available. **Modeling** refers to the process

of analysis and synthesis to determine a suitable mathematical description that captures the relevant dynamical characteristics and simulation to the techniques of setting up and experimenting with the model.

Models of three-phase synchronous and induction machines for studying electromechanical and low-frequency electrical transients are well established because of the importance of generator and load behavior in stability and fault studies. Electric machines, however, do interact with other connected components over a wide range of frequencies, from fractions of Hertz for electromechanical phenomena to millions of Hertz for electromagnetic phenomena.

Reduced models suitable for limited frequency ranges are often preferred over complex models because of the relative ease in usage — as in determining the values of model parameter and in implementing a simulation. In practice, reduced models that portray essential behavior over a limited frequency range are obtained by making judicious approximations. Hence, one has to be aware of the assumptions and limitations when deciding on the level of modeling details of other components in the simulation and when interpreting the simulation results.

Basics in Modeling

Most machine models for electromechanical transient studies are derived from a lumped-parameter circuit representation of the machine's windings. Such lumped-parameter circuit representations are adequate for low-frequency electromechanical phenomena. They are suited for dynamical studies, often times to determine the machine's performance and control behavior or to learn about the nature of interactions from electromechanical oscillations. Studies of interactions occurring at higher frequencies, such as surge or traveling waves studies, may require a distributed-parameter circuit representation of the machine windings.

A lumped-parameter model for dynamical studies typically will include the voltage equations of the windings, derived using a coupled circuit approach, and an expression for the developed electromagnetic torque. The latter is obtained from an expression of the developed electromagnetic power by considering the input power expression and allowing for losses and magnetic energy storage terms. The expression for the developed electromagnetic torque is obtained by dividing that for developed electromagnetic power by the rotor mechanical speed. The rotor speed, in turn, is determined by an equation of the rotor's dynamics that equates the rotor's inertia torque to its acceleration torque.

For example, in a reduced order model of a separately excited dc machine that ignores the details of commutation action and only portrays the average values of voltage, current, and power, the armature winding can be represented as an equivalent winding whose axis is determined by the position of the commutator brushes. The induced voltage in the armature, E_a, due to field flux can be expressed as $k_a\omega\phi$, k_a being a machine constant; ω, the rotor speed; and ϕ the flux per pole. When armature reaction is ignored, ϕ will be the flux produced by the field winding. (See Fig. 66.30).

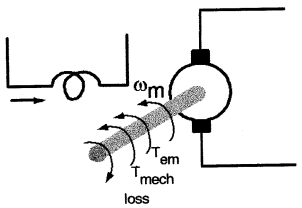


FIGURE 66.30 dc machine.

Using motoring convention, the voltage equations of the armature and field windings with axes that are at right angles to each other can be expressed as

$$V_a = E_a + R_a I_a + L_{aq}\frac{dI_a}{dt} \quad \text{V}$$
$$v_f = R_f i_f + L_f \frac{di_f}{dt} \tag{66.42}$$

where V_a is the terminal voltage of the armature winding, R_a its effective resistance including brush drops, L_{aq} its inductance, v_f the applied field voltage, R_f and L_f, the field circuit resistance and inductance. For motoring, positive I_a will flow into the positive terminal of V_a, as power flows from the external voltage source into the armature winding. Like the physical device, the model is capable of motoring and generating, and the transition from one mode to the other will take place naturally.

Equating the acceleration torque of the rotor to its inertia torque, one obtains:

$$T_{em} - T_{loss} + T_{mech} = J\frac{d\omega_m}{dt} \quad \text{N-m} \tag{66.43}$$

where T_{em} is the electromagnetic torque developed by the machine; T_{loss}, the equivalent torque representing friction and windage and stray load losses; and T_{mech}, the externally applied mechanical torque on the shaft in the direction of rotation. As shown in Fig. 66.30, T_{em} is positive for motoring and negative for generating; T_{mech} is negative for motoring and positive for generating.

Like the derivation of E_a, the developed torque, T_{em}, can be shown to be equal to $k_a\phi I_a$ by considering first the total power flow into the windings, that is,

$$\begin{aligned} V_a I_a &= E_a I_a + R_a I_a^2 + \frac{d\left(L_{aq} I_a^2/2\right)}{dt} \quad \text{W} \\ v_f i_f &= R_f i_f^2 + \frac{d\left(L_f i_f^2/2\right)}{dt} \end{aligned} \tag{66.44}$$

Summing the input powers to both windings, dropping the resistive losses and magnetic energy storage terms, and equating the remaining term to developed power, one will obtain the following relationships from which an expression of the developed torque can be written.

$$T_{em}\omega_m = P_{em} = E_a I_a \quad \text{W} \tag{66.45}$$

Figure 66.31 shows the flowchart for a simple dc machine simulation. The required inputs are V_a, v_f, and T_{mech}. Solving the windings' voltage equations with the required inputs, we can obtain the winding currents, I_a and i_f. The magnetizing curve block contains open-circuit test data of the machine to translate i_f to $k_a\phi$ or the ratio of the open-circuit armature voltage to some fixed speed, that is E_a/ω_{mo}. The simulation yields the output of the two winding currents, the field flux, the developed torque, and the rotor speed.

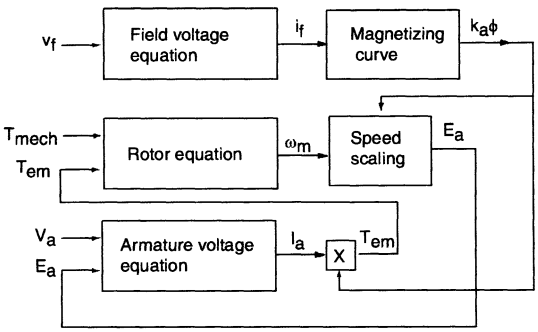
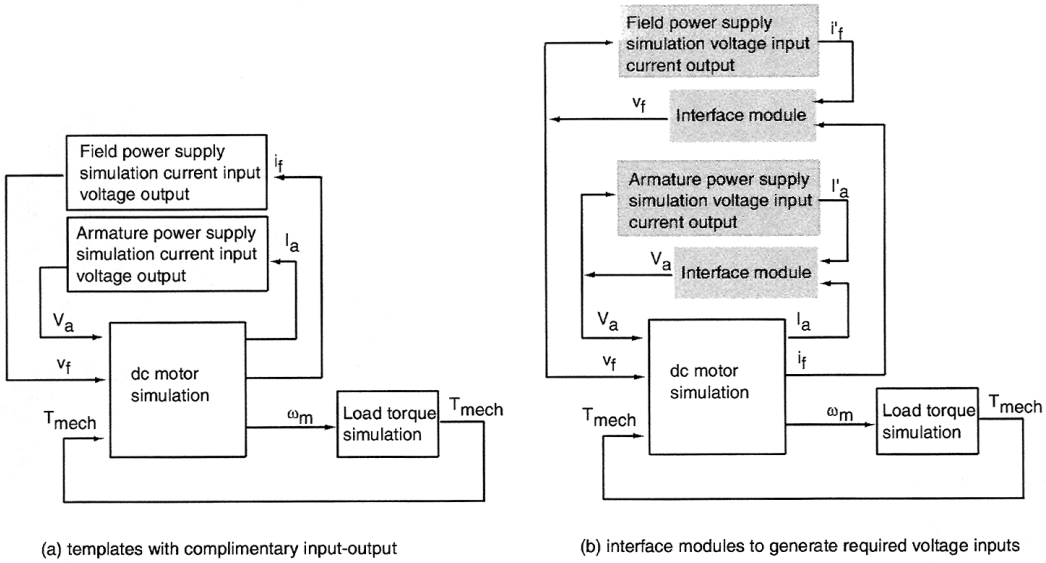


FIGURE 66.31 dc machine simulation flowchart.

Modular Approach

Simulation of larger systems consisting of electric machines can be assembled directly from basic equations of the individual components and their connections. On a higher level of integration, it will be more convenient and advantageous to utilize templates of subsystems to construct the full system in a modular manner. Subsystem templates once verified can be reused with confidence for studies that are within the scope of the models implemented. The tasks of constructing and debugging a simulation using the modular approach can be much easier than building the same simulation from elementary representations.

Proper considerations to matching inputs to outputs of the connected templates are required when using templates. Take, for example, a template of the above dc motor simulation. Such a template will require inputs of V_a, v_f, and T_{mech} to produce outputs of I_a, i_f, flux, and rotor speed. On the mechanical side, the motor template has to be interfaced to the simulation of the mechanical prime mover or load for its remaining input of T_{mech}. In the case of a simple load, the load torque, T_{mech}, could be constant or some simple function of rotor speed, as shown in Fig. 66.32. On the electrical side, the motor template has to be interfaced to the templates of the power supplies to the armature and field windings for its inputs of V_a and v_f. These voltages can come from



(a) templates with complimentary input-output

(b) interface modules to generate required voltage inputs

FIGURE 66.32 Interface for dc motor simulation.

the simulations of the power supply circuits if they provide outputs of these voltages. If not, as in the case where the templates of the power supply circuits also require voltages V_a and v_f as their inputs, the interconnection of the motor and power supply circuits templates will require an interface module with current as input and voltage as output as shown in Fig. 66.32(b).

In practice, the interface module can be of physical or fictitious origin — the latter essentially a convenient but acceptable approximation. Referring again to Fig. 66.32 of the power supply connected to the motor, examples of an interface module of physical origin would be where shunt filtering capacitors or bleeding resistors are actually present at the terminals of the motor windings. Written with current as input and voltage as output, the equations for simulating the shunt capacitor and the shunt bleeding resistor are

$$V = \frac{1}{C}\int i\,dt \quad V = Ri \tag{66.46}$$

where i is the net current flowing into the branch in both cases.

But if the actual system does not have elements with equations that can be written to accept current as input and voltage as output, shunt R and C of fictitious origin can be inserted at the cost of some loss in accuracy to fulfill the necessary interface requirement. For good accuracy, the current in the introduced fictitious branch element should be kept very small relative to the currents of the actual branches by using very large R or very small C. In practice, loop instability in analog computation and numerical stiffness in digital computation will determine the lower bound on how small an error can be attained. In spite of the small error introduced, this technique can be very useful, as evident from trying to use the above dc motor simulation to simulate an open-circuit operation of the motor. While it is possible to reformulate the equations to handle the open-circuit condition, the equations as given with voltage input and current output can be used for the open-circuit condition if one is willing to accept a small inaccuracy of introducing a very large resistor to approximate the open-circuit. On the other hand, the simulation of short-circuit operation with the above **model** can be easily implemented using a V_a of zero.

Mathematical Transformations

Mathematical transformations are used in the analysis and simulation of three-phase systems, mostly to decouple variables. For example, the transformations to symmetrical components or $\alpha\beta 0$ components are used

in the analysis of unbalanced three-phase network analysis. Transformations to decouple variables, to facilitate the solution of difficult equations with time-varying coefficients, or to refer variables from one frame of reference to another are employed in the analysis and simulation of three-phase ac machines.

For example, in the analysis and simulation of a three-phase synchronous machine with a salient pole rotor, transformation of stator quantities onto a frame of reference attached to the asymmetrical rotor results in a much simpler model. In this rotor reference model, the inductances are not dependent on rotor position and, in steady-state operation, the stator voltages and currents are not time varying.

Park's transformation decouples and rotates the stator variables of a synchronous machine onto a dq reference frame that is fixed to the rotor. The positive d-axis of the dq frame is aligned with the magnetic axis of the field winding, that of the positive q-axis is ahead in the direction of rotation or lead the positive d-axis by $\pi/2$. Defined in this manner, the internal excitation voltage given by $E_f = \omega L_{af} i_f$ is in the direction of the positive q-axis. Park's original dq0 transformation [1929] was expressed in terms of the angle, θ_d, between the rotor's d-axis and the axis of the stator's a-phase winding.

The so-called qd0 transformation in more recent publications is Park's transformation expressed in terms of the angle, θ_q, between the rotor's q-axis and the axis of the stator's a-phase winding. The row order of components in the qd0 transformation matrix from top of bottom is q-, d-, and then 0. As evident from Fig. 66.33, the angle θ_q is equal to $\theta_d + \pi/2$. The transformation from abc variables to a qd0 reference frame is accomplished with

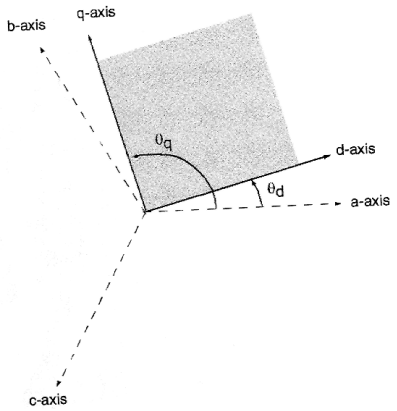


FIGURE 66.33 qd0 transformation.

$$[\mathbf{f}_{qd0}] = [\mathbf{T}_{qd0}(\theta_q)][\mathbf{f}_{abc}] \tag{66.47}$$

where **f** can be voltage, current, or flux, and

$$\left[\mathbf{T}_{qd0}\left(\theta_q\right)\right] = \frac{2}{3}\begin{bmatrix} \cos\theta_q & \cos\left(\theta_q - \frac{2\pi}{3}\right) & \cos\left(\theta_q + \frac{2\pi}{3}\right) \\ \sin\theta_q & \sin\left(\theta_q - \frac{2\pi}{3}\right) & \sin\left(\theta_q + \frac{2\pi}{3}\right) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{66.48}$$

Transforming back from qd0 to abc is accomplished by premultiplying both sides of Eq. (66.47) with the inverse

$$\left[\mathbf{T}_{qd0}\left(\theta_q\right)\right]^{-1} = \begin{bmatrix} \cos\theta_q & \sin\theta_q & 1 \\ \cos\left(\theta_q - \frac{2\pi}{3}\right) & \sin\left(\theta_q - \frac{2\pi}{3}\right) & 1 \\ \cos\left(\theta_q + \frac{2\pi}{3}\right) & \sin\left(\theta_q + \frac{2\pi}{3}\right) & 1 \end{bmatrix} \tag{66.49}$$

The angle θ_q can be determined from

$$\theta_d\left(t\right) = \int_0^t \omega\left(t\right)dt + \theta_q\left(0\right) \quad \text{rad} \tag{66.50}$$

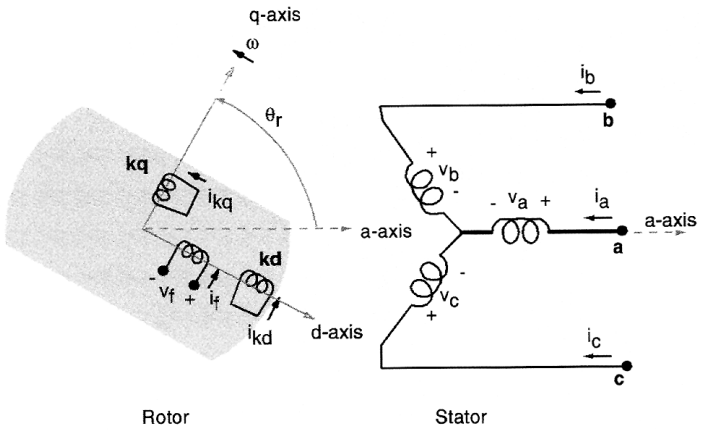


FIGURE 66.34 Circuit representation of idealized synchronous machine.

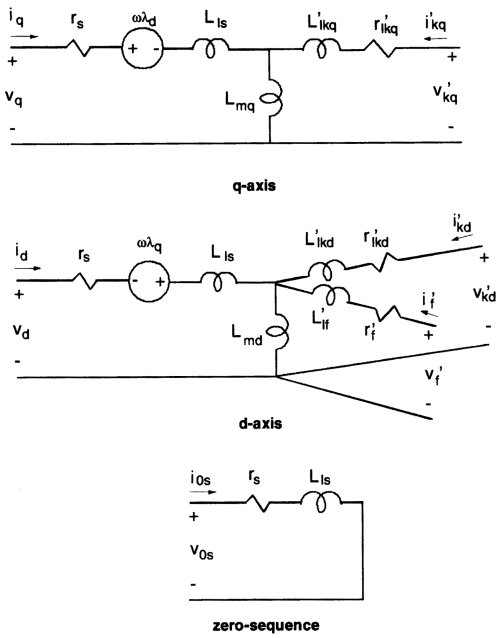


FIGURE 66.35 Equivalent qd0 circuits of synchronous machine.

where $\omega(t)$ is the rotational speed of the qd reference frame and $\theta_q(0)$, the initial value of θ_q at $t = 0$. In the case of the rotor's qd reference frame, $\omega(t)$ is equal to the rotor's speed in electrical radians (per second), that is, $\omega(t) = \omega_r(t)$.

Figure 66.34 shows a circuit representation of an idealized synchronous machine with damper windings, *kd* and *kq*, and field winding, *f*, on the rotor. The equivalent circuit representation and equations of the machine in its own rotor qd0 reference frame and in motor convention are shown in Fig. 66.35 and Table 66.3, respectively.

Base Quantities

Oftentimes, the machine equations are expressed in terms of the flux linkages per second, ψ, and reactances, x, instead of λ and L. These are related simply by the base or rated value of angular frequency, ω_b, that is,

$$\psi = \omega_b \lambda \quad \text{and} \quad x = \omega_b L \tag{66.51}$$

TABLE 66.3 qd0 Model of Synchronous Machine

Voltage equations

$$v_q = r_s i_q + \frac{d\lambda_q}{dt} + \lambda_d \frac{d\theta_r}{dt} \qquad v'_f = r'_f i'_f + \frac{d\lambda_f}{dt}$$
$$v_d = r_s i_d + \frac{d\lambda_d}{dt} - \lambda_q \frac{d\theta_r}{dt} \qquad v'_{kd} = r'_{kd} i'_{kd} + \frac{d\lambda'_{kd}}{dt}$$
$$v_0 = r_s i_0 + \frac{d\lambda_0}{dt} \qquad v'_{kq} = r'_{kq} i'_{kq} + \frac{d\lambda'_{kq}}{dt}$$

Flux linkage equations

$$\lambda_q = L_q i_q + L_{mq} i'_{kq} \qquad \lambda'_f = L_{md} i_d + L_{md} i'_{kd} + L'_{ff} i'_f$$
$$\lambda_d = L_d i_d + L_{md} i'_f + L_{md} i'_{kd} \qquad \lambda'_{kd} = L_{md} i_d + L_{md} i'_f + L'_{kdkd} i'_{kd}$$
$$\lambda_0 = L_{ls} i_0 \qquad \lambda'_{kq} = L_{mq} i_q + L'_{kqkq} i'_{kq}$$

Torque equation

$$T_{em} = \frac{3}{2}\frac{P}{2}\left(\lambda_d i_q - \lambda_q i_d\right) \quad \text{N-m}$$

where $\omega_b = 2\pi f_{rated}$ electrical radian per second, f_{rated} being the rated frequency in Hertz of the machine. When dealing with complex waveforms, it is logical to use peak rather than the rms value as the base value. The base quantities with peak rather rms value of a P-pole, three-phase induction machines with rated line-to-line rms voltage, V_{rated}, and rated volt-ampere, S_{rated}, area as follows:

base voltage, $V_b = \sqrt{2/3}V_{rated}$ base volt-ampere, $S_b = S_{rated}$ base current, $I_b = 2S_b/3V_b$
base impedance, $Z_b = V_b/I_b$ base torque, $T_b = S_b/\omega_{bm}$ base speed, $\omega_{bm} = 2\omega_b/P$

Simulation of Synchronous Machines

Table 66.4 shows the main steps in a simulation of a three-phase synchronous machine with T_{mech}, v_f, and *abc* stator voltages as input. The rotor's speed and angle, $\omega_r(t)$ and δ, are determined by the rotor's equation of motion.

$$T_{em} + T_{mech} - T_{damp} = J\frac{d\omega_{rm}(t)}{dt} \quad \text{N-m} \tag{66.52}$$

The developed torque, T_{em}, is positive for motoring operation and negative for generating operation.

The rotor angle, δ, is defined as the angle of the q_r axis of the rotor with respect to the q_e axis of the synchronously **rotating reference frame,** that is,

$$\begin{aligned}\delta(t) &= \theta_r(t) - \theta_e(t) \quad \text{elect. rad} \\ &= \int_0^t \left(\omega_r(t) - \omega_e\right)dt + \theta_r(0) - \theta_e(0)\end{aligned} \tag{66.53}$$

Since the synchronous speed, ω_e, is a constant,

$$\frac{d\left(\omega_r(t) - \omega_e\right)}{dt} = \frac{d\omega_r(t)}{dt} \tag{66.54}$$

TABLE 66.4 Simulation of Synchronous Machine

Transform input stator *abc* voltages to the qd reference frame attached to the rotor using $[\mathbf{v}_{qd0}] = [\mathbf{T}_{qd0}(\theta_r)][\mathbf{v}_{abc}]$ where $\theta_r(t) = \int_0^t \omega_r(t)dt + \theta_r(0)$. The currents or flux linkages of the cut set of three inductors in both the q- and d-axis circuits of Fig. 66.35, are not independent. Using the winding flux linkages per second as states, the mutual flux linkages per second are expressed as

$$\psi_{mq} = \omega_b L_{mq}\left(i_q + i'_{kq}\right) = x_{MQ}\left(\frac{\psi_q}{x_{ls}} + \frac{\psi'_{kq}}{x'_{lkq}}\right) \quad \psi_{md} = \omega_b L_{md}\left(i_d + i'_{kd} + i'_f\right) = x_{MD}\left(\frac{\psi_d}{x_{ls}} + \frac{\psi'_{kd}}{x'_{lkd}} + \frac{\psi'_f}{x'_{lf}}\right)$$

where

$$\frac{1}{x_{MQ}} = \frac{1}{x_{mq}} + \frac{1}{x'_{lkq}} + \frac{1}{x_{ls}} \quad \frac{1}{x_{MD}} = \frac{1}{x_{md}} + \frac{1}{x'_{lkd}} + \frac{1}{x'_{lf}} + \frac{1}{x_{ls}}$$

Solve winding flux linkages using the following integral form of the winding voltage equations:

$$\psi_q = \omega_b \int \left(v_q - \frac{\omega_r}{\omega_b}\psi_d + \frac{r_s}{x_{ls}}\left(\psi_{mq} - \psi_q\right)\right)dt \qquad \psi'_{kq} = \frac{\omega_b r'_{kq}}{x'_{lkq}} \int \left(\psi_{mq} - \psi'_{kq}\right)dt$$

$$\psi_d = \omega_b \int \left(v_d + \frac{\omega_r}{\omega_b}\psi_q + \frac{r_s}{x_{ls}}\left(\psi_{md} - \psi_d\right)\right)dt \qquad \psi'_{kd} = \frac{\omega_b r'_{kd}}{x'_{lkd}} \int \left(\psi_{md} - \psi'_{kd}\right)dt$$

$$\psi_0 = \omega_b \int \left(v_0 - \frac{r_s}{x_{ls}}\psi_0\right)dt \qquad \psi'_f = \frac{\omega_b r'_f}{x_{md}} \int \left(E_f + \frac{x_{md}}{x'_{lf}}\left(\psi_{md} - \psi'_f\right)\right)dt$$

where $E_f = x_{md}\frac{v'_f}{\tau_f}$, and

$$\psi_q = x_{ls} i_q + \psi_{mq} \qquad \psi_d = x_{ls} i_d + \psi_{md} \qquad \psi_0 = x_{ls} i_0$$

$$\psi'_f = x'_{lf} i'_f + \psi_{md} \qquad \psi'_{kd} = x'_{lkd} i'_{kd} + \psi_{md} \qquad \psi'_{kq} = x'_{lkq} i'_{kq} + \psi_{mq}$$

Determine qd0 winding currents from winding flux linkages.

$$i_q = \frac{\psi_q - \psi_{mq}}{x_{ls}} \qquad i_d = \frac{\psi_d - \psi_{md}}{x_{ls}} \qquad i_0 = \frac{\psi_0}{x_{ls}}$$

$$i'_{kd} = \frac{\psi'_{kd} - \psi_{md}}{x'_{lkd}} \qquad i'_{kq} = \frac{\psi'_{kq} - \psi_{mq}}{x'_{lkq}} \qquad i'_f = \frac{\psi'_f - \psi_{md}}{x'_{lf}}$$

Transform qd0 currents to *abc* using $[\mathbf{i}_{abc}] = [\mathbf{T}_{qd0}^{-1}(\theta_r)][\mathbf{i}_{qd0}]$.

Using $(2/P)\,\omega_r(t)$ in place of $\omega_{rm}(t)$ and Eq. (66.54) to replace $d\omega_r(t)/dt$, Eq. (66.52) can be rewritten in terms of the slip speed:

$$\omega_r(t) - \omega_e = \frac{P}{2J}\int_0^t \left(T_{em} + T_{mech} - T_{damp}\right)dt \quad \text{elect. rad/s} \tag{66.55}$$

The angles $\theta_r(t)$ and $\theta_e(t)$ are the respective angles of the q_r and q_e axes of the rotor and synchronously **rotating reference frames** measured with respect to the stationary axis of the a-phase stator winding. Note that δ is the angle between the q_r axis of the rotor and the q_e axis of the reference synchronously rotating frame. For multi-machine systems, the rotor angles of the machines could all be referred to a common synchronously rotating reference frame at some bus or to the q_r axis of the rotor of a chosen reference machine.

A flowchart showing the main blocks for the above simulation is given in Fig. 66.36. As shown, the input voltages and output currents are in abc phase quantities. For some studies, the representation of the supply

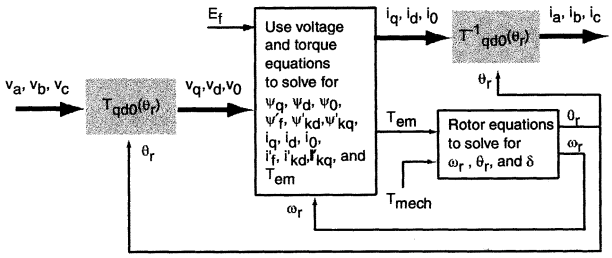


FIGURE 66.36 Block diagram of synchronous machine simulation.

network connected to the machine may not be in phase variables. For example, in linearized analysis, and also in transient stability of power systems, the network representation is usually in a synchronously **rotating reference frame.** In linearized analysis, the small-signal representation of the system is obtained by making small perturbations about an operating point. When the machines are in their respective rotor qd0 reference frames and the power network is in a synchronously rotating qd0 reference frame, the qd0 variables of the network and machines are in steady-state; thus linearized analysis about an operating point can be performed.

In transient stability, the main interest is the stability of the system after some large disturbances. The **models** employed are to portray the transient behavior of the power flows in the network and the electromechanical response of the machines. When dealing with large networks, the fast electromagnetic transients of the network are usually ignored and a static network representation is used. At each new time step of the **dynamic simulation**, an update of the network condition can be obtained by solving the phasor equations of the static network along with power or current injections from the machines. Because the phasor quantities of the network can be expressed as qd components of a synchronously rotating qd0 reference [Ong, 1998], the exchange of voltage variables between network and machine at the bus will require a rotational transformation given below

$$\begin{bmatrix} v_q \\ v_d \\ v_0 \end{bmatrix} = \begin{bmatrix} \cos\delta & -\sin\delta & 0 \\ \sin\delta & \cos\delta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_q^e \\ v_d^e \\ v_0 \end{bmatrix} \tag{66.56}$$

where δ is the rotor angle of the q_r of the machines qd0 rotor reference frame measured with respect to the q_e axis of the network's synchronously rotating reference frame. The above transformation is also applicable to the exchange of current variables between network and machine.

Other synchronous machine models, besides that given in Table 66.3, are used in power system analysis. Typically, when the network is large and the phenomenon of interest is somewhat localized in nature, machines further away from the action can be represented by simpler models to save computation time. On the other hand, certain phenomena may require an even more sophisticated model than that given in Table 66.3. Canay [1993] described refinements in both the rotor circuit representation and the method of parameter determination to obtain a closer fit of the rotor variables. For studying shaft torsion, the damper circuit representation should not be ignored. In transient stability studies, machines beyond the first two neighborhoods of the disturbance can be represented by progressively simpler models with distance from the disturbance. Ignoring just the $p\psi_q$ and $p\psi_d$ terms and also setting $\omega = \omega_e$ in the stator equations will yield a so-called subtransient model of two orders less than the above given model. Further simplification by setting $p\psi'_{kq}$ and $p\psi'_{kd}$ to zero or omitting the damper winding equations will yield a so-called transient model of another two orders less. Finally, setting $p\psi'_f$ to zero and holding field flux linkage constant yields the constant field flux linkage model.

Significant savings in computing time can also be made by neglecting the subtransient and transient saliency of the machine. When rotor saliency is ignored, the effective stator impedances along the rotor's q_r and d_r axes are equal. In other words, the stator impedance in the synchronously rotating reference frame of the network will not be a function of rotor angle. Because its value need not be updated with the rotor angle at each time step of the dynamic simulation, the constant stator impedance of this model can be absorbed into the network's admittance or impedance representation.

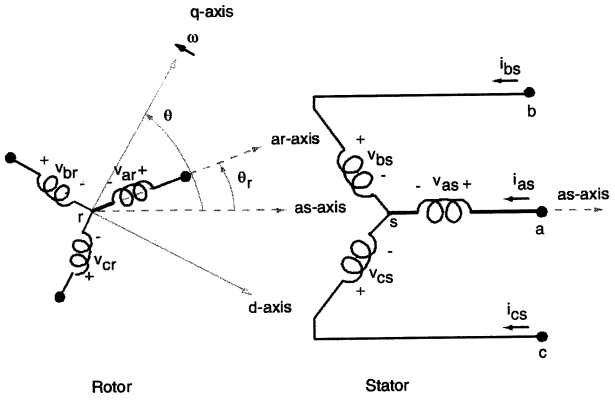


FIGURE 66.37 Circuit representation of induction machine.

Three-Phase Induction Machines

Figure 66.37 shows a circuit representation of a symmetrical three-phase induction machine with uniform airgap. The axes of the qd0 reference frames are assumed to be rotating at an arbitrary angular speed of ω. The angles $\theta(t)$ and $\theta_r(t)$, in electrical radians, can be determined from

$$\theta(t) = \int_0^t \omega(t)dt + \theta(0) \quad \theta_r(t) = \int_0^t \omega_r dt + \theta_r(0) \tag{66.57}$$

where $\theta(0)$ and $\theta_r(0)$ are their respective initial values at time $t = 0$.

As before, the voltage equations of the stator and rotor windings can be written using the coupled circuit approach. Corresponding voltage equations in the arbitrary qd0 reference frame can be obtained by applying the transformation $T_{qd0}(\theta)$ to the stator variables and the transformation $T_{qd0}(\theta - \theta_r)$ to the rotor variables. The equations of a symmetrical induction machine in the arbitrary reference frame in terms of the flux linkages per second and reactances are summarized in Table 66.5.

Seldom is there a need to simulate an induction machine in the arbitrary **rotating reference frame.** Induction machine loads are often simulated on the network's synchronously rotating reference frame in power system studies. However, in transient studies of adjustable speed drives, it is usually more convenient to simulate the induction machine and its converter on a stationary reference frame. Equations of the machine in the stationary and synchronously rotating reference frames can be obtained by setting the speed of the arbitrary reference frame, ω, to zero and ω_e, respectively.

Often the stator windings are connected to the supply by a three-wire connection, as shown in Fig. 66.38. With a three-wire connection, the stator zero-sequence current, i_{0s}, or $(i_{as} + i_{bs} + i_{cs})/3$, is zero by physical constraint, irrespective of whether the phase currents are balanced or not. The phase currents could be unbalanced, as in single-phasing operation. The stator neutral is free-floating. Its voltage, v_{sg}, measured with respect to some ground point g, need not be zero. Where the applied voltages are non-sinusoidal, as in the case when the supply is from a bridge inverter, v_{sg} is not zero.

In general, the input stator phase voltages, v_{ag}, v_{bg}, and v_{cg}, for the simulation of the induction machine can be established from the following relationships:

$$v_{as} = v_{ag} - v_{sg} \quad v_{bs} = v_{bg} - v_{sg} \quad v_{cs} = v_{cg} - v_{sg} \tag{66.58}$$

When point *s* is solidly connected to point, *g*, v_{sg} will be zero. Otherwise, if R_{sg} and L_{sg} are the resistance and inductance of the connection between points *s* and *g*, v_{sg} can be determined from

TABLE 66.5 Model of induction machine in arbitrary qdo

Voltage equations

$$v_{qs} = \frac{p}{\omega_b}\psi_{qs} + \frac{\omega}{\omega_b}\psi_{ds} + r_s i_{qs} \qquad v'_{qr} = \frac{p}{\omega_b}\psi'_{qr} + \left(\frac{\omega - \omega_r}{\omega_b}\right)\psi'_{dr} + r'_r i'_{qr}$$

$$v_{ds} = \frac{p}{\omega_b}\psi_{ds} - \frac{\omega}{\omega_b}\psi_{qs} + r_s i_{ds} \qquad v'_{dr} = \frac{p}{\omega_b}\psi'_{dr} - \left(\frac{\omega - \omega_r}{\omega_b}\right)\psi'_{qr} + r'_r i'_{dr}$$

$$v_{0s} = \frac{p}{\omega_b}\psi_{0s} + r_s i_{0s} \qquad v'_{0r} = \frac{p}{\omega_b}\psi'_{0r} + r'_r i'_{0r}$$

Flux linkage equations

$$\begin{bmatrix} \psi_{qs} \\ \psi_{ds} \\ \psi_{0s} \\ \psi'_{qr} \\ \psi'_{dr} \\ \psi'_{0r} \end{bmatrix} = \begin{bmatrix} x_{ls} + x_m & 0 & 0 & x_m & 0 & 0 \\ 0 & x_{ls} + x_m & 0 & 0 & x_m & 0 \\ 0 & 0 & x_{ls} & 0 & 0 & 0 \\ x_m & 0 & 0 & x'_{lr} + x_m & 0 & 0 \\ 0 & x_m & 0 & 0 & x'_{lr} + x_m & 0 \\ 0 & 0 & 0 & 0 & 0 & x'_{lr} \end{bmatrix} \begin{bmatrix} i_{qs} \\ i_{ds} \\ i_{0s} \\ i'_{qr} \\ i'_{dr} \\ i'_{0r} \end{bmatrix}$$

Torque equation

$$T_{em} = \frac{3}{2}\frac{P}{2\omega_r}\left[\frac{\omega}{\omega_b}\left(\psi_{ds} i_{qs} - \psi_{qs} i_{ds}\right) + \frac{\omega - \omega_r}{\omega_b}\left(\psi'_{dr} i'_{qr} - \psi'_{qr} i'_{dr}\right)\right] \quad \text{N-m}$$

$$= \frac{3}{2}\frac{P}{2\omega_b}\left(\psi'_{qr} i'_{dr} - \psi'_{dr} i'_{qr}\right) = \frac{3}{2}\frac{P}{2\omega_b}\left(\psi_{ds} i_{qs} - \psi_{qs} i_{ds}\right) = \frac{3}{2}\frac{P}{2\omega_b} x_m\left(i'_{dr} i_{qs} - i'_{qr} i_{ds}\right)$$

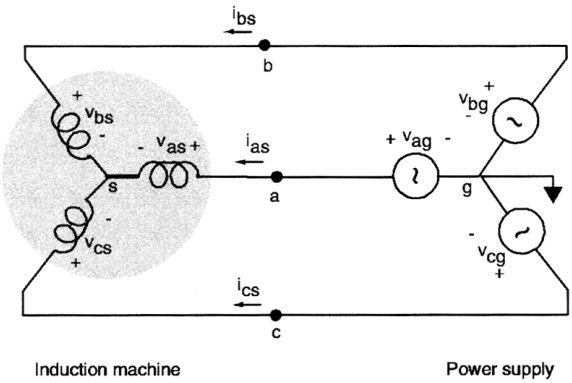


FIGURE 66.38 Three-wire power supply connection.

$$v_{sg} = R_{sg}\left(i_{as} + i_{bs} + {}_{cs}\right) + L_{sg}\frac{d}{dt}\left(i_{as} + i_{bs} + i_{cs}\right) = 3\left(R_{sg} + L_{sg}\frac{d}{dt}\right)i_{0s} \tag{66.59}$$

Where the stator windings' neutral is free-floating, v_{sg} can be determined from an open-circuit approximation of the form shown in Eq. (66.46).

Defining Terms

Model of an electric machine: Differential algebraic equations describing the dynamic behavior of the electric machine.

Dynamic simulation: Setting up of a model capable of portraying certain dynamic behavior of the real device and performing experiments on the model.

Rotating reference frame: A rotating qd plane. For example, a synchronously rotating reference frame is a qd plane that is rotating at synchronous speed as defined by the fundamental excitation frequency.

References

Canay, I. M. (1993) Modelling of Alternating-Current Machines Having Multiple Rotor Circuits, *IEEE Trans. on Energy Conversion*, Vol. 8, No. 2, June 1993, pp. 280–296.

Demerdash, N. A. O. and Alhamadi M. A. (1995) Three-Dimensional Finite Element Analysis of Permanent Magnet Brushless DC Motor Drives – Status of the State of Art, *IEEE Trans. on Industrial Electronics*, Vol. 43, No. 2, April 1995, pp. 268–275.

Ong, C. M. (1998) *Dynamic Simulation of Electric Machinery*, Prentice-Hall PTR, New Jersey.

Park, R. H. (1929) Two-Reaction Theory of Synchronous Machines Generalized Method of Analysis. Part I, *A.I.E.E. Transactions*, Vol. 48, 1929, pp. 716–727.

Preston, T. W., Reece, A. B. J., and Sangha, P. C. (1988) Induction Motor Analysis by Time-Stepping Techniques, *IEEE Trans. on Magnetics*, Vol. 24, No. 1, Jan. 1998, pp. 471–474.

Rahman, M. A. and Little, T. A. (1984) Dynamic Performance Analysis of Permanent Magnet Synchronous Magnet Motors, *IEEE Trans. on Power Apparatus and Systems*, Vol. 103, No. 6, June 1984, pp. 1277–1282.

Salon, S. J. (1995) *Finite Element Analysis of Electrical Machines*, Kluwer Academic Publishers, Boston.

Shen, J. (1995) *Computational Electromagnetics Using Boundary Element: Advances in Modeling Eddy Currents*, Computational Mechanics Publication, Southampton, UK.

Further Information

The above chapter section has briefly described some of the techniques of the coupled-circuit approach and qd0 transformation in modeling, and the treatment of interface and floating neutral conditions in implementing a simulation. For more information on modeling and implementation of machine simulations, see [Ong, 1998].

Some techniques of modeling permanent magnet machines are described in [Rahman and Little, 1984; Ong, 1998].

Problems concerning effects of local saturation, anisotropic magnetic properties, and eddy-currents in machines require detailed modeling of the field region. Two- and three-dimensional models of the field region can be solved using finite-element [Salon, 1995] and boundary-element [Shen, 1995] techniques. Although the field models are not as amenable as the circuit models for use in large system studies, they have been successfully integrated with lumped circuit element models in dynamic simulations [Demerdash and Alhamadi, 1995; Preston et al., 1988].

Stanton, K.N., Giri, J.C., Bose, A.J. “Energy Management”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

67

Energy Management

K. Neil Stanton
Stanton Associates

Jay C. Giri
Cegelec ESCA Corporation

Anjan Bose
Washington State University

67.1 Introduction
67.2 Power System Data Acquisition and Control
67.3 Automatic Generation Control
Load Frequency Control • Economic Dispatch • Reserve Monitoring • Interchange Transaction Scheduling
67.4 Load Management
67.5 Energy Management
67.6 Security Control
67.7 Operator Training Simulator
Energy Control System • Power System Dynamic Simulation • Instructional System

67.1 Introduction

Energy management is the process of monitoring, coordinating, and controlling the generation, transmission, and distribution of electrical energy. The physical plant to be managed includes generating plants that produce energy fed through transformers to the high-voltage transmission network (grid), interconnecting generating plants and load centers. Transmission lines terminate at substations that perform switching, voltage transformation, measurement, and control. Substations at load centers transform to subtransmission and distribution levels. These lower-voltage circuits typically operate radially, i.e., no normally closed paths between substations through subtransmission or distribution circuits. (Underground cable networks in large cities are an exception.)

Since transmission systems provide negligible energy storage, supply and demand must be balanced by either generation or load. Production is controlled by turbine governors at generating plants, and automatic generation control is performed by control center computers remote from generating plants. Load management, sometimes called demand-side management, extends remote supervision and control to subtransmission and distribution circuits, including control of residential, commercial, and industrial loads.

Events such as lightning strikes, short circuits, equipment failure, or accidents may cause a system fault. Protective relays actuate rapid, local control through operation of circuit breakers before operators can respond. The goal is to maximize safety, minimize damage, and continue to supply load with the least inconvenience to customers. Data acquisition provides operators and computer control systems with status and measurement information needed to supervise overall operations. **Security** control analyzes the consequences of faults to establish operating conditions that are both robust and economical.

Energy management is performed at control centers (see Fig. 67.1), typically called system control centers, by computer systems called *energy management systems* (EMS). Data acquisition and remote control is performed by computer systems called *supervisory control and data acquisition* (SCADA) systems. These latter systems may be installed at a variety of sites including system control centers. An EMS typically includes a SCADA "front-end" through which it communicates with generating plants, substations, and other remote devices.

Figure 67.2 illustrates the **applications** layer of modern EMS as well as the underlying layers on which it is built: the operating system, a database manager, and a utilities/services layer.

FIGURE 67.1 Central dispatch operation arena of Entergy Corporation's Beaumont Control Center (Beaumont, Texas) which includes a modern EMS.

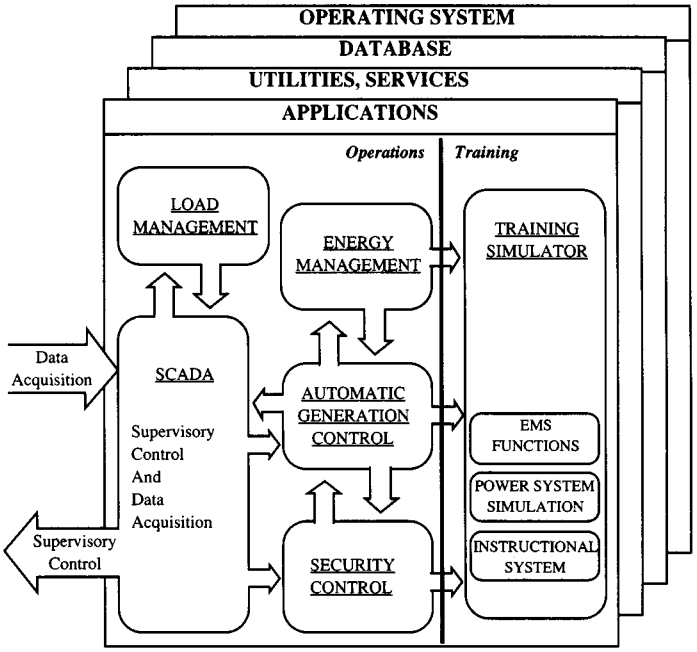


FIGURE 67.2 Layers of a modern EMS.

67.2 Power System Data Acquisition and Control

A SCADA system consists of a master station that communicates with **remote terminal units** (RTUs) for the purpose of allowing operators to observe and control physical plants. Generating plants and transmission substations certainly justify RTUs, and their installation is becoming more common in distribution substations as costs decrease. RTUs transmit device status and measurements to, and receive control commands and setpoint data from, the master station. Communication is generally via dedicated circuits operating in the range of 600 to 4800 bits/s with the RTU responding to periodic requests initiated from the master station (polling) every 2 to 10 s, depending on the criticality of the data.

The traditional functions of SCADA systems are summarized:

- Data acquisition: Provides telemetered measurements and status information to operator.
- Supervisory control: Allows operator to remotely control devices, e.g., open and close circuit breakers. A "select before operate" procedure is used for greater safety.
- Tagging: Identifies a device as subject to specific operating restrictions and prevents unauthorized operation.

- Alarms: Informs operator of unplanned events and undesirable operating conditions. Alarms are sorted by criticality, area of responsibility, and chronology. Acknowledgment may be required.
- Logging: Logs all operator entry, all alarms, and selected information.
- Load shed: Provides both automatic and operator-initiated tripping of load in response to system emergencies.
- Trending: Plots measurements on selected time scales.

Since the master station is critical to power system operations, its functions are generally distributed among several computer systems depending on specific design. A dual computer system configured in primary and standby modes is most common. SCADA functions are listed below without stating which computer has specific responsibility.

- Manage communication circuit configuration
- Downline load RTU files
- Maintain scan tables and perform polling
- Check and correct message errors
- Convert to engineering units
- Detect status and measurement changes
- Monitor abnormal and out-of-limit conditions
- Log and time-tag sequence of events
- Detect and annunciate alarms
- Respond to operator requests to:
  - Display information
  - Enter data
  - Execute control action
  - Acknowledge alarms
- Transmit control action to RTUs
- Inhibit unauthorized actions
- Maintain historical files
- Log events and prepare reports
- Perform load shedding

67.3 Automatic Generation Control

Automatic generation control (AGC) consists of two major and several minor functions that operate on-line in real time to adjust the generation against load at minimum cost. The major functions are load frequency control and economic **dispatch,** each of which is described below. The minor functions are reserve monitoring, which assures enough reserve on the system, **interchange** scheduling, which initiates and completes scheduled interchanges, and other similar monitoring and recording functions.

Load Frequency Control

Load frequency control (LFC) has to achieve three primary objectives which are stated below in priority order:

1. To maintain frequency at the scheduled value
2. To maintain net power interchanges with neighboring control areas at the scheduled values
3. To maintain power allocation among units at economically desired values

The first and second objectives are met by monitoring an error signal, called *area control error* (ACE), which is a combination of net interchange error and frequency error and represents the power imbalance between

generation and load at any instant. This ACE must be filtered or smoothed such that excessive and random changes in ACE are not translated into control action. Since these excessive changes are different for different systems, the filter parameters have to be tuned specifically for each control area. The filtered ACE is then used to obtain the proportional plus integral control signal. This control signal is modified by limiters, deadbands, and gain constants that are tuned to the particular system. This control signal is then divided among the generating units under control by using participation factors to obtain *unit control errors* (UCE).

These participation factors may be proportional to the inverse of the second derivative of the cost of unit generation so that the units would be loaded according to their costs, thus meeting the third objective. However, cost may not be the only consideration because the different units may have different response rates and it may be necessary to move the faster generators more to obtain an acceptable response. The UCEs are then sent to the various units under control and the generating units monitored to see that the corrections take place. This control action is repeated every 2 to 6 s.

In spite of the integral control, errors in frequency and net interchange do tend to accumulate over time. These time errors and accumulated interchange errors have to be corrected by adjusting the controller settings according to procedures agreed upon by the whole interconnection. These accumulated errors as well as ACE serve as performance measures for LFC.

The main philosophy in the design of LFC is that each system should follow its own load very closely during normal operation, while during emergencies each system should contribute according to its relative size in the interconnection without regard to the locality of the emergency. Thus, the most important factor in obtaining good control of a system is its inherent capability of following its own load. This is guaranteed if the system has adequate regulation margin as well as adequate response capability. Systems that have mainly thermal generation often have difficulty in keeping up with the load because of the slow response of the units.

The design of the controller itself is an important factor, and proper tuning of the controller parameters is needed to obtain "good" control without "excessive" movement of units. Tuning is system-specific, and although system simulations are often used as aids, most of the parameter adjustments are made in the field using heuristic procedures.

Economic Dispatch

Since all the generating units that are on-line have different costs of generation, it is necessary to find the generation levels of each of these units that would meet the load at the minimum cost. This has to take into account the fact that the cost of generation in one generator is not proportional to its generation level but is a nonlinear function of it. In addition, since the system is geographically spread out, the transmission losses are dependent on the generation pattern and must be considered in obtaining the optimum pattern.

Certain other factors have to be considered when obtaining the optimum generation pattern. One is that the generation pattern provide adequate reserve margins. This is often done by constraining the generation level to a lower boundary than the generating capability. A more difficult set of constraints to consider are the transmission limits. Under certain real-time conditions it is possible that the most economic pattern may not be feasible because of unacceptable line flows or voltage conditions. The present-day economic dispatch (ED) algorithm cannot handle these security constraints. However, alternative methods based on optimal power flows have been suggested but have not yet been used for real-time dispatch.

The minimum cost dispatch occurs when the incremental cost of all the generators is equal. The cost functions of the generators are nonlinear and discontinuous. For the equal marginal cost algorithm to work it is necessary for them to be convex. These incremental cost curves are often represented as monotonically increasing piecewise-linear functions. A binary search for the optimal marginal cost is conducted by summing all the generation at a certain marginal cost and comparing it with the total power demand. If the demand is higher, a higher marginal cost is needed, and vice versa. This algorithm produces the ideal setpoints for all the generators for that particular demand, and this calculation is done every few minutes as the demand changes.

The losses in the power system are a function of the generation pattern, and they are taken into account by multiplying the generator incremental costs by the appropriate penalty factors. The penalty factor for each generator is a reflection of the sensitivity of that generator to system losses, and these sensitivities can be obtained from the transmission loss factors (Section 67.6).

This ED algorithm generally applies to only thermal generation units that have cost characteristics of the type discussed here. The hydro units have to be dispatched with different considerations. Although there is no cost for the water, the amount of water available is limited over a period, and the displacement of fossil fuel by this water determines its worth. Thus, if the water usage limitation over a period is known, say from a previously computed hydro optimization, the water worth can be used to dispatch the hydro units.

LFC and the ED functions both operate automatically in real time but with vastly different time periods. Both adjust generation levels, but LFC does it every few seconds to follow the load variation, while ED does it every few minutes to assure minimal cost. Conflicting control action is avoided by coordinating the control errors. If the unit control errors from LFC and ED are in the same direction, there is no conflict. Otherwise, a logic is set to either follow load (permissive control) or follow economics (mandatory control).

Reserve Monitoring

Maintaining enough reserve capacity is required in case generation is lost. Explicit formulas are followed to determine the spinning (already synchronized) and ready (10 min) reserves required. The availability can be assured by the operator manually, or, as mentioned previously, the ED can also reduce the upper dispatchable limits of the generators to keep such generation available.

Interchange Transaction Scheduling

The contractual exchange of power between utilities has to be taken into account by the LFC and ED functions. This is done by calculating the net interchange (sum of all the buy and sale agreements) and adding this to the generation needed in both the LFC and ED. Since most interchanges begin and end on the hour, the net interchange is ramped from one level to the new over a 10- or 20-min period straddling the hour. The programs achieve this automatically from the list of scheduled transactions.

67.4 Load Management

SCADA, with its relatively expensive RTUs installed at distribution substations, can provide status and measurements for distribution feeders at the substation. Distribution automation equipment is now available to measure and control at locations dispersed along distribution circuits. This equipment can monitor sectionalizing devices (switches, interruptors, fuses), operate switches for circuit reconfiguration, control voltage, read customers' meters, implement time-dependent pricing (on-peak, off-peak rates), and switch customer equipment to manage load. This equipment requires significantly increased functionality at distribution control centers.

Distribution control center functionality varies widely from company to company, and the following list is evolving rapidly.

- Data acquisition: Acquires data and gives the operator control over specific devices in the field. Includes data processing, quality checking, and storage.
- Feeder switch control: Provides remote control of feeder switches.
- Tagging and alarms: Provides features similar to SCADA.
- Diagrams and maps: Retrieves and displays distribution maps and drawings. Supports device selection from these displays. Overlays telemetered and operator-entered data on displays.
- Preparation of switching orders: Provides templates and information to facilitate preparation of instructions necessary to disconnect, isolate, reconnect, and reenergize equipment.
- Switching instructions: Guides operator through execution of previously prepared switching orders.
- Trouble analysis: Correlates data sources to assess scope of trouble reports and possible dispatch of work crews.
- Fault location: Analyzes available information to determine scope and location of fault.
- Service restoration: Determines the combination of remote control actions which will maximize restoration of service. Assists operator to dispatch work crews.

- Circuit continuity analysis: Analyzes circuit topology and device status to show electrically connected circuit segments (either energized or deenergized).
- Power factor and voltage control: Combines substation and feeder data with predetermined operating parameters to control distribution circuit power factor and voltage levels.
- Electrical circuit analysis: Performs circuit analysis, single-phase or three-phase, balanced or unbalanced.
- Load management: Controls customer loads directly through appliance switching (e.g., water heaters) and indirectly through voltage control.
- Meter reading: Reads customers' meters for billing, peak demand studies, time of use tariffs. Provides remote connect/disconnect.

67.5 Energy Management

Generation control and ED minimize the current cost of energy production and transmission within the range of available controls. Energy management is a supervisory layer responsible for economically scheduling production and transmission on a global basis and over time intervals consistent with cost optimization. For example, water stored in reservoirs of hydro plants is a resource that may be more valuable in the future and should, therefore, not be used now even though the cost of hydro energy is currently lower than thermal generation. The global consideration arises from the ability to buy and sell energy through the interconnected power system; it may be more economical to buy than to produce from plants under direct control. Energy accounting processes transaction information and energy measurements recorded during actual operation as the basis of payment for energy sales and purchases.

Energy management includes the following functions:

- System load forecast: Forecasts system energy demand each hour for a specified forecast period of 1 to 7 days.
- Unit commitment: Determines start-up and shut-down times for most economical operation of thermal generating units for each hour of a specified period of 1 to 7 days.
- Fuel scheduling: Determines the most economical choice of fuel consistent with plant requirements, fuel purchase contracts, and stockpiled fuel.
- Hydro-thermal scheduling: Determines the optimum schedule of thermal and hydro energy production for each hour of a study period up to 7 days while ensuring that hydro and thermal constraints are not violated.
- Transaction evaluation: Determines the optimal incremental and production costs for exchange (purchase and sale) of additional blocks of energy with neighboring companies.
- Transmission loss minimization: Recommends controller actions to be taken in order to minimize overall power system network losses.
- Security constrained dispatch: Determines optimal outputs of generating units to minimize production cost while ensuring that a network security constraint is not violated.
- Production cost calculation: Calculates actual and economical production costs for each generating unit on an hourly basis.

67.6 Security Control

Power systems are designed to survive all probable contingencies. A contingency is defined as an event that causes one or more important components such as transmission lines, generators, and transformers to be unexpectedly removed from service. Survival means the system stabilizes and continues to operate at acceptable voltage and frequency levels without loss of load. Operations must deal with a vast number of possible conditions experienced by the system, many of which are not anticipated in planning. Instead of dealing with the impossible task of analyzing all possible system states, security control starts with a specific state: the current state if executing the real-time network sequence; a postulated state if executing a study sequence. Sequence means sequential execution of programs that perform the following steps:

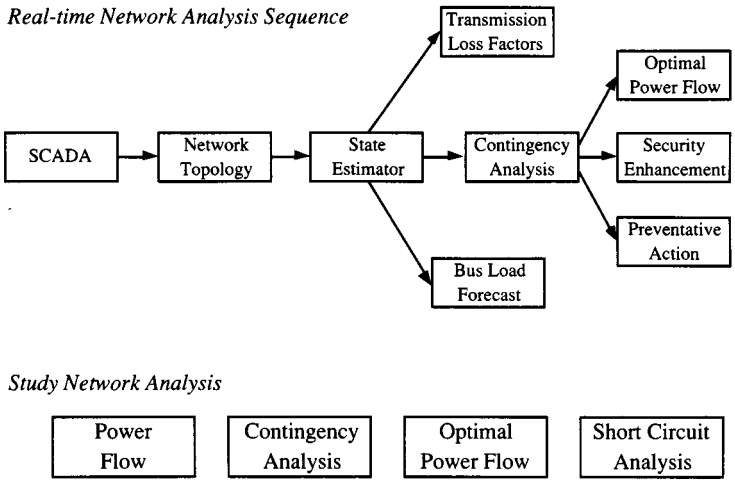


FIGURE 67.3 Real-time and study network analysis sequences.

1. Determine the state of the system based on either current or postulated conditions.
2. Process a list of contingencies to determine the consequences of each contingency on the system in its specified state.
3. Determine preventive or corrective action for those contingencies which represent unacceptable risk.

Real-time and study network analysis sequences are diagramed in Fig. 67.3.

Security control requires topological processing to build network models and uses large-scale ac network analysis to determine system conditions. The required applications are grouped as a network subsystem which typically includes the following functions:

- Topology processor: Processes real-time status measurements to determine an electrical connectivity (bus) model of the power system network.
- State estimator: Uses real-time status and analog measurements to determine the "best" estimate of the state of the power system. It uses a redundant set of measurements; calculates voltages, phase angles, and power flows for all components in the system; and reports overload conditions.
- Power flow: Determines the steady-state conditions of the power system network for a specified generation and load pattern. Calculates voltages, phase angles, and flows across the entire system.
- Contingency analysis: Assesses the impact of a set of contingencies on the state of the power system and identifies potentially harmful contingencies that cause operating limit violations.
- Optimal power flow: Recommends controller actions to optimize a specified objective function (such as system operating cost or losses) subject to a set of power system operating constraints.
- Security enhancement: Recommends corrective control actions to be taken to alleviate an existing or potential overload in the system while ensuring minimal operational cost.
- Preventive action: Recommends control actions to be taken in a "preventive" mode before a contingency occurs to preclude an overload situation if the contingency were to occur.
- Bus load forecasting: Uses real-time measurements to adaptively forecast loads for the electrical connectivity (bus) model of the power system network.
- Transmission loss factors: Determines incremental loss sensitivities for generating units; calculates the impact on losses if the output of a unit were to be increased by 1 MW.
- Short-circuit analysis: Determines fault currents for single-phase and three-phase faults for fault locations across the entire power system network.

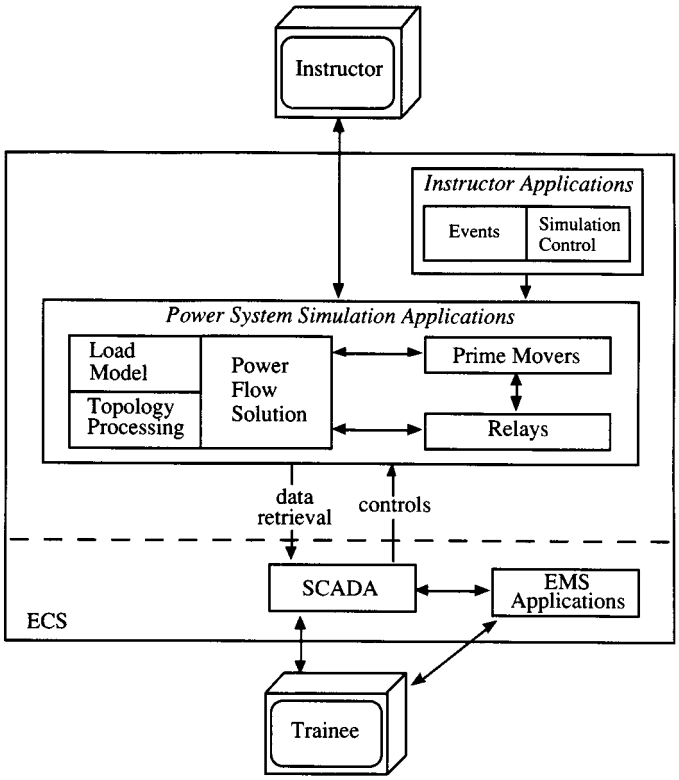


FIGURE 67.4 OTS block diagram.

67.7 Operator Training Simulator

Training simulators were originally created as generic systems for introducing operators to the electrical and dynamic behavior of power systems. Today, they model actual power systems with reasonable fidelity and are integrated with EMS to provide a realistic environment for operators and dispatchers to practice normal, every-day operating tasks and procedures as well as experience emergency operating situations. The various training activities can be safely and conveniently practiced with the simulator responding in a manner similar to the actual power system.

An operator training simulator (OTS) can be used in an investigatory manner to recreate past actual operational scenarios and to formulate system restoration procedures. Scenarios can be created, saved, and reused. The OTS can be used to evaluate the functionality and performance of new real-time EMS functions and also for tuning AGC in an off-line, secure environment.

The OTS has three main subsystems (Fig. 67.4).

Energy Control System

The energy control system (ECS) emulates normal EMS functions and is the only part of the OTS with which the trainee interacts. It consists of the supervisory control and data acquisition (SCADA) system, generation control system, and all other EMS functions.

Power System Dynamic Simulation

This subsystem simulates the dynamic behavior of the power system. System frequency is simulated using the "long-term dynamics" system model, where frequency of all units is assumed to be the same. The prime-mover dynamics are represented by models of the units, turbines, governors, boilers, and boiler auxiliaries. The network

flows and states (bus voltages and angles, topology, transformer taps, etc.) are calculated at periodic intervals. Relays are modeled, and they emulate the behavior of the actual devices in the field.

Instructional System

This subsystem includes the capabilities to start, stop, restart, and control the simulation. It also includes making savecases, retrieving savecases, reinitializing to a new time, and initializing to a specific real-time situation.

It is also used to define event schedules. Events are associated with both the power system simulation and the ECS functions. Events may be deterministic (occur at a predefined time), conditional (based on a predefined set of power system conditions being met), or probabilistic (occur at random).

Defining Terms

Application: A software function within the energy management system which allows the operator to perform a specific set of tasks to meet a specific set of objectives.

Dispatch: The allocation of generation requirement to the various generating units that are available.

Distribution system: That part of the power system network which is connected to, and responsible for, the final delivery of power to the customer; typically the part of the network that operates at 33 kV and below, to 120 V.

Interchange or transaction: A negotiated purchase or sale of power between two companies.

Remote terminal unit (RTU): Hardware that gathers system-wide real-time data from various locations within substations and generating plants for telemetry to the energy management system.

Security: The ability of the power system to sustain and survive planned and unplanned events without violating operational constraints.

Related Topics

65.3 Secondary Distribution System • 65.6 Load Characteristics • 66.1 Generators • 105.1 Introduction

References

Application of Optimization Methods for Economy/Security Functions in Power System Operations, IEEE tutorial course, IEEE Publication 90EH0328-5-PWR, 1990.

Distribution Automation, IEEE Power Engineering Society, IEEE Publication EH0280-8-PBM, 1988.

C. J. Erickson, *Handbook of Electrical Heating,* IEEE Press, 1995.

Energy Control Center Design, IEEE tutorial course, IEEE Publication 77 TU0010-9 PWR, 1977.

Fundamentals of Load Management, IEEE Power Engineering Society, IEEE Publication EH0289-9-PBM, 1988.

Fundamentals of Supervisory Controls, IEEE tutorial course, IEEE Publication 91 EH0337-6 PWR, 1991.

M. Kleinpeter, Energy Planning and Policy, New York: Wiley, 1995.

"Special issue on computers in power system operations," *Proc. IEEE,* vol. 75, no. 12, 1987.

W. C. Turner, *Energy Management Handbook,* Fairmont Press, 1997.

Further Information

Current innovations and applications of new technologies and algorithms are presented in the following publications:

- *IEEE Power Engineering Review* (monthly)
- *IEEE Transactions on Power Systems* (bimonthly)
- *Proceedings of the Power Industry Computer Application Conference* (biannual)

Arnold, C.P., Watson, N.R. “Power System Analysis Software”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

68

Power System Analysis Software

68.1 Introduction

68.2 Early Analysis Programs
Load Flow (Power Flow) • Fault Analysis • Transient Stability • Fast Transients • Reliability • Economic Dispatch and Unit Commitment

68.3 The Second Generation of Programs
Graphics • Protection • Other Uses for Load Flow Analysis • Extensions to Transient Stability Analysis • Voltage Collapse • SCADA • Power Quality • Finite Element Analysis • Grounding • Other Programs

68.4 Further Development of Programs
Program Suites

68.5 Conclusions

C.P. Arnold and
N.R. Watson
University of Canterbury, New Zealand

68.1 Introduction

Power system software can be grouped in many different ways, e.g., functionality, computer platform, etc. but here it is grouped by end user. There are four major groups of end users for the software:

- major utilities
- small utilities, and industry consumers of electricity
- consultants
- universities

Large comprehensive program packages are required by utilities. They are complex, with many different functions and must have very easy input/output (IO). They serve the needs of a single electrical system and may be tailor-made for the customer. They can be integrated with the electrical system using SCADA (Supervisory Control And Data Acquisition). It is not within the scope of this chapter to discuss the merits of these programs. Suffice to say that the component programs used in these packages usually have the same generic/development roots as the programs used by the other three end user groups.

The programs used by the other three groups have usually been initially created in the universities. They start life as research programs and later are used for teaching and/or consultancy programs. Where the consultant is also an academic, then the programs may well retain their crude research style IO. However, if they are to be used by others who are not so familiar with the algorithms, then usually they are modified to make them more user friendly. Once this is achieved, the programs become commercial and are used by consultants, industry, and utilities. These are the types of programs that are now so commonly seen in the engineering journals quite often bundled together in a generic package.

68.2 Early Analysis Programs

Two of the earliest programs to be developed for power system analysis were the fault and load flow (power flow) programs. Both were originally produced in the late 1950s. Many programs in use today are either based on these two types of program or have one or the other embedded in them.

Load Flow (Power Flow)

The need to know the flow patterns and voltage profiles in a network was the driving force behind the development of load flow programs.

Although the network is linear, load flow analysis is iterative because of nodal (busbar) constraints. At most busbars the active and reactive powers being delivered to customers are known but the voltage level is not. As far as the load flow analysis is concerned, these busbars are referred to as PQ buses. The generators are scheduled to deliver a specific active power to the system and usually the voltage magnitude of the generator terminals is fixed by automatic voltage regulation. These busbars are known as PV buses.

As losses in the system cannot be determined before the load flow solution, one generator busbar only has its voltage magnitude specified. In order to give the required two specifications per node, this bus also has its voltage angle defined to some arbitrary value, usually zero. This busbar is known as the slack bus. The slack bus is a mathematical requirement for the program and has no exact equivalent in reality. However, in operating practice, the total load plus the losses are not known. When a system is not in power balance, i.e., when the input power does not equal the load power plus losses, the imbalance modifies the rotational energy stored in the system. The system frequency thus rises if the input power is too large and falls if the input power is too little. Usually a generating station and probably one machine is given the task of keeping the frequency constant by varying the input power. This control of the power entering a node can be seen to be similar to the slack bus.

The algorithms first adopted had the advantages of simple programming and minimum storage but were slow to converge requiring many iterations. The introduction of ordered elimination, which gives implicit inversion of the network matrix, and sparsity programming techniques, which reduces storage requirements, allowed much better algorithms to be used. The Newton-Raphson method gave convergence to the solution in only a few iterations. Using Newtonian methods of specifying the problem, a Jacobian matrix containing the partial derivatives of the system at each node can be constructed. The solution by this method has quadratic convergence. This method was followed quite quickly by the Fast Decoupled Newton-Raphson method. This exploited the fact that under normal operating conditions, and providing that the network is predominately reactive, the voltage angles are not affected by reactive power flow and voltage magnitudes are not effected by real power flow. The Fast Decoupled method requires more iterations to converge but each iteration uses less computational effort than the Newton Raphson method. A further advantage of this method is the robustness of the algorithm.

Further refinements can be added to a load flow program to make it give more realistic results. Transformer on-load tap changers, voltage limits, active and reactive power limits, plus control of the voltage magnitudes at buses other than the local bus help to bring the results close to reality. Application of these limits can slow down convergence.

The problem of obtaining an accurate, load flow solution, with a guaranteed and fast convergence has resulted in more technical papers than any other analysis topic. This is understandable when it is realized that the load flow solution is required during the running of many other types of power system analyses. While improvements have been made, there has been no major breakthrough in performance. It is doubtful if such an achievement is possible as the time required to prepare the data and process the results represents a significant part of the overall time of the analysis.

Fault Analysis

A fault analysis program derives from the need to adequately rate switchgear and other busbar equipment for the maximum possible fault current that could flow through them.

Initially only three-phase faults were considered and it was assumed that all busbars were operating at unity per unit voltage prior to the fault occurring. The load current flowing prior to the fault was also neglected.

By using the results of a load flow prior to performing the fault analysis, the load currents can be added to the fault currents allowing a more accurate determination of the total currents flowing in the system.

Unbalanced faults can be included by using symmetrical components. The negative sequence network is similar to the positive sequence network but the zero sequence network can be quite different primarily because of ground impedance and transformer winding configurations.

Transient Stability

After a disturbance, due usually to a network fault, the synchronous machine's electrical loading changes and the machines speed up (under very light loading conditions they can slow down). Each machine will react differently depending on its proximity to the fault, its initial loading and its time constants. This means that the angular positions of the rotors relative to each other change. If any angle exceeds a certain threshold (usually between 140° and 160°) the machine will no longer be able to maintain synchronism. This almost always results in its removal from service.

Early work on transient stability had concentrated on the reaction of one synchronous machine coupled to a very large system through a transmission line. The large system can be assumed to be infinite with respect to the single machine and hence can be modeled as a pure voltage source. The synchronous machine is modeled by the three phase windings of the stator plus windings on the rotor representing the field winding and the eddy current paths. These are resolved into two axes, one in line with the direct axis of the rotor and the other in line with the quadrature axis situated 90° (electrical) from the direct axis. The field winding is on the direct axis. Equations can be developed which determine the voltage in any winding depending on the current flows in all the other windings. A full set of differential equations can be produced which allows the response of the machine to various electrical disturbances to be found. The variables must include rotor angle and rotor speed which can be evaluated from a knowledge of the power from the turbine into, and power to the system out of the machine. The great disadvantage with this type of analysis is that the rotor position is constantly changing as it rotates. As most of the equations involve trigonometrical functions relating to stator and rotor windings, the matrices must be constantly reevaluated. In the most severe cases of network faults the results, once the dc transients decay, are balanced. Further, on removal of the fault the network is considered to be balanced. There is thus much computational effort involved in obtaining detailed information for each of the three phases which is of little value to the power system engineer. By contrast, this type of analysis is very important to machine designers. However, programs have been written for multi-machine systems using this method.

Several power system catastrophes in the U.S. and Europe in the 1960s gave a major boost to developing transient stability programs. What was required was a simpler and more efficient method of representing the machines in large power systems.

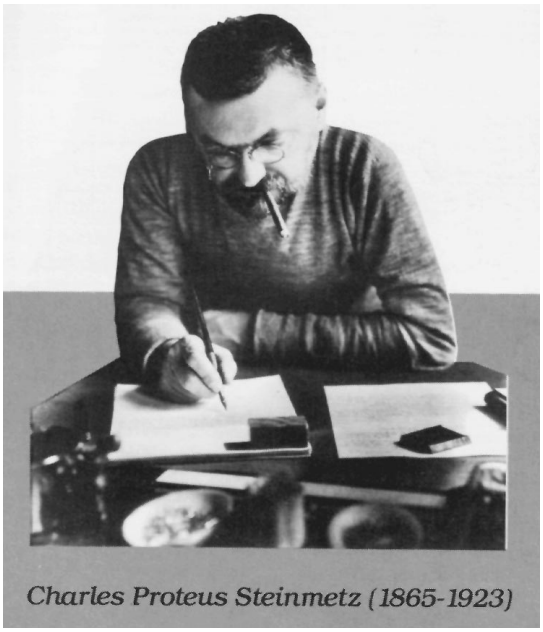
Initially, transient stability programs all ran in the time domain. A set of differential equations is developed to describe the dynamic behavior of the synchronous machines. These are linked together by algebraic equations for the network and any other part of the system that has a very fast response, i.e., an insignificant time constant, relative to the synchronous machines. All the machine equations are written in the direct and quadrature axes of the rotor so that they are constant regardless of the rotor position. The network is written in the real and imaginary axes similar to that used by the load flow and faults programs. The transposition between these axes only requires knowledge of the rotor angle relative to the synchronously rotating frame of reference of the network.

Later work involved looking at the response of the system, not to major disturbances but to the build-up of oscillations due to small disturbances and poorly set control systems. As the time involved for these disturbances to occur can be large, time domain solutions are not suitable and frequency domain models of the system were produced. Lyapunov functions have also been used, but good models have been difficult to produce. However, they are now of sufficiently good quality to compete with time domain models where quick estimates of stability are needed such as in the day to day operation of a system.

Charles Proteus Steinmetz (1865–1923)

Charles Steinmetz (1865–1923) came to the United States in 1889 from Breslau, Germany, where he was a student at the University of Breslau. He joined the inventor Rudolf Eickemeyer in building electric apparatus at Yonkers, New York, and at age 27 he formulated the law of hysteresis, which made it possible to reduce the loss of efficiency in electrical apparatus. When Eickemeyer's firm was bought by General Electric, Steinmetz joined the new company, beginning a 31-year relationship that ended only with his death.

His improvements in methods of making calculations of current in alternating current circuits revolutionized power engineering, and his theory of electrical transients stood as another important contribution. In the midst of his GE career, Steinmetz was also a professor at Union College and a vocal champion of civic and political causes. (Courtesy of the IEEE Center for the History of Electrical Engineering.)



Charles Proteus Steinmetz (1865-1923)

Fast Transients

While the transient stability program assumed a fast transient response was equivalent to an instantaneous response and only concentrated on the slower response of the synchronous machines, the requirement to model the fast transient response of traveling waves on transmission lines brought about the development of programs that treated variables with large time constants as if they were constants and modeled the variables with very small time constants by differential equations.

The program is based on the equations governing voltage and current wave propagation along a lossless line. Attenuation is then included using suitable lumped resistances. A major feature of the method is that inductance and capacitance can both be represented by resistance in parallel with a current source. This allows a purely resistive network to be formed.

Whereas, with the most other programs, source code was treated as intellectual property, the development of the fast transient program was done by many different researchers who pooled their ideas and programs. An electromagnetic transient program developed quickly and it probably became the first power systems analysis tool to be used for many different purposes throughout the world. From this base, numerous commercial packages have been developed.

In parallel with the development of electromagnetic transient programs, several state variable programs were produced to examine the fast transient behavior of parts of the electrical system, such as ac transmission lines and HVdc transmission systems. As these programs were specifically designed for the purpose they were intended, it gave them certain advantages over the general purpose electromagnetic transient program.

Reliability

Of constant concern to the operators of power systems is the reliability of equipment. This has become more important as systems are run harder. In the past, reliability was ensured by building in reserve equipment which was either connected in parallel with other similar devices or could be easily connected in the event of a failure. Not only that, knowledge of the capabilities of materials has increased so that equipment can be built with a more certain level of reliability.

However, reliability of a system is governed by the reliability of all the parts and their configuration. Much work has been done on the determination of the reliability of power systems but work is still being done to fully model power system components and integrate them into system reliability models.

The information that is obtained from reliability analysis is very much governed by the nature of the system. The accepted breakdown of a power system containing generation, transmission, and distribution is into three hierarchical levels. The first level is for the generation facilities alone, the second level contains generation and transmission, while the third level contains generation, transmission, and distribution facilities. Much of the early work was focused on the generation facilities. The reasons for this was that, first, more information was available about the generation; second, the size of the problem was smaller; and, third, the emphasis of power systems was placed in generation. With the onset of deregulation, distribution and customer requirements are now considered paramount.

At the generation and transmission levels, the loss of load expectation and frequency and duration evaluation are prime reliability indicators. A power system component may well have several derated states along with the fully operational and non-operational states. Recursive techniques are available to construct the system models and they can include multi-state components.

The usual method for evaluating reliability indices at the distribution level, such as the average interruption duration per customer per year, is an analytical approach based on a failure modes assessment and the use of equations for series and parallel networks.

Economic Dispatch and Unit Commitment

Many programs are devoted to power system operational problems and the minimization of the cost of production and delivery of energy is of great importance. Two types of program which deal with this problem are economic dispatch and unit commitment.

Economic dispatch uses optimization techniques to determine the level of power each generator (unit) must supply to the system in order to meet the demand. Each unit must have its generating costs, which will be nonlinear functions of energy, defined along with the units operational maximum and minimum power limits. The transmission losses of the system must also be taken into account to ensure an overall minimum cost.

Unit commitment calculates the necessary generating units that should be connected (committed) at any time in order to supply the demand and losses plus allow sufficient reserve capability to withstand a load increase or accidental loss of a generating unit. Several operating restrictions must be taken into account when determining which machines to commit or decommit. These include maximum and minimum running times for a unit and the time needed to commit a unit. Fuel availability constraints must also be considered. For example, there may be limited fuel reserves such as coal stocks or water in the dam. Other fuel constraints may be minimum water flows below the dam or agreements to purchase minimum amounts of fuel. Determining unit commitment for a specific time cannot be evaluated without consideration of the past operational configuration or the future operating demands.

68.3 The Second Generation of Programs

It is not the intention to suggest that only the above programs were being produced initially. However, most of the other programs remained as either research tools or one-off analysis programs. The advent of the PC gave a universal platform on which most users and programs could come together. This process was further assisted when windowing reduced the need for such a high level of computer literacy on the part of users. For

example, electromagnetic transient program's generality, which made it so successful, is also a handicap and it requires good programming skill to utilize it fully. This has lead to several commercial programs that are loosely based on the methods of analysis first used in by the electromagnetic transient program. They have the advantage of a much improved user interface.

Not all software is run on PCs. Apart from the Macintosh, which has a similar capability to a PC but which is less popular with engineers, more powerful workstations are available usually based on the Unix operating system. Mini computers and mainframe computers are also still in general use in universities and industry even though it had been thought that they would be totally superseded.

Hardware and software for power system operation and control required at utility control centers is usually sold as a total package. These systems, although excellent, can only be alluded to here as the information is proprietary. The justification for a particular configuration requires input from many diverse groups within the utility.

Graphics

Two areas of improvement that stand out in this second wave of generally available programs are both associated with the graphical capabilities of computers. A good diagram can be more easily understood than many pages of text or tables.

The ability to produce graphical output of the results of an analysis has made the use of computers in all engineering fields, not just power system analysis, much easier. Tabulated results are never easy to interpret. They are also often given to a greater degree of accuracy than the input data warrants. A graph of the results, where appropriate, can make the results very easy to interpret and if there is also an ability to graph any variable with any other, or two if three dimensions can be utilized, then new and possibly significant information can be quickly assimilated.

New packages became available for business and engineering which were based on either the spreadsheet or database principle. These also had the ability to produce graphical output. It was no longer essential to know a programming language to do even quite complex engineering analysis. The programming was usually inefficient and obtaining results was more laborious, e.g., each iteration had to be started by hand. But, as engineers had to use these packages for other work, they became very convenient tools.

A word of caution here—be careful that the results are graphed in an appropriate manner. Most spreadsheet packages have very limited x-axis (horizontal) manipulation. Provided the x-axis data comes in regular steps, the results are acceptable. However, we have seen instances where very distorted graphs have been presented because of this problem.

Apart from the graphical interpretation of results, there are now several good packages that allow the analyst to enter the data graphically. It is a great advantage to be able to develop a one-line, or three-phase, diagram of a network directly with the computer. All the relevant system components can be included. Parameter data still require entry in a more orthodox manner but by merely clicking on a component, a data form for that component can be made available. The chances of omitting a component are greatly reduced with this type of data entry. Further, the same system diagram can be used to show the results of some analyses.

An extension of the network diagram input is to make the diagram relate to the actual topography. In these cases, the actual routes of transmission lines are shown and can be superimposed on computer generated geographical maps. The lines in these cases have their lengths automatically established and, if the line characteristics are known, the line parameters can be calculated.

These topographical diagrams are an invaluable aid for power reticulation problems, for example, the minimum route length of reticulation given all the points of supply and the route constraints. Other optimization algorithms include determination of line sizes and switching operations.

The analysis techniques can be either linear or nonlinear. If successful, the nonlinear algorithm is more accurate but these techniques suffer from larger data storage requirements, greater computational time, and possible divergence. There are various possible optimization techniques that can and have been applied to this problem. There is no definitive answer and each type of problem may require a different choice.

The capability chart represents a method of graphically displaying power system performance. These charts are drawn on the complex power plane and define the real and reactive power that may be supplied from a

point in the system during steady state operation. The power available is depicted as a region on the plane and the boundaries of the region represent the critical operating limits of the system. The best known example of a capability chart is the operating chart of a synchronous machine. The power available from the generator is restricted by limiting values of the rotor current, stator current, turbine power (if a generator), and synchronous stability limits. Capability charts have been produced for transmission lines and HVdc converters.

Where the capability chart is extended to cover more than one power system component, the two-dimensional capability chart associated with a single busbar can be regarded as being a single slice of an overall $2n$ dimensional capability chart for the n busbars that make up a general power system. If the system is small, a contour plotting approach can be used to gradually trace out the locus on the complex power plane. A load flow algorithm is used to iteratively solve the operating equations at each point on the contour, without having to resort to an explicit closed form solution.

The good contour behavior near the operating region has allowed a faster method to be adopted. A seed load flow solution, corresponding to the nominal system state, is obtained to begin drawing the chart. A region growing process is then used to locate the region in which all constrained variables are less than 10% beyond their limits. This process is similar to a technique used in computer vision systems to recognize shapes of objects. The region grows by investigating the six nearest lattice vertices to any unconstrained vertex. Linear interpolation along the edges between vertices is then used to estimate the points of intersection between the contour and the lattice. This method has a second advantage in that it can detect holes and islands in the chart. However, it should be noted that these regions are purely speculative and have not been found in practice.

Protection

The need to analyze protection schemes has resulted in the development of protection coordination programs. Protection schemes can be divided into two major groupings: unit and non-unit schemes.

The first group contains schemes that protect a specific area of the system, i.e., a transformer, transmission line, generator, or busbar. The most obvious example of unit protection schemes is based on Kirchhoff's current law—the sum of the currents entering an area of the system must be zero. Any deviation from this must indicate an abnormal current path. In these schemes, the effects of any disturbance or operating condition outside the area of interest are totally ignored and the protection must be designed to be stable above the maximum possible fault current that could flow through the protected area. Schemes can be made to extend across all sides of a transformer to account for the different currents at different voltage levels. Any analysis of these schemes are thus of more concern to the protection equipment manufacturers.

The non-unit schemes, while also intended to protect specific areas, have no fixed boundaries. As well as protecting their own designated areas, the protective zones can overlap into other areas. While this can be very beneficial for backup purposes, there can be a tendency for too great an area to be isolated if a fault is detected by different non-unit schemes. The most simple of these schemes measures current and incorporates an inverse time characteristic into the protection operation to allow protection nearer to the fault to operate first. While this is relatively straightforward for radial schemes, in networks, where the current paths can be quite different depending on operating and maintenance strategies, protection can be difficult to set and optimum settings are probably impossible to achieve. It is in these areas where protection software has become useful to manufacturers, consultants, and utilities.

The very nature of protection schemes has changed from electromechanical devices, through electronic equivalents of the old devices, to highly sophisticated system analyzers. They are computers in their own right and thus can be developed almost entirely by computer analysis techniques.

Other Uses for Load Flow Analysis

It has already been demonstrated that load flow analysis is necessary in determining the economic operation of the power system and it can also be used in the production of capability charts. Many other types of analyses require load flow to be embedded in the program.

As a follow on from the basic load flow analysis, where significant unbalanced load or unbalanced transmission causes problems, a three-phase load flow may be required to study their effects. These programs require each phase to be represented separately and mutual coupling between phases to be taken into account. Transformer winding connections must be correctly represented and the mutual coupling between transmission lines on the same tower or on the same right-of-way must also be included.

Motor starting can be evaluated using a transient stability program but in many cases this level of analysis is unnecessary. The voltage dip associated with motor start up can be determined very precisely by a conventional load flow program with a motor starting module.

Optimal power system operation requires the best use of resources subject to a number of constraints over any specified time period. The problem consists of minimizing a scalar objective function (normally a cost criterion) through the optimal control of a vector of control parameters. This is subject to the equality constraints of the load flow equations, inequality constraints on the control parameters, and inequality constraints of dependent variables and dependent functions. The programs to do this analysis are usually referred to as optimal power flow (OPF) programs.

Often optimal operation conflicts with the security requirements of the system. Load flow studies are used to assess security (security assessment). This can be viewed as two separate functions. First, there is a need to detect any operating limit violations through continuous monitoring of the branch flows and nodal voltages. Second, there is a need to determine the effects of branch outages (contingency analysis). To reduce this to a manageable level, the list of contingencies is reduced by judicial elimination of most of the cases that are not expected to cause violations. From this the possible overloading of equipment can be forecast. The program should be designed to accommodate the condition where generation cannot meet the load because of network islanding.

The conflicting requirements of system optimization and security require that they be considered together. The more recent versions of OPF interface with contingency analysis and the computational requirements are enormous.

Extensions to Transient Stability Analysis

Transient stability programs have been extended to include many other system components, including FACTS (flexible ac transmission systems) and dc converters.

FACTS may be either shunt or branch devices. Shunt devices usually attempt to control busbar voltage by varying their shunt susceptance. The device is therefore relatively simple to implement in a time domain program. Series devices may be associated with transformers. Stability improvement is achieved by injecting a quadrature component of voltage derived from the other two phases rather than by a tap changer which injects a direct component of voltage. Fast acting power electronics can inject either or a combination of both direct and quadrature voltage to help maintain voltage levels and improve stability margins.

Dc converters for HVdc links and rectifier loads have received much attention. The converter controls are very fast acting and therefore a quasi steady state (QSS) model can be considered accurate. That is, the model of the converter terminals contains no dynamic equations and in effect the link behaves as if it was in steady state for every time solution of the ac system. While this may be so some time after a fault has been removed, during and just after a fault the converters may well suffer from commutation failure or fire through. These events cannot be predicted or modeled with a QSS model. In this case, an appropriate method of analysis is to combine a state variable model of the converter, which can model the firing of the individual valves, with a conventional multi-machine transient stability program containing a QSS model. During the period of maximum disturbance, the two models can operate together. Information about the overall system response is passed to the state variable model at regular intervals. Similarly the results from the detailed converter model are passed to the multi machine model overriding its own QSS model. As the disturbance reduces, the results from the two different converter models converge and it is then only necessary to run the computationally inexpensive QSS model within the multi machine transient stability program.

Voltage Collapse

Steady state analysis of the problem of voltage instability and voltage collapse are often based on load flow analysis programs. However, time solutions can provide further insight into the problem.

A transient stability program can be extended to include induction machines which are associated with many of the voltage collapse problems. In these studies, it is the stability of the motors that are examined rather than the stability of the synchronous machines. The asynchronous nature of the induction machine means that rotor angle is not a concern, but instead the capability of the machines to recover after a fault has depressed the voltage and allowed the machines to slow down. The re-accelerating machines draw more reactive current which can hold the terminal voltage down below that necessary to allow recovery. Similarly starting a machine will depress the voltage which affects other induction machines which further lowers the voltage.

However, voltage collapse can also be due to longer term problems. Transient stability programs then need to take into account controls that are usually ignored. These include automatic transformer tap adjustment and generator excitation limiters which control the long-term reactive power output to keep the field currents within their rated values.

The equipment that can contribute to voltage collapse must also be more carefully modeled. Simple impedance models for loads ($P = P_oV^2$; $Q = Q_oV^2$) are no longer adequate. An improvement can be obtained by replacing the (mathematical) power 2 in the equations by more suitable values. Along with the induction machine models, the load characteristics can be further refined by including saturation effects.

SCADA

SCADA (Supervisory Control And Data Acquisition) has been an integral part of system control for many years. A control center now has much real time information available so that human and computer decisions about system operation can be made with a high degree of confidence.

In order to achieve high quality input data, algorithms have been developed to estimate the state of a system based on the available on-line data (state estimation). These methods are based on weighted least squares techniques to find the best state vector to fit the scatter of data. This becomes a major problem when conflicting information is received. However, as more data becomes available, the reliability of the estimate can be improved.

Power Quality

One form of poor power quality which has received a large amount of attention is the high level of harmonics that can exist and there are numerous harmonic analysis programs now available.

Recently, the harmonic levels of both currents and voltages have increased considerably due to the increasing use of non-linear loads such as arc furnaces, HVdc converters, FACTS equipment, dc motor drives, and ac motor speed control. Moreover, commercial sector loads now contain often unacceptable levels of harmonics due to widespread use of equipment with rectifier-fed power supplies with capacitor output smoothing (e.g., computer power supplies and fluorescent lighting). The need to conserve energy has resulted in energy efficient designs that exacerbate the generation of harmonics. Although each source only contributes a very small level of harmonics, due to their small power ratings, widespread use of small non-linear devices may create harmonic problems which are more difficult to remedy than one large harmonic source.

Harmonic analysis algorithms vary greatly in their algorithms and features; however, almost all use the frequency domain. The most common technique is the direct method (also known as current injection method). Spectral analysis of the current waveform of the non-linear components is performed and entered into the program. The network data is used to assemble a system admittance matrix for each frequency of interest. This set of linear equations is solved for each frequency to determine the node voltages and, hence, current flow throughout the system. This method assumes the non-linear component is an ideal harmonic current source. The next more advanced technique is to model the relationship between the harmonic currents injected by a component to its terminal voltage waveform. This then requires an iterative algorithm, which does require

excursion into the time domain for modeling this interaction. When the fundamental (load flow) is also included, thus simulating the interaction between fundamental and harmonic frequencies, it is termed a harmonic power flow. The most advanced technique, which is still only a research tool, is the harmonic domain. In this iterative technique one Jacobian is built-up that represents all harmonic frequencies. This allows coupling between harmonics, which occurs, for example, in salient synchronous machines, to be represented.

There are many other features that need to be considered, such as whether the algorithm uses symmetrical components or phase coordinates, or whether it is single- or three-phase. Data entry for single-phase typically requires the electrical parameters, whereas three-phase analysis normally requires the physical geometry of the overhead transmission lines and cables and conductor details so that a transmission line parameter program or cable parameter program can calculate the line or cable electrical parameters.

The communication link between the monitoring point and the control center can now be very sophisticated and can utilize satellites. This technology has led to the development of systems to analyze the power quality of a system. Harmonic measurement and analysis has now reached a high level of maturity. Many different pieces of information can be monitored and the results over time stored in a database. Algorithms based on the fast Fourier transform can then be used to convert this data from the time domain to the frequency domain. Computing techniques coupled with fast and often parallel computing allows this information to be displayed in real time. By utilizing the time stamping capability of the global positioning system (GPS), information gathered at remote sites can be linked together. Using the GPS time stamp, samples taken exactly simultaneously can be feed to a harmonic state estimator which can even determine the position and magnitude of harmonics entering the system as well as the harmonic voltages and currents at points not monitored (provided enough initial monitoring points exist).

One of the most important features of harmonic analysis software is the ability to display the results graphically. The refined capabilities of present three-dimensional graphics packages has simplified the analysis considerably.

Finite Element Analysis

Finite element analysis is not normally used by power system engineers although it is a common tool of high voltage and electrical machine engineers. It is necessary, for example, where accurate machine representation is required. For example, in a unit connected HVdc terminal the generators are closely coupled to the rectifier bridges. The ac system at the rectifier end is isolated from all but its generator. There is no need for costly filters to reduce harmonics. Models of the synchronous machine suitable for a transient stability study can be obtained from actual machine tests. For fast transient analysis, a three-phase generator model can be used but it will not account for harmonics. A finite element model of the generator provides the means of allowing real time effects such as harmonics and saturation to be directly included. Any geometric irregularities in the generator can be accounted for and the studies can be done at the design stage rather than having to rely on measurements or extrapolation from manufactured machines to obtain circuit parameters. There is no reliance on estimated machine parameters. The disadvantages are the cost and time to run a simulation and it is not suitable at present to integrate with existing transient stability programs as it requires a high degree of expertise. As the finite element model is in this case used in a time simulation, part of the air gap is left unmeshed in the model. At each time step the rotor is placed in the desired position and the missing elements in the air gap region formed using the nodes on each side of the gap.

Grounding

The safe grounding of power system equipment is very important, especially as the short circuit capability of power systems continues to grow. Programs have been developed to evaluate and design grounding systems in areas containing major power equipment, such as substations and to evaluate the effects of fault current on remote, separately grounded equipment.

The connection to ground may consist of a ground mat of buried conductors, electrodes (earth rods), or both. The shape and dimensions of the electrodes, their locations, and the layout of an ground mat, plus the resistivity of the ground at different levels must be specified in order to evaluate the ground resistance. A grid of buried conductors and electrodes is usually considered to be all at the same potential. Where grid sections are joined by buried or aerial links, these links can have resistance allowing the grid sections to have different potentials. It is usual to consider a buried link as capable of radiating current into the soil.

Various methods of representing the fault current are available. The current can be fixed or it can be determined from the short circuit MVA and the busbar voltage. A more complex fault path may need to be constructed for faults remote from the site being analyzed.

From the analysis, the surface potential over the affected area can be evaluated and, from that, step and touch potentials calculated. Three-dimensional graphics of the surface potentials are very useful in highlighting problem areas.

Other Programs

There are too many other programs available to be discussed. For example, neither automatic generator control nor load forecasting have been included. However, an example of a small program that can stand alone or fit into other programs is given here.

In order to obtain the electrical parameters of overhead transmission lines and underground cables, utility programs have been developed. Transmission line parameter programs use the physical geometry of the conductors, the conductor type, and ground resistivity to calculate the electrical parameters of the line. Cable parameter programs use the physical dimensions of the cable, its construction, and its position in the ground. The results of these programs are usually fed directly to network analysis programs such as load flow or faults. The danger of errors introduced during transfer are thus minimized. This is particularly true for three-phase analyses due to the volume of data involved.

68.4 Further Development of Programs

Recently there has been a shift in emphasis in the types of program being constructed. Deregulation (a misnomer of grand proportions) is making financial considerations a prime operating constraint. New programs are now being developed which assist in the buying and selling of energy through the electrical system.

Following on from the solution of the economic dispatch, "time of use" pricing has been introduced into some power system operations. Under this system, the price of electricity at a given time reflects the marginal cost of generation at that time. As the marginal generator changes over time, so does the price of electricity.

The next stage is to price electricity not only on time but also on the place of use (nodal pricing). Thus, the cost of transportation of the energy from the producer to the user is included in the price. This can be a serious problem at present when power is exchanged between utilities. It will become increasingly common as the individual electricity producers and users set up contractual agreements for supply and use. A major problem at present is the lack of common agreement as to whether nodal pricing is the most appropriate mechanism for a deregulated wholesale electricity market. Clarification will occur as the structure of the industry changes.

Nodal pricing also takes into account other commercial and financial factors. These include the pricing of both generation and transmission constraints, the setting of a basis for transmission constraint hedges and for the economic dispatch of generation. The programs must be designed to give both the suppliers and consumers of energy the full opportunity costs of the operation of the power system.

Inherent in nodal pricing must be such factors as marginal cost pricing, short run price, and whether the price is *ex ante* (before) or *ex post* (after) the event. Thus far, the programming effort has concentrated on real power pricing but the cost of reactive power should also eventually be included.

The changes in the operation of power systems, which are occurring throughout the world at present, will inevitably force changes to many of the programs in use today and, as shown above, new programs will emerge.

These programs are an example of direct transfer of university programs to major utilities. However, because the number of organizations involved in the industry is increasing, these and other programs will become more generally available.

Program Suites

As more users become involved with a program, its quirks become less acceptable and it must become easy to use, i.e., user friendly. Second, with the availability of many different types of program, it became important to be able to transfer the results of one program to the input of another. If the user has access to the source code, this can often be done relatively quickly by generating an output file in a suitable format for the input of the second program. There has, therefore, been a great deal of attention devoted to creating common formats for data transfer as well as producing programs with easy data entry formats and good result processing capabilities.

Many good "front end" programs are now available which allow the user to quickly write an analysis program and utilize the built in IO features of the package. There are also several good general mathematical packages available. Much research work can now be done using tools such as these. The researcher is freed from the chore of developing algorithms and IO routines by using these standard packages. Not only that, extra software is being developed which can turn these general packages into specialist packages. It may well be that before long all software will be made to run on sophisticated developments of these types of package and the stand alone program will fall into oblivion.

68.5 Conclusions

There are many more programs available than can be discussed here. Those that have been discussed are not necessarily more significant than those omitted. There are programs to help you with almost every power system problem you have and new software is constantly becoming available to solve the latest problems.

Make sure that programs you use are designed to do the job you require. Some programs make assumptions which give satisfactory results in most cases but may not be adequate for your particular case. No matter how sophisticated and friendly the program may appear, it is the algorithm and processing of data which are the most important parts. As programs become more complex and integrated, new errors (regressions) can be introduced. Wherever possible check the answers and always make sure they feel right.

Related Topics

110.3 The Bathtub Curve • 110.4 Mean Time to Failure (MTTF) • 110.22 Reliability and Economics

Further Information

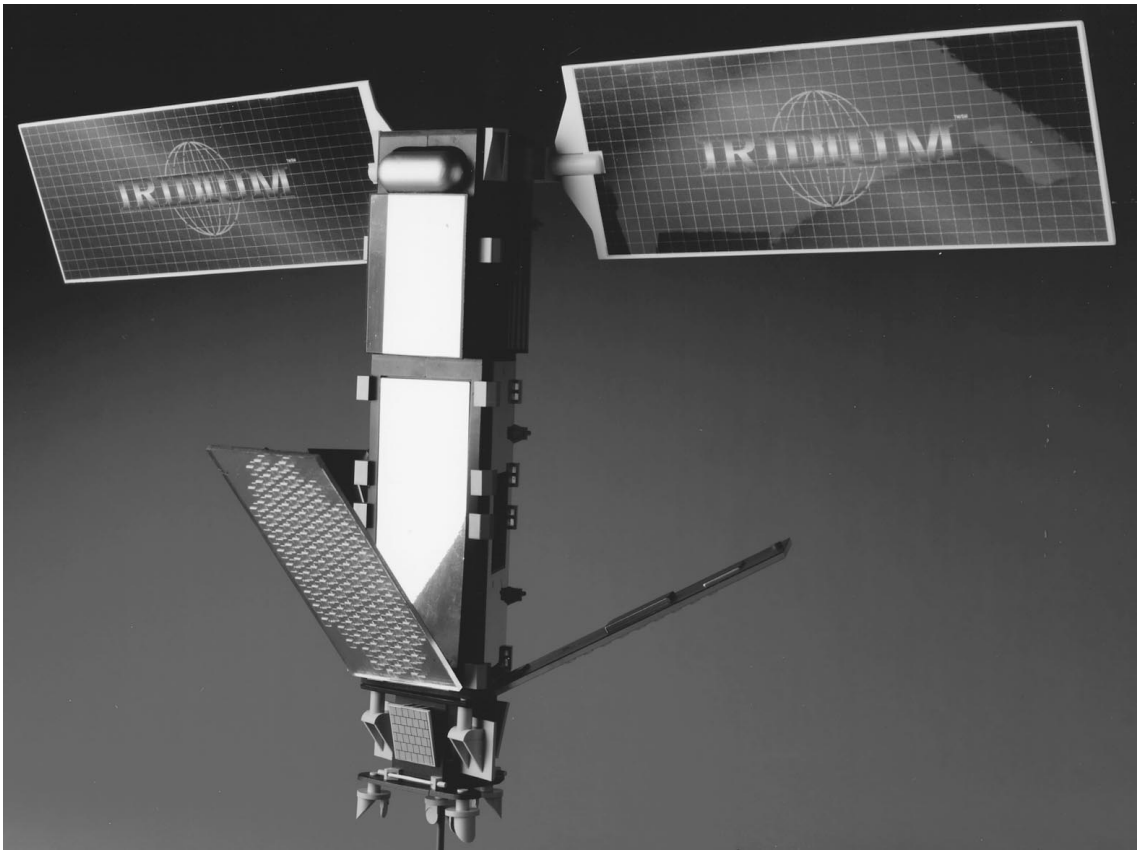
There are several publications that can keep engineers up to date with the latest developments in power system analysis. The *IEEE Spectrum* (U.S.) and the *IEE Review* (U.K.) are the two most well respected, general interest, English language journals that report on the latest development in electrical engineering. The *Power Engineering Journal* produced by the IEE regularly runs tutorial papers, many of which are of direct concern to power systems analysts. However, for magazine-style coverage of the developments in power system analysis, the *IEEE Computer Applications in Power* is in the authors' opinion, the most useful.

Finally, a few text books that provide a much greater insight into the programs discussed in the chapter have been included below.

J. Arrillaga and C.P. Arnold, *Computer Analysis of Power Systems,* London: John Wiley & Sons, 1990.
R. Billinton and R.N. Allan, *Reliability Evaluation of Power Systems,* New York: Plenum Press, 1984.
A.S. Debs, *Modern Power Systems Control and Operation,* New York: Kluwer Academic Publishers, 1988.
C.A. Gross, *Power System Analysis,* New York: John Wiley & Sons, 1986.
B.R. Gungor, *Power Systems,* New York: Harcourt Brace Jovanovich, 1988.

G.T. Heydt, *Computer Analysis Methods for Power Systems,* Stars in a Circle Publications, 1996.
IEEE Brown Book—Power Systems Analysis, IEEE, 1990.
G.L. Kusic, *Computer-Aided Power System Analysis,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
B.M. Weedy, *Electric Power Systems,* New York: John Wiley & Sons, 1987.
A.J. Wood and B.F. Wollenberg, *Power Generation, Operation and Control,* New York: John Wiley & Sons, 1984.

Shaw, L. "Section VII – Communications"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



The IRIDIUM system is a satellite-based, wireless personal communications network designed to permit any type of telephone transmission—voice, data, fax, or paging—to reach its destination anywhere on earth. The IRIDIUM constellation will consist of 66 interconnected satellites like the one shown above.

Each IRIDIUM satellite, weighing approximately 689 kg (1,500 lb), will orbit above the earth at an altitude of 420 nautical miles and communicate directly with subscriber equipment. Each of the satellites will function like extremely tall cellular towers and will project tightly focused beams over the ground. The low-earth orbit of IRIDIUM satellites as well as recent advances in microelectronics make it possible to communicate with a handheld phone.

The IRIDIUM system was conceived in 1987 by engineers at Motorola's Satellite Communications Division. It is being financed by a private international consortium of telecommunications and industrial companies and is expected to begin operation in 1998. (Photo courtesy of Motorola.)

VII

Communications

69 Broadcasting *R.C. Dorf, Z. Wan, J.F. Lindsey III, D.F. Doelitzsch, J. Whitaker, M.S. Roden, S. Salek, A.H. Clegg*
Modulation and Demodulation • Radio • Television Systems • High-Definition Television • Digital Audio Broadcasting

70 Digital Communication *R.C. Dorf, Z. Wan, L.B. Millstein, M.K. Simon*
Error Control Coding • Equalization • Spread Spectrum Communications

71 Optical Communication *T.E. Darcie, J.C. Palais, I.P. Kaminow*
Lightwave Technology for Video Transmission • Long Distance Fiber Optic Communications • Photonic Networks

72 Networks *M.N. Huber, J.N. Daigle, J. Bannister, M. Gerla, R.B. Robrock II*
B-ISDN • Computer Communication Networks • Local-Area Networks • The Intelligent Network

73 Information Theory *H.V. Poor, C.G. Looney, R.J. Marks II, S. Verdú, J.A. Thomas, T.M. Cover*
Signal Detection • Noise • Stochastic Processes • The Sampling Theorem • Channel Capacity • Data Compression

74 Satellites and Aerospace *D.F. DiFonzo*
Satellite Applications • Satellite Functions • Satellite Orbits and Pointing Angles • Communications Link • System Noise Temperature and *G/T* • Digital Links • Interference • Some Particular Orbits • Access and Modulation • Frequency Allocations • Satellite Subsystems • Trends

75 Personal and Office *W.C.Y. Lee, R.E. Ziemer, M. Ovan, G.D. Mandyam*
Mobile Radio and Cellular Communications • Facsimile • Wireless Local-Area Networks for the 1990s • Wireless PCS

76 Phase-Locked Loop *S.L. Maddy*
Loop Filter • Noise • PLL Design Procedures • Components • Applications

77 Telemetry *C.H. Hoeppner*
Introduction to Telemetry • Measuring and Transmitting • Applications of Telemetry • Limitations of Telemetry • Transmitters and Batteries • Receivers and Discriminators • Antennas and Total System Operation • Calibration • Telemetry Frequency Allocations • Telemetry Antennas • Measuring and Transmitting • Modulating and Multiplexing • Passive Telemeters • The Receiving Station

78 Computer-Aided Design and Analysis of Communication Systems *W.H. Tranter, K.L. Kosbar*
The Role of Simulation • Motivation for the Use of Simulation • Limitations of Simulation • Simulation Structure • The Interdisciplinary Nature of Simulation • Model Design • Low-Pass Models • Pseudorandom Signal and Noise Generators • Transmitter, Channel, and Receiver Modeling • Symbol Error Rate Estimation • Validation of Simulation Results • A Simple Example Illustrating Simulation Products

Leonard Shaw
Polytechnic University, New York

ELECTRICAL TECHNOLOGY has been involved in aiding communication between a sender and a receiver of information since the advent of the electrical telegraph. The evolution of electrical communications technology has been influenced by both advances in devices for processing and transmitting electrical signals, as well as the growth and variety of communications applications that have become essential to modern society. A large fraction of electrical engineers are involved with some aspect of communications, as evidenced by the size of the IEEE Communications Society, which is second only to the Computer Society. In fact, communication between computers makes up a large part of communication system traffic, and communication technology is playing an increasing role *within* computers as they employ multiple processors and processors that are geographically distributed.

This section presents an overview of a variety of communication systems that have been developed to overcome the constraints of physical communication channels by exploiting the capabilities of the electronic and optoelectronic devices that are described elsewhere in this handbook. As a reflection of the dual influences of electrotechnology and user applications, some of the following chapters have application themes (broadcasting, satellite and aerospace, personal and office, and telemetry), while the rest have themes related to systems techniques (digital, optical, network, information theory, phase-locked loop, and computer-aided design).

The conventional radio station is a prototype of a broadcasting system in which a single transmitter sends the same message to multiple receivers. Chapter 69 reviews the basic notions of modulation needed to match the transmitted signal to the propagation and noise characteristics of the transmission medium and outlines recent developments in systems for high-definition television (HDTV) and digital audio broadcasting (DAB).

The chapter on digital techniques emphasizes the coding techniques used to detect and correct transmission errors (which are inevitable even if systems can be designed to reduce their frequency of occurrence). Since the rate of pulse transmission over a channel can be maximized by having an accurate model for the channel, such systems are improved by continually readjusting the channel model as the characteristics change with time. This chapter also discusses adaptive equalizers that match electrical pulse shapes to changing channels.

The development of fiber-optic cables and efficient solid-state lasers has revolutionized telephone communications. Chapter 71 describes some of the related developments in signal design and transmission for optical systems that carry voice, video, and computer data messages.

Traditional telephone switching has evolved into a huge field of telecommunication networks, with the advent of new media such as fiber-optic cables and satellites and the rapidly growing digital traffic such as that between computers, and supporting e-mail and the World Wide Web. Chapter 72 describes switching and transmission protocols and other standards that are being developed to coordinate the design of equipment that sends and receives messages over the networks.

The chapter on information theory uses that term in a broad sense to describe mathematical models and techniques for describing and simplifying both deterministic and random signals. These techniques can be used for efficient communication by removing inessential information and by showing how a receiver can distinguish useful information from noise disturbances.

Satellite and aerospace applications, described in Chapter 74, provide dramatic examples of challenging communication environments where, due to equipment weight limitations and great distances, signals are weak compared to the associated noise, and propagation characteristics are nonlinear.

Personal and office innovations related to communication systems are as dramatic to the ordinary citizen as those in entertainment applications such as HDTV and digital audio. Chapter 75 describes how facsimile systems, which are especially useful for rapid transmission of graphical information, exploit standardized techniques for compressing black-and-white images. Also presented are new developments in modulation techniques and propagation modeling that have been stimulated by mobile telephone and wireless network applications.

Phase-locked loops are presented in Chapter 76 as good examples of electronic systems that are able to detect weak signals whose characteristics change with time in environments where there is strong interference from noise and from competing transmitters.

Telemetry systems are dedicated to collection and transmission of data from many sensors, often in hostile or distant environments. Chapter 77 describes how constraints on equipment size, weight and power often lead to novel methods for data multiplexing and transmission.

This section concludes with a chapter on computer-aided design methods that are being exploited to design communication systems more rapidly and effectively. Many of the problems, such as best location of a large number of nodes in a network where the construction costs and performance measures are a complex function of design parameters, are best solved by a designer who works interactively with computer algorithms.

Nomenclature

Symbol	Quantity	Unit	Symbol	Quantity	Unit
A_{eff}	effective area of antenna	m^2	M	detector gain	
B	bit rate	Mbytes/s	μ	rms modulation index	
B	channel bandwidth	Hz	n	effective input current noise density	
C	capacitance	F			
CIR	carrier-to-interference ratio		N	number of equalizer coefficients	
CNR	carrier-to-noise ratio		NF	noise ratio	
D	propagation delay	s	P	power density	W/m
$\Delta\lambda$	spectral width	Hz	P	probability of error	
$\Delta\tau$	pulse spread	s	PE	preemphasis factor	
E	electric field intensity	V/m	q	interference reduction factor	
f	carrier frequency	Hz	r	distance	m
F	noise figure		R_L	input impedance	Ω
$g(t)$	complex envelope		ρ	correlation coefficient	
G	power gain of antenna	dB	$s(t)$	modulated signal	
$H(x)$	entropy	bit	S	throughput	terabit/s
η	quantum efficiency		SNR	signal-to-noise ratio	
I	polarization isolation	dB	σ^2	variance of noise samples	
K	loop gain		t_R	rise time	s
$m(t)$	modulating signal		t_o	sample time	s
M	bit rate delay product		Z_{fs}	impedance of free space	$120\pi\,\Omega$

Dorf, R.C., Wan, Z., Lindsey III, J.F., Doelitzsch, D.F., Whitaker J., Roden, M.S., Salek, S., Clegg, A.H. “Broadcasting”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

69

Broadcasting

Richard C. Dorf
University of California, Davis

Zhen Wan
University of California, Davis

Jefferson F. Lindsey III
Southern Illinois University at Carbondale

Dennis F. Doelitzsch
3-D Communications

Jerry Whitaker
Technical Press

Martin S. Roden
California State University

Stanley Salek
Hammett & Edison

Almon H. Clegg
CCi

69.1 Modulation and Demodulation
Modulation • Superheterodyne Technique • Pulse-Code Modulation • Frequency-Shift Keying • *M*-ary Phase-Shift Keying • Quadrature Amplitude Modulation

69.2 Radio
Standard Broadcasting (Amplitude Modulation) • Frequency Modulation

69.3 Television Systems
Scanning Lines and Fields • Interlaced Scanning Fields • Synchronizing Video Signals • Television Industry Standards • Transmission Equipment • Television Reception

69.4 High-Definition Television
Proposed Systems

69.5 Digital Audio Broadcasting
The Need for DAB • DAB System Design Goals • Historical Background • Technical Overview of DAB • Audio Compression and Source Encoding • System Example:Eureka-147/DAB

69.1 Modulation and Demodulation

Richard C. Dorf and Zhen Wan

Modulation is the process of impressing the source information onto a bandpass signal with a carrier frequency f_c. This bandpass signal is called the modulated signal $s(t)$, and the baseband source signal is called the modulating signal $m(t)$. The modulated signal could be represented by

$$s(t) = \text{Re}\{g(t)e^{j\omega ct}\} \tag{69.1}$$

or, equivalently,

$$s(t) = R(t)\cos[\omega_c t + \theta(t)] \tag{69.2}$$

and

$$s(t) = x(t)\cos\omega_c t - y(t)\sin\omega_c t \tag{69.3}$$

where $\omega_c = 2\pi f_c$. The complex envelope is

$$g(t) = R(t)e^{j\theta(t)} = x(t) + jy(t) \tag{69.4}$$

and $g(t)$ is a function of the modulating signal $m(t)$. That is,

$$g(t) = g[m(t)]$$

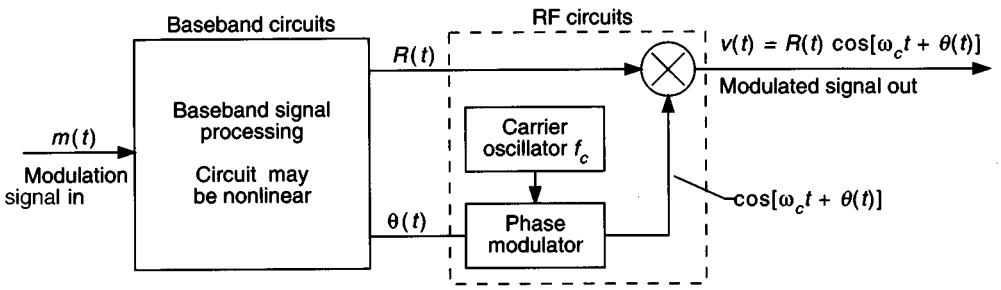


FIGURE 69.1 Generalized transmitter using the AM-PM generation technique.

Thus $g[\cdot]$ performs a mapping operation on $m(t)$. The particular relationship that is chosen for $g(t)$ in terms of $m(t)$ defines the type of modulation used.

In Table 69.1, examples of the mapping function $g(m)$ are given for the following types of modulation:

- AM: amplitude modulation
- DSB-SC: double-sideband suppressed-carrier modulation
- PM: phase modulation
- FM: frequency modulation
- SSB-AM-SC: single-sideband AM suppressed-carrier modulation
- SSB-PM: single-sideband PM
- SSB-FM: single-sideband FM
- SSB-EV: single-sideband envelope-detectable modulation
- SSB-SQ: single-sideband square-law-detectable modulation
- QM: quadrature modulation

Modulation

In Table 69.1, a generalized approach may be taken to obtain universal transmitter models that may be reduced to those used for a particular modulation type. We also see that there are equivalent models which correspond to different circuit configurations, yet they may be used to produce the same type of modulated signal at their outputs. It is up to communication engineers to select an implementation method that will optimize performance, yet retain low cost based on the state of the art in circuit development.

There are two canonical forms for the generalized transmitter. Figure 69.1 is an AM-PM type circuit as described in Eq. (69.2). In this figure, the baseband signal processing circuit generates $R(t)$ and $\theta(t)$ from $m(t)$. The R and θ are functions of the modulating signal $m(t)$ as given in Table 69.1 for the particular modulation type desired.

Figure 69.2 illustrates the second canonical form for the generalized transmitter. This uses in-phase and quadrature-phase (IQ) processing. Similarly, the formulas relating $x(t)$ and $y(t)$ are shown in Table 69.1, and the baseband signal processing may be implemented by using either analog hardware or digital hardware with software. The remainder of the canonical form utilizes radio frequency (RF) circuits as indicated.

Any type of signal modulation (AM, FM, SSB, QPSK, etc.) may be generated by using either of these two canonical forms. Both of these forms conveniently separate baseband processing from RF processing.

Superheterodyne Technique

Most receivers employ the **superheterodyne receiving** technique (see Fig. 69.3). This technique consists of either down-converting or up-converting the input signal to some convenient frequency band, called the *intermediate frequency* (IF) band, and then extracting the information (or modulation) by using the appropriate detector. This basic receiver structure is used for the reception of all types of bandpass signals, such as television, FM, AM, satellite, and radar signals.

TABLE 69.1 Complex Envelope Functions for Various Types of Modulation

Type of Modulation	Mapping Functions $g[m]$	Corresponding Quadrature Modulation $x(t)$	Corresponding Quadrature Modulation $y(t)$	Corresponding Amplitude and Phase Modulation $R(t)$	Corresponding Amplitude and Phase Modulation $\theta(t)$	Linearity	Remarks
AM	$1 + m(t)$	$1 + m(t)$	0	$\lvert 1 + m(t)\rvert$	$\begin{cases} 0, & m(t) > -1 \\ 180°, & m(t) < -1 \end{cases}$	L[b]	$m(t) > -1$ required for envelope detection.
DSB-SC	$m(t)$	$m(t)$	0	$\lvert m(t)\rvert$	$\begin{cases} 0, & m(t) > 0 \\ 180°, & m(t) < 0 \end{cases}$	L	Coherent detection required.
PM	$e^{jD_p m(t)}$	$\cos[D_p m(t)]$	$\sin[D_p m(t)]$	1	$D_p m(t)$	NL	D_p is the phase deviation constant (radian/volts).
FM	$e^{jDf\int_{-\infty}^{t} m(\sigma)d\sigma}$	$\cos\left[D_f \int_{-\infty}^{t} m(\sigma)d\sigma\right]$	$\sin\left[D_f \int_{-\infty}^{t} m(\sigma)d\sigma\right]$	1	$D_f \int_{-\infty}^{t} m(\sigma)d\sigma$	NL	D_f is the frequency deviation constant (radian/volt-sec).
SSB-AM-SC[a]	$m(t) \pm j\hat{m}(t)$	$m(t)$	$\pm\, \hat{m}(t)$	$\sqrt{[m(t)^2 + [\hat{m}(t)]^2}$	$\tan^{-1}[\pm\, \hat{m}(t)/m(t)]$	L	Coherent detection required.
SSB-PM[a]	$e^{jD_p[m(t)\pm j\hat{m}(t)]}$	$e^{\mp D_p \hat{m}(t)}\cos[D_p m(t)]$	$e^{\mp D_p \hat{m}(t)}\sin[D_p m(t)]$	$e^{\mp D_p \hat{m}(t)}$	$D_p m(t)$	NL	
SSB-FM[a]	$e^{jDf\int_{-\infty}^{t}[m(\sigma)\pm j\hat{m}(\sigma)]d\sigma}$	$e^{\mp D_f \int_{-\infty}^{t} \hat{m}(\sigma)d\sigma}\cos\left[D_f \int_{-\infty}^{t} m(\sigma)d\sigma\right]$	$e^{\mp D_f \int_{-\infty}^{t} \hat{m}(\sigma)d\sigma}\sin\left[D_f \int_{-\infty}^{t} m(\sigma)d\sigma\right]$	$e^{\mp D_f \int_{-\infty}^{t} \hat{m}(\sigma)d\sigma}$	$D_f \int_{-\infty}^{t} m(\sigma)d\sigma$	NL	
SSB-EV[a]	$e^{\{\ln[1 + m(t)]\pm j\hat{\ln}[1 + m(t)]\}}$	$[1 + m(t)]\cos\{\hat{\ln}[1 + m(t)]\}$	$\pm[1 + m(t)]\sin\{\hat{\ln}[1 + m(t)]\}$	$1 + m(t)$	$\pm\hat{\ln}[1 + m(t)]$	NL	$m(t) > -1$ is required so that the ln will have a real value.
SSB-SQ[a]	$e^{(1/2)\{\ln[1 + m(t)]\pm j\hat{\ln}[1 + m(t)]\}}$	$\sqrt{1 + m(t)}\cos\left\{\frac{1}{2}\hat{\ln}[1 = m(t)]\right\}$	$\pm\sqrt{1 + m(t)}\sin\left\{\frac{1}{2}\hat{\ln}[1 + m(t)]\right\}$	$\sqrt{1 + m(t)}$	$\pm\frac{1}{2}\hat{\ln}[1 + m(t)]$	NL	$m(t) > -1$ is required so that the ln will have a real value.
QM	$m_1(t) + jm_2(t)$	$m_1(t)$	$m_2(t)$	$\sqrt{m_1^2(t) + m_2^2(t)}$	$\tan^{-1}[m_2(t)/m_1(t)]$	L	Used in NTSC color television: requires coherent detection.

L = linear, NL = nonlinear, [ˆ.] is the Hilbert transform (i.e., –90° phase-shifted version) of [·]. The Hilbert transform is $\hat{x}(t) \triangleq x(t) * \frac{1}{\pi t} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(\lambda)}{t - \lambda}\, d\lambda$

[a]Use upper signs for upper sideband signals and lower signs for lower sideband signals.

[b]In the strict sense, AM signals are not linear because the carrier term does not satisfy the linearity (superposition) condition.

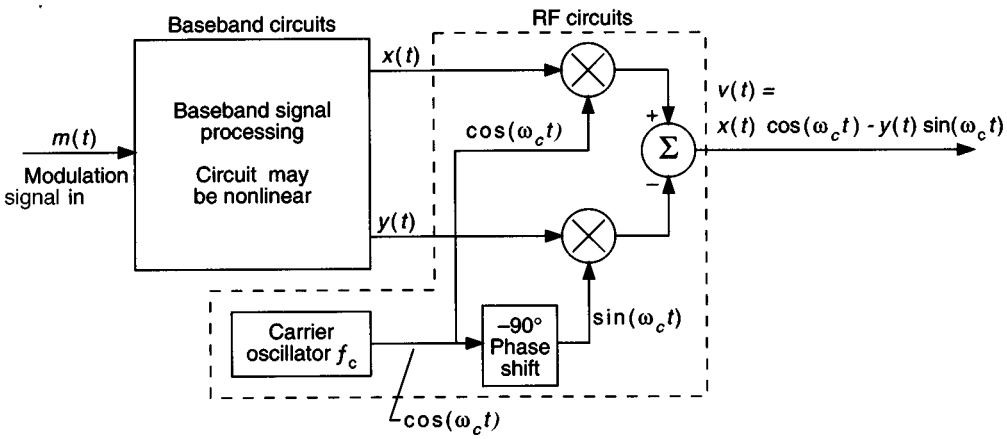


FIGURE 69.2 Generalized transmitter using the quadrature generation technique.

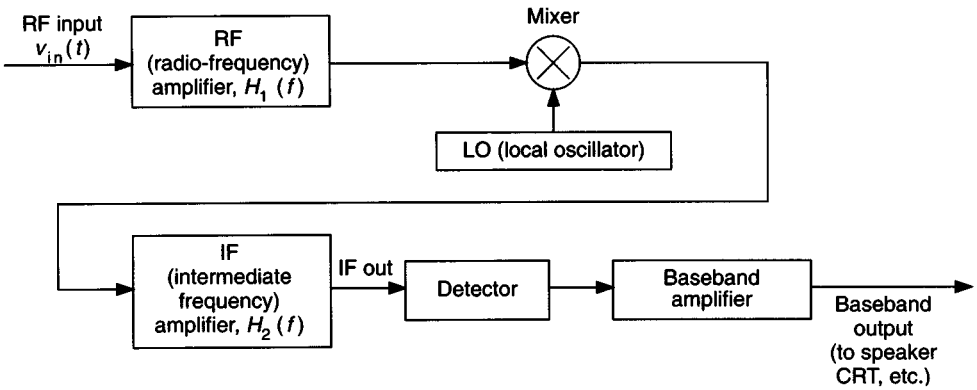


FIGURE 69.3 Superheterodyne receiver.

If the complex envelope $g(t)$ is desired for generalized signal detection or for optimum reception in digital systems, the $x(t)$ and $y(t)$ quadrature components, where $x(t) + jy(t) = g(t)$, may be obtained by using quadrature product detectors, as illustrated in Fig. 69.4. $x(t)$ and $y(t)$ could be fed into a signal processor to extract the modulation information. Disregarding the effects of noise, the signal processor could recover $m(t)$ from $x(t)$ and $y(t)$ (and, consequently, demodulate the IF signal) by using the inverse of the complex envelope generation functions given in Table 69.1.

The generalized modulation techniques are shown in Table 69.1. In digital communication systems, discrete modulation techniques are usually used to modulate the source information signal. Discrete modulation includes:

- PCM = pulse-code modulation
- DM = differential modulation
- DPCM = differential pulse-code modulation
- FSK = frequency-shift keying
- PSK = phase-shift keying
- DPSK = differential phase-shift keying
- MPSK = *M*-ary phase-shift keying
- QAM = quadrature amplitude modulation

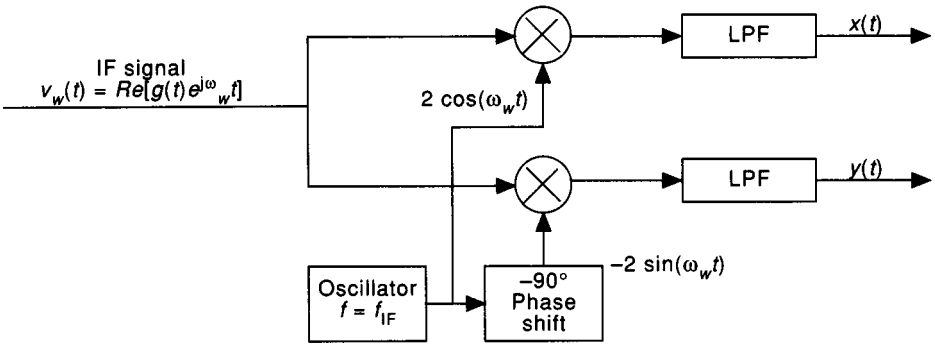


FIGURE 69.4 IQ (in-phase and quadrature-phase) detector.

TABLE 69.2 Performance of a PCM System with Uniform Quantizing and No Channel Noise

Number of Quantizer Levels Used, M	Length of the PCM Word, n (bits)	Bandwidth of PCMSignal (First Null Bandwidth)[a]	Recovered Analog Signal Power-to-Quantizing Noise Power Ratios	
			$(S/N)_{pk\,out}$	$(S/N)_{out}$
2	1	$2B$	10.8	6.0
4	2	$4B$	16.8	12.0
8	3	$6B$	22.8	18.1
16	4	$8B$	28.9	24.1
32	5	$10B$	34.9	30.1
64	6	$12B$	40.9	36.1
128	7	$14B$	46.9	42.1
256	8	$16B$	52.9	48.2
512	9	$18B$	59.0	54.2
1024	10	$20B$	65.0	60.2

[a] B is the absolute bandwidth of the input analog signal.

Pulse-Code Modulation

PCM is essentially analog-to-digital conversion of a special type, where the information contained in the instantaneous samples of an analog signal is represented by digital words in a serial bit stream. The PCM signal is generated by carrying out three basic operations: sampling, quantizing, and encoding (see Fig. 69.5). The sampling operation generates a flat-top pulse amplitude modulation (PAM) signal. The quantizing converts the actual sampled value into the nearest of the M amplitude levels. The PCM signal is obtained from the quantized PAM signal by encoding each quantized sample value into a digital word.

Frequency-Shift Keying

The FSK signal can be characterized as one of two different types. One type is called *discontinuous-phase* FSK since $\theta(t)$ is discontinuous at the switching times. The discontinuous-phase FSK signal is represented by

$$s(t) = \begin{cases} A_c \cos(\omega_1 t + \theta_1) \text{ for } t \text{ in time interval when a binary 1 is sent} \\ A_c \cos(\omega_2 t + \theta_2) \text{ for } t \text{ in time interval when a binary 0 is sent} \end{cases} \quad (69.5)$$

Radio Distance and Direction Indicator

Luis W. Alvarez
Patented August 30, 1949
#2,480,208

An excerpt from Luis Alvarez's patent application:

This invention relates to a communications system and more particularly to a system for presenting in panoramic form the location and disposition of objects as they might be seen from the air. In particular, the system hereinafter described is a radar or radio echo detection system presenting objects and targets principally on the ground lying in the path of flight of an airplane.

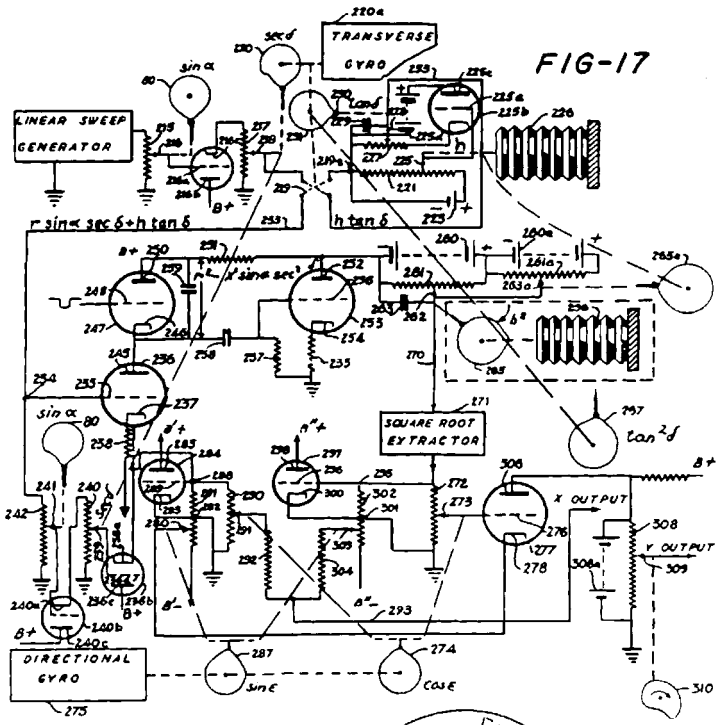


FIG-17

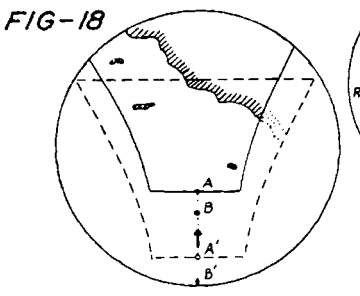


FIG-18

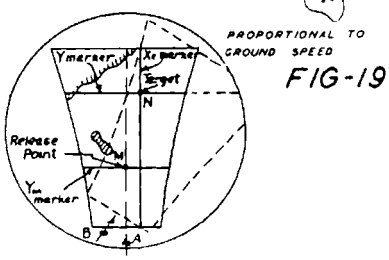


FIG-19

Ground radar systems were already known and used by the military. These involved a highly directional antenna alternately coupled to a transmitter and receiver with the antenna swept in a radial fashion. The display consisted of a cathode ray tube with targets represented by radial sweeps from the center of the screen. Dr. Alvarez took on the special problem of panoramic presentation of ground targets from aircraft. He solved the computation and display problems associated with the hyperbolic shape of the radar beams as transmitted and received from a moving aircraft. He also described handling pitch, roll, yaw, and other disturbances. (Copyright © 1995, DewRay Products, Inc. Used with permission.)

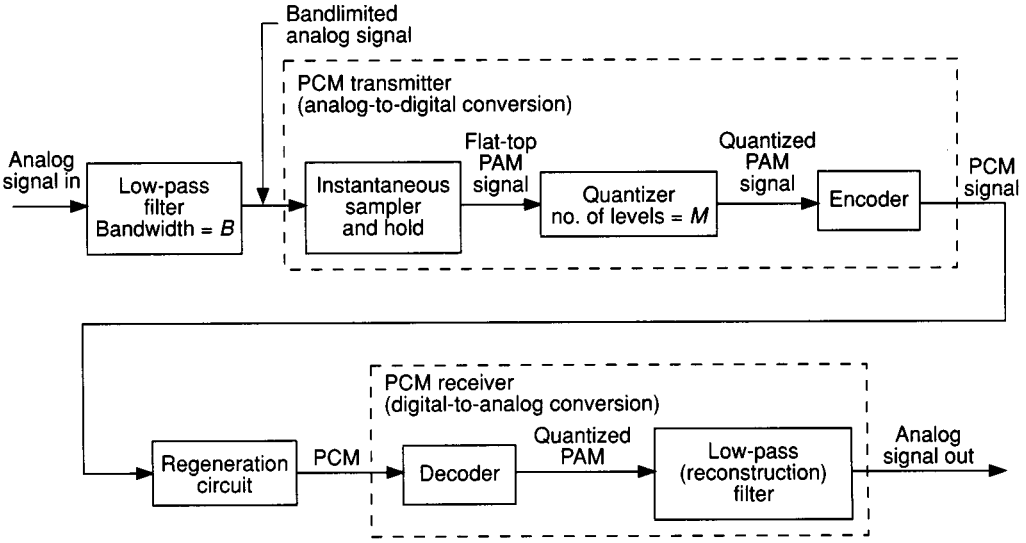


FIGURE 69.5 A PCM transmission system.

where f_1 is called the mark (binary 1) frequency and f_2 is called the space (binary 0) frequency. The other type is continuous-phase FSK. The continuous-phase FSK signal is generated by feeding the data signal into a frequency modulator, as shown in Fig. 69.6(b). This FSK signal is represented by

$$s(t) = A_c \cos\left[\omega_c t + D_f \int_{-\infty}^{t} m(\lambda)d\lambda\right]$$

or

$$s(t) = \mathrm{Re}\{g(t)e^{j\omega ct}\} \tag{69.6}$$

where

$$g(t) = A_c e^{j\theta(t)} \tag{69.7}$$

$$\theta(t) = D_f \int_{-\infty}^{t} m(\lambda)d\lambda \quad \text{for FSK} \tag{69.8}$$

Detection of FSK is illustrated in Fig. 69.7.

M-ary Phase-Shift Keying

If the transmitter is a PM transmitter with an *M*-level digital modulation signal, MPSK is generated at the transmitter output. A plot of the permitted values of the complex envelope, $g(t) = A_c e^{j\theta(t)}$, would contain *M* points, one value of *g* (a complex number in general) for each of the *M* multilevel values, corresponding to the *M* phases that θ is permitted to have.

MPSK can also be generated using two quadrature carriers modulated by the *x* and *y* components of the complex envelope (instead of using a phase modulator)

$$g(t) = A_c e^{j\theta(t)} = x(t) + jy(t) \tag{69.9}$$

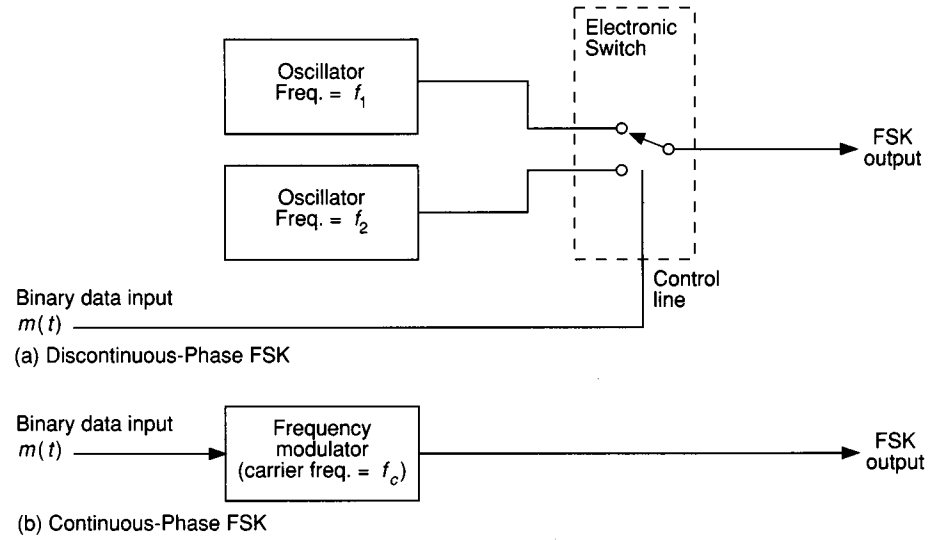


FIGURE 69.6 Generation of FSK.

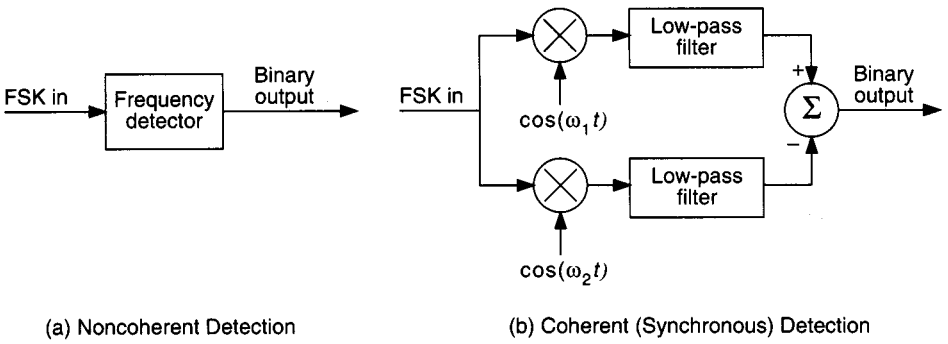


FIGURE 69.7 Detection of FSK.

where the permitted values of x and y are

$$x_i = A_c \cos \theta_i \tag{69.10}$$

$$y_i = A_c \sin \theta_i \tag{69.11}$$

for the permitted phase angles θ_i, $i = 1, 2, ..., M$, of the MPSK signal. This is illustrated by Fig. 69.8, where the signal processing circuit implements Eqs. (69.10) and (69.11).

MPSK, where $M = 4$, is called quadrature-phase-shift-keyed (QPSK) signaling.

Quadrature Amplitude Modulation

Quadrature carrier signaling is called quadrature amplitude modulation (QAM). In general, QAM signal constellations are not restricted to having permitted signaling points only on a circle (of radius A_c, as was the case for MPSK). The general QAM signal is

$$s(t) = x(t) \cos \omega_c t - y(t) \sin \omega_c t \tag{69.12}$$

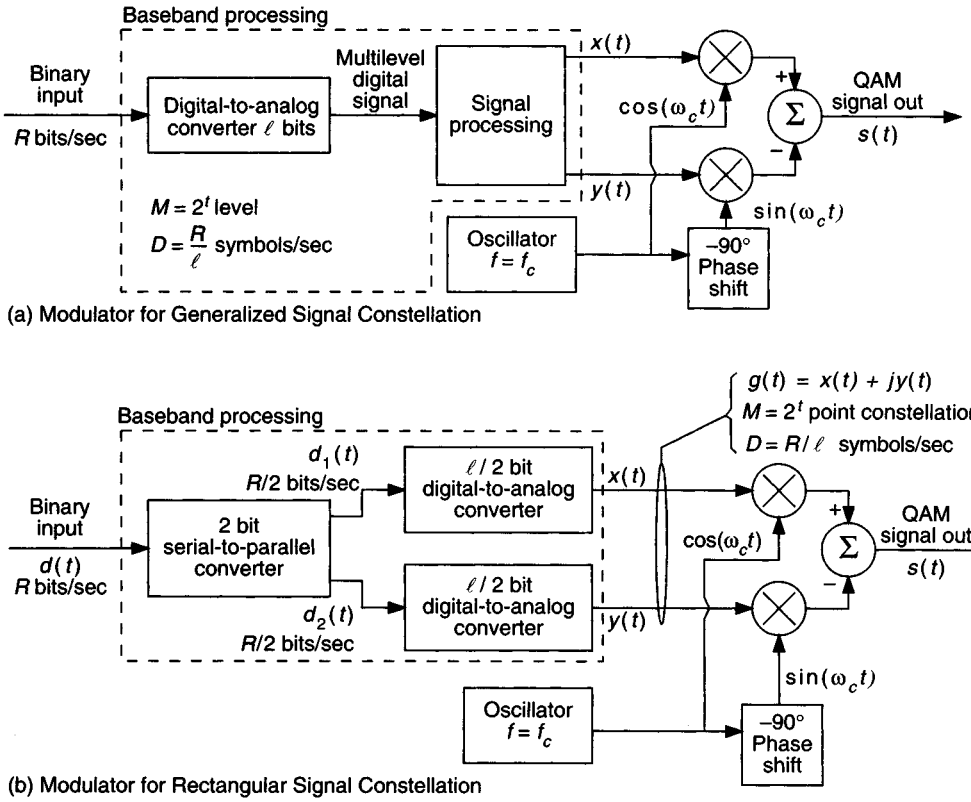


FIGURE 69.8 Generation of QAM signals.

TABLE 69.3 Spectral Efficiency for QAM Signaling with Raised Cosine-Roll-Off Pulse Shaping

Number of Levels, M (symbols)	Size of DAC, ℓ (bits)	$\eta = \frac{R}{B_T}$ bits/s/Hz					
		$r = 0.0$	$r = 0.1$	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$
2	1	1.00	0.909	0.800	0.667	0.571	0.500
4	2	2.00	1.82	1.60	1.33	1.14	1.00
8	3	3.00	2.73	2.40	2.00	1.71	1.50
16	4	4.00	3.64	3.20	2.67	2.29	2.00
32	5	5.00	4.55	4.0	3.33	2.86	2.50

DAC = digital-to-analog converter.
$\eta = R/B_T = \ell/2$ bits/s per hertz.
r is the roll-off factor of the filter characteristic.

where

$$g(t) = x(t) + jy(t) = R(t)e^{j\theta(t)} \tag{69.13}$$

The generation of QAM signals is shown in Fig. 69.8. The spectral efficiency for QAM signaling is shown in Table 69.3.

Defining Terms

Modulation: The process of impressing the source information onto a bandpass signal with a carrier frequency f_c. It can be expressed as

$$s(t) = \mathrm{Re}\{g(t)\ e^{j\omega ct}\}$$

where $g(t)$ is a function of the modulating signal $m(t)$. That is,

$$g(t) = g[m(t)]$$

$g[\cdot]$ performs a mapping operation on $m(t)$. The particular relationship that is chosen for $g(t)$ in terms of $m(t)$ defines the type of modulation used.

Superheterodyne receiver: Most receivers employ the superheterodyne receiving technique, which consists of either down-converting or up-converting the input signal to some convenient frequency band, called the intermediate frequency band, and then extracting the information (or modulation) by using an appropriate detector. This basic receiver structure is used for the reception of all types of bandpass signals, such as television, FM, AM, satellite, and radar signals.

Related Topics

69.2 Radio Broadcasting • 70.1 Coding

References

L. W. Couch, *Digital and Analog Communication Systems,* New York: Prentice-Hall, 1995.

F. Dejager, "Delta modulation of PCM transmission using a 1-unit code," Phillips Res. Rep., no. 7, pp. 442–466, Dec. 1952.

J.H. Downing, *Modulation Systems and Noise,* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

J. Dunlop and D.G. Smith, *Telecommunications Engineering,* London: Van Nostrand, 1989.

B.P. Lathi, *Modern Digital and Analog Communication Systems,* New York: CBS College, 1983.

J.H. Park, Jr., "On binary DPSK detection," *IEEE Trans. Commun.,* COM-26, pp. 484–486, 1978.

M. Schwartz, *Information Transmission, Modulation and Noise,* New York: McGraw-Hill, 1980.

Further Information

The monthly journal *IEEE Transactions on Communications* describes telecommunication techniques. The performance of *M*-ary QAM schemes is evaluated in its March 1991 issue, pp. 405–408. The IEEE magazine *IEEE Communications* is a valuable source.

Another source is *IEEE Transactions on Broadcasting,* which is published quarterly by The Institute of Electrical and Electronics Engineers, Inc.

The biweekly magazine *Electronics Letters* investigates the error probability of coherent PSK and FSK systems with multiple co-channel interferences in its April 11, 1991, issue, pp. 640–642. Another relevant source regarding the coherent detection of MSK is described on pp. 623–625 of the same issue. All subscriptions inquiries and orders should be sent to IEE Publication Sales, P.O. Box 96, Stevenage, Herts, SG1 2SD, United Kingdom.

69.2 Radio Broadcasting

Jefferson F. Lindsey III and Dennis F. Doelitzsch

Standard Broadcasting (Amplitude Modulation)

Standard broadcasting refers to the transmission of voice and music received by the general public in the 535- to 1705-kHz frequency band. Amplitude modulation is used to provide service ranging from that needed for small communities to higher-power broadcast stations needed for larger regional areas. The ***primary service***

The Revolutionary Technology of Radio

The beginning of the present century saw the birth of several technologies that were to be revolutionary in their impact. The most exciting of these was radio or, as it was generally called at the time, "wireless". No other technology would seem to obliterate the barriers of distance in human communication or to bring individuals together with such immediacy and spontaneity. And seldom had there emerged an activity that seemed so mysterious and almost magical to most of the population.

Radio was mysterious not only to the layman, but also to many engineers and technically informed individuals. The mystery lay largely in radio's application of principles and phenomena only recently identified by physicists and engineers working at the frontiers of their specialties. The existence of electromagnetic waves that traveled like light had been predicted by the brilliant physicist James Clerk Maxwell in the 1860s and proven by the young German Heinrich Hertz in the 1880s. The possible use of these waves for communicating through space without wires occurred to many; however, the first practical steps to making radio useful are generally attributed to Oliver Lodge in England, Guglielmo Marconi in Italy, and Aleksandr Popov in Russia. Marconi's broadcast of Morse code across the Atlantic in 1901 first showed the world just what enormous potential radio had for changing the whole concept of long-distance communication. The next few years saw feverish activity everywhere as men tried to translate the achievements of the pioneers into the foundations of a practical technology.

By 1912, radio technology had attracted a small number of dedicated individuals who identified their own future with the progress of their chosen field. Some of these had organized themselves into small, localized societies, but it was clear to many that a broader vision was needed if radio practitioners were to achieve the recognition and respect of technical professionals. It was with such a vision in mind that representatives of two of these local societies met in New York City in May 1912 to form the Institute of Radio Engineers. The IRE was to be an international society dedicated to the highest professional standards and to the advancement of the theory and practice of radio technology.

The importance of radio lay not simply in its expansion of the means of human communication over distances, but also in its exploitation and expansion of very novel scientific and technical capabilities. As the century progressed, radio would give rise to the 20th century's most revolutionary technology of all — electronics. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

area is defined as the area in which the groundwave signal is not subject to objectionable interference or objectionable fading. The ***secondary service*** *area* refers to an area serviced by skywaves and not subject to objectionable interference. *Intermittent service area* refers to an area receiving service from either a groundwave or a skywave but beyond the primary service area and subject to some interference and fading.

Frequency Allocations

The carrier frequencies for standard broadcasting in the United States (referred to internationally as medium-wave broadcasting) are designated in the Federal Communications Commission (FCC) Rules and Regulations, Vol. III, Part 73. A total of 117 carrier frequencies are allocated from 540 to 1700 kHz in 10-kHz intervals. Each carrier frequency is required by the FCC rules to deviate no more than ±20 Hz from the allocated frequency, to minimize heterodyning from two or more interfering stations. Double-sideband full-carrier modulation, commonly called *amplitude modulation* (AM), is used in standard broadcasting for sound transmission. Typical modulation frequencies for voice and music range from 50 Hz to 10 kHz. Each channel is generally thought of as 10 kHz in width, and thus the frequency band is designated from 535 to 1705 kHz; however, when the modulation frequency exceeds 5 kHz, the radio frequency bandwidth of the channel exceeds 10 kHz and

adjacent channel interference may occur. To improve the high-frequency performance of transmission and to compensate for the high-frequency roll-off of many consumer receivers, FCC rules require that stations boost the high-frequency amplitude of transmitted audio using preemphasis techniques. In addition stations may also use multiplexing to transmit stereophonic programming. The FCC adopted Motorola's C-QUAM compatible quadrature amplitude modulation in 1994. Approximately 700 AM stations transmit in stereo.

Channel and Station Classifications

In standard broadcast (AM), stations are classified according to their operating power, protection from interference, and hours of operation. A Class A station operates with 10 to 50 kW of power servicing a large area with primary, secondary, and intermittent coverage and is protected from interference both day and night. These stations are called "clear channel" stations because the channel is cleared of nighttime interference over a major portion of the country. Class B stations operate full time with transmitter powers of 0.25 to 50 kW and are designed to render primary service only over a principal center of population and the rural area contiguous thereto. While nearly all Class A stations operate with 50 kW, most Class B stations must restrict their power to 5 kW or less to avoid interfering with other stations. Class B stations operating in the 1605 to 1705 kHz band are restricted to a power level of 10 kW daytime and 1 kW nighttime. Class C stations operate on six designated channels (1230, 1240, 1340, 1400, 1450, and 1490) with a maximum power of 1 kW or less full time and render primarily local service to smaller communities. Class D stations operate on Class A or B frequencies with Class B transmitter powers during daytime, but nighttime operation, if permitted at all, must be at low power (less than 0.25 kW) with no protection from interference.

Although Class A stations cover large areas at night, approximately in a 1220-km (750-mi) radius, the nighttime coverage of Class B, C, and D stations is limited by interference from other stations, electrical devices, and atmospheric conditions to a relatively small area. Class C stations, for example, have an interference-free nighttime coverage radius of approximately 8 to 16 km. As a result, there may be large differences in the area that the station covers daytime versus nighttime. With over 5200 AM stations licensed for operation by the FCC, interference, both day and night, is a factor that significantly limits the service which stations may provide. In the absence of interference, a daytime signal strength of 2 mV/m is required for reception in populated areas of more than 2500, while a signal of 0.5 mV/m is generally acceptable in less populated areas. Secondary nighttime service is provided in areas receiving a 0.5-mV/m signal 50% or more of the time without objectionable interference. Table 69.4 indicates the daytime contour overlap limits. However, it should be noted that these limits apply to new stations and modifications to existing stations. Nearly every station on the air was allocated prior to the implementation of these rules when the interference criteria were less restrictive.

Field Strength

The field strength produced by a standard broadcast station is a key factor in determining the primary and secondary service areas and interference limitations of possible future radio stations. The field strength limitations are specified as field intensities by the FCC with the units volts per meter; however, measuring devices may read volts or decibels referenced to 1 mW (dBm), and a conversion may be needed to obtain the field intensity. The power received may be measured in dBm and converted to watts. Voltage readings may be converted to watts by squaring the root mean square (rms) voltage and dividing by the field strength meter input resistance, which is typically on the order of 50 or 75 Ω. Additional factors needed to determine **electric field intensity** are the power gain and losses of the field strength receiving antenna system. Once the power gain and losses are known, the effective area with loss compensation of the field strength receiver antenna may be obtained as

$$A_{\text{eff}} = G \frac{\lambda^2}{4\pi} L \tag{69.14}$$

where A_{eff} = effective area including loss compensation, m^2; G = power gain of field strength antenna, W/W; λ = wavelength, m; and L = mismatch loss and cable loss factor, W/W.

From this calculation, the power density in watts per square meter may be obtained by dividing the received power by the effective area, and the electric field intensity may be calculated as

TABLE 69.4 Protected Service Signal Intensities for Standard Broadcasting (AM)

Class of Station	Power (kW)	Class of Channel Used	Signal Strength Contour of Area Protected from Objectionable Interference* (μV/m)		Permissible Interfering Signal	
			Day†	Night	Day†	Night‡
A	10–50	Clear	SC 100 AC 500	SC 500 50% SW AC 500 GW	SC 5 AC 250	SC 25 AC 250
B	0.25–50	Clear Regional	500	2000†	25 AC 250	25 250
C	0.25–1	Local	500	Not precise§	SC 25	Not precise
D	0.25–50	Clear Regional	500	Not precise	SC 25 AC 250	Not precise

*When a station is already limited by interference from other stations to a contour of higher value than that normally protected for its class, this higher-value contour shall be the established protection standard for such station. Changes proposed by Class A and B stations shall be required to comply with the following restrictions. Those interferers that contribute to another station's RSS using the 50% exclusion method are required to reduce their contribution to that RSS by 10%. Those lesser interferers that contribute to a station's RSS using the 25% exclusion method but do not contribute to that station's RSS using the 50% exclusion method may make changes not to exceed their present contribution. Interferers not included in a station's RSS using the 25% exclusion method are permitted to increase radiation as long as the 25% exclusion threshold is not equaled or exceeded. In no case will a reduction be required that would result in a contributing value that is below the pertinent value specified in the table.

†Groundwave.

‡Skywave field strength for 10% or more of the time. For Alaska, Class SC is limited to 5 μV/m.

§During nighttime hours, Class C stations in the contiguous 48 states may treat all Class B stations assigned to 1230, 1240, 1340, 1400, 1450, and 1490 kHz in Alaska, Hawaii, Puerto Rico and the U.S. Virgin Islands as if they were Class C stations.

Note: SC = same channel; AC = adjacent channel; SW = skywave; GW = groundwave; RSS = root of sum squares.

Source: FCC Rules and Regulations, Revised 1991; vol. III, pt. 73.182(a).

$$E = \sqrt{\mathcal{P} Z_{fs}} \tag{69.15}$$

where E = electric field intensity, V/m; $\mathcal{P}$ = power density, W/m^2; and $Z_{fs} = 120\pi\ \Omega$, impedance of free space.

The protected service contours and permissible interference contours for standard broadcast stations shown in Table 69.4, along with a knowledge of the field strength of existing broadcast stations, may be used in determining the potential for establishing new standard broadcast stations.

Propagation

One of the major factors in the determination of field strength is the propagation characteristic that is described by the change in electric field intensity with an increase in distance from the broadcast station antenna. This variation depends on a number of factors including frequency, distance, surface dielectric constant, surface loss tangent, polarization, local topography, and time of day. Generally speaking, groundwave propagation occurs at shorter ranges both during day and night periods. Skywave propagation permits longer ranges and occurs during night periods, and thus some stations must either reduce power or cease to operate at night to avoid causing interference. Propagation curves in the broadcast industry are frequently referred to a reference level of 100 mV/m at 1 km; however, a more general expression of groundwave propagation may be obtained by using the Bremmer series [Bremmer, 1949]. A typical groundwave propagation curve for electric field strength as a function of distance is shown in Fig. 69.9 for an operating frequency of 770–810 kHz. The ground conductivity varies from 0.1 to 5000 mS/m, and the ground relative dielectric constant is 15.

The **effective radiated power** (ERP) refers to the effective power output from the antenna in a specified direction and includes the transmitter power output, transmission line losses, and antenna power gain. The ERP in most cases exceeds the transmitter output power, since that antenna power gain is normally 2 or more. For a hypothetical perfect isotropic radiator with a power gain of 1, the ERP is found to be

$$\text{ERP} = \frac{E^2 r^2}{30} \tag{69.16}$$

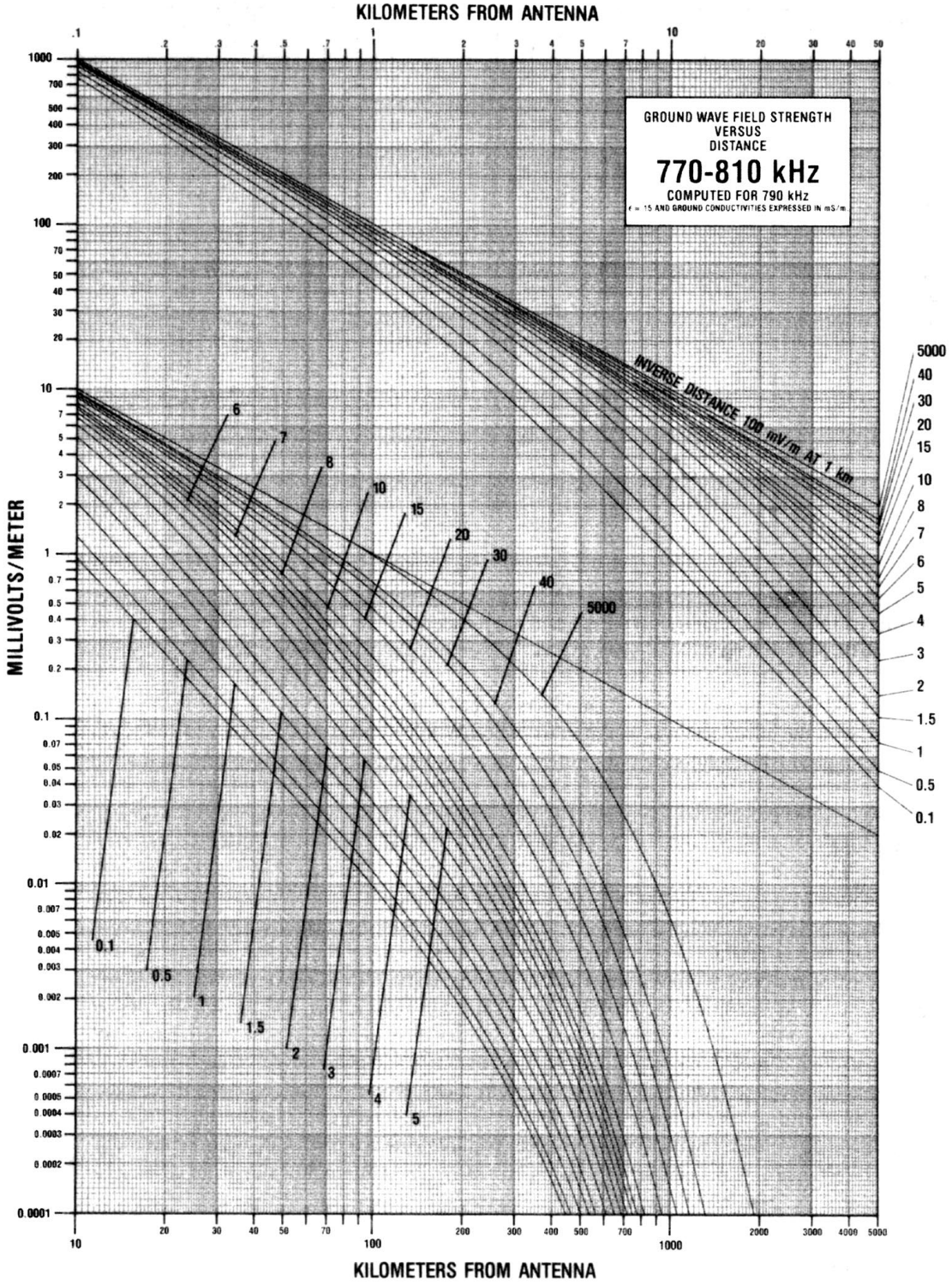


FIGURE 69.9 Typical groundwave propagation for standard AM broadcasting. (*Source:* 1986 National Association of Broadcasters.)

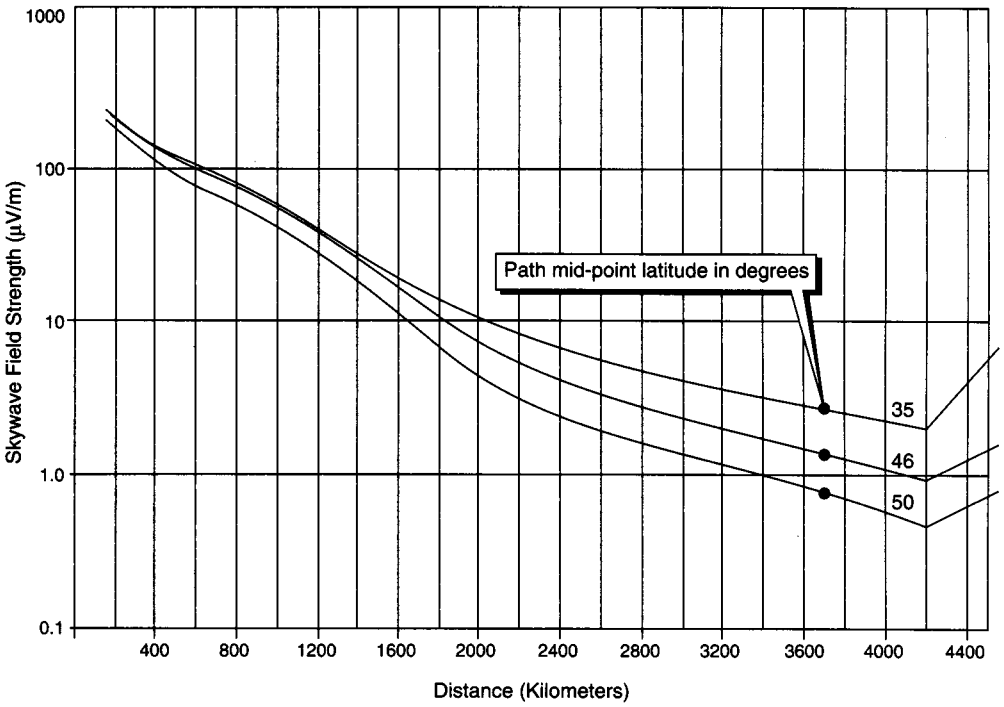


FIGURE 69.10 Skywave propagation for standard AM broadcasting. (*Source:* FCC Rules and Regulations, 1982, vol. III, pt. 73.190, fig. 2.)

where E is the electric field intensity, V/m, and r is the distance, m. For a distance of 1 km (1000 m), the ERP required to produce a field intensity of 100 mV/m is found to be 333.3 W. Since the field intensity is proportional to the square root of the power, field intensities may be determined at other powers.

Skywave propagation necessarily involves some fading and less predictable field intensities and is most appropriately described in terms of statistics or the percentage of time a particular field strength level is found. Figure 69.10 shows skywave propagation for a 100-mV/m field strength at a distance of 1 km for midpoint path latitudes of 35 to 50 degrees.

Transmitters

Standards that cover AM broadcast transmitters are given in the Electronic Industry Association (EIA) Standard TR-101A, "Electrical Performance Standard for Standard Broadcast Transmitters." Parameters and methods for measurement include the following: carrier output rating, carrier power output capability, carrier frequency range, carrier frequency stability, carrier shift, carrier noise level, magnitude of radio frequency (RF)harmonics, normal load, transmitter output circuit adjustment facilities, RF and audio interface definitions, modulation capability, audio input level for 100% modulation, audio frequency response, audio frequency harmonic distortion, rated power supply, power supply variation, operating temperature characteristics, and power input.

Standard AM broadcast transmitters range in power output from 5 W up to 50 kW units. While solid-state devices are used for many models (especially the lower-powered units), several manufacturers still retain tubes in the final amplifiers of their high-powered models. This is changing, however, with the introduction in recent years of 50-kW fully transistorized models. A block diagram of a typical 1-kW solid-state transmitter is shown in Fig. 69.11.

Antenna Systems

The antenna system for a standard AM broadcast station typically consists of a quarter-wave vertical tower, a ground system of 120 or more quarter-wave radials buried a few inches underground, and an antenna tuning

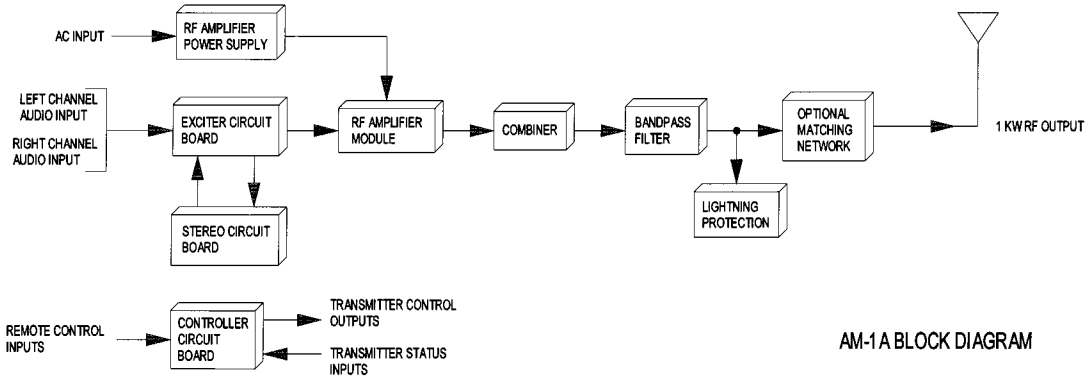


FIGURE 69.11 Block diagram of typical 1-kW solid-state AM transmitter. (*Source:* Broadcast Electronics Inc., Quincy, Ill. Reprinted with permission.)

unit to "match" the complex impedance of the antenna system to the characteristic impedance of the transmitter and transmission line so that maximum transfer of power may occur. Typical heights for AM broadcast towers range from 150 to 500 ft. When the radiated signal must be modified to prevent interference to other stations or to provide better service in a particular direction, additional towers may be combined in a phased array to produce the desired field intensity contours. For example, if a station power increase would cause interference with existing stations, a directional array could be designed that would tailor the coverage to protect the existing stations while allowing increases in other directions. The protection requirements can generally be met with arrays consisting of 4 towers or less, but complex arrays have been constructed consisting of 12 or more towers to meet stringent requirements at a particular location. An example of a directional antenna pattern is shown in Fig. 69.12. This pattern provides major coverage to the southwest and restricts radiation (and thus interference) towards the northeast.

Frequency Modulation

Frequency-modulation (FM) broadcasting refers to the transmission of voice and music received by the general public in the 88- to 108-MHz frequency band. FM is used to provide higher-fidelity reception than is available with standard broadcast AM. In 1961 stereophonic broadcasting was introduced with the addition of a double-sideband suppressed carrier for transmission of a left-minus-right difference signal. The left-plus-right sum channel is sent with use of normal FM. Some FM broadcast systems also include a **subsidiary communications authorization (SCA)** subcarrier for private commercial uses. FM broadcast is typically limited to line-of-sight ranges. As a result, FM coverage is localized to a range of 75 mi (120 km) depending on the antenna height and ERP.

Frequency Allocations

The 100 carrier frequencies for FM broadcast range from 88.1 to 107.9 MHz and are equally spaced every 200 kHz. The channels from 88.1 to 91.9 MHz are reserved for educational and noncommercial broadcasting and those from 92.1 to 107.9 MHz for commercial broadcasting. Each channel has a 200-kHz bandwidth. The maximum frequency swing under normal conditions is ±75 kHz. Stations operating with an SCA may under certain conditions exceed this level, but in no event may exceed a frequency swing of ±82.5 kHz. The carrier frequency is required to be maintained within ±2000 Hz. The frequencies used for FM broadcasting generally limit the coverage to the line-of-sight or a slightly greater distance. The actual coverage area is determined by the ERP of the station and the height of the transmitting antenna above the average terrain in the area. Either increasing the power or raising the antenna will increase the coverage area.

Station Classifications

In FM broadcast, stations are classified according to their maximum allowable ERP and the transmitting antenna height above average terrain in their service area. Class A stations provide primary service to a radius of about

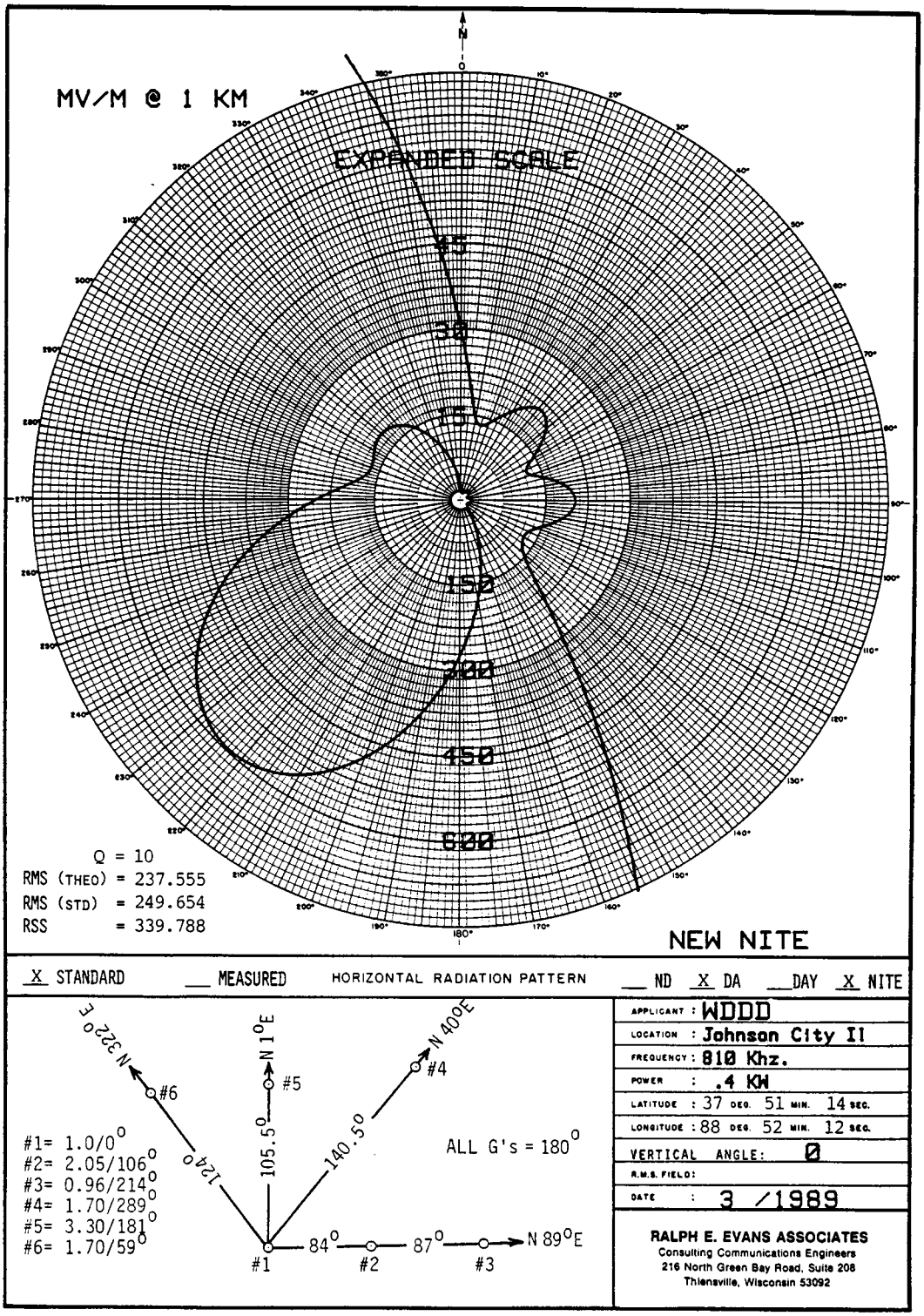


FIGURE 69.12 Directional AM antenna pattern for a six-element array. (*Source:* WDDD-AM, Marion, Ill., and Ralph Evans Associates.)

TABLE 69.5 FM Station Classifications, Powers, and Tower Heights

Station Class	Maximum ERP	HAAT, m (ft)	Distance, km
A	6 kW (7.8 dBk)	100 (328)	28
B1	25 kW (14.0 dBk)	100 (328)	39
B	50 kW (17.0 dBk)	150 (492)	52
C3	25 kW (14.0 dBk)	100 (328)	39
C2	50 kW (17.0 dBk)	150 (492)	52
C1	100 kW (20.0 dBk)	299 (981)	72
C	100 kW (20.0 dBk)	600 (1968)	92

Source: FCC Rules and Regulations, Revised 1991; vol. III, Part 73.211(b)(1).

28 km with 6000 W of ERP at a maximum height of 100 m. The most powerful class, Class C, operates with maximums of 100,000 W of ERP and heights up to 600 m with a primary coverage radius of over 92 km. The powers and heights above average terrain (HAAT) for all of the classes are shown in Table 69.5. All classes may operate at antenna heights above those specified but must reduce the ERP accordingly. Stations may not exceed the maximum power specified, even if antenna height is reduced. The classification of the station determines the allowable distance to other co-channel and adjacent channel stations.

Field Strength and Propagation

The field strength produced by an FM broadcast station depends on the ERP, antenna heights, local terrain, tropospheric scattering conditions, and other factors. From a statistical point of view, however, an estimate of the field intensity may be obtained from Fig. 69.13. A factor in the determination of new licenses for FM broadcast is the separation between allocated co-channel and adjacent channel stations, the class of station, and the antenna heights. The spacings are given in Table 69.6. The primary coverage of all classes of stations (except B and B1, which are 0.5 mV/m and 0.7 mV/m, respectively) is the 1.0 mV/m contour. The distance to the primary contour, as well as to the "city grade" or 3.16 mV/m contour may be estimated using Fig. 69.13. Although FM broadcast propagation is generally thought of as line-of-sight, larger ERPs along with the effects of diffraction, refraction, and tropospheric scatter allow coverage slightly greater than line-of-sight.

Transmitters

FM broadcast transmitters typically range in power output from 10 W to 50 kW. A block diagram of a dual FM transmitter is shown in Fig. 69.14. This system consists of two 25-kW transmitters that are operated in parallel and that provide increased reliability in the event of a failure in either the exciter or transmitter power amplifier. The highest-powered solid-state transmitters are currently 10 kW, but manufacturers are developing new devices that will make higher-power solid-state transmitters both cost-efficient and reliable.

Antenna Systems

FM broadcast antenna systems are required to have a horizontally polarized component. Most antenna systems, however, are circularly polarized, having both horizontal and vertical components. The antenna system, which usually consists of several individual radiating bays fed as a phased array, has a radiation characteristic that concentrates the transmitted energy in the horizontal plane toward the population to be served, minimizing the radiation out into space and down toward the ground. Thus, the ERP towards the horizon is increased with gains up to 10 dB. This means that a 5-kW transmitter coupled to an antenna system with a 10-dB gain would have an ERP of 50 kW. Directional antennas may be employed to avoid interference with other stations or to meet spacing requirements. Figure 69.15 is a plot of the horizontal and vertical components of a typical nondirectional circularly polarized FM broadcast antenna showing the effect upon the pattern caused by the supporting tower.

Preemphasis

Preemphasis is employed in an FM broadcast transmitter to improve the received signal-to-noise ratio. The preemphasis upper-frequency limit shown is based on a time constant of 75 μs as required by the FCC for FM

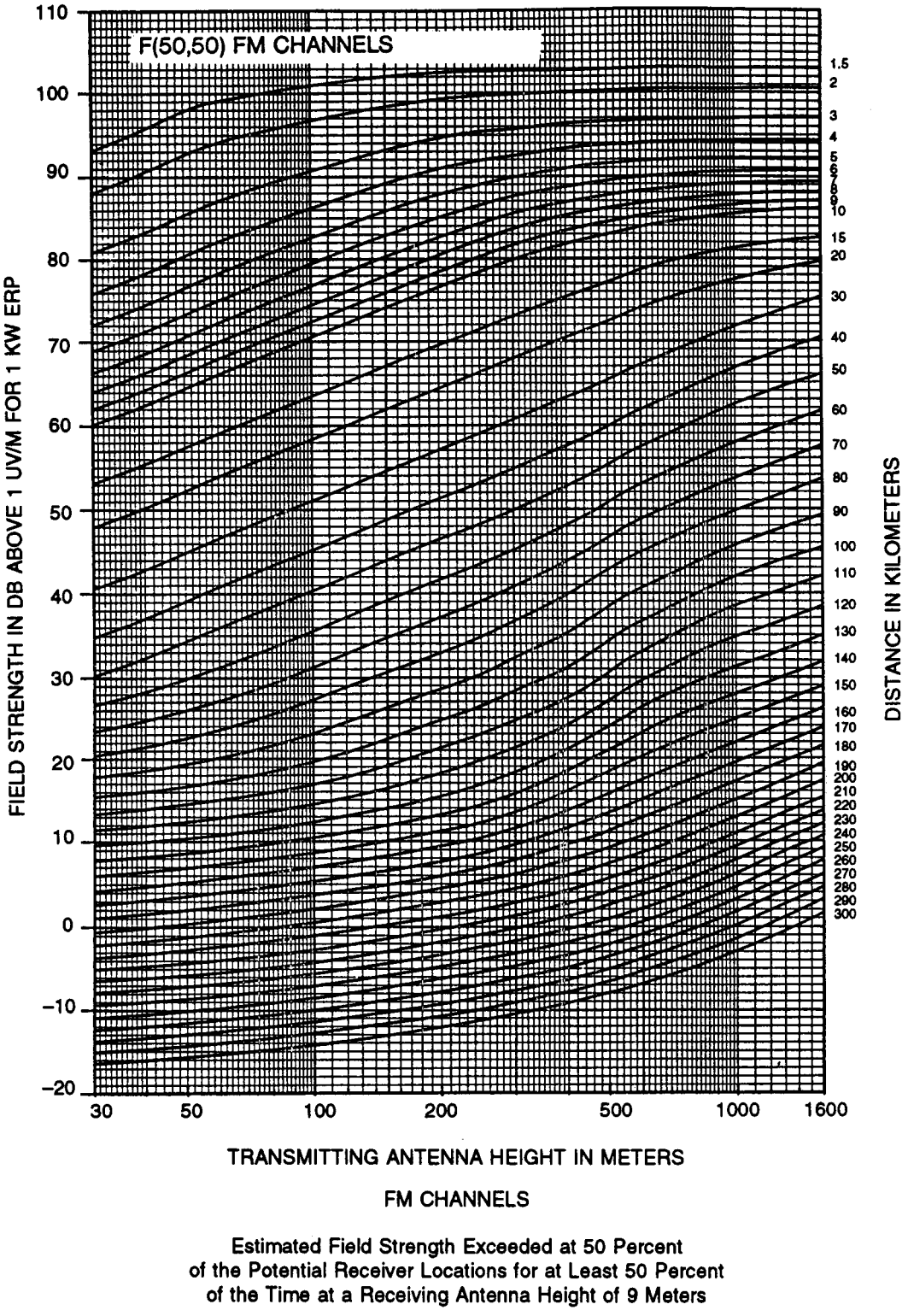


FIGURE 69.13 Propagation for FM broadcasting. (*Source:* FCC Rules and Regulations, Revised 1990; vol. III, pt. 73.333.)

TABLE 69.6 Distance Separation Requirement for FM Stations

Station Class Relation	Minimum Distance Separation Requirements, km (mi)			
	Co-Channel	200 kHz	400/600 kHz	10.6/10.8 MHz
A to A	115 (71)	72 (45)	31 (19)	10 (6)
A to B1	143 (89)	96 (60)	48 (30)	12 (7)
A to B	178 (111)	113 (70)	69 (43)	15 (9)
A to C3	142 (88)	89 (55)	42 (26)	12 (7)
A to C2	166 (103)	106 (66)	55 (34)	15 (9)
A to C1	200 (124)	133 (83)	75 (47)	22 (14)
A to C	226 (140)	165 (103)	95 (59)	29 (18)
B1 to B1	175 (109)	114 (71)	50 (31)	14 (9)
B1 to B	211 (131)	145 (90)	71 (44)	17 (11)
B1 to C3	175 (109)	114 (71)	50 (31)	14 (9)
B1 to C2	200 (124)	134 (83)	56 (35)	17 (11)
B1 to C1	233 (145)	161 (100)	77 (48)	24 (15)
B1 to C	259 (161)	193 (120)	105 (65)	31 (19)
B to B	241 (150)	169 (105)	74 (46)	20 (12)
B to C3	211 (131)	145 (90)	71 (44)	17 (11)
B to C2	211 (131)	145 (90)	71 (44)	17 (11)
B to C1	270 (168)	195 (121)	79 (49)	27 (17)
B to C	274 (170)	217 (135)	105 (65)	35 (22)
C3 to C3	153 (95)	99 (62)	43 (27)	14 (9)
C3 to C2	177 (110)	117 (73)	56 (35)	17 (11)
C3 to C1	211 (131)	144 (90)	76 (47)	24 (15)
C3 to C	237 (147)	176 (109)	96 (60)	31 (19)
C2 to C2	190 (118)	130 (81)	58 (36)	20 (12)
C2 to C1	224 (139)	158 (98)	79 (49)	27 (17)
C2 to C	237 (147)	176 (109)	96 (60)	31 (19)
C1 to C1	245 (152)	177 (110)	82 (51)	34 (21)
C1 to C	270 (168)	209 (130)	105 (65)	35 (22)
C to C	290 (180)	241 (150)	105 (65)	48 (30)

Source: FCC Rules and Regulations, Revised 1991; vol. III, pt. 73.207.

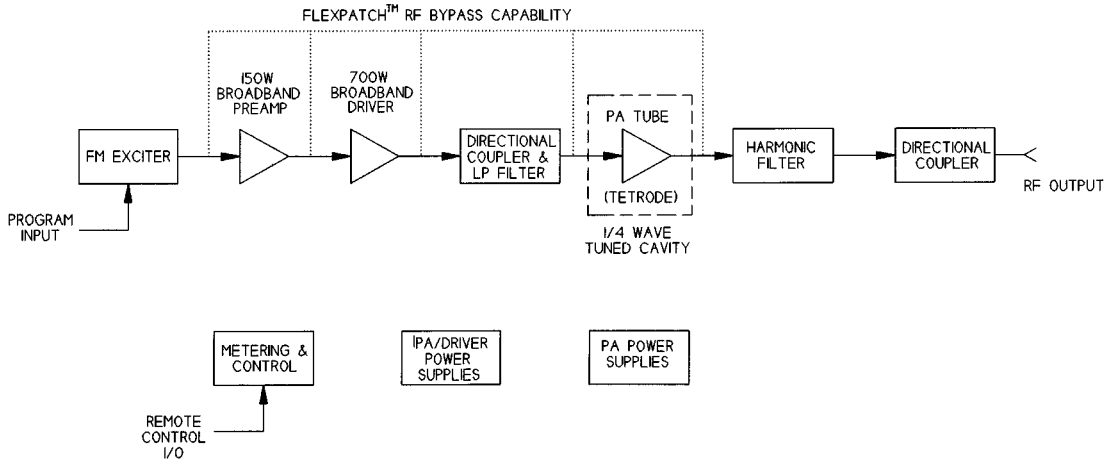


FIGURE 69.14 Block diagram of typical FM transmitter. (*Source:* Harris Corporation, Quincy, Ill.)

broadcast transmitters. Audio frequencies from 50 to 2120 Hz are transmitted with normal FM, whereas audio frequencies from 2120 Hz to 15 kHz are emphasized with a larger modulation index. There is significant signal-to-noise improvement when the receiver is equipped with a matching deemphasis circuit.

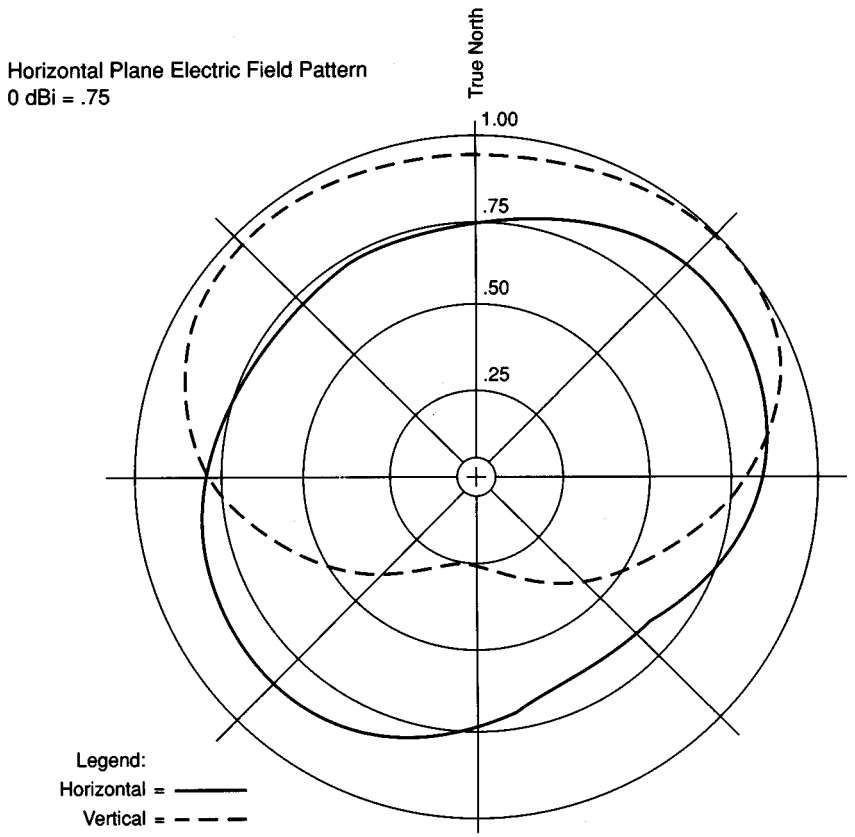


FIGURE 69.15 Typical nondirectional 92.5-MHz FM antenna characteristics showing the effect of the tower structure. (*Source:* Electronics Research, Inc., Newburgh, Ind.)

FM Spectrum

The monophonic system was initially developed to allow sound transmissions for audio frequencies from 50 to 15,000 Hz to be contained within a ±75-kHz RF bandwidth. With the development of FM stereo, the original FM signal (consisting of a left-plus-right channel) is transmitted in a smaller bandwidth to be compatible with a monophonic FM receiver, and a left-minus-right channel is frequency-multiplexed on a subcarrier of 38-kHz using double-sideband suppressed carrier. An unmodulated 19-kHz subcarrier is derived from the 38-kHz subcarrier to provide a synchronous demodulation reference for the stereophonic receiver. The synchronous detector at 38 kHz recovers the left-minus-right channel information, which is then combined with the left-plus-right channel information in sum and difference combiners to produce the original left-channel and right-channel signals. In addition stations may utilize an SCA in a variety of ways, such as paging, data transmission, specialized foreign language programs, radio reading services, utility load management, and background music. An FM stereo station may utilize multiplex subcarriers within the range of 53 to 99 kHz with up to 20% modulation of the main carrier using any form of modulation. The only requirement is that the station does not exceed its occupied bandwidth limitations.

Defining Terms

Effective radiated power: Refers to the effective power output from an antenna in a specified direction and includes transmitter output power, transmission line loss and antenna power gain.

Electric field intensity: Measure of signal strength in volts per meter used to determine channel allocation criteria and interference considerations.

Primary service: Refers to areas in which the groundwave signal is not subject to objectionable interference or objectionable fading.

SCA: Subsidiary communications authorization for paging, data transmission, specialized foreign language programs, radio readings services, utility load management and background music using multiplexed subcarriers from 53–99 kHz in connection with broadcast FM.

Secondary service: Refers to areas serviced by skywaves and not subject to objectionable interference.

Related Topics

69.1 Modulation and Demodulation • 38.1 Wire

References

A. F. Barghausen, "Medium frequency sky wave propagation in middle and low latitudes," *IEEE Trans. Broadcast,* vol. 12, pp. 1–14, June 1966.

G.W. Bartlett, Ed., *National Association of Broadcasters Engineering Handbook,* 6th ed., Washington: The National Association of Broadcasters, 1975.

H. Bremmer, *Terrestrial Radio Waves: Theory of Propagation*, Amsterdam: Elsevier, 1949.

Electronic Industries Association, Standard TR-101A, *Electrical Performance Standards for AM Broadcast Transmitters*, 1948.

Federal Communications Commission, Rules and Regulations, vol. III, parts 73 and 74, October 1982.

Further Information

Pike & Fischer, Inc., in Bethesda, Md., offers an updated FCC rule service for a fee.

Several trade journals are good sources for up-to-date information such as *Broadcast Engineering*, Overland Park, Kan., and *Radio World*, Falls Church, Va.

Application-oriented computer software is available from R.F. Systems, Shawnee Mission, Kan.

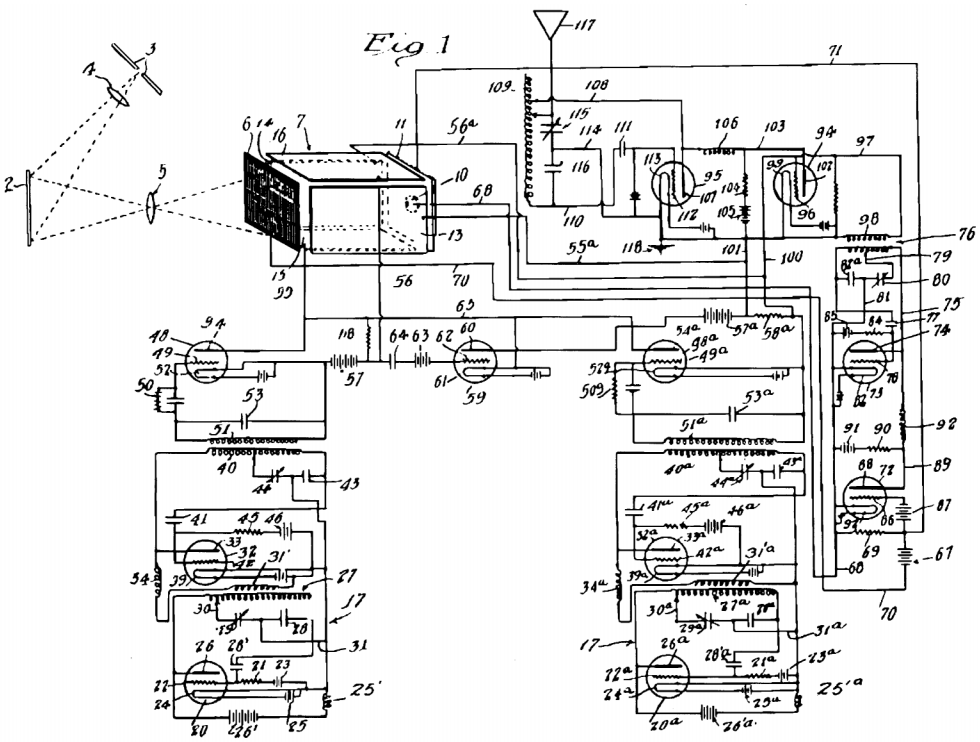
The Society of Broadcast Engineers (SBE), Indianapolis, Ind., and the National Association of Broadcasters (NAB), Washington, D.C., are sources of further information.

69.3 Television Systems

Jerry Whitaker

The technology of television is based on the conversion of light rays from still or moving scenes and pictures into electronic signals for transmission or storage, and subsequent reconversion into visual images on a screen. A similar function is provided in the production of motion picture film; however, where film records the brightness variations of a complete scene on a single frame in a short exposure no longer than a fraction of a second, the elements of a television picture must be scanned one piece at a time. In the television system, a scene is dissected into a **frame** composed of a mosaic of *picture elements* (pixels). A **pixel** is defined as the smallest area of a television image that can be transmitted within the parameters of the system. This process is accomplished by:

- Analyzing the image with a photoelectric device in a sequence of *horizontal* ***scans*** from the top to the bottom of the image to produce an electric signal in which the brightness and color values of the individual picture elements are represented as voltage levels of a video waveform
- Transmitting the values of the picture elements in sequence as voltage levels of a video signal
- Reproducing the image of the original scene in a video signal display of parallel scanning lines on a viewing screen



Television System

Philo T. Farnsworth
Patented August 26, 1930
#1,773,980

An excerpt from Philo Farnsworth's patent application:

In the process and apparatus of the present invention, light from all portions of the object whose image is to be transmitted, is focused at one time upon a light sensitive plate of a photo-electrical cell to thereby develop an electronic discharge from said plate, in which each portion of the cross-section of such electronic discharge will correspond in electrical intensity with the intensity of light imposed on that portion of the sensitive plate from which the electrical discharge originated. Such a discharge is herein termed an electrical image.

Up to this time, the television process attempted to transmit an image converted to an electrical signal by scanning with mechanically moving apparatus during the brief time period the human eye would retain a picture. Such equipment could not move at sufficient speed to provide full-shaded images to the viewer. At the age of 20, Farnsworth succeeded in producing the first all-electronic television image. It took more that two decades to be adopted for consumer use, but it is easy to see how important this invention has become in today's society.

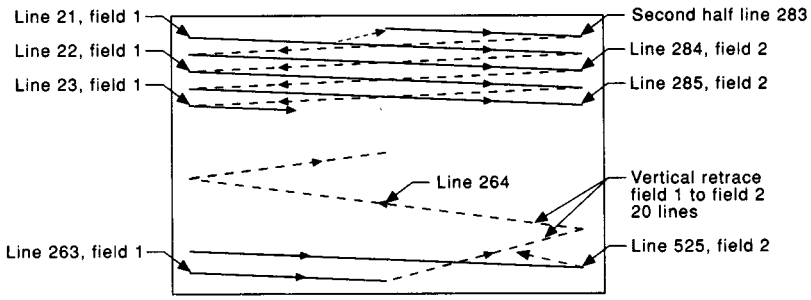


FIGURE 69.16 The interlaced scanning pattern (raster) of the television image. (*Source:* Electronic Industries Association.)

Scanning Lines and Fields

The image pattern of electrical charges on a camera tube target or CCD, corresponding to the brightness levels of a scene, are converted to a video signal in a sequential order of picture elements in the scanning process. At the end of each horizontal line sweep, the video signal is *blanked* while the beam returns rapidly to the left side of the scene to start scanning the next line. This process continues until the image has been scanned from top to bottom to complete one ***field*** *scan.*

After completion of this first field scan, at the midpoint of the last line, the beam again is blanked as it returns to the top center of the target where the process is repeated to provide a second field scan. The spot size of the beam as it impinges upon the target must be fine enough to leave unscanned areas between lines for the second scan. The pattern of scanning lines covering the area of the target, or the screen of a picture display, is called a **raster.**

Interlaced Scanning Fields

Because of the half-line offset for the start of the beam return to the top of the raster and for the start of the second field, the lines of the second field lie in-between the lines of the first field. Thus, the lines of the two are **interlaced.** The two interlaced fields constitute a single television *frame.* Figure 69.16 shows a frame scan with interlacing of the lines of two fields.

Reproduction of the camera image on a cathode ray tube (CRT) or solid-state display is accomplished by an identical operation, with the scanning beam modulated in density by the video signal applied to an element of the electron gun or control element, in the case of a solid-state display device. This control voltage to the display varies the brightness of each picture element on the screen.

Blanking of the scanning beam during the return trace is provided for in the video signal by a "blacker-than-black" pulse waveform. In addition, in most receivers and monitors another blanking pulse is generated from the horizontal and vertical scanning circuits and applied to the display system to ensure a black screen during scanning retrace. The retrace lines are shown as diagonal dashed lines in Fig. 69.16.

The interlaced scanning format, standardized for monochrome and compatible color, was chosen primarily for two partially related and equally important reasons:

- To eliminate viewer perception of the intermittent presentation of images, known as *flicker*
- To reduce video bandwidth requirements for an acceptable flicker threshold level

Perception of flicker is dependent primarily upon two conditions:

- The brightness level of an image
- The relative area of an image in a picture

The 30-Hz transmission rate for a full 525-line television frame is comparable to the highly successful 24-frame-per-second rate of motion-picture film. However, at the higher brightness levels produced on television screens, if all 483 lines (525 less blanking) of a television image were to be presented sequentially as single

frames, viewers would observe a disturbing flicker in picture areas of high brightness. For a comparison, motion-picture theaters on average produce a screen brightness of 10 to 25 ft·L (footlambert), whereas a direct-view CRT may have a highlight brightness of 50 to 80 ft·L. It should be noted also that motion-picture projectors flash twice per frame to reduce the flicker effect.

Through the use of interlaced scanning, single field images with one-half the vertical resolution capability of the 525-line system are provided at the high flicker-perception threshold rate of 60 Hz. Higher resolution of the full 483 lines of vertical detail is provided at the lower flicker-perception threshold rate of 30 Hz. The result is a relatively flickerless picture display at a screen brightness of well over 50 to 75 ft·L, more than double that of motion-picture film projection. Both 60-Hz fields and 30-Hz frames have the same horizontal resolution capability.

The second advantage of interlaced scanning, compared to progressive scanning, where the frame is constructed in one pass over the display face (rather than in two through interlace), is a reduction in video bandwidth for an equivalent flicker threshold level. Progressive scanning of 525 lines would have to be completed in 1/60 s to achieve an equivalent level of flicker perception. This would require a line scan to be completed in half the time of an interlaced scan. The bandwidth then would double for an equivalent number of pixels per line.

The standards adopted by the Federal Communications Commission (FCC) for monochrome television in the United States specified a system of 525 lines per frame, transmitted at a frame rate of 30 Hz, with each frame composed of two interlaced fields of horizontal lines. Initially in the development of television transmission standards, the 60-Hz power line waveform was chosen as a convenient reference for vertical scan. Furthermore, in the event of coupling of power line hum into the video signal or scanning/deflection circuits, the visible effects would be stationary and less objectionable than moving **hum bars** or distortion of horizontal-scanning geometry. In the United Kingdom and much of Europe, a 50-Hz interlaced system was chosen for many of the same reasons. With improvements in television receivers, the power line reference was replaced with a stable crystal oscillator, rendering the initial reason for the frame rate a moot point.

The existing 525-line monochrome standards were retained for color in the recommendations of the National Television System Committee (NTSC) for compatible color television in the early 1950s. The NTSC system, adopted in 1953 by the FCC, specifies a scanning system of 525 horizontal lines per frame, with each frame consisting of two interlaced fields of 262.5 lines at a field rate of 59.94 Hz. Forty-two of the 525 lines in each frame are blanked as black picture signals and reserved for transmission of the vertical scanning synchronizing signal. This results in 483 visible lines of picture information. Because the vertical blanking interval represents a significant amount of the total transmitted waveform, the television industry has sought ways to carry additional data during the blanking interval. Such applications include closed captioning and system test signals.

Synchronizing Video Signals

In monochrome television transmission, two basic synchronizing signals are provided to control the timing of picture-scanning deflection:

- Horizontal sync pulses at the line rate.
- Vertical sync pulses at the field rate in the form of an interval of wide horizontal sync pulses at the field rate. Included in the interval are **equalizing pulses** at twice the line rate to preserve interlace in each frame between the even and odd fields (offset by a half line).

In color transmissions, a third synchronizing signal is added during horizontal scan blanking to provide a frequency and phase reference for color signal encoding circuits in cameras and decoding circuits in receivers. These synchronizing and reference signals are combined with the picture video signal to form a **composite video** waveform.

The scanning and color-decoding circuits in receivers must follow the frequency and phase of the synchronizing signals to produce a stable and geometrically accurate image of the proper color **hue** and **saturation.** Any change in timing of successive vertical scans can impair the interlace of the even and odd fields in a frame. Small errors in horizontal scan timing of lines in a field can result in a loss of resolution in vertical line structures. Periodic errors over several lines that may be out of the range of the horizontal scan automatic frequency control circuit in the receiver will be evident as jagged vertical lines.

TABLE 69.7 Frequency Allocations for TV Channels 2 through 83 in the U.S.

Channel Designation	Frequency Band, MHz	Channel Designation	Frequency Band, MHz	Channel Designation	Frequency Band, MHz
2	54–60	30	566–572	58	734–740
3	60–66	31	572–578	59	740–746
4	66–72	32	578–584	60	746–752
5	76–82	33	584–590	61	752–758
6	82–88	34	590–596	62	758–764
7	174–180	35	596–602	63	764–770
8	180–186	36	602–608	64	770–776
9	186–192	37	608–614	65	776–782
10	192–198	38	614–620	66	782–788
11	198–204	39	620–626	67	788–794
12	204–210	40	626–632	68	794–800
13	210–216	41	632–638	69	800–806
14	470–476	42	638–644	70	806–812
15	476–482	43	644–650	71	812–818
16	482–488	44	650–656	72	818–824
17	488–494	45	656–662	73	824–830
18	494–500	46	662–668	74	830–836
19	500–506	47	668–674	75	836–842
20	506–512	48	674–680	76	842–848
21	512–518	49	680–686	77	848–854
22	518–524	50	686–692	78	854–860
23	524–530	51	692–698	79	860–866
24	530–536	52	698–704	80	866–872
25	536–542	53	704–710	81	872–878
26	542–548	54	710–716	82	878–884
27	548–554	55	716–722	83	884–890
28	554–560	56	722–728		
29	560–566	57	728–734		

Television Industry Standards

There are three primary color transmission standards in use today:

- *NTSC* (National Television Systems Committee): Used in the United States, Canada, Central America, most of South America, and Japan. In addition, NTSC is used in various countries or possessions heavily influenced by the United States.
- *PAL* (Phase Alternation each Line): Used in England, most countries and possessions influenced by the British Commonwealth, many western European countries and China. Variation exists in PAL systems.
- *SECAM* (Sequential Color with [Avec] Memory): Used in France, countries and possessions influenced by France, the USSR (generally the former Soviet Bloc nations), and other areas influenced by Russia.

The three standards are incompatible for a variety of reasons (see Benson and Whitaker, 1991).

Television transmitters in the United States operate in three frequency bands:

- Low-band VHF (very high frequency), channels 2 through 6
- High-band VHF, channels 7 through 13
- UHF (ultra-high frequency), channels 14 through 83 (UHF channels 70 through 83 currently are assigned to mobile radio services)

Table 69.7 shows the frequency allocations for channels 2 through 83. Because of the wide variety of operating parameters for television stations outside the United States, this section will focus primarily on TV transmission as it relates to the Unites States.

Maximum power output limits are specified by the FCC for each type of service. The maximum **effective radiated power** (ERP) for low-band VHF is 100 kW; for high-band VHF it is 316 kW; and for UHF it is 5 MW. The ERP of a station is a function of transmitter power output (TPO) and antenna gain. ERP is determined by multiplying these two quantities together and subtracting transmission line loss.

The second major factor that affects the coverage area of a TV station is antenna height, known in the broadcast industry as *height above average terrain* (HAAT). HAAT takes into consideration the effects of the geography in the vicinity of the transmitting tower. The maximum HAAT permitted by the FCC for a low- or high-band VHF station is 1000 ft (305 m) east of the Mississippi River and 2000 ft (610 m) west of the Mississippi. UHF stations are permitted to operate with a maximum HAAT of 2000 ft (610 m) anywhere in the United States (including Alaska and Hawaii).

The ratio of visual output power to **aural** output power can vary from one installation to another; however, the aural is typically operated at between 10 and 20% of the visual power. This difference is the result of the reception characteristics of the two signals. Much greater signal strength is required at the consumer's receiver to recover the visual portion of the transmission than the aural portion. The aural power output is intended to be sufficient for good reception at the fringe of the station's coverage area but not beyond. It is of no use for a consumer to be able to receive a TV station's audio signal but not the video.

In addition to high power stations, two classifications of low-power TV stations have been established by the FCC to meet certain community needs: They are:

- ***Translator:*** A low-power system that rebroadcasts the signal of another station on a different channel. Translators are designed to provide "fill-in" coverage for a station that cannot reach a particular community because of the local terrain. Translators operating in the VHF band are limited to 100 W power output (ERP), and UHF translators are limited to 1 kW.
- ***Low-Power Television (LPTV):*** A service established by the FCC designed to meet the special needs of particular communities. LPTV stations operating on VHF frequencies are limited to 100 W ERP, and UHF stations are limited to 1 kW. LPTV stations originate their own programming and can be assigned by the FCC to any channel, as long as sufficient protection against interference to a full-power station is afforded.

Composite Video

The composite video waveform is shown in Fig. 69.17. The actual radiated signal is inverted, with modulation extending from the synchronizing pulses at maximum carrier level (100%) to reference picture white at 7.5%. Because an increase in the amplitude of the radiated signal corresponds to a decrease in picture brightness, the polarity of modulation is termed *negative*. The term *composite* is used to denote a video signal that contains:

- Picture luminance and chrominance information
- Timing information for synchronization of scanning and color signal processing circuits

The negative-going portion of the waveform shown in Fig. 69.17 is used to transmit information for synchronization of scanning circuits. The positive-going portion of the amplitude range is used to transmit luminance information representing brightness and, for color pictures, chrominance.

At the completion of each line scan in a receiver or monitor, a horizontal synchronizing (***H****-sync*) pulse in the composite video signal triggers the scanning circuits to return the beam rapidly to the left of the screen for the start of the next line scan. During the return time, a horizontal blanking signal at a level lower than that corresponding to the blackest portion of the scene is added to avoid the visibility of the retrace lines. In a similar manner, after completion of each field, a vertical blanking signal blanks out the retrace portion of the scanning beam as it returns to the top of the picture to start the scan of the next field. The small-level difference between video reference black and blanking level is called **setup.** Setup is used as a guard band to ensure separation of the synchronizing and video-information functions and adequate blanking of the scanning retrace lines on receivers.

The waveforms of Fig. 69.18 show the various reference levels of video and sync in the composite signal. The unit of measurement for video level was specified initially by the Institute of Radio Engineers (IRE). These ***IRE*** *units* are still used to quantify video signal levels. The primary IRE values are given in Table 69.8.

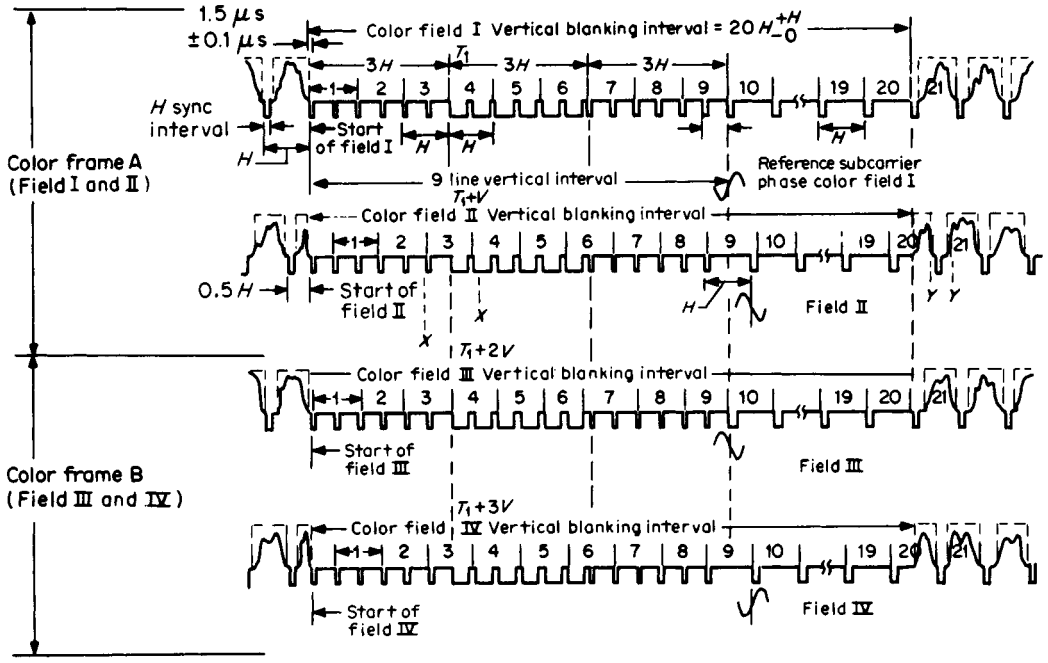


FIGURE 69.17 The principal components of the NTSC color television waveform. (*Source:* Electronic Industries Association.)

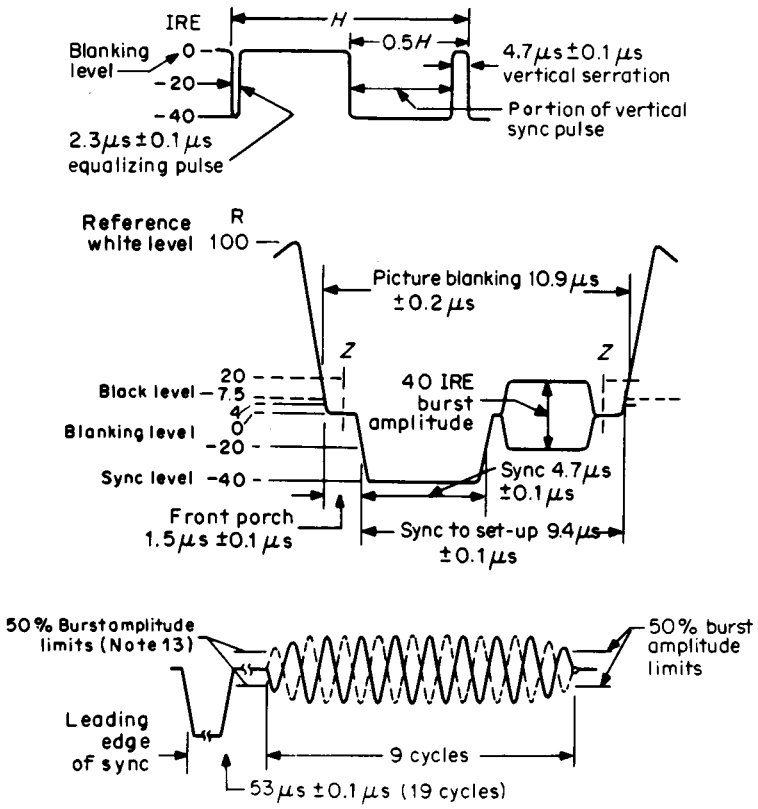


FIGURE 69.18 Sync pulse widths for the NTSC color system. (*Source:* Electronic Industries Association.)

Color Signal Encoding

To facilitate an orderly introduction of color television broadcasting in the United States and other countries with existing monochrome services, it was essential that the new transmissions be compatible. In other words, color pictures would provide acceptable quality on unmodified monochrome receivers. In addition, because of the limited availability of the RF spectrum, another related requirement was the need to fit approximately 2-MHz bandwidth of color information into the 4.2-MHz video bandwidth of the existing 6-MHz broadcasting channels with little or no modification of existing transmitters. This is accomplished by using the band-sharing color signal system developed by the NTSC and by taking advantage of the fundamental characteristics of the eye regarding color sensitivity and resolution.

The video-signal spectrum generated by scanning an image consists of energy concentrated near harmonics of the 15,734-Hz line scanning frequency. Additional lower-amplitude sideband components exist at multiples of 60 Hz (the field scan frequency) from each line scan harmonic. Substantially no energy exists halfway between the line scan harmonics, that is, at odd harmonics of one half line frequency. Thus, these blank spaces in the spectrum are available for the transmission of a signal for carrying color information and its sideband. In addition, a signal modulated with color information injected at this frequency is of relatively low visibility in the reproduced image because the odd harmonics are of opposite phase on successive scanning lines and in successive frames, requiring four fields to repeat. Furthermore, the visibility of the color video signal is reduced further by the use of a subcarrier frequency near the cutoff of the video bandpass.

TABLE 69.8 Video and Sync Levels in IRE Units

Signal Level	IRE Level
Reference white	100
Blanking level width measurement	20
Color burst sine wave peak	+20 to –20
Reference black	7.5
Blanking	0
Sync pulse width measurement	–20
Sync level	–40

In the NTSC system, color is conveyed using two elements:

- A luminance signal
- A chrominance signal

The luminance signal is derived from components of the three primary colors — red, green, and blue — in the proportions for *reference white*, E_y, as follows:

$$E_y = 0.3E_R + 0.59E_G + 0.11E_B$$

These transmitted values equal unity for white and thus result in the reproduction of colors on monochrome receivers at the proper luminance level. This is known as the *constant-luminance* principle.

The color signal consists of two chrominance components, *I* and *Q*, transmitted as amplitude-modulated sidebands of two 3.579545-MHz subcarriers in quadrature. The subcarriers are suppressed, leaving only the sidebands in the color signal. Suppression of the carriers permits demodulation of the color signal as two separate color signals in a receiver by reinsertion of a carrier of the phase corresponding to the desired color signal (**synchronous demodulation**).

I and *Q* signals are composed of red, green, and blue primary color components produced by color cameras and other signal generators. The phase relationship among the *I* and *Q* signals, the derived primary and complementary colors, and the color synchronizing burst can be shown graphically on a **vectorscope** display. The horizontal and vertical sweep signals on a vectorscope are produced from R-Y and B-Y subcarrier sine waves in quadrature, producing a circular display. The chrominance signal controls the intensity of the display. A vectorscope display of an Electronic Industries Association (EIA) standard color bar signal is shown in Fig. 69.19.

Color-Signal Decoding

Each of the two chroma signal carriers can be recovered individually by means of synchronous detection. A reference subcarrier of the same phase as the desired chroma signal is applied as a gate to a balanced demodulator. Only the modulation of the signal in the same phase as the reference will be present in the output. A

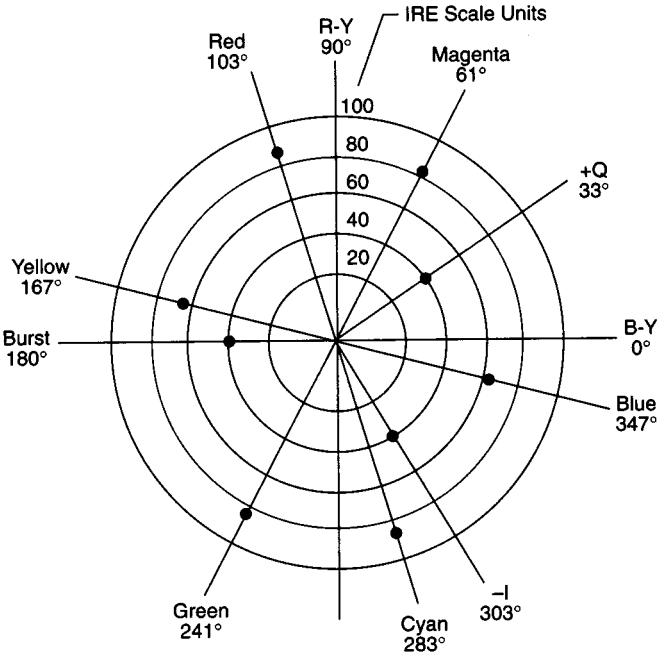


FIGURE 69.19 Vectorscope representation for chroma and vector amplitude relationships in the NTSC system. (*Source:* Electronic Industries Association.)

low-pass filter may be added to remove second harmonic components of the chroma signal generated in the process.

Transmission Equipment

Television transmitters are classified in terms of their operating band, power level, type of final amplifier stage, and cooling method. The transmitter is divided into two basic subsystems:

- The *visual* section, which accepts the video input, amplitude modulates an RF carrier, and amplifies the signal to feed the antenna system
- The *aural* section, which accepts the audio input, frequency modulates a separate RF carrier and amplifies the signal to feed the antenna system

The visual and aural signals are combined to feed a single radiating system.

Transmitter Design Considerations

Each manufacturer has a particular philosophy with regard to the design and construction of a broadcast TV transmitter. Some generalizations can, however, be made with respect to basic system design.

When the power output of a TV transmitter is discussed, the visual section is the primary consideration. Output power refers to the *peak power* of the visual section of the transmitter (*peak of sync*). The FCC-licensed ERP is equal to the transmitter power output minus feedline losses times the power gain of the antenna.

A low-band VHF station can achieve its maximum 100-kW power output through a wide range of transmitter and antenna combinations. A 35-kW transmitter coupled with a gain-of-4 antenna would work, as would a 10-kW transmitter feeding an antenna with a gain of 12. Reasonable pairings for a high-band VHF station would range from a transmitter with a power output of 50 kW feeding an antenna with a gain of 8, to a 30-kW transmitter connected to a gain-of-12 antenna. These combinations assume reasonable feedline losses. To reach the exact power level, minor adjustments are made to the power output of the transmitter, usually by a front panel power trim control.

UHF stations that want to achieve their maximum licensed power output are faced with installing a very high-power transmitter. Typical pairings include a transmitter rated for 220 kW and an antenna with a gain of 25, or a 110-kW transmitter and a gain-of-50 antenna. In the latter case, the antenna could pose a significant problem. UHF antennas with gains in the region of 50 are possible, but not advisable for most installations because of the coverage problems that can result. High-gain antennas have a narrow vertical radiation pattern that can reduce a station's coverage in areas near the transmitter site.

At first examination, it might seem reasonable and economical to achieve licensed ERP using the lowest transmitter power output possible and highest antenna gain. Other factors, however, come into play that make the most obvious solution not always the best solution. Factors that limit the use of high-gain antennas include:

- The effects of high-gain designs on coverage area and signal penetration
- Limitations on antenna size because of tower restrictions, such as available vertical space, weight, and windloading
- The cost of the antenna

The amount of output power required of a transmitter will have a fundamental effect on system design. Power levels dictate whether the unit will be of solid-state or vacuum-tube design; whether air, water, or vapor cooling must be used; the type of power supply required; the sophistication of the high-voltage control and supervisory circuitry; and many other parameters.

Solid-state devices are generally used for VHF transmitters below 35 kW and for low-power UHF transmitters (below 10 kW). Tetrodes may also be used in these ranges. As solid-state technology advances, the power levels possible in a reasonable transmitter design steadily increase. In the realm of high power UHF transmitters, the **klystron** is a common power output device. Klystrons use an *electron bunching* technique to generate high power (55 kW from a single tube is not uncommon) at microwave frequencies. The klystron, however, is relatively inefficient in its basic form. A stock klystron with no efficiency-optimizing circuitry might be only 40 to 50% efficient, depending on the type of device used. Various schemes have been devised to improve klystron efficiency, the best known of which is **beam pulsing.** Two types of pulsing are in common used:

- *Mod-anode pulsing*, a technique designed to reduce power consumption of the klystron during the color burst and video portion of the signal (and thereby improve overall system efficiency)
- *Annular control electrode* (ACE) pulsing, which accomplishes basically the same thing by incorporating the pulsing signal into a low-voltage stage of the transmitter, rather than a high-voltage stage (as with mod-anode pulsing).

Still another approach to improving UHF transmitter efficiency involves entirely new classes of vacuum tubes: the **Klystrode** (also known as the *inductive output tube,* IOT) and the **multistage depressed collector (MSDC) klystron.** (The Klystrode is a registered trademark of Varian.) The IOT is a device that essentially combines the cathode/grid structure of the tetrode with the drift tube/collector structure of the klystron. The MSDC klystron incorporates a collector assembly that operates at progressively lower voltage levels. The net effect for the MSDC is to recover energy from the electron stream rather than dissipating the energy as heat.

Elements of the Transmitter

A television transmitter can be divided into four major subsystems:

- The exciter
- Intermediate power amplifier (IPA)
- Power amplifier (PA)
- High-voltage power supply

Figure 69.20 shows the audio, video, and RF paths for a typical television transmitter.

The modulated visual intermediate frequency (IF) signal is band-shaped in a vestigial sideband filter, typically a surface-acoustic-wave (SAW) filter. Envelope-delay correction is not required for the SAW filter because of the uniform delay characteristics of the device. Envelope-delay compensation may, however, be needed for other parts of the transmitter. The SAW filter provides many benefits to transmitter designers and operators.

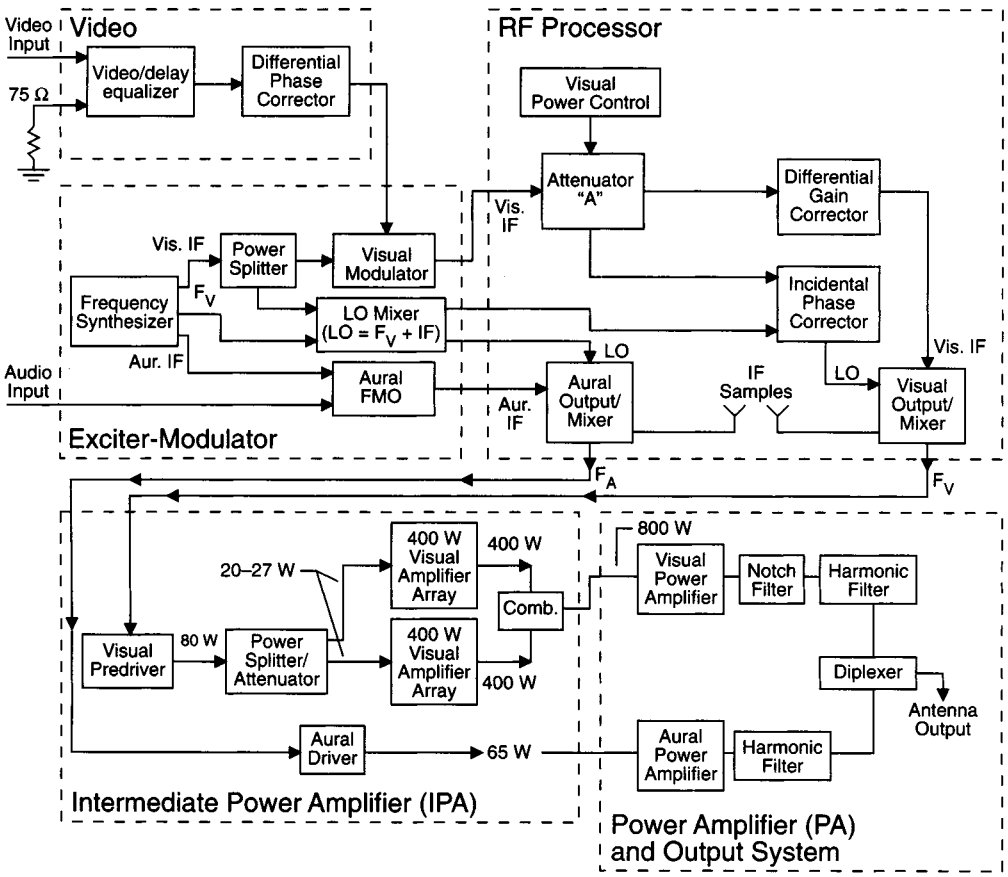


FIGURE 69.20 Simplified block diagram of a VHF television transmitter.

A SAW filter requires no adjustments and is stable with respect to temperature and time. A *color-notch filter* is required at the output of the transmitter because imperfect linearity of the IPA and PA stages introduces unwanted modulation products.

The power amplifier raises the output energy of the transmitter to the desired RF operating level. Tetrodes in television service are operated in the class B mode to obtain reasonable efficiency while maintaining a linear transfer characteristic. Class B amplifiers, when operated in tuned circuits, provide linear performance because of the flywheel effect of the resonance circuit. This allows a single tube to be used instead of two in push-pull fashion. The bias point of the linear amplifier is chosen so that the transfer characteristic at low modulation levels matches that at higher modulation levels. The plate (anode) circuit of a tetrode PA is usually built around a coaxial resonant cavity, which provides a stable and reliable tank circuit.

Solid state transmitters typically incorporate a massively parallel design to achieve the necessary power levels. So-called power blocks of 1 kW or greater are combined as required to meet the target transmitter power output. Most designs use MOSFETs running in a class D (or higher) switching mode. Any one of several combiner schemes may be used to couple the power blocks to the load. Depending on the design, high-reliability features may be incorporated into the transmitter, including automatic disconnection of failed power blocks and hot-changing of defective modules.

UHF transmitters using a klystron in the final output stage must operate class A, the most linear but also most inefficient operating mode for a vacuum tube. Two types of klystrons have traditionally been used: *integral cavity* and *external cavity* devices. The basic theory of operation is identical for each tube, but the mechanical approach is radically different. In the **integral cavity klystron,** the cavities are built into the device to form a

single unit. In the **external cavity klystron,** the cavities are outside the vacuum envelope and are bolted around the tube when the klystron is installed in the transmitter. A number of factors come into play in a discussion of the relative merits of integral vs. external cavity designs. Primary considerations include operating efficiency, purchase price, and life expectancy.

Transmitters based on IOT or MSDC klystron final tubes have much in common with traditional klystron-based systems. There are, however, a number of significant differences, including:

- Low-level video waveform precorrection circuitry
- Drive power requirements
- Power supply demands and complexity
- Fault/arc suppression and protection
- Cooling system design and complexity
- Overall system efficiency

The transmitter block diagram of Fig. 69.20 shows separate visual and aural PA stages. This configuration is normally used for high-power transmitters. Low-power designs often use a combined mode (*common amplification*) in which the aural and visual signals are added prior to the PA. This approach offers a simplified system but at the cost of additional precorrection of the input video signal.

PA stages often are configured so that the circuitry of the visual and aural amplifiers is identical, providing backup protection in the event of a visual PA failure. The aural PA can then be reconfigured to amplify both the aural and the visual signals at reduced power.

The aural output stage of a television transmitter is similar in basic design to a frequency modulated (FM) broadcast transmitter. Tetrode output devices generally operate class C; solid-state devices operate in one of many possible switching modes for high efficiency. The aural PA for a UHF transmitter may use a klystron, IOT, MSDC, tetrode, or a group of solid-state power blocks.

Harmonic filters are employed to attenuate out-of-band radiation of the aural and visual signals to ensure compliance with FCC requirements. Filter designs vary depending upon the manufacturer; however, most are of coaxial construction utilizing L and C components housed within a prepackaged assembly. Stub filters are also used, typically adjusted to provide maximum attenuation at the second harmonic of the operating frequency of the visual carrier and the aural carrier.

The filtered visual and aural outputs are fed to a hybrid diplexer where the two signals are combined to feed the antenna. For installations that require dual-antenna feedlines, a hybrid combiner with quadrature-phased outputs is used. Depending upon the design and operating power, the color-notch filter, aural and visual harmonic filters, and diplexer may be combined into a single mechanical unit.

Antenna System

Broadcasting is accomplished by the emission of coherent electromagnetic waves in free space from one or more radiating-antenna elements that are excited by modulated RF currents. Although, by definition, the radiated energy is composed of mutually dependent magnetic and electric vector fields, conventional practice in television engineering is to measure and specify radiation characteristics in terms of the electric field only.

The field vectors may be polarized horizontally, vertically, or circularly. Television broadcasting, however, has used horizontal polarization for the majority of installations worldwide. More recently interest in the advantages of circular polarization has resulted in an increase in this form of transmission, particularly for VHF channels. Both horizontal and circular polarization designs are suitable for tower-top or side-mounted installations. The latter option is dictated primarily by the existence of a previously installed tower-top antenna. On the other hand, in metropolitan areas where several antennas must be located on the same structure, either a stacking or candelabra-type arrangement is feasible. Another approach to TV transmission involves combining the RF outputs of two or more stations and feeding a single wideband antenna. This approach is expensive and requires considerable engineering analysis to produce a combiner system that will not degrade the performance of either transmission system.

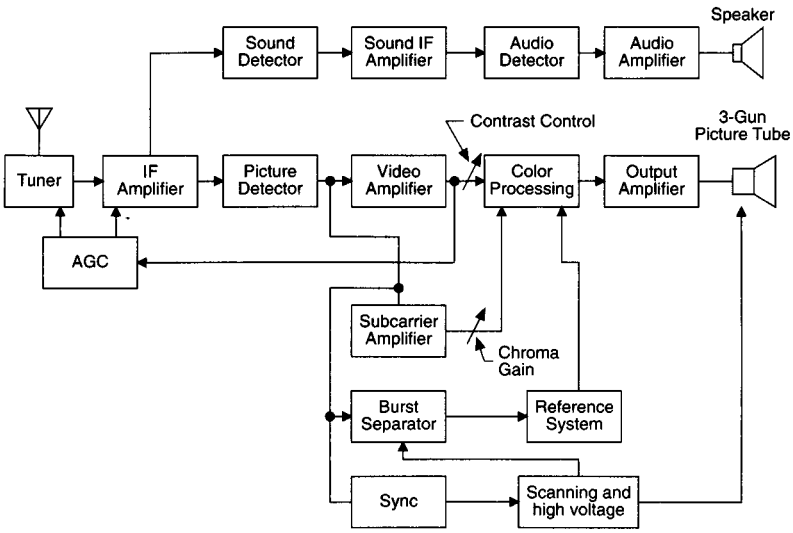


FIGURE 69.21 Simplified schematic block diagram of a color television receiver.

Television Reception

The broadcast channels in the United States are 6 MHz wide for transmission on conventional 525-line standards. The minimum signal level at which a television receiver will provide usable pictures and sound is called the *sensitivity level*. The FCC has set up two standard signal level classifications, Grades A and B, for the purpose of licensing television stations and allocating coverage areas. Grade A refers to urban areas relatively near the transmitting tower; Grade B use ranges from suburban to rural and other fringe areas a number of miles from the transmitting antenna.

Many sizes and form factors of receivers are manufactured. Portable personal types include pocket-sized or hand-held models with picture sizes of 2 to 4 in. diagonal for monochrome and 5 to 6 in. for color. Large screen sizes are available in monochrome where low cost and light weight are prime requirements. However, except where portability is important, the majority of television program viewing is in color. The 19- and 27-in. sizes dominate the market.

Television receiver functions may be broken down into several interconnected blocks. With the increasing use of large-scale integrated circuits, the isolation of functions has become less obvious in the design of receivers. The typical functional configuration of a receiver using a trigun picture tube is shown in Fig. 69.21.

Display Systems

Color video displays may be classified under the following categories:

- Direct-view CRT
- Large-screen display, optically projected from a CRT
- Large-screen display, projected from a modulated light beam
- Large-area display of individually driven light-emitting CRTs or incandescent picture elements
- Flat-panel matrix of transmissive or reflective picture elements
- Flat-panel matrix of light-emitting picture elements

The CRT remains the dominant type of display for both consumer and professional 525-/625-line television applications. The Eidophor and light-valve systems using a modulated light source have found wide application for presentations to large audiences in theater environments, particularly where high screen brightness is required. Matrix-driven flat-panel displays are used in increasing numbers for small-screen personal television receivers and for portable projector units. Video and data projectors using LCD technology have gained wide acceptance.

Cathode Ray Tube Display

The direct-view CRT is the dominant display device in television. The attributes offered by CRTs include the following:

- High brightness
- High resolution
- Excellent gray-scale reproduction
- Low cost compared to other types of displays

From the standpoint of television receiver manufacturing simplicity and low cost, packaging of the display device as a single component is attractive. The tube itself is composed of only three basic parts: an electron gun, an envelope, and a shadow-mask phosphor screen. The luminance efficiency of the electron optical system and the phosphor screen is high. A peak beam current of under 1 μA in a 25-in. tube will produce a highlight brightness of up to 100 ft·L. The major drawback is the power required to drive the horizontal sweep circuit and the high accelerating voltage necessary for the electron beam. This requirement is partially offset through generation of the screen potential and other lower voltages by rectification of the scanning flyback voltage.

As consumer demands drive manufacturers to produce larger picture sizes, the weight and depth of the CRT and the higher power and voltage requirements become serious limitations. These are reflected in sharply increasing receiver costs. To withstand the atmospheric pressures on the evacuated glass envelope, CRT weight increases exponentially with the viewable diagonal. Nevertheless, manufacturers have continued to meet the demand for increased screen sizes with larger direct-view tubes. Improved versions of both tridot delta and in-line guns have been produced. The tridot gun provides small spot size at the expense of critical convergence adjustments for uniform resolution over the full-tube faceplate. In-line guns permit the use of a self-converging deflection yoke that will maintain dynamic horizontal convergence over the full face of the tube without the need for correction waveforms. The downside is slightly reduced resolution.

Defining Terms

Aural: The sound portion of a television signal.

Beam pulsing: A method used to control the power output of a klystron in order to improve the operating efficiency of the device.

Blanking: The portion of a television signal that is used to blank the screen during the horizontal and vertical retrace periods.

Composite video: A single video signal that contains luminance, color, and synchronization information. NTSC, PAL, and SECAM are all examples of composite video formats.

Effective radiated power: The power supplied to an antenna multiplied by the relative gain of the antenna in a given direction.

Equalizing pulses: In an encoded video signal, a series of 2X line frequency pulses occurring during vertical blanking, before and after the vertical synchronizing pulse. Different numbers of equalizing pulses are inserted into different fields to ensure that each field begins and ends at the right time to produce proper interlace. The 2X line rate also serves to maintain horizontal synchronization during vertical blanking.

External cavity klystron: A klystron device in which the resonant cavities are located outside the vacuum envelope of the tube.

Field: One of the two (or more) equal parts of information into which a frame is divided in interlace video scanning. In the NTSC system, the information for one picture is divided into two fields. Each field contains one-half the lines required to produce the entire picture. Adjacent lines in the picture are contained in alternate fields.

Frame: The information required for one complete picture in an interlaced video system. For the NTSC system, there are two fields per frame.

H (horizontal): In television signals, *H* may refer to any of the following: the horizontal period or rate, horizontal line of video information, or horizontal sync pulse.

Hue: One of the characteristics that distinguishes one color from another. Hue defines color on the basis of its position in the spectrum (red, blue, green, yellow, etc.). Hue is one of the three characteristics of television color. Hue is often referred to as *tint.* In NTSC and PAL video signals, the hue information at any particular point in the picture is conveyed by the corresponding instantaneous phase of the active video subcarrier.

Hum bars: Horizontal black and white bars that extend over the entire TV picture and usually drift slowly through it. Hum bars are caused by an interfering power line frequency or one of its harmonics.

Integral cavity klystron: A klystron device in which the resonant cavities are located inside the vacuum envelope of the tube.

Interlaced: A shortened version of *interlaced scanning* (also called *line interlace*). Interlaced scanning is a system of video scanning whereby the odd- and even-numbered lines of a picture are transmitted consecutively as two separate interleaved fields.

IRE: A unit equal to 1/140 of the peak-to-peak amplitude of a video signal, which is typically 1 V. The 0 IRE point is at blanking level, with the sync tip at –40 IRE and white extending to +100 IRE. IRE stands for *Institute of Radio Engineers*, an organization preceding the IEEE, which defined the unit.

Klystrode: An amplifier device for UHF-TV signals that combines aspects of a tetrode (grid modulation) with a klystron (velocity modulation of an electron beam). The result is a more efficient, less expensive device for many applications. (Klystrode is a trademark of EIMAC, a division of Varian Associates.) The term *Inductive Output Tube* (IOT) is a generic name for this class of device.

Klystron: An amplifier device for UHF and microwave signals based on velocity modulation of an electron beam. The beam is directed through an input cavity, where the input RF signal polarity initializes a *bunching effect* on electrons in the beam. The bunching effect excites subsequent cavities, which increase the bunching through an energy flywheel concept. Finally, the beam passes an output cavity that couples the amplified signal to the load (antenna system). The beam falls onto a collector element that forms the return path for the current and dissipates the heat resulting from electron beam bombardment.

Low-power TV (LPTV): A television service authorized by the FCC to serve specific confined areas. An LPTV station may typically radiate between 100 and 1000 W of power, covering a geographic radius of 10 to 15 mi.

Multistage depressed collector (MSDC) klystron: A specially designed klystron in which decreasing voltage zones cause the electron beam to be reduced in velocity before striking the collector element. The effect is to reduce the amount of heat that must be dissipated by the device, improving operating efficiency.

Pixel: The smallest distinguishable and resolvable area in a video image. A pixel is a single point on the screen. The word pixel is derived from *picture element.*

Raster: A predetermined pattern of scanning the screen of a CRT. *Raster* may also refer to the illuminated area produced by scanning lines on a CRT when no video is present.

Saturation: The intensity of the colors in the active picture, the voltage levels of the colors. Saturation relates to the degree by which the eye perceives a color as departing from a gray or white scale of the same brightness. A 100% saturated color does not contain any white; adding white reduces saturation. In NTSC and PAL video signals, the color saturation at any particular instant in the picture is conveyed by the corresponding instantaneous amplitude of the active video subcarrier.

Scan: One sweep of the target area in a camera tube or of the screen in a picture tube.

Setup: A video term relating to the specified base of an active picture signal. In NTSC, the active picture signal is placed 7.5 IRE units above blanking (0 IRE). Setup is the separation in level between the *video blanking* and *reference black* levels.

Synchronous detection: A demodulation process in which the original signal is recovered by multiplying the modulated signal by the output of a synchronous oscillator locked to the carrier.

Translator: An unattended television or FM broadcast repeater that receives a distant signal and retransmits the picture and/or audio locally on another channel.

Vectorscope: An oscilloscope-type device used to display the color parameters of a video signal. A vectorscope decodes color information into R-Y and B-Y components, which are then used to drive the *X* and *Y* axis of the scope. The total lack of color in a video signal is displayed as a dot in the center of the vectorscope. The angle, distance around the circle, magnitude, and distance away from the center indicate the phase and amplitude of the color signal.

Related Topics

69.2 Radio • 69.4 High-Definition Television

References

K. B. Benson and J. Whitaker, Eds., *Television Engineering Handbook,* rev. ed., New York: McGraw-Hill, 1991.
K. B. Benson and J. Whitaker, *Television and Audio Handbook for Technicians and Engineers*, New York: McGraw-Hill, 1990.
J. Whitaker, *Radio Frequency Transmission Systems: Design and Operation*, New York: McGraw-Hill, 1991.
J. Whitaker, *Maintaining Electronic Systems*, Boca Raton: CRC Press, 1991.

Further Information

Additional information on the topic of television system technology is available from the following sources:

Broadcast Engineering magazine, a monthly periodical dealing with television technology. The magazine, published by Intertec Publishing, located in Overland Park, Kan., is free to qualified subscribers.

The Society of Motion Picture and Television Engineers, which publishes a monthly journal and holds conferences in the fall and winter. The SMPTE is headquartered in White Plains, N.Y.

The Society of Broadcast Engineers, which holds an annual technical conference in the spring. The SBE is located in Indianapolis, Ind.

The National Association of Broadcasters, which holds an annual engineering conference and trade show in the spring. The NAB is headquartered in Washington, D.C.

In addition, the following books are recommended:

K.B. Benson and J. Whitaker, Eds., *Television Engineering Handbook*, rev. ed., New York: McGraw-Hill, 1991.

K.B. Benson and J. Whitaker, Eds., *Television and Audio Handbook for Technicians and Engineers*, New York: McGraw-Hill, 1990.

National Association of Broadcasters Engineering Handbook, 8th ed., Washington, D.C.: NAB, 1992.

69.4 High-Definition Television

Martin S. Roden

When standards were developed for television, few people dreamed of its evolution into a type of universal communication terminal. While these traditional standards are acceptable for entertainment video, they are not adequate for many emerging applications, such as videotext. We must evolve into a high-resolution standard. High-definition TV (HDTV) is a term applied to a broad class of new systems whose developments have received worldwide attention.

We begin with a brief review of the current television standards. The reader is referred to Section 69.3 for a more detailed treatment of conventional television.

Japan and North America use the National Television Systems Committee (NTSC) standard that specifies 525 scanning lines per picture, a field rate of 59.94 per second (nominally 60 Hz), and 2:1 **interlaced scanning** (although there are about 60 fields per second, there are only 30 new frames per second). The **aspect ratio** (ratio of width to height) is 4:3. The bandwidth of the television signal is 6 MHz, including the sound signal. In Europe and some other countries, the phase-alternation line (PAL) or the sequential color and memory (SECAM) standard is used. This specifies 625 scanning lines per picture and a field rate of 50 per second. The bandwidth of this type of television signal is 8 MHz.

HDTV systems nominally double the number of scan lines in a frame and change the aspect ratio to 16:9. Of course, if we were willing to start from scratch and abandon all existing television systems, we could set the bandwidth of each channel to a number greater than 6 (or 8) MHz, thereby achieving higher resolution. The Japan Broadcasting Corporation (NHK) has done just this in their HDTV system. This system permits 1125 lines per frame with 30 frames per second and 60 fields per second (2:1 interlaced scanning). The aspect ratio

is 16:9. The system is designed for a bandwidth of 10 MHz per channel. With the 1990 launching of the BS-3 satellite, two channels were devoted to this form of HDTV. To fit the channel within a 10-MHz bandwidth (instead of the approximately 50 MHz that would be needed to transmit using traditional techniques), bandwidth compression was required. It should be noted that the Japanese system is primarily analog frequency modulation (FM) (the sound is digital). The approach to decreasing bandwidth is multiple sub-Nyquist encoding (**MUSE**). The sampling below Nyquist lowers the bandwidth requirement, but moving images suffer from less resolution.

Europe began its HDTV project in mid-1986 with a joint initiative involving West Germany (Robert Bosch GmbH), the Netherlands (NV Phillips), France (Thomson SA), and the United Kingdom (Thorn/EMI Plc.). The system, termed **Eureka 95** or D2-MAC, has 1152 lines per frame, 50 fields per second, 2:1 interlaced scanning, and a 16:9 aspect ratio. A more recent European proposed standard is for 1250 scanning lines at 50 fields per second. This is known as the **Eureka EU95.** It is significant to note that the number of lines specified by Eureka EU95 is exactly twice that of the PAL and SECAM standard currently in use. The field rate is the same, so it is possible to devise compatible systems that would permit reception of HDTV by current receivers (of course, with adapters and without enhanced definition). The HDTV signal requires nominally 30 MHz of bandwidth.

In the United States, the FCC has ruled (in March 1990) that any new HDTV system must permit continuation of service to contemporary NTSC receivers. This significant constraint applies to terrestrial broadcasting (as opposed to videodisk, videotape, and cable television). The HDTV signals will be sent on **"taboo channels,"** those that are not used in metropolitan areas to provide adequate separation. Thus, these currently unused channels would be used for simulcast signals. Since the proposed HDTV system for the United States uses digital transmission, transmitter power can be less than that used for conventional television — this reduces interference with adjacent channels. Indeed, in heavily populated urban areas (where many stations are licensed for broadcast), the HDTV signals will have to be severely limited in power.

When a color television signal is converted from analog to digital (A/D), the luminance, hue, and saturation signals must each be digitized using 8 bits of A/D per sample. Digital transmission of conventional television therefore requires a nominal bit rate of about 216 megabits/s, while uncompressed HDTV nominally requires about 1200 megabits/s. If we were to use a digital modulation system that transmits 1 bit per hertz of bandwidth, we see that the HDTV signal requires over 1 GHz of bandwidth, yet only 6 MHz is allocated. Clearly significant data compression is required!

Proposed Systems

In the early 1990s, four digital HDTV approaches were submitted for FCC testing. The four were proposed by General Instrument Corporation, the Advanced Television Research Consortium (composed of NBC, David Sarnoff Research Center, Philips Consumer Electronics, and Thomson Consumer Electronics, Inc.), Zenith Electronics in cooperation with AT&T Bell Labs and AT&T Microelectronics, and the American Television Alliance (General Instrument Corporation and MIT). There were many common aspects to the four proposals, but major differences existed in the data compression approaches. The data compression techniques can be viewed as two-dimensional extensions of techniques used in voice encoding.

Something unprecedented happened in Spring 1993. The various competing parties decided, with some encouragement from an FCC advisory committee, to merge to form a **Grand Alliance**. The Alliance consists of seven members: AT&T, General Instrument Corp., MIT, Philips, Sarnoff, Thomson, and Zenith. This permitted the selection of the "best" features of each of the proposals. The advisory committee was then able to spend Fall 1995 on completion of the proposed HDTV standard. In the following, we describe a generic system. The reader is referred to the references for details.

Figure 69.22 shows a general block diagram of a digital HDTV transmitter. Each frame from the camera is digitized, and the system has the capability of storing one entire frame. Thus the processor works with two inputs—the current frame (*A*) and the previous frame (*B*). The current frame and the previous frame are compared in a **motion detector** that generates coded motion information (*C*). Algorithms used for motion estimation attempt to produce three-dimensional parameters from sequential two-dimensional information. Parameters may include velocity estimates for blocks of the picture.

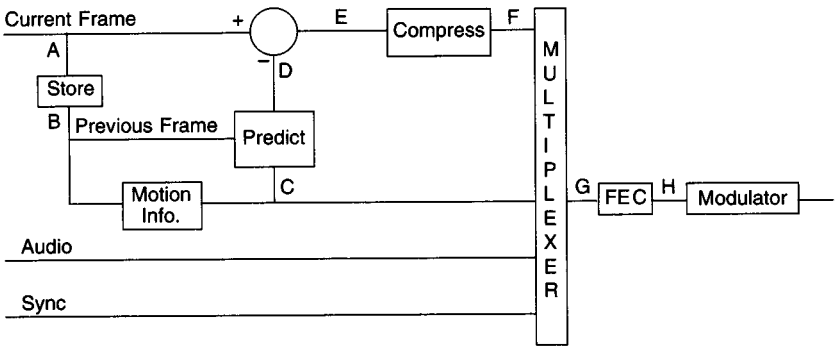


FIGURE 69.22 Block diagram of HDTV transmitter.

The parameters from the motion detector are processed along with the previous frame to produce a *prediction* of the current frame (*D*). Since the motion detector parameters are transmitted, the receiver can perform a similar prediction of the current frame.

The predicted current frame is compared to the actual current frame, and a difference signal (*E*) is generated. This difference signal will generally have a smaller dynamic range than the original signal. For example, if the television image is static (is not changing with time), the difference signal will be zero.

The difference signal is *compressed* to form the transmitted video signal (*F*). This compression is performed both in the time and transform domains. **Entropy coding** of the type used in facsimile can be incorporated to take spatial continuity into account (i.e., a picture usually does not change over the span of a single picture element, so variations of "run length" coding can often compress the data). The compression technique incorporates the **MPEG-2** syntax. The actual compression algorithms (based on the **discrete cosine transform**) are adaptive so a variety of formats can be accommodated (e.g., 1080-line interlaced scanning, 720-line progressive, bi-directional). The main feature is that the data rate is decreased by extracting essential parameters that describe the waveform.

Four data streams are asynchronously multiplexed to form the information to be transmitted (*G*). These four signals consist of the coded differential video, the motion detector parameters, the digital audio signal (using Dolby Labs' AC-3 digital audio), and the synchronizing signals. Other information can be multiplexed, including various control signals that may be needed by cable operators.

Forward error correction is applied to the multiplexed digital signal to produce an encoded signal (*H*) that makes the transmission less susceptible to uncorrected bit errors. This is needed because of the anticipated low transmission power rates. Error control is also important because compression can amplify error effects—a single bit error can affect many picture elements.

The encoded data signal forms the input to the modulator. To further conserve bandwidth, a type of quadrature modulation is employed. The actual form is 8-VSB, a variation of **digital vestigial sideband** that includes **trellis coding**. This possesses many of the advantages of guadrature amplitude modulation (QAM).

The corresponding receiver is shown in Fig. 69.23. The receiver simply forms the inverse of each transmitter operation. The received signal is first demodulated. The resulting data signal is decoded to remove the redundancy and correct errors. A demultiplexer separates the signal into the original four (or more) data signals. The audio and synchronization signals need no further processing.

The demultiplexed video signal is, hopefully, the same as the transmitted signal ("*F*"). We use letters with quotation marks to indicate that the signals are estimates of their transmitted counterpart. This reproduced video signal is decompressed, using the inverse algorithm of that used in the transmitter, to yield an estimate of the original differential picture signal ("*E*"). The predict block in the receiver implements the same algorithm as that of the transmitter. Its inputs are the reconstructed motion signal ("*C*") and the previous reconstructed frame ("*B*"). When the predictor output ("*D*") is added to the reconstructed differential picture signal ("*E*"), the result is a reconstructed version of the current frame.

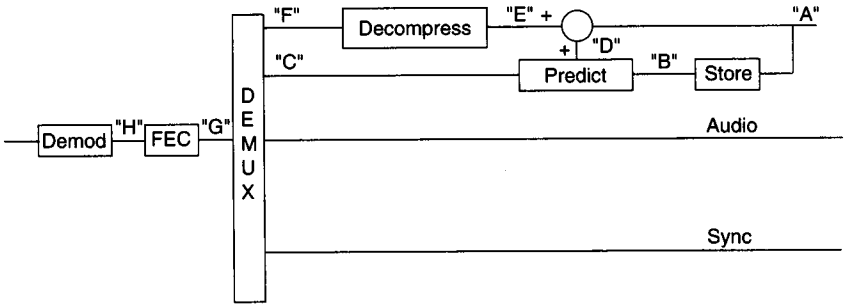


FIGURE 69.23 Block diagram of HDTV receiver.

Defining Terms

Aspect ratio: Ratio of frame width to height.

Digital vestigial sideband: A form of digital modulation where a portion of one of the sidebands is partially suppressed.

Discrete cosine transform: A popular format for video compression. The spatial signal is expanded in a cosine series, where the higher frequencies represent increased video resolution.

Entropy coding: A form of data compression that reduces a transmission to a shorter length by reducing signal redundancy.

Eureka 95 and EU95: European proposed HDTV systems.

Grand Alliance: A consortium formed of seven of the organizations proposing HDTV systems.

Interlaced scanning: A bandwidth reduction technique wherein every other scan line is first transmitted followed by the "in between" lines.

Motion detector: A system that compares two adjacent frames to detect differences.

MPEG-2: Video compression standard devised by the Moving Picture Experts Group.

MUSE: Multiple sub-Nyquist encoding, a technique used in Japanese HDTV system.

Taboo channels: Channels that the FCC does not currently assign in order to avoid interference from adjacent channels.

Trellis coding: A form of digital encoding which provides a constraint (i.e., a structure) to a stream of digital data.

Related Topic

69.3 Television Systems

References

G.W. Beakley, "Channel coding for digital HDTV terrestrial broadcasting," *IEEE Transactions on Broadcasting*, vol. 37, no. 4, 1991.

Grand Alliance, "Proposed HDTV standard". May be obtained as ftp from *ga-doc.sarnoff.com.* May also be obtained by sending an e-mail to *grand_alliance@sarnoff.com.*

R. Hopkins, "Digital HDTV broadcasting," *IEEE Transactions on Broadcasting*, vol. 37, no. 4, 1991.

R.K. Jurgen, Ed., "High-definition television update," *IEEE Spectrum*, April 1988.

R.K. Jurgen, Ed., "Consumer electronics," *IEEE Spectrum*, January 1989.

R.K. Jurgen, Ed., "The challenges of digital HDTV," *IEEE Spectrum*, April 1991.

J.C. McKinney, "HDTV approaches the end game," *IEEE Transactions on Broadcasting*, vol. 37, no. 4, 1991.

S. Prentiss, *HDTV*, Blue Ridge Summit, Pa.: TAB Books, 1990.

M.S. Roden, *Analog and Digital Communication Systems*, 4th ed., Englewood Cliffs, N.J.: Prentice-Hall, 1996.

W.Y. Zou, "Digital HDTV compression techniques," IEEE Transactions on Broadcasting, vol. 37, no. 4, 1991.

Further Information

As HDTV transitions from a proposed system to a commercially available product, you can expect information to appear in a variety of places from the most esoteric research publications to popular business and entertainment publications. During the development process, the best places to look are the IEEE publications (IEEE, NY) and the broadcasting industry journals. The *IEEE Transactions on Broadcasting* and the *IEEE Transactions on Consumer Electronics* continue to have periodic articles relating to the HDTV standards and implementation of these standards. Another source of information, though not overly technical, is the periodical *Broadcasting and Cable* (Cahners Publishing, NY).

69.5 Digital Audio Broadcasting

Stanley Salek and Almon H. Clegg

Digital audio broadcasting (DAB) is a developing technology that promises to give consumers a new and better aural broadcast system. DAB will offer dramatically better reception quality over existing AM and FM broadcasts by better audio quality and by superior resistance to interference in stationary and mobile/portable reception environments. Additionally, the availability of a digital data stream direct to consumers will open the prospects of providing extra services to augment basic sound delivery.

As of this writing, seven proponents have announced DAB transmission and reception systems. From the data available describing these potential systems, it is clear that there is only partial agreement on which transmission method will provide the best operational balance. This chapter provides a general overview of the common aspects of DAB systems, as well as a description of one of the proposed transmission methods.

The Need for DAB

In the years since the early 1980s, the consumer marketplace has undergone a great shift toward digital electronic technology. The explosion of personal computer use has led to greater demands for information, including multimedia integration. Over the same time period, compact disc (CD) digital audio technology has overtaken long-playing records (and has nearly overtaken analog tape cassettes) as the consumer audio playback media of choice. Similar digital transcription methods and effects also have been incorporated into commonly available audio and video equipment. Additionally, it is virtually certain that the upcoming transition to a high-definition television broadcast system will incorporate full digital methods for video and audio transmission. Because of these market pressures, the radio broadcast industry has determined that the existing analog methods of broadcasting must be updated to keep pace with the advancing audio marketplace.

In addition to providing significantly enhanced audio quality, DAB systems are being developed to overcome the technical deficiencies of existing AM and FM analog broadcast systems. The foremost problem of current broadcast technology, as perceived by the industry, is its susceptibility to interference. AM medium-wave broadcasts, operating in the 530- to 1700-kHz frequency range, are prone to disruption by fluorescent lighting and by power system distribution networks, as well as by numerous other manufactured unintentional radiators, including computer and telephone systems. Additionally, natural effects, such as nighttime skywave propagation interference between stations and lightning, cause irritating service disruption to AM reception. FM broadcast transmissions in the 88- to 108-MHz band are much more resistant to these types of interference. However, multipath propagation and abrupt signal fading, especially found in urban and mountainous areas containing a large number of signal reflectors and shadowers (e.g., buildings and terrain), can seriously degrade FM reception, particularly in automobiles.

DAB System Design Goals

DAB systems are being designed with several technical goals in mind. The first goal is to create a service that delivers compact disc quality stereo sound for broadcast to consumers. The second is to overcome the interference problems of current AM and FM broadcasts, especially under portable and mobile reception conditions.

Third, DAB must be spectrally efficient in that total bandwidth should be no greater than that currently used for FM broadcasts. Fourth, the DAB system should provide space in its data stream to allow for the addition of ancillary services, such as program textual information display or software downloading. Finally, DAB receivers must not be overly cumbersome, complex, or expensive, to foster rapid consumer acceptance.

In addition to these goals, desired features include the reduced RF transmission power requirements (when compared to AM and FM broadcast stations with the same signal coverage), the mechanism to seamlessly fill in coverage areas that are shadowed from the transmitted signal, and the ability to easily integrate DAB receivers into personal, home, and automotive sound systems.

Historical Background

DAB development work began in Europe in 1986, with the initial goal to provide high-quality audio services to consumers directly by satellite. Companion terrestrial systems were developed to evaluate the technology being considered, as well as to provide fill-in service in small areas where the satellite signals were shadowed. A consortium of European technical organizations known as Eureka-147/DAB demonstrated the first working terrestrial DAB system in Geneva in September 1988. Subsequent terrestrial demonstrations of the system followed in Canada in the summer of 1990, and in the United States in April and September of 1991.

For the demonstrations, VHF and UHF transmission frequencies between 200 and 900 MHz were used with satisfactory results. Because most VHF and UHF frequency bands suitable for DAB are already in use (or reserved for high-definition television and other new services), an additional Canadian study in 1991 evaluated frequencies near 1500 MHz (L-band) for use as a potential worldwide DAB allocation. This study concluded that L-band frequencies would support a DAB system such as Eureka-147, while continuing to meet the overall system design goals.

In early 1992, the World Administrative Radio Conference (WARC-92) was held, during which frequency allocations for many different radio systems were debated. As a result of WARC-92, a worldwide L-band standard of 1452 to 1492 MHz was designated for both satellite and terrestrial digital radio broadcasting. However, because of existing government and military uses of L-band, the United States was excluded from the standard. Instead, an S-band allocation of 2310 to 2360 MHz was substituted. Additionally, Asian nations including Japan, China, and CIS opted for an extra S-band allocation in the 2535- to 2655-MHz frequency range.

In mid-1991, because of uncertainty as to the suitability of using S-band frequencies for terrestrial broadcasting, most DAB system development work in the United States shifted from out-band (i.e., UHF, L-band, and S-band) to in-band. In-band terrestrial systems would merge DAB services with existing AM and FM broadcasts, using novel adjacent- and co-channel modulating schemes. Since 1992, two system proponents have demonstrated proprietary methods of extracting a compatible digital RF signal from co-channel analog FM broadcast transmissions. Thus, in-band DAB could permit a logical transition from analog to digital broadcasting for current broadcasters, within the current channel allocation scheme.

In 1991, a digital radio broadcasting standards committee was formed by the Electronic Industries Association (EIA). Present estimates are that the committee may complete its testing and evaluation of the various proposed systems by 1997. As of mid-1996, laboratory testing of several proponent systems had been completed, and field testing of some of those systems, near San Francisco, Calif. was getting underway.

Technical Overview of DAB

Regardless of the actual signal delivery system used, all DAB systems share a common overall topology. Figure 69.24 presents a block diagram of a typical DAB transmission system.

To maintain the highest possible audio quality, program material would be broadcast from digital sources, such as CD players and digital audio recorders, or digital audio feeds from network sources. Analog sources, such as microphones, are converted to a digital audio data stream using an analog-to-digital (A/D) converter, prior to switching or summation with the other digital sources.

The linear digital audio data stream from the studio is then applied to the input of a **source encoder.** The purpose of this device is to reduce the required bandwidth of the audio information, helping to produce a spectrally efficient RF broadcast signal. For example, 16-bit linear digital audio sampled at 48 kHz (the standard

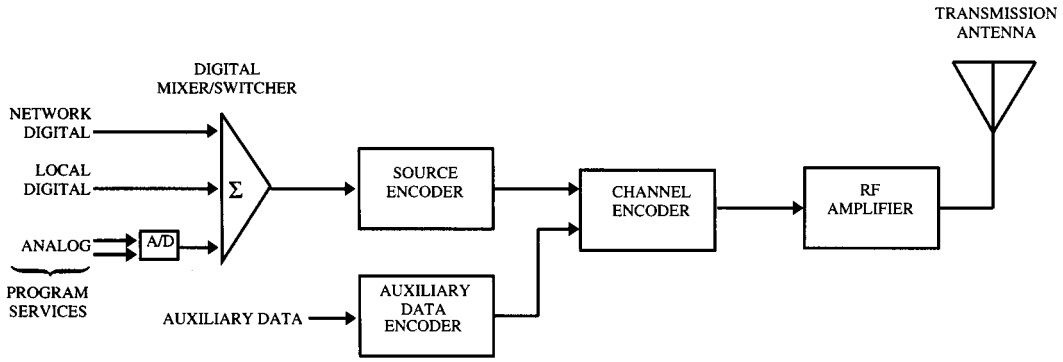


FIGURE 69.24 An example DAB transmission system. (*Source:* Hammett & Edison, Inc., Consulting Engineers.)

professional rate) requires a data stream of 1.536 megabits/s to transmit a stereo program in a serial format. This output represents a bandwidth of approximately 1.5 MHz, much greater than that used by an equivalent analog audio modulating signal [Smyth, 1992]. Source encoders can reduce the data rate by factors of 8:1 or more, yielding a much more efficient modulating signal.

Following the source encoder, the resulting serial digital signal is applied to the input of the **channel encoder,** a device that modulates the transmitted RF wave with the reduced-rate audio information. Auxiliary serial data, such as program information and/or receiver control functions, also can be input to the channel encoder for simultaneous transmission.

The channel encoder uses sophisticated modulating techniques to accomplish the goals of interference cancellation and high spectral efficiency. Methods of interference cancellation include expansion of time and frequency diversity of the transmitted information, as well as the inclusion of error correction codes in the data stream. Time diversity involves transmitting the same information multiple times by using a predetermined time interval. Frequency diversity, such as that produced by spread-spectrum, multiple-carrier, or frequency-hopping systems, provides the means to transmit identical data on several different frequencies within the bandwidth of the system. At the receiver, real-time mathematical processes are used to locate the required data on a known frequency at a known time. If the initial information is found to be unusable because of signal interference, the receiver simply uses the same data found on another frequency and/or at another time, producing seamless demodulation.

Spectral efficiency is a function of the modulation system used. Among the modulation formats that have been proposed for DAB transmission are QPSK, M-ary QAM, and MSK [Springer, 1992]. Using these and other formats, digital transmission systems that use no more spectrum than their analog counterparts have been designed.

The RF output signal of the channel encoder is amplified to the appropriate power level for transmission. Because the carrier-to-noise (C/N) ratio of the modulated waveform is not generally so critical as that required for analog communications systems, relatively low transmission power often can be used. Depending on the sophistication of the data recovery circuits contained in the DAB receiver, the use of C/N ratios as low as 6 dB are possible, without causing a degradation to the received signal.

DAB reception is largely the inverse of the transmission process, with the inclusion of sophisticated error correction circuits. Fig. 69.25 shows a typical DAB receiver.

DAB reception begins in a similar manner as is used in virtually all receivers. A receiving antenna feeds an appropriate stage of RF selectivity and amplification from which a sample of the coded DAB signal is derived. This signal then drives a channel decoder, which reconstructs the audio and auxiliary data streams. To accomplish this task, the channel decoder must demodulate and de-interleave the data contained on the RF carrier and then apply appropriate computational and statistical error correction functions.

The source decoder converts the reduced bit-rate audio stream back to pseudolinear at the original sampling rate. The decoder computationally expands the mathematically reduced data and fills the gaps left from the extraction of irrelevant audio information with averaged code or other masking data. The output of the source

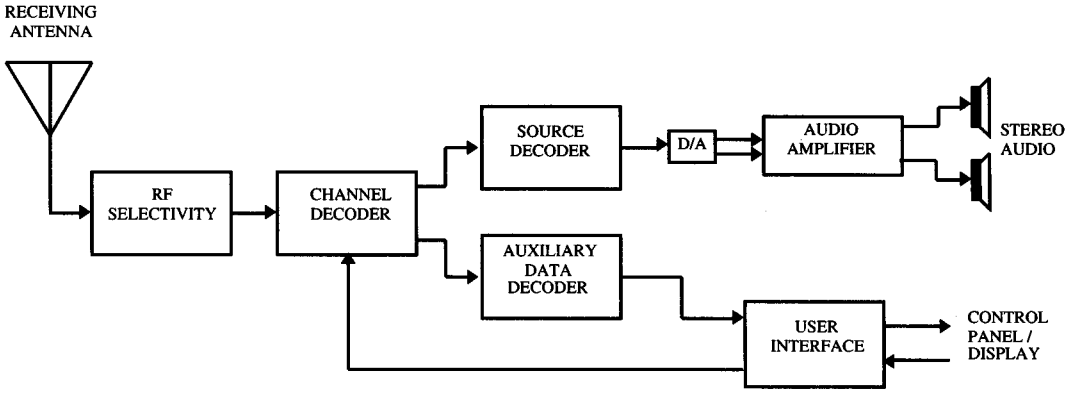


FIGURE 69.25 An example DAB receiver. (*Source:* Hammett & Edison, Inc., Consulting Engineers.)

decoder feeds audio digital-to-analog (D/A) converters, and the resulting analog stereo audio signal is amplified for the listener.

In addition to audio extraction, DAB receivers likely will be capable of decoding auxiliary data. This data can be used in conjunction with the user interface to control receiver functions, or for a completely separate purpose. A typical user interface could contain a data display screen in addition to the usual receiver tuning and audio controls. This data screen could be used to obtain information about the programming, news reports, sports scores, advertising, or any other useful data sent by the station or an originating network. Also, external interfaces could be used to provide a software link to personal computer systems.

Audio Compression and Source Encoding

The development of digital audio encoding started with research into pulse-code modulation (PCM) in the late 1930s and evolved, shortly thereafter, to include work on the principles of digital PCM coding. Linear predictive coding (LPC) and adaptive delta pulse-code modulation (ADPCM) algorithms had evolved in the early 1970s and later were adopted into standards such as C.721 (published by the CCITT) and CD-I (Compact Disc-Interactive). At the same time, algorithms were being invented for use with phoneme-based speech coding. Phonetic coding, a first-generation "model-based" speech-coding algorithm, was mainly implemented for low bit-rate speech and text-to-speech applications. These classes of algorithms for speech further evolved to include both CELP (Code Excited Linear Predictive) and VSELP (Vector Selectable Excited Linear Predictive) algorithms by the mid-1980s. In the late 1980s, these classes of algorithms were also shown to be useful for high-quality audio music coding. These audio algorithms were put to commercial use from the late 1970s to the latter part of the 1980s.

Subband coders evolved from the early work on quadrature mirror filters in the mid-1970s and continued with polyphase filter-based schemes in the mid-1980s. Hybrid algorithms employing both subband and ADPCM coding were developed in the latter part of the 1970s and standardized (e.g., CCITT G.722) in the mid- to late 1980s. Adaptive transform coders for audio evolved in the mid-1980s from speech coding work done in the late 1970s.

By employing psychoacoustic noise-masking properties of the human ear, perceptual encoding evolved from early work of the 1970s and where high-quality speech coders were employed. Music quality bit-rate reduction schemes such as MPEG (Motion Picture Expert Group), PASC (Precision Adaptive Subband Coding), and ATRAC (Adaptive TRansform Acoustic Coding) have been developed. Further refinements to the technology will focus attention on novel approaches such as wavelet-based coding and the use of entropy coding schemes. However, recent progress has been significant, and the various audio coding schemes that have been demonstrated publicly over the time period from 1990 to 1995 have shown steady increases in compression ratios at given audio quality levels.

Audio coding for digital broadcasting will likely use one of the many perceptual encoding schemes previously mentioned or some variation thereof. Fundamentally, they all depend on two basic psychoacoustic phenomena:

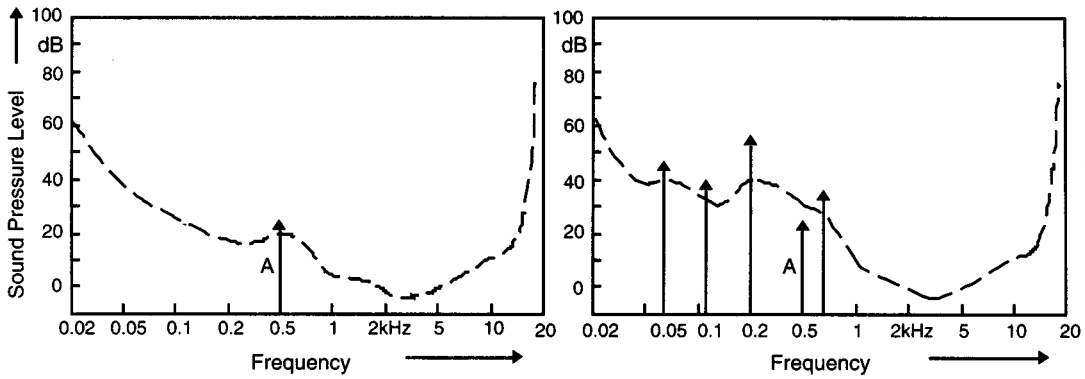


FIGURE 69.26 An example of the masking effect. Based on the hearing threshold of the human ear (dashed line), a 500-Hz sinusoidal acoustic waveform, shown at *A* on the left graph, is easily audible at relatively low levels. However, it can be masked by adding nearby higher-amplitude components, as shown on the right. (*Source:* CCi.)

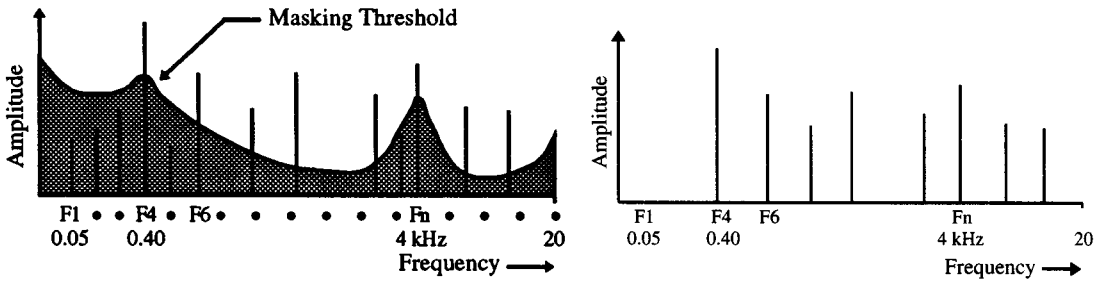


FIGURE 69.27 Source encoders use an empirically derived masking threshold to determine which audio components can be discarded (left). As shown on the right, only the audio components with amplitudes above the masking threshold are retained. (*Source:* CCi.)

(1) the threshold of human hearing, and (2) masking of nearby frequency components. In the early days of hearing research, Harvey Fletcher, a researcher at Bell Laboratories, measured the hearing of many human beings and published the well-known Fletcher-Munson threshold-of-hearing chart. Basically it states that, depending on the frequency, audio sounds below certain levels cannot be heard by the human ear. Further, the masking effect, simply stated, is when two frequencies are very close to each other and one is a higher level than the other, the weaker of the two is masked and will not be heard. These two principles allow for as much as 80% of the data representing a musical signal to be discarded.

Figure 69.26 shows how introduction of frequency components affects the ear's threshold of hearing versus frequency. Figure 69.27 shows how the revised envelope of audibility results in the elimination of components that would not be heard.

The electronic implementation of these algorithms employs a digital filter that breaks the audio spectrum into many subbands, and various coefficient elements are built into the program to decide when it is permissible to remove one or more of the signal components. The details of how the bands are divided and how the coefficients are determined are usually proprietary to the individual system developers. Standardization groups have spent many worker-hours of evaluation attempting to determine the most accurate coding system.

System Example: Eureka-147/DAB

As of this writing, Eureka-147/DAB is the only fully developed DAB system that has demonstrated a capability to meet virtually all the described system goals. Developed by a European consortium, it is an out-band system

in that its design is based on the use of a frequency spectrum outside the AM and FM radio broadcast bands. Out-band operation is required because the system packs up to 16 stereophonic broadcast channels (plus auxiliary data) into one contiguous band of frequencies, which can occupy a total bandwidth of up to 4 MHz. Thus, overall efficiency is maintained, with 16 digital program channels occupying about the same total bandwidth as 16 equivalent analog FM broadcast channels. System developers have promoted Eureka-147/DAB for satellite transmission, as well as for terrestrial applications in locations that have a suitable block of unused spectrum in the L-band frequency range or below.

In recent tests and demonstrations, the ISO/MPEG-2 source encoding/decoding system has been used. Originally developed by IRT (Institut für Rundfunktecknik) in Germany as MUSICAM (Masking pattern-adapted Universal Subband Integrated Coding And Multiplexing), the system works by dividing the original digital audio source into 32 subbands. As with the source encoders described earlier, each of the bands is digitally processed to remove redundant information and sounds that are not perceptible to the human ear. Using this technique, the original audio, sampled at a rate of 768 kilobits/s per channel, is reduced to as little as 96 kilobits/s per channel, representing a compression ratio of 8:1.

The Eureka-147/DAB channel encoder operates by combining the transmitted program channels into a large number of adjacent narrowband RF carriers, which are each modulated using QPSK and grouped in a way that maximizes spectrum efficiency known as orthogonal frequency-division multiplex (OFDM). The information to be transmitted is distributed among the RF carriers and is also time-interleaved to reduce the effects of selective fading. A guard interval is inserted between blocks of transmitted data to improve system resistance to intersymbol interference caused by multipath propagation. Convolutional coding is used in conjunction with a Viterbi maximum-likelihood decoding algorithm at the receiver to make constructive use of echoed signals and to correct random errors [Alard and Lassalle, 1988].

RF power levels of just a few tens of watts per program channel have been used in system demonstrations, providing a relatively wide coverage area, depending on the height of the transmitting antenna above surrounding terrain. This low power level is possible because the system can operate at a C/N ratio of less than 10 dB, as opposed to the more than 30 dB that is required for high-fidelity demodulation of analog FM broadcasts.

Another demonstrated capability of the system is its ability to use "**gap filler**" transmitters to augment signal coverage in shadowed areas. A gap filler is simply a system that directly receives the DAB signal at an unobstructed location, provides RF amplification, and retransmits the signal, on the same channel, into the shadowed area. Because the system can make constructive use of signal reflections (within a time window defined by the guard interval and other factors), the demodulated signal is uninterrupted on a mobile receiver when it travels between an area served by the main signal into the service area of the gap filler.

Defining Terms

Channel encoder: A device that converts source-encoded digital information into an analog RF signal for transmission. The type of modulation used depends on the particular digital audio broadcasting (DAB) system, although most modulation techniques employ methods by which the transmitted signal can be made more resistant to frequency-selective signal fading and multipath distortion effects.

Gap filler: A low-power transmitter that boosts the strength of transmitted DAB RF signals in areas which normally would be shadowed due to terrain obstruction. Gap fillers can operate on the same frequency as DAB transmissions or on alternate channels that can be located by DAB receivers using automatic switching.

Source encoder: A device that substantially reduces the data rate of linearly digitized audio signals by taking advantage of the psychoacoustic properties of human hearing, eliminating redundant and subjectively irrelevant information from the output signal. Transform source encoders work entirely within the frequency domain, while time-domain source encoders work primarily in the time domain. Source decoders reverse the process, using various masking techniques to simulate the properties of the original linear data.

Related Topics

69.2 Radio • 73.6 Data Compression

References

M. Alard and R. Lassalle, "Principles of modulation and channel coding for digital broadcasting for mobile receivers," in *Advanced Digital Techniques for UHF Satellite Sound Broadcasting* (collected papers), European Broadcasting Union, pp. 47–69, 1988.

R. Bruno, "Digital audio and video compression, present and future," presented to the Delphi Club, Tokyo, Japan, July 1992.

G. Chouinard and F. Conway, "Broadcasting systems concepts for digital sound," in *Proceedings of the 45th Annual Broadcast Engineering Conference,* National Association of Broadcasters, 1991, pp. 257–266.

F. Conway, R. Voyer, S. Edwards, and D. Tyrie, "Initial experimentation with DAB in Canada," in *Proceedings of the 45th Annual Broadcast Engineering Conference,* National Association of Broadcasters, 1991, pp. 281–290.

S. Kuh and J. Wang, "Communications systems engineering for digital audio broadcast," in *Proceedings of the 45th Annual Broadcast Engineering Conference*, National Association of Broadcasters, 1991, pp. 267–272.

P. H. Moose and J.M. Wozencraft, "Modulation and coding for DAB using multi-frequency modulation," in *Proceedings of the 45th Annual Broadcast Engineering Conference,* National Association of Broadcasters, 1991, pp. 405–410.

M. Rau, L. Claudy, and S. Salek, *Terrestrial Coverage Considerations for Digital Audio Broadcasting Systems*, National Association of Broadcasters, 1990.

S. Smyth, "Digital audio data compression," *Broadcast Engineering Magazine,* pp. 52–60, Feb. 1992.

K.D. Springer, *Interference Between FM and Digital M-PSK Signals in the FM Band*, National Association of Broadcasters, 1992.

Further Information

The National Association of Broadcasters publishes periodic reports on the technical, regulatory, and political status of DAB in the United States. Additionally, their Broadcast Engineering Conference proceedings published since 1990 contain a substantial amount of information on emerging DAB technologies.

IEEE Transactions on Broadcasting, published quarterly by the Institute of Electrical and Electronics Engineers, Inc., periodically includes papers on digital broadcasting.

Additionally, the biweekly newspaper publication *Radio World* provides continuous coverage of DAB technology, including proponent announcements, system descriptions, field test reports, and broadcast industry reactions.

Dorf, R.C., Wan, Z., Millstein, L.B., Simon, M..K. “Digital Communication”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

70

Digital Communication

Richard C. Dorf
University of California, Davis

Zhen Wan
University of California, Davis

L. B. Milstein
University of California

M. K. Simon
Jet Propulsion Laboratory

70.1 Error Control Coding
Block Codes • Convolutional Codes • Code Performance • Trellis-Coded Modulation

70.2 Equalization
Linear Transversal Equalizers • Nonlinear Equalizers • Linear Receivers • Nonlinear Receivers

70.3 Spread Spectrum Communications
A Brief History • Why Spread Spectrum? • Basic Concepts and Terminology • Spread Spectrum Techniques • Applications of Spread Spectrum

70.1 Error Control Coding

Richard C. Dorf and Zhen Wan

Error correcting codes may be classified into two broad categories: **block codes** and **tree codes.** A block code is a mapping of k input binary symbols into n output binary symbols. Consequently, the block coder is a *memoryless* device. Since $n > k$, the code can be selected to provide redundancy, such as *parity bits*, which are used by the decoder to provide some error detection and error correction. The codes are denoted by (n, k), where the code rate R is defined by $R = k/n$. Practical values of R range from 1/4 to 7/8, and k ranges from 3 to several hundred [Clark and Cain, 1981]. Some properties of block codes are given in Table 70.1.

A tree code is produced by a coder that has *memory.* **Convolutional codes** are a subset of tree codes. The convolutional coder accepts k binary symbols at its input and produces n binary symbols at its output, where the n output symbols are affected by $v + k$ input symbols. Memory is incorporated since $v > 0$. The code rate is defined by $R = k/n$. Typical values for k and n range from 1 to 8, and the values for v range from 2 to 60. The range of R is between 1/4 and 7/8 [Clark and Cain, 1981].

Block Codes

In block code, the n code digits generated in a particular time unit depend only on the k message digits within that time unit. Some of the errors can be detected and corrected if $d \geq s + t + 1$, where s is the number of errors that can be detected, t is the number of errors that can be corrected, and d is the hamming distance. Usually, $s \geq t$, thus, $d \geq 2t + 1$. A general code word can be expressed as $a_1, a_2,...,a_k, c_1, c_2,...,c_r$. k is the number of information bits and r is the number of check bits. Total word length is $n = k + r$.

In Fig. 70.1, the gain h_{ij} $(i = 1, 2,..., r, j = 1, 2,..., k)$ are elements of the parity check matrix **H**. The k data bits are shifted in each time, while $k + r$ bits are simultaneously shifted out by the commutator.

Cyclic Codes

Cyclic codes are block codes such that another code word can be obtained by taking any one code word, shifting the bits to the right, and placing the dropped-off bits on the left. An encoding circuit with $(n - k)$ shift registers is shown in Fig. 70.2.

TABLE 70.1 Properties of Block Codes

	Code[a]			
Property	BCH	Reed–Solomon	Hamming	Maximal Length
Block length	$n = 2^m - 1$, $m = 3, 4, 5, \ldots$	$n = m(2^m - 1)$ bits	$n = 2^m - 1$	$n = 2^m - 1$
Number of parity bits		$r = m2t$ bits	$r = m$	
Minimum distance	$d \geq 2t + 1$	$d = m(2t + 1)$ bits	$d = 3$	$d = 2^m - 1$
Number of information bits	$k \geq n - mt$			$k = m$

[a] m is any positive integer unless otherwise indicated; n is the block length; k is the number of information bits; t is the number of errors that can be corrected; r is the number of parity bits; d is the distance.

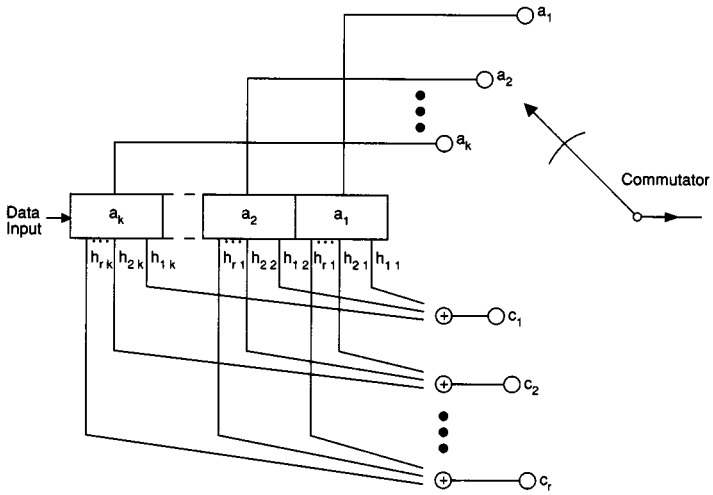


FIGURE 70.1 An encoding circuit of (n, k) block code.

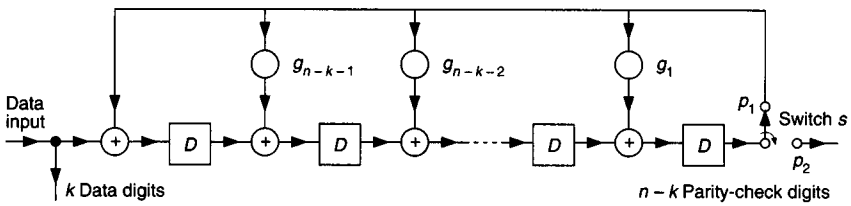


FIGURE 70.2 An encoder for systematic cyclic code. (*Source:* B.P. Lathi, *Modern Digital and Analog Communications,* New York: CBS College Publishing, 1983. With permission.)

In Fig. 70.2, the gain g_ks are the coefficients of the generator polynomial $g(x) = x^{n-k} + g_1x^{n-k-1} + \ldots + g_{n-k-1}x + 1$. The gains g_k are either 0 or 1. The k data digits are shifted in one at a time at the input with the switch s held at position p_1. The symbol D represents a one-digit delay. As the data digits move through the encoder, they are also shifted out onto the output lines, because the first k digits of code word are the data digits themselves. As soon as the last (or kth) data digit clears the last $(n - k)$ register, all the registers contain the parity-check digits. The switch s is now thrown to position p_2, and the $n - k$ parity-check digits are shifted out one at a time onto the line.

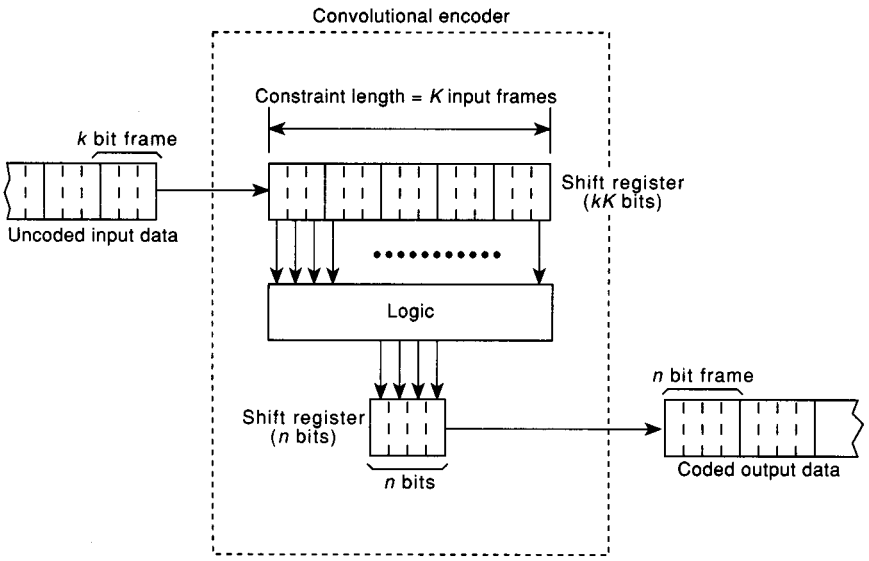


FIGURE 70.3 Convolutional encoding ($k = 3$, $n = 4$, $K = 5$, and $R = 3/4$).

Examples of cyclic and related codes are

1. Bose–Chaudhuri–Hocquenhem (BCH)
2. Reed–Solomon
3. Hamming
4. Maximal length
5. Reed–Muller
6. Golay codes

Convolutional Codes

In convolutional code, the block of n code digits generated by the encoder in a particular time unit depends not only on the block of k message digits within that time unit but also on the block of data digits within a previous span of $N - 1$ time units ($N > 1$). A convolutional encoder is illustrated in Fig. 70.3.

Here k bits (one input frame) are shifted in each time, and concurrently n bits (the output frame) are shifted out, where $n > k$. Thus, every k-bit input frame produces an n-bit output frame. Redundancy is provided in the output, since $n > k$. Also, there is memory in the coder, since the output frame depends on the previous K input frames where $K > 1$. The *code rate* is $R = k/n$, which is 3/4 in this illustration. The *constraint length*, K, is the number of input frames that are held in the kK bit shift register. Depending on the particular convolutional code that is to be generated, data from the kK stages of the shift register are added (modulo 2) and used to set the bits in the n-stage output register.

Code Performance

The improvement in the performance of a digital communication system that can be achieved by the use of coding is illustrated in Fig. 70.4. It is assumed that a digital signal plus channel noise is present at the receiver input. The performance of a system that uses binary-phase-shift-keyed (BPSK) signaling is shown both for the case when coding is used and for the case when there is no coding. For the BPSK no code case, $P_e = Q(\sqrt{2(E_b/N_o)})$. For the coded case a (23,12) Golay code is used; P_e is the *probability of bit error*—also called the *bit error rate* (BER)—that is measured at the receiver output.

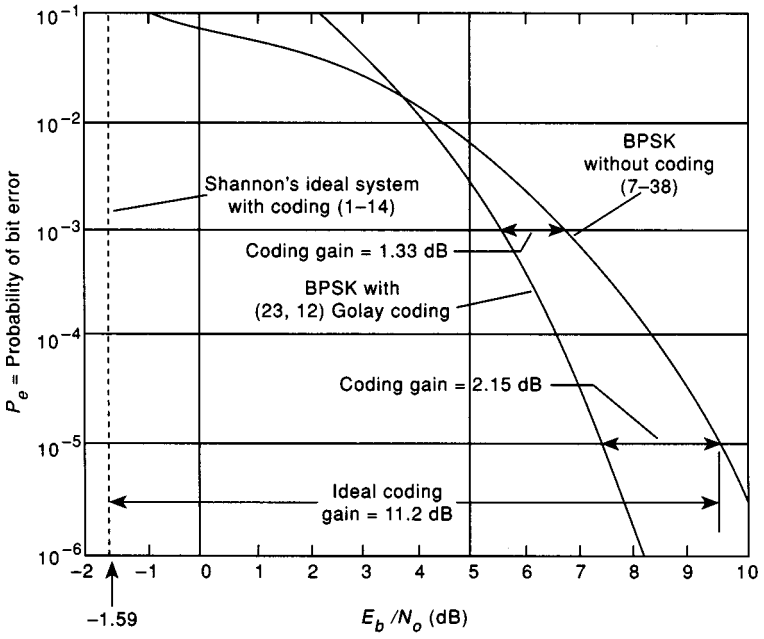


FIGURE 70.4 Performance of digital systems—with and without coding. E_b is the energy-per-bit to noise-density at the receiver input. The function $Q(x)$ is $Q(x) = (1/\sqrt{2\pi x})e^{-x2/2}$.

TABLE 70.2 Coding Gains with BPSK or QPSK

Coding Technique Used	Coding Gain (dB) at 10^{-5} BER	Coding Gain (dB) at 10^{-8} BER	Data Rate Capability
Ideal coding	11.2	13.6	
Concatenated Reed–Solomon and convolution (Viterbi decoding)	6.5–7.5	8.5–9.5	Moderate
Convolutional with sequential decoding (soft decisions)	6.0–7.0	8.0–9.0	Moderate
Block codes (soft decisions)	5.0–6.0	6.5–7.5	Moderate
Concatenated Reed–Solomon and short block	4.5–5.5	6.5–7.5	Very high
Convolutional with Viterbi decoding	4.0–5.5	5.0–6.5	High
Convolutional with sequential decoding (hard decisions)	4.0–5.0	6.0–7.0	High
Block codes (hard decisions)	3.0–4.0	4.5–5.5	High
Block codes with threshold decoding	2.0–4.0	3.5–5.5	High
Convolutional with threshold decoding	1.5–3.0	2.5–4.0	Very high

BPSK: modulation technique—binary phase-shift keying; QPSK: modulation technique—quadrature phase-shift keying; BER: bit error rate.

Source: V.K. Bhargava, "Forward error correction schemes for digital communications," *IEEE Communication Magazine*, 21, 11–19, © 1983 IEEE. With permission.

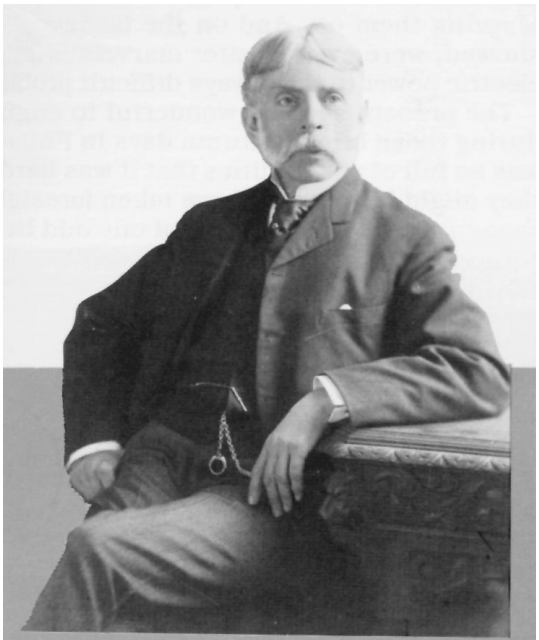
Trellis-Coded Modulation

Trellis-coded modulation (TCM) combines multilevel modulation and coding to achieve coding gain without bandwidth expansion [Ungerboeck, 1982, 1987]. TCM has been adopted for use in the new CCITT V.32 modem that allows an information data rate of 9600 b/s (bits per second) to be transmitted over VF (voice frequency) lines. The TCM has a coding gain of 4 dB [Wei, 1984]. The combined modulation and coding operation of TCM is shown in Fig. 70.5(b). Here, the serial data from the source, $m(t)$, are converted into parallel (m-bit)

George Anson Hamilton (1843–1935)

Telegraphy captivated George Hamilton's interest while he was still a boy — to the extent that he built a small telegraph line himself, from sinking the poles to making the necessary apparatus. By the time he was 17, he was the manager of the telegraph office of the Atlantic & Great Western Railroad at Ravenna, Ohio. Hamilton continued to hold managerial positions with telegraph companies until 1873 when he became assistant to Moses G. Farmer in his work on general electrical apparatus and machinery.

In 1875, Hamilton joined Western Union as assistant electrician and, for the next two years, worked with Gerritt Smith in establishing and maintaining the first quadruplex telegraph circuits in both America and England. He then focused on the development of the Wheatstone high-speed automatic system and was also the chief electrician on the Key West–Havana cable repair expedition. Hamilton left Western Union in 1889, however, to join Western Electric, where he was placed in charge of the production of fine electrical instruments until the time of his retirement. (Courtesy of the IEEE Center for the History of Electrical Engineering.)



George Anson Hamilton (1843-1935)

data, which are partitioned into k-bit and $(m - k)$-bit words where $k \geq m$. The k-bit words (frames) are convolutionally encoded into $(n = k + 1)$-bit words so that the code rate is $R = k/(k + 1)$. The amplitude and phase are then set jointly on the basis of the coded n-bit word and the uncoded $(m - k)$-bit word. Almost 6 dB of coding gain can be realized if coders of constraint length 9 are used.

Defining Terms

Block code: A mapping of k input binary symbols into n output binary symbols.

Convolutional code: A subset of tree codes, accepting k binary symbols at its input and producing n binary symbols at its output.

Cyclic code: Block code such that another code word can be obtained by taking any one code word, shifting the bits to the right, and placing the dropped-off bits on the left.

Tree code: Produced by a coder that has memory.

Related Topics

69.1 Modulation • 70.2 Equalization

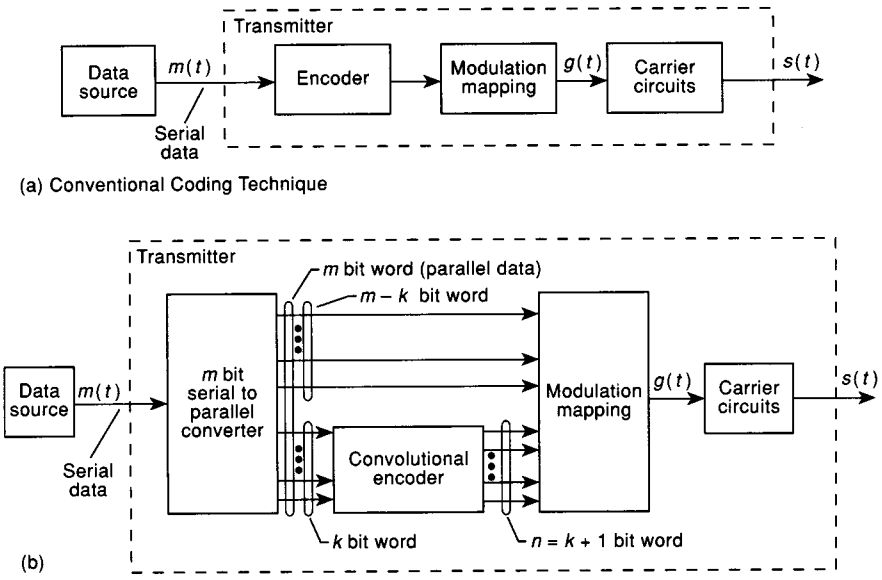


FIGURE 70.5 Transmitters for conventional coding and for TCM.

References

V.K. Bhargava, "Forward error correction schemes for digital communications," *IEEE Communication Magazine*, 21, 1983.

G.C. Clark and J.B. Cain, *Error-Correction Coding for Digital Communications*, New York: Plenum, 1981.

L.W. Couch, *Digital and Analog Communication Systems*, New York: Macmillan, 1990.

B.P. Lathi, *Modern Digital and Analog Communication*, New York: CBS College Publishing, 1983.

G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, vol. IT-28 (January), pp. 55–67, 1982.

G. Ungerboeck, "Trellis-coded modulation with redundant signal sets," Parts 1 and 2, *IEEE Communications Magazine*, vol. 25, no. 2 (February), pp. 5–21, 1987.

L. Wei, "Rotationally invariant convolutional channel coding with expanded signal space—Part II: Nonlinear codes," *IEEE Journal on Selected Areas in Communications*, vol. SAC-2, no. 2, pp. 672–686, 1984.

Further Information

For further information refer to *IEEE Communications* and *IEEE Journal on Selected Areas in Communications*.

70.2 Equalization

Richard C. Dorf and Zhen Wan

In bandwidth-efficient digital communication systems the effect of each symbol transmitted over a time dispersive channel extends beyond the time interval used to represent that symbol. The distortion caused by the resulting overlap of received symbols is called **intersymbol interference** (ISI) [Lucky et al., 1968]. ISI arises in all pulse-modulation systems, including frequency-shift keying (FSK), phase-shift keying (PSK), and quadrature amplitude modulation (QAM) [Lucky et al., 1968]. However, its effect can be most easily described for a baseband PAM system.

The purpose of an **equalizer,** placed in the path of the received signal, is to reduce the ISI as much as possible to maximize the probability of correct decisions.

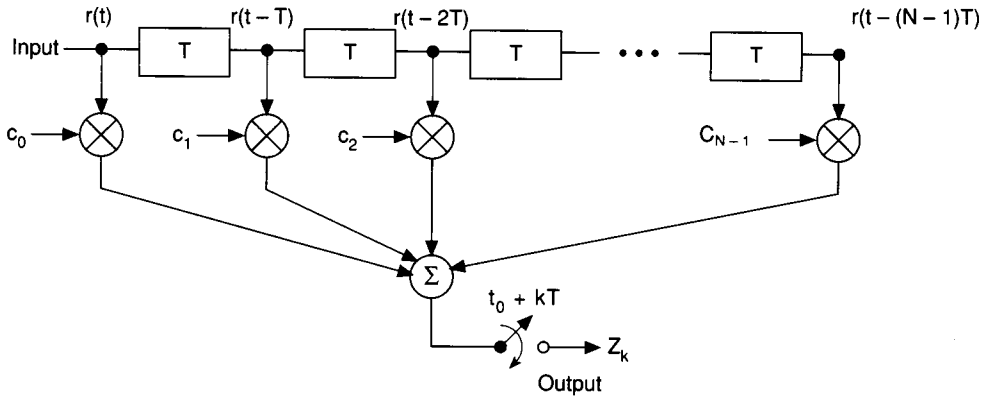


FIGURE 70.6 Linear transversal equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 648. With permission.)

Linear Transversal Equalizers

Among the many structures used for equalization, the simplest is the transversal (tapped delay line or nonrecursive) equalizer shown in Fig. 70.6. In such an equalizer the current and past values $r(t - nT)$ of the received signal are linearly weighted by equalizer coefficients (tap gains) c_n and summed to produce the output. In the commonly used digital implementation, samples of the received signal at the symbol rate are stored in a digital shift register (or memory), and the equalizer output samples (sums of products) $z(t_0 + kT)$ or z_k are computed digitally, once per symbol, according to

$$z_k = \sum_{n=0}^{N-1} c_n r(t_0 + kT - nt)$$

where N is the number of equalizer coefficients and t_0 denotes sample timing.

The equalizer coefficients, c_n, $n = 0, 1, \ldots, N-1$, may be chosen to force the samples of the combined channel and equalizer impulse response to zero at all but one of the NT-spaced instants in the span of the equalizer. Such an equalizer is called a *zero-forcing* (ZF) equalizer [Lucky, 1965].

If we let the number of coefficients of a ZF equalizer increase without bound, we would obtain an infinite-length equalizer with zero ISI at its output. An infinite-length zero-ISI equalizer is simply an inverse filter, which inverts the folded frequency response of the channel. Clearly, the ZF criterion neglects the effect of noise altogether. A finite-length ZF equalizer is approximately inverse to the folded frequency response of the channel. Also, a finite-length ZF equalizer is guaranteed to minimize the peak distortion or worst-case ISI only if the peak distortion before equalization is less than 100% [Lucky, 1965].

The *least-mean-squared* (LMS) equalizer [Lucky et al.,1968] is more robust. Here the equalizer coefficients are chosen to minimize the mean squared error (MSE)—the sum of squares of all the ISI terms plus the noise power at the output of the equalizer. Therefore, the LMS equalizer maximizes the signal-to-distortion ratio (S/D) at its output within the constraints of the equalizer time span and the delay through the equalizer.

Automatic Synthesis

Before regular data transmission begins, automatic synthesis of the ZF or LMS equalizers for unknown channels may be carried out during a training period. During the training period, a known signal is transmitted and a synchronized version of this signal is generated in the receiver to acquire information about the channel characteristics. The automatic adaptive equalizer is shown in Fig. 70.7. A noisy but unbiased estimate:

$$\frac{\delta e_k^2}{\delta c_n(k)} = 2e_k r(t_0 + kT - nT)$$

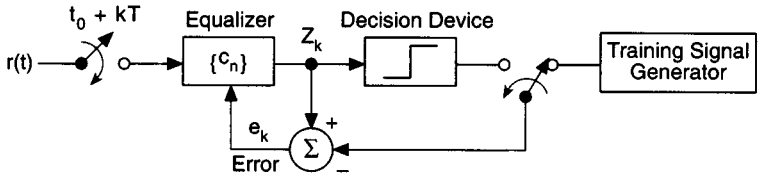


FIGURE 70.7 Automatic adaptive equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 651. With permission.)

is used. Thus, the tap gains are updated according to

$$c_n(k+1) = c_n(k) - \Delta e_k r(t_0 + kT - nT), \quad n = 0, 1, \ldots, N-1$$

where $c_n(k)$ is the nth tap gain at time k, e_k is the error signal, and Δ is a positive adaptation constant or step size, error signals $e_k = z_k - q_k$ can be computed at the equalizer output and used to adjust the equalizer coefficients to reduce the sum of the squared errors. Note $q_k = \hat{x}_k$.

The most popular equalizer adjustment method involves updates to each tap gain during each symbol interval. The adjustment to each tap gain is in a direction opposite to an estimate of the gradient of the MSE with respect to that tap gain. The idea is to move the set of equalizer coefficients closer to the unique optimum set corresponding to the minimum MSE. This symbol-by-symbol procedure developed by Widrow and Hoff [Feher, 1987] is commonly referred to as the *stochastic gradient* method.

Adaptive Equalization

After the initial training period (if there is one), the coefficients of an adaptive equalizer may be continually adjusted in a *decision-directed* manner. In this mode the error signal $e_k = z_k - q_k$ is derived from the final (not necessarily correct) receiver estimate $\{q_k\}$ of the transmitted sequence $\{x_k\}$ where q_k is the estimate of x_k. In normal operation the receiver decisions are correct with high probability, so that the error estimates are correct often enough to allow the adaptive equalizer to maintain precise equalization. Moreover, a decision-directed adaptive equalizer can track slow variations in the channel characteristics or linear perturbations in the receiver front end, such as slow jitter in the sampler phase.

Nonlinear Equalizers

Decision-Feedback Equalizers

A decision-feedback equalizer (DFE) is a simple nonlinear equalizer [Monsen, 1971], which is particularly useful for channels with severe amplitude distortion and uses decision feedback to cancel the interference from symbols which have already been detected. Fig. 70.8 shows the diagram of the equalizer.

The equalized signal is the sum of the outputs of the forward and feedback parts of the equalizer. The forward part is like the linear transversal equalizer discussed earlier. Decisions made on the equalized signal are fed back via a second transversal filter. The basic idea is that if the values of the symbols already detected are known (past decisions are assumed to be correct), then the ISI contributed by these symbols can be canceled exactly, by subtracting past symbol values with appropriate weighting from the equalizer output.

The forward and feedback coefficients may be adjusted simultaneously to minimize the MSE. The update equation for the forward coefficients is the same as for the linear equalizer. The feedback coefficients are adjusted according to

$$b_m(k+1) = b_m(k) + \Delta e_k \hat{x}_{k-m} \quad m = 1, \ldots, M$$

where $\hat{x}_k$ is the kth symbol decision, $b_m(k)$ is the mth feedback coefficient at time k, and there are M feedback coefficients in all. The optimum LMS settings of b_m, $m = 1, \ldots, M$, are those that reduce the ISI to zero, within the span of the feedback part, in a manner similar to a ZF equalizer.

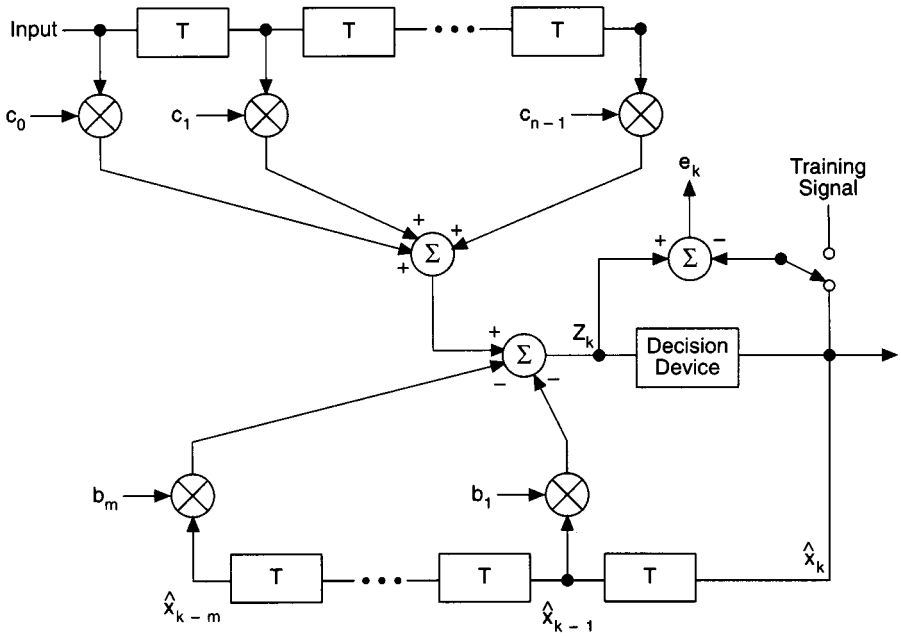


FIGURE 70.8 Decision-feedback equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 655. With permission.)

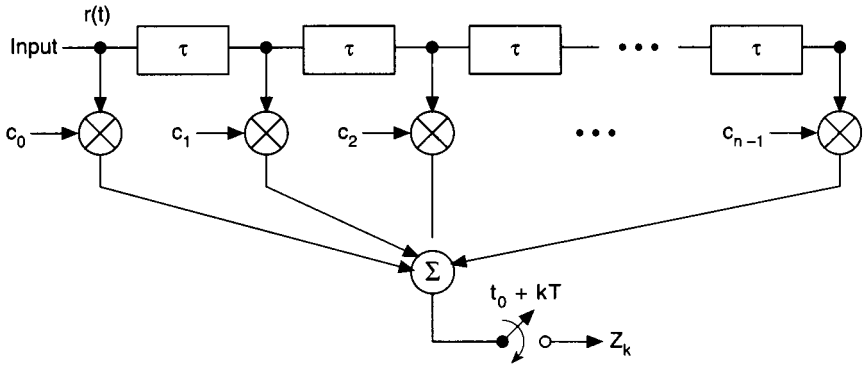


FIGURE 70.9 Fractionally spaced equalizer. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, p. 656. With permission.)

Fractionally Spaced Equalizers

The optimum receive filter in a linear modulation system is the cascade of a filter matched to the actual channel, with a transversal *T*-spaced equalizer [Forney, 1972]. The fractionally spaced equalizer (FSE), by virtue of its sampling rate, can synthesize the best combination of the characteristics of an adaptive matched filter and a *T*-spaced equalizer, within the constraints of its length and delay. A *T*-spaced equalizer, with symbol-rate sampling at its input, cannot perform matched filtering. A *fractionally spaced equalizer* can effectively compensate for more severe delay distortion and deal with amplitude distortion with less noise enhancement than a *T*-equalizer.

A fractionally spaced transversal equalizer [Monsen, 1971] is shown in Fig. 70.9. The delay-line taps of such an equalizer are spaced at an interval τ, which is less than, or a fraction of, the symbol interval *T*. The tap spacing τ is typically selected such that the bandwidth occupied by the signal at the equalizer input is $|f| <$

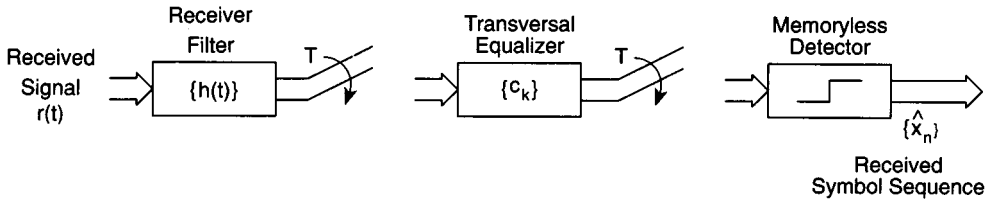


FIGURE 70.10 Conventional linear receiver.

$1/2\tau$: that is, τ-spaced sampling satisfies the sampling theorem. In an analog implementation, there is no other restriction on τ, and the output of the equalizer can be sampled at the symbol rate. In a digital implementation τ must be KT/M, where K and M are integers and $M > K$. (In practice, it is convenient to choose $\tau = T/M$, where M is a small integer, e.g., 2.) The received signal is sampled and shifted into the equalizer delay line at a rate M/T, and one input is produced each symbol interval (for every M input sample). In general, the equalizer output is given by

$$z_k = \sum_{n=0}^{N-1} c_n r\left(t_0 + kT - \frac{nKT}{M}\right)$$

The coefficients of a KT/M equalizer may be updated once per symbol based on the error computed for that symbol according to

$$c_n(k+1) = c_n(k) - \Delta e_k r\left(t_0 + kT - \frac{nKT}{M}\right), \qquad n = 0, 1, \ldots, N-1$$

Linear Receivers

When the channel does not introduce any amplitude distortion, the linear receiver is optimum with respect to the ultimate criterion of minimum probability of symbol error. The *conventional linear receiver* consists of a matched filter, a symbol-rate sampler, an infinite-length T-spaced equalizer, and a memoryless detector. The linear receiver structure is shown in Fig. 70.10.

In the conventional linear receiver, a memoryless threshold detector is sufficient to minimize the probability of error; the equalizer response is designed to satisfy the zero-ISI constraint, and the matched filter is designed to minimize the effect of the noise while maximizing the signal.

Matched Filter

The matched filter is the linear filter that maximizes $(S/N)_{\text{out}} = s_0^2(t)/n_0^2(t)$ of Fig. 70.11 and has a transfer function given by

$$H(f) = K \frac{S^*(f)}{P_n(f)} e^{-j\omega t_0}$$

where $S(f) = F[s(t)]$ is the Fourier transform of the known input signal $s(t)$ of duration T sec. $P_n(f)$ is the PSD of the input noise, t_0 is the sampling time when $(S/N)_{\text{out}}$ is evaluated, and K is an arbitrary real nonzero constant.

A general representation for a matched filter is illustrated in Fig. 70.11. The input signal is denoted by $s(t)$ and the output signal by $s_0(t)$. Similar notation is used for the noise.

Nonlinear Receivers

When amplitude distortion is present in the channel, a memoryless detector operating on the output of this receiver filter no longer minimizes symbol error probability. Recognizing this fact, several authors have investigated optimum or approximately optimum nonlinear receiver structures subject to a variety of criteria [Lucky, 1973].

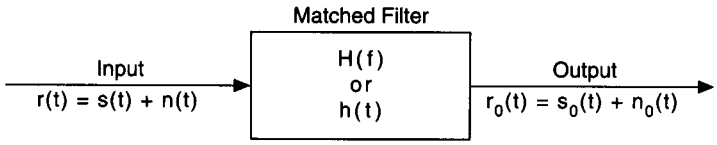


FIGURE 70.11 Matched filter.

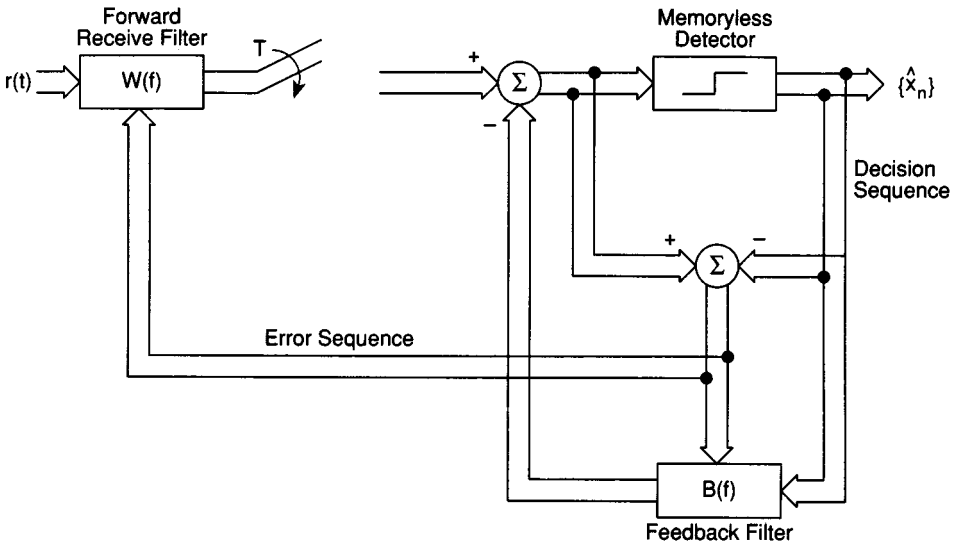


FIGURE 70.12 Conventional decision-feedback receiver. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 675. With permission.)

Decision-Feedback Equalizers

A DFE takes advantage of the symbols that have already been detected (correctly with high probability) to cancel the ISI due to these symbols without noise enhancement. A DFE makes memoryless decisions and cancels all trailing ISI terms. Even when the whitened matched filter (WMF) is used as the receive filter for the DFE, the DFE suffers from a reduced effective signal-to-noise ratio, and error propagation, due to its inability to defer decisions.

An infinite-length DFE receiver takes the general form (shown in Fig. 70.12) of a forward linear receive filter, symbol-rate sampler, canceler, and memoryless detector. The symbol-rate output of the detector is then used by the feedback filter to generate future outputs for cancellation.

Adaptive Filters for MLSE

For unknown and/or slowly time-varying channels, the receive filter must be adaptive in order to obtain the ultimate performance gain from MLSE (maximum-likelihood sequence estimation). Secondly, the complexity of the MLSE becomes prohibitive for practical channels with a large number of ISI terms. Therefore, in a practical receiver, an adaptive receive filter may be used prior to Viterbi detection to limit the time spread of the channel as well as to track slow time variation in the channel characteristics [Falconer and Magee, 1973].

Several adaptive receive filters are available that minimize the MSE at the input to the Viterbi algorithm. These methods differ in the form of constraint [Falconer and Magee, 1973] on the desired impulse response (DIR) which is necessary in this optimization process to exclude the selection of the null DIR corresponding to no transmission through the channel. The general form of such a receiver is shown in Fig. 70.13.

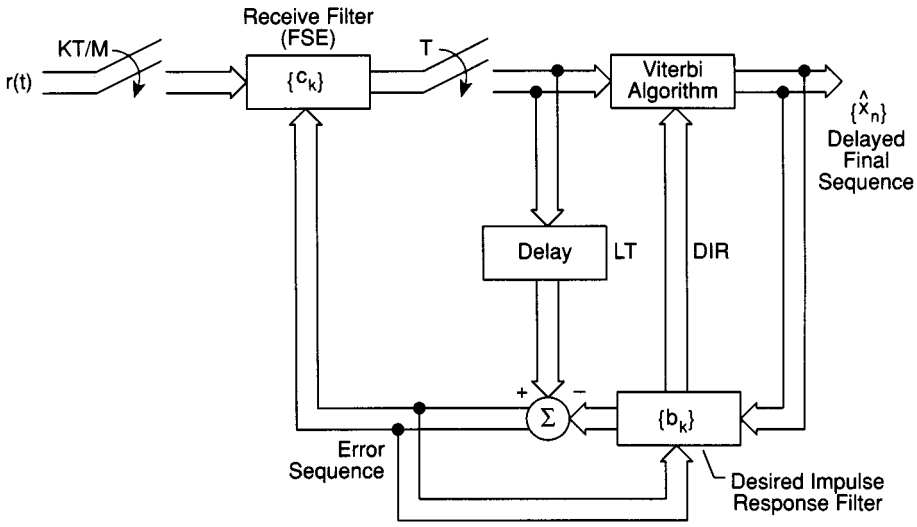


FIGURE 70.13 General form of adaptive MLSE receiver with finite-length DIR. (*Source:* K. Feher, *Advanced Digital Communications*, Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 684. With permission.)

One such constraint is to restrict the DIR to be causal and to restrict the first coefficient of the DIR to be unity. In this case the delay (LT) in Fig. 70.13 is equal to the delay through the Viterbi algorithm and the first coefficient of $\{b_k\}$ is constrained to be unity.

The least restrictive constraint on the DIR is the unit energy constraint proposed by Falconer and Magee [1973]. This leads to yet another form of the receiver structure as shown in Fig. 70.13. However, the adaptation algorithm for updating the DIR coefficients $\{b_k\}$ is considerably more complicated [Falconer and Magee, 1973]. Note that the fixed predetermined WMF and T-spaced prefilter combination of Falconer and Magee [1973] has been replaced in Fig. 70.13 by a general fractionally spaced adaptive filter.

Defining Terms

Equalizer: A filter used to reduce the effect of intersymbol interference.
Intersymbol interference: The distortion caused by the overlap (in time) of adjacent symbols.

Related Topic

70.1 Coding

References

L.W. Couch, *Digital and Analog Communication Systems,* New York: Macmillan, 1990.

D.D. Falconer and F.R. Magee, Jr., "Adaptive channel memory truncation for maximum likelihood sequence estimation," *Bell Syst. Technical Journal,* vol. 5, pp. 1541–1562, November 1973.

K. Feher, *Advanced Digital Communications,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

G.D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Information Theory,* vol. IT-88, pp. 363–378, May 1972.

R.W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. Journal,* vol. 44, pp. 547–588, April 1965.

R.W. Lucky, "A survey of the communication theory literature: 1968–1973," *IEEE Trans. Information Theory,* vol. 52, pp. 1483–1519, November 1973.

R.W. Lucky, J. Salz, and E.J. Weldon, Jr., *Principles of Data Communication,* New York: McGraw-Hill, 1968.

P. Monsen, "Feedback equalization for fading dispersive channels," *IEEE Trans. Information Theory,* vol. IT-17, pp. 56–64, January 1971.

70.3 Spread Spectrum Communications[1]

L.B. Milstein and M.K. Simon

A Brief History

Spread spectrum (SS) has its origin in the military arena where the friendly communicator is (1) susceptible to detection/interception by the enemy and (2) vulnerable to intentionally introduced unfriendly interference (jamming). Communication systems that employ spread spectrum to reduce the communicator's detectability and combat the enemy-introduced interference are respectively referred to as **low probability of intercept (LPI)** and **antijam** (AJ) **communication systems.** With the change in the current world political situation wherein the U.S. Department of Defense (DOD) has reduced its emphasis on the development and acquisition of new communication systems for the original purposes, a host of new commercial applications for SS has evolved, particularly in the area of cellular mobile communications. This shift from military to commercial applications of SS has demonstrated that the basic concepts that make SS techniques to useful in the military can also be put to practical peacetime use. In the next section, we give a simple description of these basic concepts using the original military application as the basis of explanation. The extension of these concepts to the mentioned commercial applications will be treated later on in the chapter.

Why Spread Spectrum?

Spread spectrum is a communication technique wherein the transmitted modulation is *spread* (increased) in bandwidth prior to transmission over the channel and then *despread* (decreased) in bandwidth by the same amount at the receiver. If it were not for the fact that the communication channel introduces some form of narrowband (relative to the spread bandwidth) interference, the receiver performance would be transparent to the spreading and despreading operations (assuming that they are identical inverses of each other). That is, after **despreading** the received signal would be identical to the transmitted signal prior to **spreading.** In the presence of narrowband interference, however, there is a significant advantage to employing the spreading/despreading procedure described. The reason for this is as follows. Since the interference is introduced after the transmitted signal is spread, then, whereas the despreading operation at the receiver shrinks the desired signal back to its original bandwidth, at the same time it spreads the undesired signal (interference) in bandwidth by the same amount, thus reducing its power spectral density. This, in turn, serves to diminish the effect of the interference on the receiver performance, which depends on the amount of interference power in the spread bandwidth. It is indeed this very simple explanation, which is at the heart of all spread spectrum techniques.

Basic Concepts and Terminology

To describe this process analytically and at the same time introduce some terminology that is common in spread spectrum parlance, we proceed as follows. Consider a communicator that desires to send a message using a transmitted power S Watts (W) at an information rate R_b bits/s (bps). By introducing a SS modulation, the bandwidth of the transmitted signal is increased from R_b Hz to W_{ss} Hz where $W_{ss} \gg R_b$ denotes the **spread spectrum bandwidth.** Assume that the channel introduces, in addition to the usual thermal noise (assumed to have a single-sided power spectral density (PSD) equal to N_0 W/Hz), an additive interference (jamming) having power J distributed over some bandwidth W_J. After despreading, the desired signal bandwidth is once again now equal to R_b Hz and the interference PSD is now $N_J = J/W_{ss}$. Note that since the thermal noise is assumed to be white, i.e., it is uniformly distributed over all frequencies, its PSD is unchanged by the despreading operation and, thus, remains equal to N_0. Regardless of the signal and interferer waveforms, the equivalent bit energy-to-total noise ratio is, in terms of the given parameters,

[1]The material in this article was previously published by CRC Press in *The Mobile Communications Handbook,* Jerry P. Gibson, Editor-in-Chief, 1996.

$$\frac{E_b}{N_t} = \frac{E_b}{N_0 + N_J} = \frac{S/R_b}{N_0 + J/W_{ss}} \tag{70.1}$$

For most practical scenarios, the jammer limits performance and, thus, the effects of receiver noise in the channel can be ignored. Thus, assuming $N_J \gg N_0$, we can rewrite Eq. (70.1) as

$$\frac{E_b}{N_t} \cong \frac{E_b}{N_J} = \frac{S/R_b}{J/W_{ss}} = \frac{S}{J}\frac{W_{ss}}{R_b} \tag{70.2}$$

where the ratio J/S is the *jammer-to-signal power ratio* and the ratio W_{ss}/R_b is the **spreading ratio** and is defined as the **processing gain** of the system. Since the ultimate error probability performance of the communication receiver depends on the ratio E_b/N_J, we see that from the communicator's viewpoint his goal should be to minimize J/S (by choice of S) and maximize the processing gain (by choice of W_{ss} for a given desired information rate). The possible strategies for the jammer will be discussed in the section on military applications dealing with AJ communications.

Spread Spectrum Techniques

By far the two most popular spreading techniques are **direct sequence (DS) modulation** and **frequency hopping (FH) modulation.** In the following subsections, we present a brief description of each.

Direct Sequence Modulation

A direct sequence modulation $c(t)$ is formed by linearly modulating the output sequence $\{c_n\}$ of a pseudorandom number generator onto a train of pulses, each having a duration T_c called the **chip time.** In mathematical form,

$$c(t) = \sum_{n=-\infty}^{\infty} c_n p(t - nT_c) \tag{70.3}$$

where $p(t)$ is the basic pulse shape and is assumed to be of rectangular form. This type of modulation is usually used with binary phase-shift-keyed (BPSK) information signals, which have the complex form $d(t)\exp\{j(2\pi f_c t + \theta_c)\}$, where $d(t)$ is a binary-valued data waveform of rate $1/T_b$ bit/s and f_c and θ_c are the frequency and phase of the data-modulated carrier, respectively. As such, a DS/BPSK signal is formed by multiplying the BPSK signal by $c(t)$ (see Fig. 70.14), resulting in the real transmitted signal

$$x(t) = \mathrm{Re}\left\{c(t)d(t)\exp\left[j(2\pi f_c t + \theta_c)\right]\right\} \tag{70.4}$$

Since T_c is chosen so that $T_b \gg T_c$, then relative to the bandwidth of the BPSK information signal, the bandwidth of the DS/BPSK signal[2] is effectively increased by the ratio $T_b/T_c = W_{ss}/2R_b$, which is one-half the spreading factor or processing gain of the system. At the receiver, the sum of the transmitted DS/BPSK signal and the channel interference $I(t)$ (as discussed before, we ignore the presence of the additive thermal noise) are ideally multiplied by the identical DS modulation (this operation is known as despreading), which returns the DS/BPSK signal to its original BPSK form whereas the real interference signal is now the real wideband signal $\mathrm{Re}\{(t)c(t)\}$. In the previous sentence, we used the word ideally, which implies that the PN waveform used for despreading at the receiver is identical to that used for spreading at the transmitter. This simple implication covers up a

[2]For the usual case of a rectangular spreading pulse $p(t)$, the PSD of the DS/BPSK modulation will have $(\sin x/x)^2$ form with first zero crossing at $1/T_c$, which is nominally taken as one-half the spread spectrum bandwidth W_{ss}.

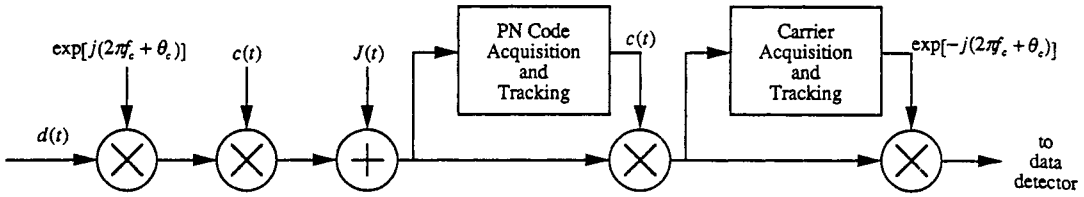


FIGURE 70.14 A DS-BPSK system (complex form).

multitude of tasks that a practical DS receiver must perform. In particular, the receiver must first acquire the PN waveform. That is, the local PN random generator that generates the PN waveform at the receiver used for despreading must be aligned (synchronized) to within one chip of the PN waveform of the received DS/BPSK signal. This is accomplished by employing some sort of **search algorithm** which typically steps the local PN waveform sequentially in time by a fraction of a chip (e.g., half a chip) and at each position searches for a high degree of correlation between the received and local PN reference waveforms. The search terminates when the correlation exceeds a given threshold, which is an indication that the alignment has been achieved. After bringing the two PN waveforms into **coarse alignment**, a **tracking algorithm** is employed to maintain **fine alignment**. The most popular forms of tracking loops are the continuous time **delay-locked loop** and its time-multiplexed version of the **tau-dither loop**. It is the difficulty in synchronizing the receiver PN generator to subnanosecond accuracy that limits PN chip rates to values on the order of hundreds of Mchips/s, which implies the same limitation on the DS spread spectrum bandwidth W_{ss}.

Frequency Hopping Modulation

A **frequency hopping (FH) modulation** $c(t)$ is formed by nonlinearly modulating a train of pulses with a sequence of pseudorandomly generated frequency shifts $\{f_n\}$. In mathematical terms, $c(t)$ has the complex form

$$c(t) = \sum_{n=-\infty}^{\infty} \exp\{j(2\pi f_n + \phi_n)\} p(t - nT_h) \tag{70.5}$$

where $p(t)$ is again the basic pulse shape having a duration T_h, called the **hop time** and $\{\phi_n\}$ is a sequence of random phases associated with the generation of the hops. FH modulation is traditionally used with multiple-frequency-shift-keyed (MFSK) information signals, which have the complex form $\exp\{j[2\pi(f_c + d(t))t]\}$, where $d(t)$ is an M-level digital waveform (M denotes the symbol alphabet size) representing the information frequency modulation at a rate $1/T_s$ symbols/s (sps). As such, an FH/MFSK signal is formed by complex multiplying the MFSK signal by $c(t)$ resulting in the real transmitted signal

$$x(t) = \mathrm{Re}\{c(t) \exp\{j[2\pi(f_c + d(t))t]\}\} \tag{70.6}$$

In reality, $c(t)$ is never generated in the transmitter. Rather, $x(t)$ is obtained by applying the sequence of pseudorandom frequency shifts $\{f_n\}$ directly to the frequency synthesizer that generates the carrier frequency f_c (see Fig. 70.15). In terms of the actual implementation, successive (not necessarily disjoint) k-chip segments of a PN sequence drive a frequency synthesizer, which hops the carrier over 2^k frequencies. In view of the large bandwidths over which the frequency synthesizer must operate, it is difficult to maintain phase coherence from hop to hop, which explains the inclusion of the sequence $\{\phi_n\}$ in the Eq. (70.5) model for $c(t)$. On a short term basis, e.g., within a given hop, the signal bandwidth is identical to that of the MFSK information modulation, which is typically much smaller than W_{ss}. On the other hand, when averaged over many hops, the signal bandwidth is equal to W_{ss}, which can be on the order of several GHz, i.e., an order of magnitude larger than that of implementable DS bandwidths. The exact relation between W_{ss}, T_h, T_s and the number of frequency shifts in the set $\{f_n\}$ will be discussed shortly.

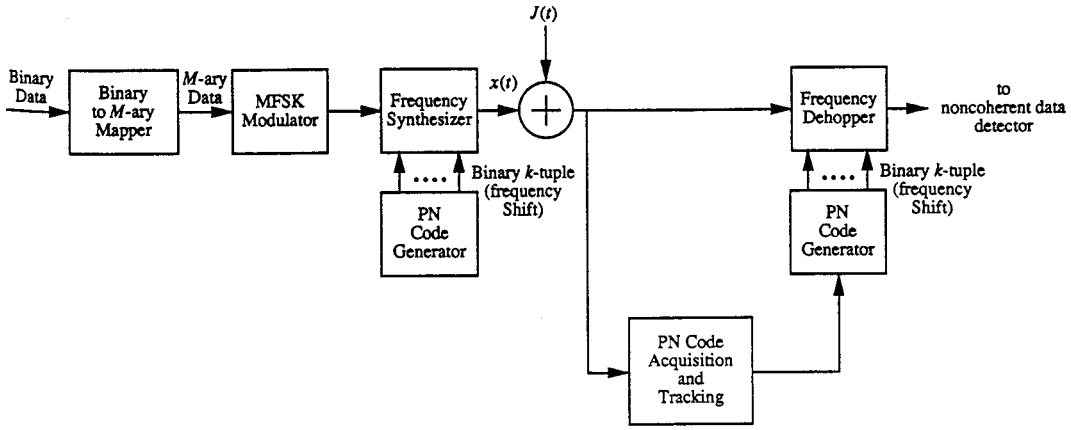


FIGURE 70.15 An FH-MFSK system.

At the receiver, the sum of the transmitted FH/MFSK signal and the channel interference $I(t)$ is ideally complex multiplied by the identical FH modulation (this operation is known as **dehopping**), which returns the FH/MFSK signal to its original MFSK form, whereas the real interference signal is now the wideband (in the average sense) signal $\mathrm{Re}\{I(t)c(t)\}$. Analogous to the DS case, the receiver must acquire and track the FH signal so that the dehopping waveform is as close to the hopping waveform $c(t)$ as possible.

FH systems are traditionally classified in accordance with the relationship between T_h and T_s. **Fast frequency-hopped (FFH)** systems are ones in which there exists one or more hops per data symbol, that is, $T_s = NT_h$ (N an integer) whereas **slow frequency-hopped (SFH)** systems are ones in which there exists more than one symbol per hop, that is, $T_h = NT_s$. It is customary in SS parlance to refer to the FH/MFSK tone of shortest duration as a "chip", despite the same usage for the PN chips associated with the code generator that drives the frequency synthesizer. Keeping this distinction in mind, in an FFH system where, as already stated, there are multiple hops per data symbol, a chip is equal to a hop. For SFH, where there are multiple data symbols per hop, a chip is equal to an MFSK symbol. Combining these two statements, the chip rate R_c in an FH system is given by the larger of $R_h = 1/T_h$ and $R_s = 1/T_s$ and, as such, is the highest system clock rate.

The frequency spacing between the FH/MFSK tones is governed by the chip rate R_c and is, thus, dependent on whether the FH modulation is FFH or SFH. In particular, for SFH where $R_c = R_s$, the spacing between FH/MFSK tones is equal to the spacing between the MFSK tones themselves. For noncoherent detection (the most commonly encountered in FH/MFSK systems), the separation of the MFSK symbols necessary to provide orthogonality[3] is an integer multiple of R_s. Assuming the minimum spacing, i.e., R_s, the entire spread spectrum band is then partitioned into a total of $N_t = W_{ss}/R_s = W_{ss}/R_c$ equally spaced FH tones. One arrangement, which is by far the most common, is to group these N_t tones into $N_b = N_t/M$ contiguous, nonoverlapping bands, each with bandwidth $M\,R_s = M\,R_c$; see Fig. 70.16(a). Assuming symmetric MFSK modulation around the carrier frequency, then the center frequencies of the $N_b = 2^k$ bands represent the set of hop carriers, each of which is assigned to a given k-tuple of the PN code generator. In this fixed arrangement, each of the N_t FH/MFSK tones corresponds to the combination of a unique hop carrier (PN code k-tuple) and a unique MFSK symbol. Another arrangement, which provides more protection against the sophisticated interferer (jammer), is to overlap adjacent M -ary bands by an amount equal to R_c; see Fig. 70.16(b). Assuming again that the center frequency of each band corresponds to a possible hop carrier, then since all but $M - 1$ of the N_t tones are available as center frequencies, the number of hop carriers has been increased from N_t/M to $N_t - (M - 1)$, which for $N_t \gg M$ is approximately an increase in randomness by a factor of M.

[3]An optimum noncoherent MFSK detector consists of a bank of energy detectors each matched to one of the M frequencies in the MFSK set. In terms of this structure, the notion of *orthogonality* implies that for a given transmitted frequency there will be no crosstalk (energy spillover) in any of the other M-1 energy detectors.

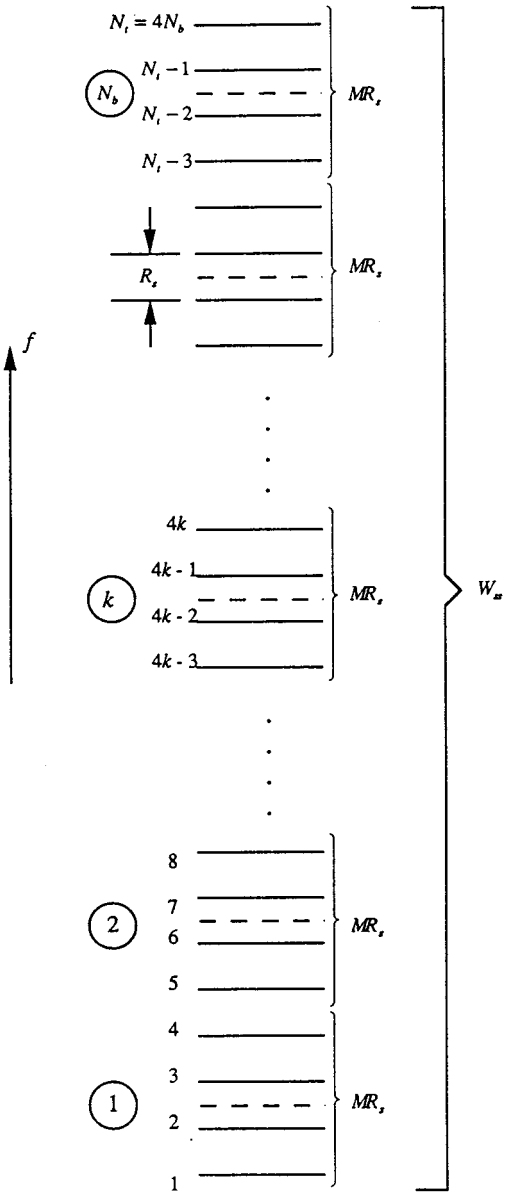


FIGURE 70.16(a) Frequency distribution for FH-4FSK —nonoverlapping bands. Dashed lines indicate lo cation of hop frequencies.

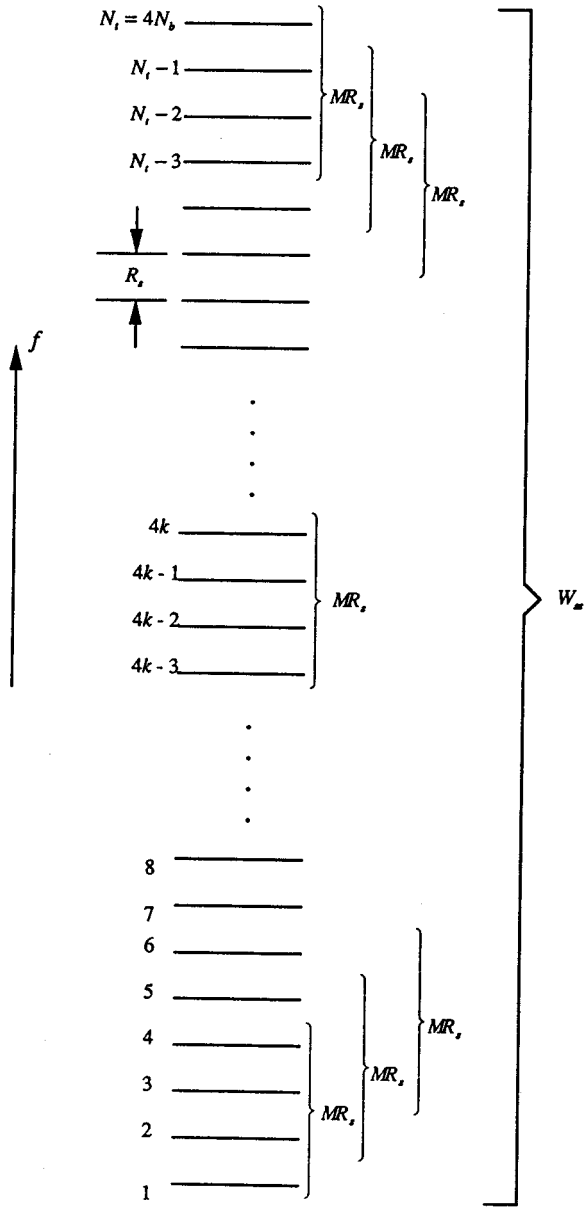


FIGURE 70.16(b) Frequency distribution for FH-4FSK — overlapping bands.

For FFH, where $R_c = R_h$, the spacing between FH/MFSK tones is equal to the hop rate. Thus, the entire spread spectrum band is partitioned into a total of $N_t = W_{ss}/R_h = W_{ss}/R_c$ equally spaced FH tones, each of which is assigned to a unique k-tuple of the PN code generator that drives the frequency synthesizer. Since for FFH there are R_h/R_s hops per symbol, then the metric used to make a noncoherent decision on a particular symbol is obtained by summing up R_h/R_s detected chip (hop) energies, resulting in a so-called *noncoherent combining loss.*

Time Hopping Modulation

Time hopping (TH) is to spread spectrum modulation what pulse position modulation (PPM) is to information modulation. In particular, consider segmenting time into intervals of T_f seconds and further segment each T_f interval into M_T increments of width T_f/M_T. Assuming a pulse of maximum duration equal to T_f/M_T, then a **time hopping spread spectrum** modulation would take the form

$$c(t) = \sum_{n=-\infty}^{\infty} p\left[t - \left(n + \frac{a_n}{M_T}\right)T_f\right] \tag{70.7}$$

where a_n denotes the pseudorandom position (one of M_T uniformly spaced locations) of the pulse within the T_r-second interval.

For DS and FH, we saw that *multiplicative* modulation, that is the transmitted signal is the product of the SS and information signals, was the natural choice. For TH, *delay* modulation is the natural choice. In particular, a TH-SS modulation takes the form

$$x(t) = \mathrm{Re}\left\{c\left(t - d(t)\right)\exp\left[j\left(2\pi f_c + \phi_T\right)\right]\right\} \tag{70.8}$$

where $d(t)$ is a digital information modulation at a rate $1/T_s$. Finally, the dehopping procedure at the receiver consists of removing the sequence of delays introduced by $c(t)$, which restores the information signal back to its original form and spreads the interferer.

Hybrid Modulations

By blending together several of the previous types of SS modulation, one can form **hybrid** modulations that, depending on the system design objectives, can achieve a better performance against the interferer than can any of the SS modulations acting alone. One possibility is to multiply several of the $c(t)$ wideband waveforms [now denoted by $c^{(i)}(t)$ to distinguish them from one another] resulting in a SS modulation of the form

$$c(t) = \prod_i c^{(i)}(t) \tag{70.9}$$

Such a modulation may embrace the advantages of the various $c^{(i)}(t)$, while at the same time mitigating their individual disadvantages.

Applications of Spread Spectrum

Military

Antijam (AJ) Communications. As already noted, one of the key applications of spread spectrum is for antijam communications in a hostile environment. The basic mechanism by which a **direct sequence spread spectrum** receiver attenuates a noise jammer was illustrated in Sec. 70.3. Therefore, in this section, we will concentrate on tone jamming.

Assume the received signal, denoted $r(t)$, is given by

$$r(t) = Ax(t) + I(t) + n_w(t) \tag{70.10}$$

where $x(t)$ is given in Eq. (70.4), A is a constant amplitude,

$$I(t) = \alpha\cos\left(2\pi f_c t + \theta\right) \tag{70.11}$$

and $n_w(t)$ is additive while Gaussian noise (AWGN) having two sided spectral density $N_0/2$. In Eq. (70.11), α is the amplitude of the tone jammer and θ is a random phase uniformly distributed in $[0, 2\pi]$.

If we employ the standard correlation receiver of Fig. 70.17, it is straightforward to show that the final test statistic out of the receiver is given by

$$g(T_b) = AT_b + \alpha\cos\theta\int_0^{T_b} c(t)\mathrm{d}t + N(T_b) \tag{70.12}$$

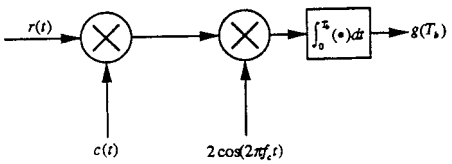


FIGURE 70.17

where $N(T_b)$ is the contribution to the test statistic due to the AWGN. Noting that, for rectangular chips, we can express

$$\int_0^{T_b} c(t)\,dt = T_c \sum_{i=1}^{M} c_i \tag{70.13}$$

where

$$M \triangleq \frac{T_b}{T_c} \tag{70.14}$$

is one-half of the processing gain. it is straightforward to show that, for a given value of θ, the signal-to-noise-plus-interference ratio, denoted by S/N_{total}, is given by

$$\frac{S}{N_{\text{total}}} = \frac{1}{\dfrac{N_0}{2E_b} + \left(\dfrac{J}{MS}\right)\cos^2\theta} \tag{70.15}$$

In Eq. (70.15), the jammer power is

$$J \triangleq \frac{\alpha^2}{2} \tag{70.16}$$

and the signal power is

$$S \triangleq \frac{A^2}{2} \tag{70.17}$$

If we look at the second term in the denominator of Eq. (70.15), we see that the ratio *J/S* is divided by *M*. Realizing that *J/S* is the ratio of the jammer power to the signal power before despreading, and *J/MS* is the ratio of the same quantity after despreading, we see that, as was the case for noise jamming, the benefit of employing direct sequence spread spectrum signalling in the presence of tone jamming is to reduce the effect of the jammer by an amount on the order of the processing gain.

Finally, one can show that an estimate of the average probability of error of a system of this type is given by

$$P_e = \frac{1}{2\pi}\int_0^{2\pi} \phi\left(-\sqrt{\frac{S}{N_{\text{total}}}}\right) d\theta \tag{70.18}$$

where

$$\phi(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy \tag{70.19}$$

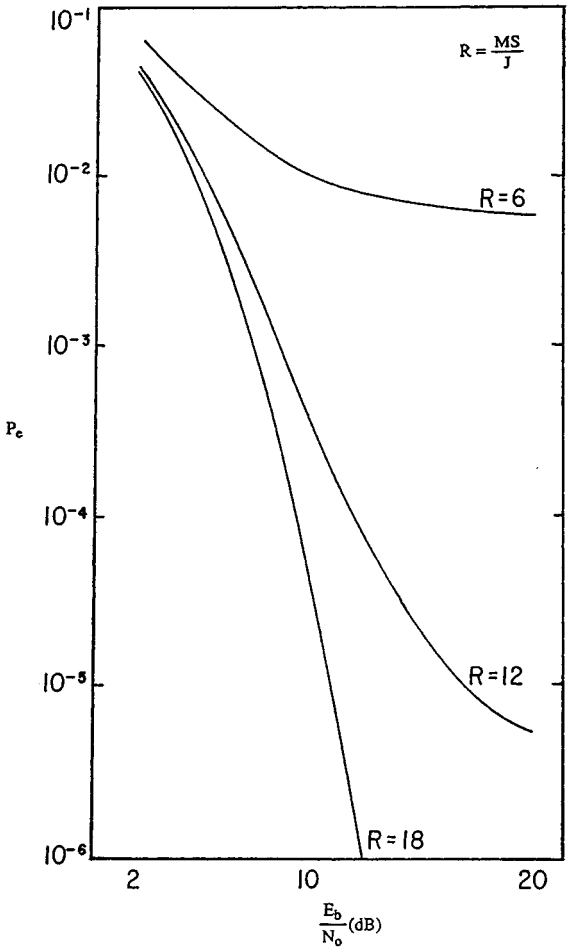


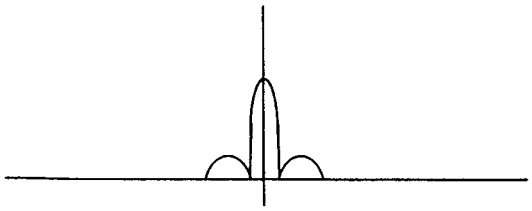
FIGURE 70.18

If Eq. (70.18) is evaluated numerically and plotted, the results are as shown in Fig. 70.18. It is clear from this figure that a large initial power advantage of the jammer can be overcome by a sufficiently large value of the processing gain.

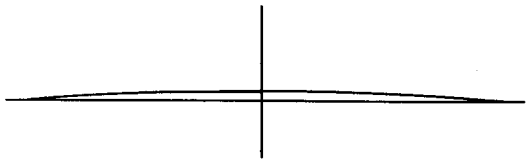
Low-Probability of Intercept (LPI). The opposite side of the AJ problem is that of LPI, that is, the desire to hide your signal from detection by an intelligent adversary so that your transmissions will remain unnoticed and, thus, neither jammed nor exploited in any manner. This idea of designing an LPI system is achieved in a variety of ways, including transmitting at the smallest possible power level, and limiting the transmission time to as short an interval in time as is possible. The choice of signal design is also important, however, and it is here that spread spectrum techniques become relevant.

The basic mechanism is reasonably straightforward; if we start with a conventional narrowband signal, say a BPSK waveform having a spectrum as shown in Fig. 70.19(a), and then spread it so that its new spectrum is as shown in Fig. 70.19(b), the peak amplitude of the spectrum after spreading has been reduced by an amount on the order of the processing gain relative to what it was before spreading. Indeed, a sufficiently large processing gain will result in the spectrum of the signal after spreading falling below the ambient thermal noise level. Thus, there is no easy way for an unintended listener to determine that a transmission is taking place.

That is not to say the spread signal cannot be detected, however, merely that it is more difficult for an adversary to learn of the transmission. Indeed, there are many forms of so-called intercept receivers that are specifically designed to accomplish this very task. By way of example, probably the best known and simplest to implement is a **radiometer,** which is just a device that measures the total power present in the received signal.



(a)



(b)

FIGURE 70.19

In the case of our intercept problem, even though we have lowered the power spectral density of the transmitted signal so that it falls below the noise floor, we have not lowered its power (i.e., we have merely spread its power over a wider frequency range). Thus, if the radiometer integrates over a sufficiently long period of time, it will eventually determine the presence of the transmitted signal buried in the noise. The key point, of course, is that the use of the spreading makes the interceptor's task much more difficult, since he has no knowledge of the spreading code and, thus, cannot despread the signal.

Commercial

Multiple Access Communications. From the perspective of commercial applications, probably the most important use of spread spectrum communications is as a multiple accessing technique. When used in this manner, it becomes an alternative to either frequency division multiple access (FDMA) or time division multiple access (TDMA) and is typically referred to as either code division multiple access (CDMA) or spread spectrum multiple access (SSMA). When using CDMA, each signal in the set is given its own spreading sequence. As opposed to either FDMA, wherein all users occupy disjoint frequency bands but are transmitted simultaneously in time, or TDMA, whereby all users occupy the same bandwidth but transmit in disjoint intervals of time, in CDMA, all signals occupy the same bandwidth and are transmitted simultaneously in time; the different waveforms in CDMA are distinguished from one another at the receiver by the specific spreading codes they employ.

Since most CDMA detectors are correlation receivers, it is important when deploying such a system to have a set of spreading sequences that have relatively low-pairwise cross-correlation between any two sequences in the set. Further, there are two fundamental types of operation in CDMA, synchronous and asynchronous. In the former case, the symbol transition times of all of the users are aligned; this allows for orthogonal sequences to be used as the spreading sequences and, thus, eliminates interference from one user to another. Alternatively, if no effort is made to align the sequences, the system operates asynchronously; in this latter mode, multiple access interference limits the ultimate channel capacity, but the system design exhibits much more flexibility.

CDMA has been of particular interest recently for applications in wireless communications. These applications include cellular communications, personal communications services (PCS), and wireless local area networks. The reason for this popularity is primarily due to the performance that spread spectrum waveforms display when transmitted over a multipath fading channel.

To illustrate this idea, consider DS signalling. As long as the duration of a single chip of the spreading sequence is less than the multipath delay spread, the use of DS waveforms provides the system designer with

one or two options. First, the multipath can be treated as a form of interference, which means the receiver should attempt to attenuate it as much as possible. Indeed, under this condition, all of the multipath returns that arrive at the receiver with a time delay greater than a chip duration from the multipath return to which the receiver is synchronized (usually the first return) will be attenuated because of the processing gain of the system.

Alternately, the multipath returns that are separated by more than a chip duration from the main path represent independent "looks" at the received signal and can be used constructively to enhance the overall performance of the receiver. That is, because all of the multipath returns contain information regarding the data that is being sent, that information can be extracted by an appropriately designed receiver. Such a receiver, typically referred to as a RAKE receiver, attempts to resolve as many individual multipath returns as possible and then to sum them coherently. This results in an *implicit* diversity gain, comparable to the use of *explicit* diversity, such as receiving the signal with multiple antennas.

The condition under which the two options are available can be stated in an alternate manner. If one envisions what is taking place in the frequency domain, it is straightforward to show that the condition of the chip duration being smaller than the multipath delay spread is equivalent to requiring that the spread bandwidth of the transmitted waveform exceed what is called the coherence bandwidth of the channel. This latter quantity is simply the inverse of the multipath delay spread and is a measure of the range of frequencies that fade in a highly correlated manner. Indeed, anytime the coherence bandwidth of the channel is less than the spread bandwidth of the signal, the channel is said to be *frequency selective* with respect to the signal. Thus, we see that to take advantage of DS signalling when used over a multipath fading channel, that signal should be designed such that it makes the channel appear frequency selective.

In addition to the desirable properties that spread spectrum signals display over multipath channels, there are two other reasons why such signals are of interest in cellular-type applications. The first has to do with a concept known as the reuse factor. In conventional cellular systems, either analog or digital, in order to avoid excessive interference from one cell to its neighbor cells, the frequencies used by a given cell are not used by its immediate neighbors (i.e., the system is designed so that there is a certain spatial separation between cells that use the same carrier frequencies). For CDMA, however, such spatial isolation is typically not needed, so that so-called *universal reuse* is possible.

Further, because CDMA systems tend to be interference limited, for those applications involving voice transmission, an additional gain in the capacity of the system can be achieved by the use of *voice activity detection.* That is, in any given two-way telephone conversation, each user is typically talking only about 50% of the time. During the time when a user is quiet, he is not contributing to the instantaneous interference. Thus, if a sufficiently large number of users can be supported by the system, statistically only about one-half of them will be active simultaneously, and the effective capacity can be doubled.

Interference Rejection. In addition to providing multiple accessing capability, spread spectrum techniques are of interest in the commercial sector for basically the same reasons they are in the military community, namely their AJ and LPI characteristics. However, the motivations for such interest differ. For example, whereas the military is interested in ensuring that systems they deploy are robust to interference generated by an intelligent adversary (i.e., exhibit jamming resistance), the interference of concern in commercial applications is unintentional. It is sometimes referred to as co-channel interference (CCI) and arises naturally as the result of many services using the same frequency band at the same time. And while such scenarios almost always allow for some type of spatial isolation between the interfering waveforms, such as the use of narrow-beam antenna patterns, at times the use of the inherent interference suppression property of a spread spectrum signal is also desired. Similarly, whereas the military is very much interested in the LPI property of a spread spectrum waveform, as indicated in Sec. 70.3, there are applications in the commercial segment where the same characteristic can be used to advantage.

To illustrate these two ideas, consider a scenario whereby a given band of frequencies is somewhat sparsely occupied by a set of conventional (i.e., nonspread) signals. To increase the overall spectral efficiency of the band, a set of spread spectrum waveforms can be overlaid on the same frequency band, thus forcing the two sets of users to share common spectrum. Clearly, this scheme is feasible only if the mutual interference that one set of users imposes on the other is within tolerable limits. Because of the interference suppression properties

of spread spectrum waveforms, the despreading process at each spread spectrum receiver will attenuate the components of the final test statistic due to the overlaid narrowband signals. Similarly, because of the LPI characteristics of spread spectrum waveforms, the increase in the overall noise level as seen by any of the conventional signals, due to the overlay, can be kept relatively small.

Defining Terms

Antijam communication system: A communication system designed to resist intentional jamming by the enemy.

Chip time (interval): The duration of a single pulse in a direct sequence modulation; typically much smaller than the formation symbol interval.

Coarse alignment: The process whereby the received signal and the despreading signal are aligned to within a single chip interval.

Dehopping: Despreading using a frequency-hopping modulation.

Delay-locked loop: A particular implementation of a closed-loop technique for maintaining fine alignment.

Despreading: The notion of decreasing the bandwidth of the received (spread) signal back to its information bandwidth.

Direct sequence modulation: A signal formed by linearly modulating the output sequence of a pseudorandom number generator onto a train of pulses.

Direct sequence spread spectrum: A spreading technique achieved by multiplying the information signal by a direct sequence modulation.

Fast frequency-hopping: A spread spectrum technique wherein the hop time is less than or equal to the information symbol interval, i.e., there exist one or more hops per data symbol.

Fine alignment: The state of the system wherein the received signal and the despreading signal are aligned to within a small fraction of a single chip interval.

Frequency-hopping modulation: A signal formed by nonlinearly modulating a train of pulses with a sequence of pseudorandomly generated frequency shifts.

Hop time (interval): The duration of a single pulse in a frequency-hopping modulation.

Hybrid spread spectrum: A spreading technique formed by blending together several spread spectrum techniques, e.g., direct sequence, frequency-hopping, etc.

Low-probability-of-intercept communication system: A communication system designed to operate in a hostile environment wherein the enemy tries to detect the presence and perhaps characteristics of the friendly communicator's transmission.

Processing gain (spreading ratio): The ratio of the spread spectrum bandwidth to the information data rate.

Radiometer: A device used to measure the total energy in the received signal.

Slow frequency-hopping: A spread spectrum technique wherein the hop time is greater than the information symbol interval, i.e., there exists more than one data symbol per hop.

Spread spectrum bandwidth: The bandwidth of the transmitted signal after spreading.

Spreading: The notion of increasing the bandwidth of the transmitted signal by a factor far in excess of its information bandwidth.

Search algorithm: A means for coarse aligning (synchronizing) the despreading signal with the received spread spectrum signal.

Tau-dither loop: A particular implementation of a closed-loop technique for maintaining fine alignment.

Time-hopping spread spectrum: A spreading technique that is analogous to pulse position modulation.

Tracking algorithm: An algorithm (typically closed loop) for maintaining fine alignment.

Related Topics

69.1 Modulation and Demodulation • 73.2 Noise

Reference

J.D. Gibson, *The Mobile Communications Handbook,* Boca Raton, FL: CRC Press, 1996.

Further Information

M.K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook,* New York: McGraw Hill, 1994 (previously published as *Spread Spectrum Communications,* Computer Science Press, 1985).

R.E. Ziemer and R. L. Peterson, *Digital Communications and Spread Spectrum Techniques,* New York: Macmillan, 1985.

J.K. Holmes, *Coherent Spread Spectrum Systems,* New York: John Wiley & Sons, 1982.

R.C. Dixon, *Spread Spectrum Systems,* 3rd ed., New York: John Wiley & Sons, 1994.

C.F. Cook, F. W. Ellersick, L. B. Milstein, and D. L. Schilling, *Spread Spectrum Communications,* IEEE Press, 1983.

Darcie, T.E., Palais, J.C., Kaminow, I.P. “Optical Communication”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

71

Optical Communication

71.1 Lightwave Technology for Video Transmission
Video Formats and Applications • Intensity Modulation • Noise Limitations • Linearity Requirements • Laser Linearity • Clipping • External Modulation • Miscellaneous Impairments • Summary

71.2 Long Distance Fiber Optic Communications
Fiber • Modulator • Light Source • Source Coupler • Isolator • Connectors and Splices • Optical Amplifier • Repeater • Photodetector • Receiver • Other Components • System Considerations • Error Rates andSignal-to-Noise Ratio • System Design

71.3 Photonic Networks
Data Links • Token Ring: FDDI, FFOL • Active Star Networks: Ethernet, Datakit® • New Approaches to Optical Networks

T.E. Darcie
AT&T Bell Laboratories

Joseph C. Palais
Arizona State University

Ivan P. Kaminow
AT&TBell Laboratories

71.1 Lightwave Technology for Video Transmission

T. E. Darcie

Lightwave technology has revolutionized the transmission of analog and, in particular, video information. Because the light output intensity from a semiconductor laser is linearly proportional to the injected current, and the current generated in a photodetector is linearly proportional to the incident optical intensity, analog information is transmitted as modulation of the optical intensity. The lightwave system is analogous to a **linear** electrical link, where current or voltage translates linearly into optical intensity. High-speed semiconductor lasers and photodetectors enable intensity-modulation bandwidths greater than 10 GHz. Hence, a wide variety of radio frequency (RF) and microwave applications have been developed [Darcie, 1990].

Converting microwaves into intensity-modulated (IM) light allows the use of optical fiber for transmission in place of bulky inflexible coaxial cable or microwave waveguide. Since the fiber attenuation is 0.2–0.4 dB/km, compared with several decibels per meter for waveguide, entirely new applications and architectures are possible. In addition, the signal is confined tightly to the core of single-mode fiber, where it is immune to electromagnetic interference, cross talk, or spectral regulatory control.

To achieve these advantages, several limitations must be overcome. The conversion of current to light intensity must be linear. Several nonlinear mechanisms must be avoided by proper laser design or by the use of various linearization techniques. Also, because the photon energy is much larger than in microwave systems, the signal fidelity is limited by quantum or **shot noise**.

This section describes the basic technology for the transmission of various video formats. We begin by describing the most common video formats and defining transmission requirements for each. Sources of noise, including shot noise, **relative intensity noise** (RIN), and receiver noise are then quantified. Limitations imposed by source nonlinearity, for both **direct modulation** of the laser bias current and **external modulation** using an interferometric $LiNbO_3$ modulator, are compared. Finally, several other impairments caused by **fiber nonlinearity** or **fiber dispersion** are discussed.

Video Formats and Applications

Each video format represents a compromise between transmission bandwidth and robustness or immunity to impairment. With the exception of emerging digital formats, each is also an entrenched standard that often reflects the inefficiencies of outdated technology.

FM Video

Frequency-modulated (FM) video has served for decades as the basis for satellite video transmission [Pratt and Bostian, 1986], where high signal-to-noise ratios (SNRs) are difficult to achieve. Video information with a bandwidth of $B_v = 4.2$ MHz is used to FM modulate an RF carrier. The resulting channel bandwidth B is given by

$$B \sim \Delta f_{pp} + 2f_m \tag{71.1}$$

where Δf_{pp} is the frequency deviation (22.5 MHz) and f_m is the audio subcarrier frequency (6.8 MHz). As a result of this bandwidth expansion to typically 36 MHz, a high SNR can be obtained for the baseband video bandwidth B_v even if the received carrier-to-noise ratio (CNR) over the FM bandwidth B is small. The SNR is given by

$$\text{SNR} = \text{CNR} + 10 \log \left[\frac{3B}{2B_v} \left(\frac{\Delta f_{pp}}{B_v} \right) \right] + \text{W} + \text{PE} \tag{71.2}$$

where W is a weighting factor (13.8 dB) that accounts for the way the eye responds to noise in the video bandwidth, and PE is a pre-emphasis factor (0–5 dB) that is gained by emphasizing the high-frequency video components to improve the performance of the FM modulator. High-quality video (SNR = 55 dB) requires a CNR of only 16 dB. This is achieved easily in a lightwave transmission system.

Applications for lightwave FM video transmission include links to satellite transmission facilities, transport of video between cable television company head-ends (super-trunking), and perhaps delivery of video to subscribers over large fiber distribution networks [Way et al., 1988; Olshansky et al., 1988].

AM-VSB Video

The video format of choice, both for broadcast and cable television distribution, is AM-VSB. Each channel consists of an RF carrier that is amplitude modulated (AM) by video information. Single-sideband vestigial (VSB) filtering is used to minimize the bandwidth of the modulated spectrum. The resultant RF spectrum is dominated by the remaining RF carrier, which is reduced by typically 5.6 dB by the AM, and contains relatively low-level signal information, including audio and color subcarriers. An AM-VSB channel requires a bandwidth of only 6 MHz, but CNRs must be at least 50 dB.

For cable distribution, many channels are frequency-division multiplexed (FDM), separated nominally by 6 MHz (8 MHz in Europe), over the bandwidth supported by the coaxial cable. A typical 60-channel cable system operates between 55.25 and 439.25 MHz. Given the large dynamic range required to transmit both the remaining RF carrier and the low-level sidebands, transmission of this multichannel spectrum is a challenge for lightwave technology.

The need for such systems in cable television distribution systems has motivated the development of suitable high-performance lasers. Before the availability of lightwave AM-VSB systems, cable systems used long (up to 20 km) trunks of coaxial cable with dozens of cascaded electronic amplifiers to overcome cable loss. Accumulations of distortion and noise, as well as inherent reliability problems with long cascades, were serious limitations.

Fiber AM-VSB trunk systems can replace the long coaxial trunks so that head-end quality video can be delivered deep within the distribution network [Chiddix et al., 1990]. Inexpensive coaxial cable extends from the optical receivers at the ends of the fiber trunks to each home. Architectures in which the number of electronic amplifiers between each receiver and any home is approximately three or fewer offer a good compromise between cost and performance. The short spans of coaxial cable support bandwidths approaching 1 GHz, two

or three times the bandwidth of the outdated long coaxial cable trunks. With fewer active components, reliability is improved. The cost of the lightwave components can be small compared to the overall system cost. These compelling technical and economic advantages resulted in the immediate demand for lightwave AM-VSB systems.

Compressed Digital Video

The next generation of video formats will be the product of compressed digital video (CDV) technology [Netravali and Haskel, 1988]. For years digital "NTSC-like" video required a bit rate of approximately 100 Mbps. CDV technology can reduce the required bit rate to less than 5 Mbps. This compression requires complex digital signal processing and large-scale circuit integration, but advances in chip and microprocessor design have made inexpensive implementation of the compression algorithms feasible.

Various levels of compression complexity can be used, depending on the ultimate bit rate and quality required. Each degree of complexity removes different types of redundancy from the video image. The image is broken into blocks of pixels, typically 8×8. By comparing different blocks and transmitting only the differences (DPCM), factors of 2 reduction in bit rate can be obtained. No degradation of quality need result. Much of the information within each block is imperceptible to the viewer. Vector quantization (VQ) or discrete-cosine transform (DCT) techniques can be used to eliminate bits corresponding to these imperceptible details. This intraframe coding can result in a factor of 20 reduction in the bit rate, although the evaluation of image quality becomes subjective. Finally, stationary images or moving objects need not require constant retransmission of every detail. Motion compression techniques have been developed to eliminate these interframe redundancies. Combinations of these techniques have resulted in coders that convert NTSC-like video (100 Mbps uncompressed) into a few megabits per second and HDTV images (1 Gbps uncompressed) into less than 20 Mbps.

CDV can be transmitted using time-division multiplexing (TDM) and digital lightwave systems or by using each channel to modulate an RF carrier and transmitting using analog lightwave systems. There are numerous applications for both alternatives. TDM systems for CDV are no different from any other digital transmission system and will not be discussed further.

Using RF techniques offers an additional level of RF compression, wherein advanced multilevel modulation formats are used to maximize the number of bits per hertz of bandwidth [Feher, 1987]. Quadrature-amplitude modulation (QAM) is one example of multilevel digital-to-RF conversion. For example, 64-QAM uses 8 amplitude and 8 phase levels and requires only 1 Hz for 5 bits of information. As the number of levels, hence the number of bits per hertz, increases, the CNR of the channel must increase to maintain error-free transmission. A 64-QAM channel requires a CNR of approximately 30 dB.

A synopsis of the bandwidth and CNR requirements for FM, AM-VSB, and CDV is shown in Fig. 71.1. AM-VSB requires high CNR but low bandwidth. FM is the opposite. Digital video can occupy a wide area, depending on the degree of digital and RF compression. The combination of CDV and QAM offers the possibility of squeezing a high-quality video channel into 1 MHz of bandwidth, with a required CNR of 30 dB. This drastic improvement over AM-VSB or FM could have tremendous impact on future video transmission systems.

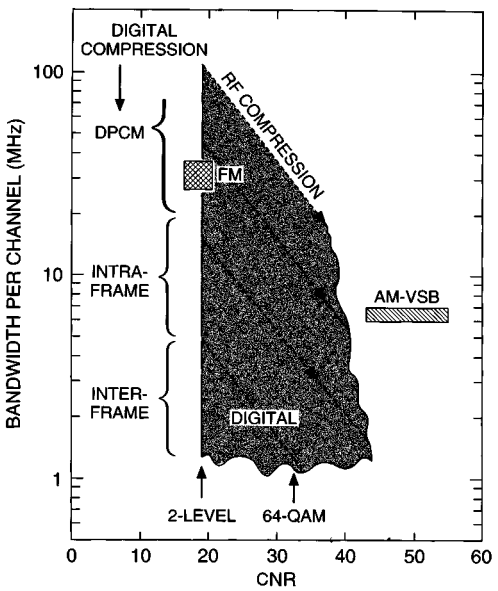


FIGURE 71.1 Bandwidth versus carrier-to-noise ratio (CNR) required for AM-VSB, FM, and digital video. Increasingly complex digital compression techniques reduce the bit rate required for NTSC-like video from 100 Mbps to less than 5 Mbps. Bandwidth efficient RF techniques like QAM minimize the bandwidth required for each bit rate but require greater CNRs.

Intensity Modulation

As mentioned in the introduction, the light output from the laser should be linearly proportional to the injected current. The laser is prebiased to an average output intensity L_0. Many video channels are combined electronically, and the total RF signal is added directly to the laser current. The optical modulation depth (m) is defined as the ratio of the peak modulation L_0 for one channel, divided by L_0. For 60-channel AM-VSB systems, m is typically near 4%.

The laser (optical carrier) is modulated by the sum of the video channels that are combined to form the total RF signal spectrum. The resultant optical spectrum contains sidebands from the IM superimposed on unintentional frequency modulation, or **chirp**, that generally accompanies IM. This complex optical spectrum must by understood if certain subtle impairments are to be avoided.

A photodetector converts the incident optical power into current. Broadband InGaAs photodetectors with responsivities (R_0) of nearly 1.0 A/W and bandwidths greater than 10 GHz are available. The detector generates a dc current corresponding to the average received optical power L_r and the complete RF modulation spectrum that was applied at the transmitter. An ac-coupled electronic preamplifier is used to remove the dc component and boost the signal to usable levels.

Noise Limitations

The definition of CNR deserves clarification. Depending on the video format and RF modulation technique, the RF power spectrum of the modulated RF carrier varies widely. For AM-VSB video the remaining carrier is the dominant feature in the spectrum. It is thereby convenient to define the CNR as the ratio of the power remaining in the carrier to the integrated noise power in a 4-MHz bandwidth centered on the carrier frequency. For FM or digitally modulated carriers, the original carrier is not generally visible in the RF spectrum. It is then necessary to define the CNR as the ratio of the integrated signal power within the channel bandwidth to the integrated noise power.

Shot Noise

Shot noise is a consequence of the statistical nature of the photodetection process. It results in a noise power spectral density, or electrical noise power per unit bandwidth (dBm/Hz) that is proportional to the received photocurrent I_r ($= R_0 L_r$). The total shot noise power in a bandwidth B is given by

$$N_s = 2eI_rB \tag{71.3}$$

where e is the electronic charge.

With small m, the detected signal current is a small fraction of the total received current. The root mean square (rms) signal power for one channel is

$$P_s = \frac{1}{2}(mI_r)^2 \tag{71.4}$$

The total shot noise power then limits the CNR (P_s/N_s) to a level referred to as the quantum limit. Received powers near 1 mW are required if CNRs greater than 50 dB are to be achieved for 40- to 80-channel AM-VSB systems.

Receiver Noise

Receiver noise is generated by the electronic amplifier used to boost the detected photocurrent to usable levels. The easiest receiver to build consists of a *pin* photodiode connected directly to a low-noise 50- to 75-Ω amplifier, as shown in Fig. 71.2(a). The effective input current noise density, (n), for this simple receiver is given by

$$n^2 = \frac{4kTF}{R_L} \tag{71.5}$$

where k is the Boltzmann constant, T is the absolute temperature, F is the **noise figure** of the amplifier, and R_L is the input impedance. For a 50-Ω input impedance and $F = 2$, $n = 20$ pA/$\sqrt{Hz}$.

A variety of more complicated receiver designs can reduce the noise current appreciably [Kasper, 1988]. The example shown in Fig. 71.2(b) uses a high-speed FET. R_L can be increased to maximize the voltage developed by the signal current at the FET input. Input capacitance becomes a limitation by shunting high-frequency components of signal current. High-frequency signals are then reduced with respect to the noise generated in the FET, resulting in poor high-frequency performance. Various impedance matching techniques have been proposed to maximize the CNR for specific frequency ranges.

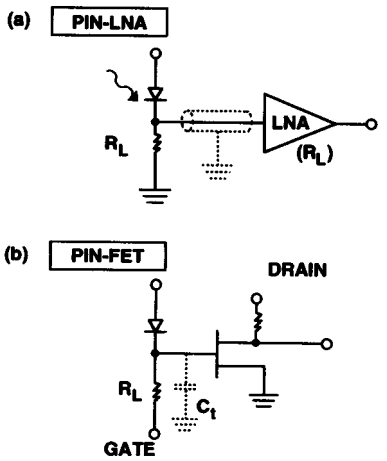


FIGURE 71.2 Receivers for broadband analog lightwave systems. Coupling a *pin* to a low-noise amplifier (a) is simple, but improved performance can be obtained using designs like the *pin* FET (b). C_t is the undesirable input capacitance.

Relative Intensity Noise

Relative intensity noise (RIN) can originate from the laser or from reflections and **Rayleigh backscatter** in the fiber. In the laser, RIN is caused by spontaneous emission in the active layer. Spontaneous emission drives random fluctuations in the number of photons in the laser which appear as a random modulation of the output intensity, with frequency components extending to tens of gigahertz. The noise power spectral density from RIN is I_r^2 RIN, where RIN is expressed in decibels per hertz.

RIN is also caused by component reflections and double-Rayleigh backscatter in the fiber, by a process called multipath interference. Twice-reflected signals arriving at the detector can interfere coherently with the unreflected signal. Depending on the modulated optical spectrum of the laser, this interference results in noise that can be significant [Darcie et al., 1991].

The CNR, including all noise sources discussed, is given by

$$\text{CNR} = \frac{m^2 I_r^2}{2B[n^2 + 2eI_r + I_r^2 \text{RIN}]} \tag{71.6}$$

All sources of intensity noise are combined into RIN. Increasing m improves the CNR but increases the impairment caused by nonlinearity, as discussed in the following subsection. The optimum operating value for m is then a balance between noise and distortion.

Figure 71.3 shows the noise contributions from shot noise, receiver noise, and RIN. For FM or digital systems, the low CNR values required allow operation with small received optical powers. Receiver noise is then generally the limiting factor. Much larger received powers are required if AM-VSB noise requirements are to be met. Although detecting more optical power helps to overcome shot and receiver noise, the ratio of signal to RIN remains constant. RIN can be dominant in high-CNR systems, when the received power is large. AM-VSB systems require special care to minimize all sources of RIN. The dominant noise source is then shot noise, with receiver noise and RIN combining to limit CNRs to within a few decibels of the quantum limit.

Linearity Requirements

Source linearity limits the depth of modulation that can be applied. Linearity, in this case, refers to the linearity of the current-to-light-intensity (I-L) conversion in the laser or voltage-to-light (V-L) transmission for an external modulator. Numerous nonlinear mechanisms must be considered for direct modulation, and no existing external modulator has a linear transfer function.

A Taylor-series expansion of the I-L or V-L characteristic, centered at the bias point, results in linear, quadratic, cubic, and higher-order terms. The linear term describes the efficiency with which the applied signal is converted

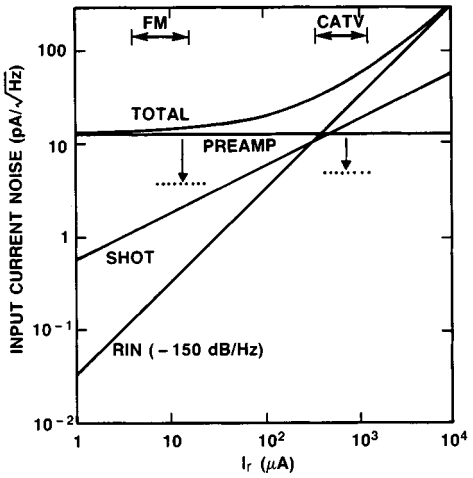


FIGURE 71.3 Current noise densities from receivers, RIN, and shot noise as a function of total received photocurrent. Receiver noise is dominant in FM or some digital systems where the total received power is small. The solid line for receiver noise represents the noise current for a typical 50-Ω low-noise amplifier. More sophisticated receiver designs could reduce the noise to the levels shown approximately by the dotted lines. RIN and shot noise are more important in AM-VSB systems.

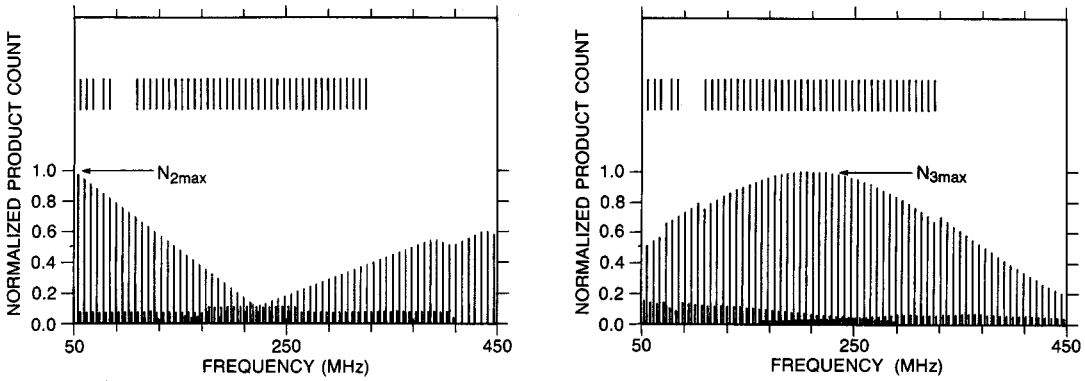


FIGURE 71.4 Second-order (a) and third-order (b) distortion products for 42-channel AM-VSB system. The maximum number of second-order products occurs at the lowest frequency channel, where 30 products contribute to the CSO. The maximum number of third-order products occurs near the center channel, where 530 products contribute to the CTB.

to linear intensity modulation. The quadratic term results in second-order distortion, the cubic produces third-order distortion, and so on.

Requirements on linearity can be derived by considering the number and spectral distribution of the distortion products generated by the nonlinear mixing between carriers in the multichannel signal. Second-order nonlinearity results in sum and difference ($f_i \pm f_j$) mixing products for every combination of the two channels. This results in as many as 50 second-order products within a single channel, in a 60-channel AM-VSB system with the standard U.S. frequency plan. Similarly, for third-order distortion, products result from mixing among all combinations of three channels. However, since the number of combinations of three channels is much larger than for two, up to 1130 third-order products can interfere with one channel. The cable industry defines the **composite second-order (CSO)** distortion as the ratio of the carrier to the largest group of second-order products within each channel. For third-order distortion, the **composite triple beat (CTB)** is the ratio of the carrier to the total accumulation of third-order distortion at the carrier frequency in each channel.

The actual impairment from these distortion products depends on the spectrum of each RF channel and on the exact frequency plan used. A typical 42-channel AM-VSB frequency plan, with carrier frequencies shown as the vertical bars on Fig. 71.4, results in the distributions of second- and third-order products shown in Fig. 71.4(a) and (b), respectively. Since the remaining carrier is the dominant feature in the spectrum of each channel, the distortion products are dominated by the mixing between these carriers. Because high-quality video requires that the CSO is –60 dBc (dB relative to the carrier), each sum or difference product must be less than –73 dBc. Likewise, for the CTB to be less than 60 dB, each product must be less than approximately –90dB.

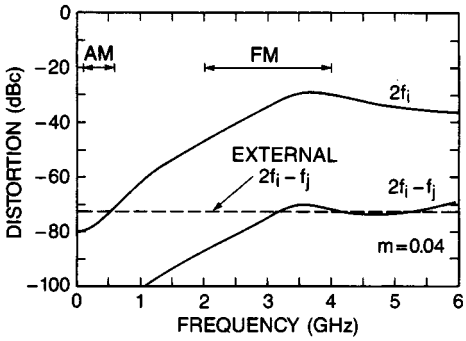


FIGURE 71.5 Resonance distortion for directly modulated laser with resonance frequency of 7 GHz. Both the second-harmonic $2f_i$ and two-tone third-order $2f_i \pm f_j$ distortion peak near half the resonance frequency and are small at low frequency. Also shown is the same third-order distortion for an external modulator biased at the point of zero second-order distortion.

FM or CDV systems have much less restrictive linearity requirements, because of the reduced sensitivity to impairment. Distortion products must be counted, as with the AM-VSB example described previously, but each product is no longer dominated by the remaining carrier. Because the carrier is suppressed entirely by the modulation, each product is distributed over more than the bandwidth of each channel. The impairment resulting from the superposition of many uncorrelated distortion products resembles noise. Quantities analogous to the CSO and CTB can be defined for these systems.

Laser Linearity

Several factors limit the light-versus-current (*L-I*) linearity of directly modulated lasers. Early work on laser dynamics led to a complete understanding of resonance-enhanced distortion (RD). RD arises from the same carrier-photon interaction within the laser that is responsible for the relaxation-oscillation resonance.

The second-harmonic distortion ($2f_i$) and two-tone third-order distortion ($2f_i - f_j$) for a typical 1.3-μm wavelength directly modulated semiconductor laser are shown in Fig. 71.5 [Darcie et al., 1986]. Both distortions are small at low frequencies but rise to maxima at half the relaxation resonance frequency. AM-VSB systems are feasible only within the low-frequency window. FM or uncompressed digital systems require enough bandwidth per channel that multichannel systems must operate in the region of large RD. Fortunately, the CNR requirements allow for the increased distortion. The large second-order RD can be avoided entirely by operating within a one-octave frequency band (e.g., 2–4 GHz), such that all second-order products are out of band.

Within the frequency range between 50 and 500 MHz, nonlinear gain and loss, intervalence-band absorption, and, more importantly, spatial-hole burning (SHB) and carrier leakage can all be significant. Carrier leakage prevents all of the current injected in the laser bond wire from entering the active layer. This leakage must be reduced to immeasurable levels for AM-VSB applications.

SHB results from the nonuniform distribution of optical power along the length of the laser. In DFB lasers, because of the grating feedback, the longitudinal distribution of optical power can be highly nonuniform. This results in distortion [Takemoto et al., 1990] that can add to or cancel other distortion, making it, in some cases, a desirable effect.

Clipping

Even if all nonlinear processes were eliminated, the allowable modulation would be limited by the fact that the minimum output power is zero. Typical operating conditions with, for example, 60 channels, each with an average modulation depth (m) near 4%, result in a peak modulation of 240%. Although improbable, modulations of more than 100% result in clipping.

The effects of clipping were first approximated by Saleh [1989], who calculated the modulation level at which the total power contained in all orders of distortion became appreciable. Even for perfectly linear lasers, the modulation depth is bounded to values beyond which all orders of distortion increase rapidly. Assuming that half the total power in all orders of distortion generated by clipping is distributed evenly over each of N channels, clipping results in a carrier-to-interference ratio (CIR) given by

$$\mathrm{CIR} = \sqrt{2\pi}\,\frac{(1+6\mu^2)}{\mu^3}\,e^{1/2\mu^2} \tag{71.7}$$

where the rms modulation index μ is

$$\mu = m\sqrt{N/2} \tag{71.8}$$

External Modulation

Laser-diode-pumped YAG lasers with low RIN and output powers greater than 200 mW have been developed recently. Combined with linearized external $LiNbO_3$ modulators, these lasers have become high-performance competitors to directly modulated lasers. YAG lasers with external modulation offer a considerable increase in launched power, and the low RIN of the YAG laser translates into a slight CNR improvement. The most challenging technical hurdle is to develop a linear low-loss optical intensity modulator.

Low-loss $LiNbO_3$ Mach–Zehnder modulators are available with insertion losses less than 3 dB, modulation bandwidths greater than a few gigahertz, and switching voltages near 5 V. The output intensity of these modulators is a sinusoidal function of the bias voltage. By prebiasing to 50% transmission, modulation applied to the Mach–Zehnder results in the most linear intensity modulation. This bias point, which corresponds to the point of inflection in the sinusoidal transfer function, produces zero second-order distortion. Unfortunately, the corresponding third-order distortion is approximately 30 dB worse than a typical directly modulated DFB laser, at low frequencies. This comparison is shown on Fig. 71.5. For high-frequency applications where RD is important, external modulators can offer improved linearity. A means of linearizing the third-order nonlinearity is essential for AM-VSB applications.

Various linearization techniques have been explored. The two most popular approaches are feedforward and predistortion. Feedforward requires that a portion of the modulated output signal be detected and compared to the original applied voltage signal to provide an error signal. This error signal is then used to modulate a second laser, which is combined with the first laser such that the total instantaneous intensity of the two lasers is a replica of the applied voltage. In principle, this technique is capable of linearizing any order of distortion and correcting RIN from the laser.

Predistortion requires less circuit complexity than feedforward. A carefully designed nonlinear circuit is placed before the nonlinear modulator, such that the combined transfer function of the predistorter-modulator is linear. Various nonlinear electronic devices or circuits can act as second- or third-order predistorters. Difficulties include matching the frequency dependence of the predistorter with that of the modulator, hence achieving good linearity over a wide frequency range. Numerous circuit designs can provide reductions in third-order distortion by 15 dB.

Miscellaneous Impairments

Laser chirp can cause problems with direct laser modulation. Chirp is modulation of the laser frequency caused by modulation of the refractive index of the laser cavity in response to current modulation. The interaction of chirp and chromatic dispersion in the fiber can cause unacceptable CSO levels for AM-VSB systems as short as a few kilometers. Dispersion converts the FM into IM, which mixes with the signal IM to produce second-order distortion [Phillips et al., 1991]. These systems must operate at wavelengths corresponding to low fiber dispersion, or corrective measures must be taken.

Chirp also causes problems with any optical component that has a transmission that is a function of optical frequency. This can occur if two optical reflections conspire to form a weak interferometer or in an **erbium-doped fiber amplifier** (EDFA) that has a frequency-dependent gain [Kuo and Bergmann, 1991]. Once again, the chirp is converted to IM, which mixes with the signal IM to form second-order distortion.

Although externally modulated systems are immune to chirp-related problems, fiber nonlinearity, in the form of stimulated Brillouin scattering (SBS), places a limit on the launched power. SBS, in which light is scattered from acoustic phonons in the fiber, causes a rapid decrease in CNR for launched powers greater than approximately 10 mW [Mao et al., 1991]. Since the SBS process requires high optical powers within a narrow optical spectral width (20 MHz), it is a problem only in low-chirp externally modulated systems. Chirp in DFB systems broadens the optical spectrum so that SBS is unimportant.

Summary

A wide range of applications for transmission of video signals over optical fiber has been made possible by refinements in lightwave technology. Numerous technology options are available for each application, each with advantages or disadvantages that must be considered in context with specific system requirements. Evolution of these video systems continues to be driven by development of new and improved photonic devices.

Defining Terms

Chirp: Modulation of the optical frequency that occurs when a laser is intensity modulated.

Composite second order (CSO): Ratio of the power in the second-order distortion products to power in the carrier in a cable television channel.

Composite triple beat (CTB): Same as CSO but for third-order distortion.

Direct modulation: Modulation of the optical intensity output from a semiconductor diode laser by direct modulation of the bias current.

Erbium-doped fiber amplifier: Fiber doped with erbium that provides optical gain at wavelengths near 1.55 μm when pumped optically at 0.98 or 1.48 μm.

External modulation: Modulation of the optical intensity using an optical intensity modulator to modulate a constant power (cw) laser.

Fiber dispersion: Characteristic of optical fiber by which the propagation velocity depends on the optical wavelength.

Fiber nonlinearity: Properties of optical fibers by which the propagation velocity, or other characteristic, depends on the optical intensity.

Lightwave technology: Technology based on the use of optical signals and optical fiber for the transmission of information.

Linear: Said of any device for which the output is linearly proportional to the input.

Noise figure: Ratio of the output signal-to-noise ratio (SNR) to the input SNR in an amplifier.

Rayleigh backscatter: Optical power that is scattered in the backwards direction by microscopic inhomogeneities in the composition of optical fibers.

Relative intensity noise: Noise resulting from undesirable fluctuations of the optical power detected in an optical communication system.

Shot noise: Noise generated by the statistical nature of current flowing through a semiconductor *p-n* junction or photodetector.

Related Topics

42.1 Lightwave Waveguides • 69.1 Modulation and Demodulation • 73.6 Data Compression

References

T.E. Darcie, "Subcarrier multiplexing for lightwave networks and video distribution systems," *IEEE J. Selected Areas in Communications,* vol. 8, p. 1240, 1990.

T. Pratt and C.W. Bostian, *Satellite Communications,* New York: Wiley, 1986.

W. Way, C. Zah. C. Caneau, S. Menmocal, F. Favire, F. Shokoochi, N. Cheung, and T.P. Lee, "Multichannel FM video transmission using traveling wave amplifiers for subscriber distribution," *Electron. Lett.,* vol. 24, p. 1370, 1988.

R. Olshansky, V. Lanzisera, and P. Hill, "Design and performance of wideband subcarrier multiplexed lightwave systems," in *Proc. ECOC '88,* Brighton, U.K., Sept. 1988, pp. 143–146.

J.A. Chiddix, H. Laor, D.M. Pangrac, L.D. Williamson, and R.W. Wolfe, "AM video on fiber in CATV systems, need and implementation," *IEEE J. Selected Areas in Communications,* vol. 8, p. 1229, 1990.

A.N. Netravali and B.G. Haskel, *Digital Pictures,* New York: Plenum Press, 1988.

K. Feher, Ed., *Advanced Digital Communications,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

B.L. Kasper, "Receiver design," in *Optical Fiber Telecommunications II,* S.E. Miller and I.P. Kaminow, Eds., San Diego: Academic Press, 1988.
T.E. Darcie, G.E. Bodeep, and A.A.M. Saleh, "Fiber-reflection-induced impairments in lightwave AM-VSB CATV systems," *IEEE J. Lightwave Technol.,* vol. 9, no. 8, pp. 991–995, Aug. 1991.
T.E. Darcie, R.S. Tucker, and G.J. Sullivan, "Intermodulation and harmonic distortion in IaGaAsP lasers," *Electron. Lett.,* vol. 21, 665–666, erratum; vol 22, p. 619, 1986.
A. Takemoto, H. Watanabe, Y. Nakajima, Y. Sakakibara, S. Kakimoto, U. Yamashita, T. Hatta, and Y. Miyake, "Distributed feedback laser diode and module for CATV systems," *IEEE J. Selected Areas in Communications,* vol. 8, 1359, 1990.
A.A.M. Saleh, "Fundamental limit on number of channels in subcarrier mulitplexed lightwave CATV systems," *Electron. Lett.,* vol. 25, no. 12, pp. 776–777, 1989.
M.R. Phillips, T.E. Darcie, D. Marcuse, G.E. Bodeep, and N.J. Frigo, "Nonlinear distortion generated by dispersive transmission of chirped intensity-modulated signals," *IEEE Photonics Technol. Lett.,* vol. 3, no. 5, pp. 481–483, 1991.
C.Y. Kuo and E.E. Bergmann, "Erbium-doped fiber amplifier second-order distortion in analog links and electronic compensation," *IEEE Photonics Technol. Lett.,* vol. 3, p. 829, 1991.
X.P. Mao, G.E. Bodeep, R.W. Tkach, A.R. Chraplyvy, T.E. Darcie, and R.M. Derosier, "Brillouin scattering in lightwave AM-VSB CATV transmission systems," *IEEE Photonics Technol. Lett.,* vol. 4, no. 3, pp. 287–289, 1991.

Further Information

National Cable Television Association (NCTA), Proceedings from Technical Sessions, annual meetings, 1724 Massachusetts Ave. NW, Washington D.C., 20036, 1969.
Society of Cable Television Engineers (SCTE), Proceeding from Technical Sessions, biennial meetings, Exton Commons, Exton, Penn.
T.E. Darcie, "Subcarrier multiplexing for lightwave multiple-access lightwave networks," *J. Lightwave Technol.,* vol. LT-5, pp. 1103–1110, Aug. 1987.
T.E. Darcie and G.E. Bodeep, "Lightwave subcarrier CATV transmission systems," *IEEE Trans. Microwave Theory and Technol.,* vol. 38, no. 5, pp. 534–533, May 1990.
IEEE J. Lightwave Technol., Special Issue on "Broadband Analog Video Transmission Over Fibers," to be published Jan./Feb. 1993.

71.2 Long Distance Fiber Optic Communications

Joseph C. Palais

When the first laser was demonstrated in 1960, numerous applications of this magnificent new tool were anticipated. Some predicted that laser beams would transmit messages through the air at high data rates between distant stations. Although laser beams can indeed travel through the atmosphere, too many problems prevent this scheme from becoming practical. Included in the objections are the need for line-of-sight paths and the unpredictability of transmission through an atmosphere where weather variations randomly change path losses. Guided paths using optical fibers offer the only practical means of optical transmission over long distances.

Long-distance fiber systems tend to have the following operational characteristics: They are more than 10 km long, transmit digital signals (rather than analog), and operate at data rates above a few tens of megabits per second. This section primarily describes systems in this category.

Figure 71.6 illustrates the basic structure of a generalized long-distance fiber optic link. Each of the components will be described in the following paragraphs.

A useful figure of merit for these systems is the product of the system data rate and its length. This figure of merit is the well-known *rate-length product.* The bandwidth of the transmitting and receiving circuits (including the light source and photodetector) limits the achievable system data rate. The bandwidth of the

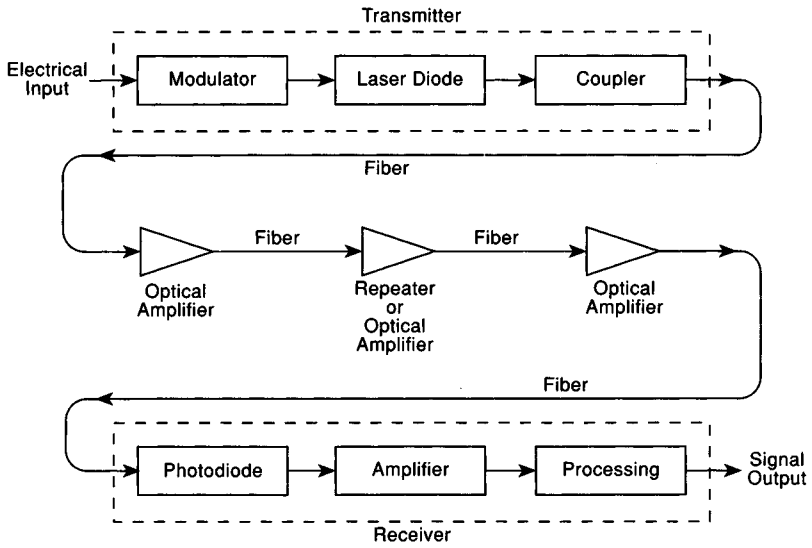


FIGURE 71.6 Long-distance fiber communication system.

fiber decreases with its length, so that the fiber itself limits the rate-length product. The losses in the system, including those in the fiber, also limit the path length. Systems are *bandwidth limited* if the rate-length figure is determined by bandwidth restraints and *loss limited* if determined by attenuation.

The first efficient fiber appeared in 1970, having a loss of 20 dB/km. Just 7 years later the first large-scale application, a link between two telephone exchanges in Chicago, was constructed. By this time the loss had been reduced to around 3 dB/km. The digital technology used could accommodate a rate of 45 Mbps over an unrepeatered length of 10 km and a total length of over 60 km with repeaters. The unrepeatered rate-length product for this initial system was a modest 0.5 Gbps × km. As fiber technology advanced, this figure steadily increased. Unrepeatered rate-length products have improved to 500 Gbps × km (e.g., 8 Gbps over a path of 60 km) and beyond. Allowing repeaters and/or optical amplifiers increases the net rate-length product considerably. Values beyond 70 Tbps × km (70,000 Gbps × km) are achievable with optical amplifiers. This latter figure allows construction of a transmission system operating at 5 Gbps over a 14,000-km path. The longest terrestrial paths are across the Atlantic and Pacific oceans, distances of about 6,000 and 9,000 km, respectively. Fibers are capable of spanning these distances with high-capacity links.

Fiber

All fibers used for long-distance communications are made of silica glass and allow only a single mode of propagation. The silica is doped with other materials to produce the required refractive index variations for the fiber core and cladding. The important fiber characteristics that limit system performance are its loss and its bandwidth. The loss limits the length of the link and the bandwidth limits the data rate.

Figure 71.7 shows the loss characteristics of single-mode silica fibers at the wavelengths of lowest attenuation. As indicated in the figure, there are three possible windows of operation. In the first window (around 820 nm), the loss is typically 3 dB/km. This is too high for long systems. In the second window (near 1300 nm), the loss is about 0.5 dB/km. In addition, the fiber bandwidth is quite high because of low pulse dispersion at this wavelength. The second window is a reasonable operating wavelength for high-capacity, long-distance systems. At 1550 nm (the third window) the loss is lowest, about 0.25 dB/km. This characteristic makes 1550 nm the optimum choice for the very longest links.

Dispersion refers to the spreading of a pulse as it travels along a **single-mode fiber**. It is due to material and waveguide effects. This spreading creates intersymbol interference if allowed to exceed about 70% of the original pulse width, causing receiver errors. The dispersion factor M is usually given in units of picoseconds of pulse spread per nanometer of spectral width of the light source and per kilometer of length of fiber.

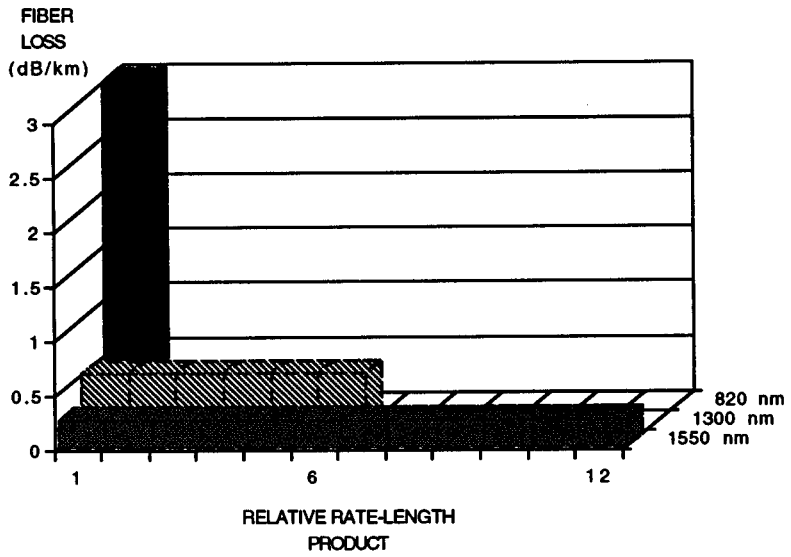


FIGURE 71.7 Fiber loss and relative unrepeatered, unamplified rate-length product.

In the range from 1200 to 1600 nm, the dispersion curve for silica can be approximated by the expression

$$M = \frac{M_0}{4}\left(\lambda - \frac{\lambda_0^4}{\lambda^3}\right) \tag{71.9}$$

where λ is the operating wavelength, λ_0 is the zero dispersion wavelength, and M_0 is the slope at the zero dispersion wavelength. M_0 is approximately 0.095 ps/(nm^2 × km). The pulse spread for a path length L, using a light source whose spectral width is $\Delta\lambda$, is then

$$\Delta\tau = ML\Delta\lambda \tag{71.10}$$

The zero dispersion wavelength, close to 1300 nm for silica fibers, makes this wavelength attractive for high-capacity links. The dispersion at 1550 nm is typically close to 20 ps/(nm × km). This is a moderate amount of dispersion. If a proposed 1550-nm system is bandwidth limited because of this spread, several alternatives are available. One solution is to use dispersion-shifted fiber, which is a special fiber with a refractive index profile designed to shift the zero dispersion wavelength from 1300 nm to 1550 nm. Another solution is to transmit soliton pulses, which use the nonlinearity of the fiber to maintain pulse shape during transmission.

Figure 71.7 includes relative unrepeatered, unamplified values of rate-length products in the three transmission windows. Because of high loss, the first window can be used only for moderate lengths (around 10 km). Because of high dispersion, the data rates are also limited in this region. In the second window, nearly zero dispersion allows high-rate transmission, but the losses limit the distance that can be covered (typically around 50 km). In the third window, the loss is about half the 1300-nm attenuation so that twice as much distance can be covered. Dispersion-shifted fiber allows the same high rates as does 1300-nm operation. Repeaters and amplifiers extend the useful distance of fiber links well beyond the distances listed here.

Modulator

A digital electrical signal modulates the light source. The driver circuit must be fast enough to operate at the system bit rate. As bit rates increase into the multigigabit per second range, this becomes increasingly difficult. Modulation can be done in the optical domain at very high speeds. In this case, the modulator follows the laser diode rather than preceding it. External modulation is usually accomplished using integrated-optic structures.

Light Source

Laser diodes or light-emitting diodes (LEDs) supply the optical carrier waves for most fiber links. LEDs cannot operate at speeds in the gigabit range, but laser diodes can. For this reason, laser diodes are normally required for high-rate, long-distance links. Laser diodes can be modulated at frequencies beyond 40 GHz.

Laser diodes emitting in the second and third fiber transmission windows are semiconductor heterojunctions made of InGaAsP. The exact emission wavelength is primarily determined by the proportions of the constituent atoms. Output powers are commonly on the order of a few milliwatts.

Typical laser diode **spectral widths** are between 1 and 5 nm when operating in more than one longitudinal mode. Single-mode laser diodes can have spectral widths of just a few tenths of a nanometer. As predicted by Eq. (71.10), narrow-spectral-width emitters minimize pulse spreading. Minimizing pulse spreading increases the fiber bandwidth and its data capacity.

Solid-state lasers other than semiconductor laser diodes may be useful in specific applications. Example of such lasers are the Nd:YAG laser and the erbium-doped fiber laser.

Source Coupler

The light emitted from the diode must be coupled as efficiently as possible into the fiber. Because the beam pattern emitted by a laser diode does not perfectly match the pattern of light propagating in the fiber, there is an inevitable mismatch loss. Good coupler designs, sometimes using miniature lenses, reduce this loss to about 3 dB when feeding a single-mode fiber.

Isolator

An optical isolator is a one-way transmission path. It allows power flow from the transmitter toward the receiver but blocks power flow in the opposite direction. It is used to protect the laser diode from back reflections, which tend to increase the laser noise.

Connectors and Splices

Connections between fibers and between the fiber and other components occur at numerous points in a long-distance link. Because there may be many splices in a long system, the loss of each one must be small. Fusion splices with an average loss of no more than 0.05 dB are often specified. Mechanical splices are also suitable. They often involve epoxy for fixing the connection. Connectors are used where remateable connections are required. Good fiber connectors introduce losses of just a few tenths of a decibel.

In addition to having low loss, good connectors and splices also minimize back reflections. This is especially important for connections near the transmitter to reduce laser noise. Fusion splices produce little reflection, but mechanical splices and all connectors must be carefully designed to keep reflected power levels low. Reflections occur because of small gaps at the interface between the mated fibers. Successful techniques for reducing reflections include the physical contact connection, where the fiber end faces are polished into hemispheres (rather than flat surfaces) so that the cores of the two mated fibers are in contact with each other. Even better performance is obtained by angling the end faces a few degrees so that reflected light is filtered out of the single propagating mode.

Optical Amplifier

Many fiber links are loss limited. One cause is the limited power available from the typical laser diode, which (together with the losses in the fiber and the other system components) restricts the length of fiber that can be used. The *fiber optic amplifier* increases the power level of the signal beam without conversion to the electrical domain. For example, gains of 30 dB are attainable at 1550 nm using the erbium-doped fiber amplifier (EDFA). Quite importantly, the EDFA has a bandwidth of over 20 nm so that several WDM or numerous OFDM channels (described later in this section) can be amplified simultaneously.

As indicated in Fig. 71.6, there are a number of possible locations for optical amplifiers in a system. An optical amplifier just following the transmitter increases the optical power traveling down the fiber. Amplifiers

along the fiber path continually keep the power levels above the system noise. An amplifier located at the fiber end acts as a receiver preamplifier, enhancing its sensitivity. Many amplifiers can be placed in a fiber network, extending the total path length to thousands of kilometers.

Repeater

The repeater is a regenerator that detects the optical signal by converting it into electrical form. It then determines the content of the pulse stream and uses this information to generate a new optical signal and launch this improved pulse train into the fiber. The new optical pulse stream is identical to the one originally transmitted. The regenerated pulses are restored to their original shape and power level by the repeater.

Many repeaters may be placed in a fiber network, extending the total path length to thousands of kilometers. The advantage of the optical amplifier over the regenerator is its lower cost and improved efficiency. The greater cost of the regenerator arises from the complexity of conversion between the optical and electrical domains. The regenerator does have the advantage of restoring the signal pulse shape, which increases the system bandwidth. This advantage is negated by a system propagating soliton pulses, which do not degrade with propagation.

Photodetector

This device converts an incoming optical beam into an electrical current. In fiber receivers, the most commonly used photodetectors are semiconductor *pin* photodiodes and avalanche photodiodes (APD). Important detector characteristics are speed of response, spectral response, internal gain, and noise. Because avalanche photodiodes have internal gain, they are preferred for highly sensitive receivers. Both Ge and InGaAs photodiodes respond in the preferred second and third fiber windows. InGaAs performs better at low signal levels because it has smaller values of dark current (that is, it is less noisy).

The current produced by a photodetector in response to incident optical power P is

$$i = G\eta eP/hf \tag{71.11}$$

where G is the detector's gain, η is its quantum efficiency (close to 0.9 for good photodiodes), h is Planck's constant (6.63×10^{-34} J s), e is the magnitude of the charge on an electron (1.6×10^{-19}), and f is the optical frequency. For *pin* photodiodes ($G = 1$), typical responsivities are on the order of 0.5 μA/μW.

Receiver

Because of the low power levels expected at the input to the receiver, an electronic amplifier is normally required following the photodetector. The remainder of the receiver includes such electronic elements as band-limiting filters, equalizers, decision-making circuitry, other amplification stages, switching networks, digital-to-analog converters, and output devices (e.g., telephones, video monitors, and computers).

Other Components

There are a number of other fiber components, not shown in Fig. 71.6, that can be found in some systems. These include passive couplers for tapping off some portion of the beam from the single fiber and wavelength-division multiplexers for coupling different optical carriers onto the transmission fiber.

System Considerations

Long-distance fiber links carry voice, video, and data information. Messages not already in digital form are first converted to it. A single voice channel is normally transmitted at a rate of 64,000 bits per second. Video requires a much higher rate. The rate could be as much as 90 Mbps or so, but video compression techniques can lower this rate significantly. Fiber systems for the telephone network operate at such high rates that many voice channels can be time-division multiplexed (TDM) onto the same fiber for simultaneous transmission. For example, a fiber operating at a rate of 2.3 Gbps could carry more than 30,000 digitized voice channels.

Several optical carriers can simultaneously propagate along the same fiber. Such wavelength-division multiplexed (WDM) links further increase the capacity of the system. Systems using two or three optical carriers are common. Adding more than a few channels (8 or so) puts strong constraints on the multiplexers and light sources. In long systems wideband optical amplifiers are preferred over regenerators for WDM systems because a single amplifier can boost all the individual carriers simultaneously while separate regenerators are needed for each carrier wavelength.

Total cable capacity is also increased by placing numerous fibers inside the cable. This is a cost-effective strategy when installing long fiber cables. The added cost of the extra fibers is small compared to the costs of actually deploying the cable itself. Fiber counts above 100 are practical. Multifiber cables can have enormous total data capacities.

Still further capacity is possible using optical frequency-division multiplexing (OFDM). In this scheme, many optical carriers very closely spaced in wavelength (maybe a few tenths of a nanometer) operate as independent channels. Hundreds of channels can be visualized in each of the two low-loss fiber windows. Systems of this type require **coherent detection** receivers to separate the closely spaced carriers.

Error Rates and Signal-to-Noise Ratio

The signal-to-noise ratio is a measure of signal quality. It determines the error rate in a digital network. At the receiver, it is given by

$$\frac{S}{N} = \frac{(G\rho P)^2 R_L}{G^n 2eR_L B(I_D + \rho P) + 4kTB} \tag{71.12}$$

where P is the received optical power, ρ is the detector's unamplified responsivity, G is the detector gain if an APD is used, n accounts for the excess noise of the APD (usually between 2 and 3), B is the receiver's bandwidth, k is Boltzmann's constant ($k = 1.38 \times 10^{-23}$ J/K), e is the magnitude of the charge on an electron (1.6×10^{-19} coulomb), T is the receiver's temperature in degrees kelvin, I_D is the detector's dark current, and R_L is the resistance of the load resistor that follows the photodetector.

The first term in the denominator of Eq. (71.12) is caused by shot noise and the second term is attributed to thermal noise in the receiver. If the shot noise term dominates (and the APD excess loss and dark current are negligible), the system is shot-noise limited. In this case the probability of error has an upper bound given by:

$$P_e = e^{-n_s} \tag{71.13}$$

where n_s is the average number of photoelectrons generated by the signal during a single bit interval when a binary 1 is received. An error rate of 10^{-9} or better requires about 21 photoelectrons per bit. Shot noise depends on the optical signal level. Because the power level is normally low at the end of a long-distance system, the shot noise is small compared to the thermal noise. Avalanche photodiodes increase the shot noise compared to the thermal noise. With APD receivers, ideal shot-noise limited operation can be approached but (because of the APD excess noise and limited gain) not reached.

If the thermal noise dominates, the error probability is given by

$$P_e = 0.5 - 0.5 \text{ erf } (0.354 \sqrt{S/N}) \tag{71.14}$$

where erf is the error function. An error rate of 10^{-9} requires a signal-to-noise ratio of nearly 22 dB.

System Design

A major part of fiber system design involves the power budget and the bandwidth budget. The next few paragraphs describe these calculations.

In a fiber system, component losses (or gains) are normally given in decibels. The decibel is defined by

$$\text{dB} = 10 \log P_2/P_1 \tag{71.15}$$

where P_2 and P_1 are the output and input powers of the component. The decibel describes relative power levels. Similarly, dBm and dBμ describe absolute power levels. They are given by

$$\text{dBm} = 10 \log P \tag{71.16}$$

where P is in milliwatts and

$$\text{dB}\mu = 10 \log P \tag{71.17}$$

where P is in microwatts.

Power budget calculations are illustrated in Table 71.1 for a system that includes an amplifier. A specific numerical example is found in the last two columns. The receiver sensitivity in dBm is subtracted from the power available from the light source in dBm. This difference is the loss budget (in decibels) for the system. All the system losses and gains are added together (keeping in mind that the losses are negative and the amplifier gains are positive). If the losses are more than the gains (as is usual), the system loss dB_{SL} will be a negative number. The loss margin is the sum of the loss budget and the system loss. It must be positive for the system to meet the receiver sensitivity requirements. The system loss margin must be specified to account for component aging and other possible system degradations. A 6-dB margin was found for the system illustrated in the table. The fiber in the table has a total loss of 24 dB. If its attenuation is 0.25 dB/km, the total length of fiber allowed would be 24/0.25 = 96 km.

TABLE 71.1 Power Budget Calculations

Source power	dBm_s	3	
Receiver sensitivity	dBm_r	–30	
Loss budget: $\text{dBm}_s - dBm_r$	dB_{LB}		33
Component efficiencies			
Connectors	dB_c	–5	
Splices	dB_s	–2	
Source coupling loss	dB_{cl}	–5	
Fiber loss	dB_f	–24	
Isolator insertion loss	dB_i	–1	
Amplifier gain	dB_a	10	
Total system loss			
$\text{dB}_c + \text{dB}_s + \text{dB}_{cl} + \text{dB}_f + \text{dB}_i + \text{dB}_a$	dB_{SL}		–27
Loss margin: $\text{dB}_{LB} + \text{dB}_{SL}$	dB_{LM}		6

In addition to providing sufficient power to the receiver, the system must also satisfy the bandwidth requirements imposed by the rate at which data are transmitted. A convenient method of accounting for the bandwidth is to combine the rise times of the various system components and compare the result with the rise time needed for the given data rate and pulse coding scheme.

The system rise time is given in terms of the data rate by the expression

$$t = 0.7/R_{NRZ} \tag{71.18}$$

for non-return-to-zero (NRZ) pulse codes and

$$t = 0.35/R_{RZ} \tag{71.19}$$

for return-to-zero (RZ) codes.

An example of bandwidth budget calculations appears in Table 71.2. The calculations are based on the accumulated rise times of the various system components.

TABLE 71.2 Bandwidth Budget Calculations[a]

Transmitter	t_t	0.8	
Fiber	t_f	0.36	
Receiver	t_r	1	
System total: $\sqrt{t_t^2 + t_f^2 + t_r^2}$	t_s		1.33
System required	t		1.4

[a]All quantities in the table are rise time values in nanoseconds.

The system in Table 71.2 runs at 500 Mbps with NRZ coding for a 100-km length of fiber. Equation (71.18) yields a required system rise time no more than 1.4 ns. The transmitter is assumed to have a rise time of 0.8 ns. The receiver rise time, taken as 1 ns in the table, is a combination of the photodetector's rise time and that of the receiver's electronics.

The fiber's rise time was calculated for a single-mode fiber operating at a wavelength of 1550 nm. Equation (71.9) shows that $M = 18$ ps/(nm × km) at 1550 nm. The light source was assumed to have a spectral width of 0.2 nm. Then, the pulse dispersion calculated from Eq. (71.10) yields a pulse spread of 0.36 ns. Because the fiber's rise time is close to its pulse spread, this value is placed in the table.

The total system rise time is the square root of the sum of the squares of the transmitter, fiber, and receiver rise times. That is:

$$t_s = \sqrt{t_t^2 + t_f^2 + t_r^2} \tag{71.20}$$

In this example, the system meets the bandwidth requirements by providing a rise time of only 1.33 ns, where as much as 1.4 ns would have been sufficient.

Defining Terms

Coherent detection: The signal beam is mixed with a locally generated laser beam at the receiver. This results in improved receiver sensitivity and in improved receiver discrimination between closely spaced carriers.
Material dispersion: Wavelength dependence of the pulse velocity. It is caused by the refractive index variation with wavelength of glass.
Quantum efficiency: A photodiode's conversion efficiency from incident photons to generated free charges.
Single-mode fiber (SMF): A fiber that can support only a single mode of propagation.
Spectral width: The range of wavelengths emitted by a light source.

Related Topics

42.2 Optical Fibers and Cables • 43.2 Amplifiers

References

E. E. Basch, Ed., *Optical-Fiber Transmission*, Indianapolis: Howard W. Sams & Co., 1987.
C. C. Chaffee, *The Rewiring of America*, San Diego: Academic Press, 1988.
M. J. F. Digonnet, *Rare Earth Doped Fiber Lasers and Amplifiers,* New York: Marcel Dekker, 1993.
R. J. Hoss, *Fiber Optic Communications Design Handbook*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.
L. B. Jeunhomme, *Single-Mode Fiber Optics,* 2nd ed., New York: Marcel Dekker, 1990.
N. Kashima, *Passive Optical Components for Optical Fiber Transmission,* Norwood, Mass.: Artech House, 1995.
G. Keiser, *Optical Fiber Communications*, 2nd ed., New York: McGraw-Hill, 1991.
R. H. Kingston, *Optical Sources, Detectors and Systems,* New York: Academic Press, 1995.
J. C. Palais, *Fiber Optic Communications*, 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1992.
S. Shimada, *Coherent Lightwave Communications Technology,* New York: Chapman and Hall, 1994.
A. Yariv, *Optical Electronics*, 4th ed., Philadelphia: Saunders College Publishing, 1991.

Further Information

Continuing information on the latest advances in long-distance fiber communications can be obtained from several professional society journals and several trade magazines including: *IEEE Journal of Lightwave Technology, IEEE Photonics Technology Letters, Lightwave*, and *Laser Focus World.*

71.3 Photonic Networks

Ivan P. Kaminow

Lightwave technology has been developed and widely utilized for local and long-distance transmission in the public telephone network (see Section 71.2 and Miller and Kaminow, 1988) and in modern CATV (cable TV) networks (Section 71.1). Computer communications have been provided utilizing copper transmission lines in private local-area networks (LAN) that cover short distances, $L < 10$ km, and involve low data throughputs $S < 10$ Mb/s. The throughput is defined as

$$S = NB$$

with N the number of simultaneous interconnections and B the communication bit-rate per user. At the higher ranges of the bit-rate-distance product, above BL = (10 Mb/s) (10 km) = 100 Mb/s·km or, equivalently, higher ranges of the bit-rate-delay product

$$M = BD = (BL)(n/c),$$

where D is the propagation delay, c/n is the (group) velocity of bits on the transmission line, and c is the velocity of light, optical fiber may be preferable to copper. For optical fibers, with refractive index $n = 1.5$, the delay is $n/c = 5 \times 10^{-9}$ s/m = 5 μs/km. Thus, with BL = 100 Mb/s·km, M is 500 bits, i.e., there are 500 bits in transit on the transmission line between transmitter and receiver.

As M gets larger, the performance of copper transmission lines—twisted pairs or coax—becomes unsatisfactory because of attenuation and pulse dispersion. The economic break-even value for M, where the added cost of lightwave technology is justified, though not precise, is in the neighborhood of 500 bits. In this section, we will cover aspects of lightwave data networks that utilize the lightwave technology discussed in Section 71.2 and some of the multiple-access methods for LANs discussed in Section 66.3. The latter section touches on commercial LAN standards that utilize optical data links for point-to-point transmission between nodes, often with multimode fiber. Here, we will first discuss some of the recent optical LAN standards and then briefly mention proposed approaches to photonic networks with terabit-per-second throughput and gigabit-per-second user access, and the novel optical components that are needed to realize this high performance. When such networks connect users separated by $L \sim$ 1000s of kilometers, $M \sim$ 10s of megabits may be in transit, requiring new approaches to congestion control for multiple access.

Data Links

A data link consists of a transmitter (T) that converts electrical pulses to optical pulses (E/O) and sends the optical pulses on an optical fiber to a receiver (R) which converts the optical pulses back to electrical pulses (O/E). The transmitter may use a light-emitting diode (LED) or a laser diode (LD) as the optical source. The LED is cheaper but has lower output power into the fiber (~10 μW vs. ~1 mW), lower modulation bandwidths (~100 Mb/s vs. ~1 Gb/s), and wider optical spectrum, leading to chromatic dispersion due to the variation of optical velocity in the fiber with wavelength. Pulse dispersion limits BL when the pulse spreading approaches a bit period. The receiver may employ a PIN (positive-intrinsic-negative) or APD (avalanche photodiode) photodetector. The former is cheaper and easier to bias but has poorer sensitivity by about 5 dB. The sensitivity of a good PIN receiver is about –50 dBm at 100 Mb/s and –35 dBm at 1 Gb/s for a bit-error-rate (BER) of 10^{-9}. Optical devices operating at a wavelength of ~0.87 μm use gallium-aluminum-arsenide materials and are less expensive than those operating at 1.3 or 1.5 μm and using indium-gallium-arsenide-phosphide materials. However, for 1.5 μm, the fiber attenuation and, for 1.3 μm, the chromatic dispersion is much less than for 0.87 μm. The fiber joining transmitter and receiver may be multimode or single mode. Multimode (with a typical core diameter of 62.5 μm) is cheaper, but since each mode travels at a different optical velocity, the modal dispersion further limits BL. The LED data links generally employ multimode fiber, and the combination of chromatic and modal dispersion limits BL to values below ~1 Gb/s·km. For an LD, single-mode fiber (core diameter ~10 μm) data link, BL of ~100 Gb/s·km is possible.

Optical data links are employed to connect electronic components of a LAN when copper is no longer feasible. However, because of its lower cost and the fact that it is often already installed, clever tricks are now being used to extend the utility of copper.

Token Ring: FDDI, FFOL

Figure 71.8 illustrates a ring network. The real topology may be a good deal more irregular than a circle, depending on the accessibility of stations. In its usual application, which uses a token-ring protocol for media access, an electronic repeater that operates at the aggregate network rate is required at each station.

A token—e.g., a "1" or a "0" bit, or a token packet—is propagated in one direction from station to station. When a station has a packet to send to another station, it adds the address of the receiving station in a header

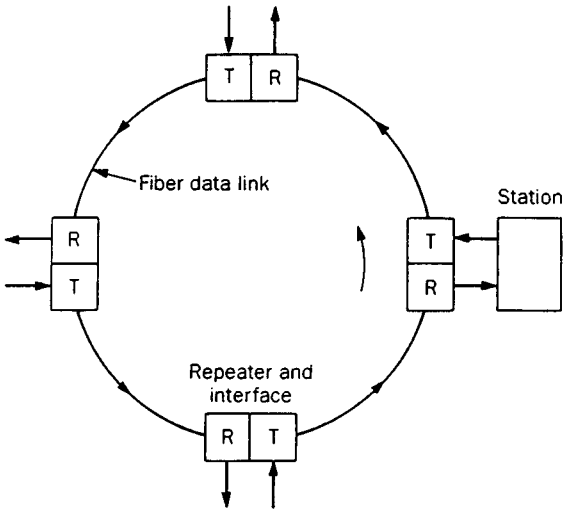


FIGURE 71.8 Undirectional ring network. R and T represent the receiver and transmitter functions, respectively.

and holds the combined packet in a buffer. The sending station reads the tokens as they go by until it receives an empty token, a "0." It then converts the "0" to a "1," a busy token, and appends the packet.

Intermediate stations repeat the bits in the packet and also "listen" for their own addresses. If a station recognizes its address in the packet header, it copies the packet. When the packet returns to the sender, it serves as an acknowledgment, and the sender removes it from the ring, after converting the token back to "0."

Commercial token rings use wire interconnections or optical data links to join stations at rates in the 10-Mb/s range. Actual network use is less than 10 Mb/s because of the time it takes an empty token to pass around the ring. This transit-time delay increases linearly with the number of stations. It includes propagation delay between stations and processing delay at each station, which must examine the header of every packet before repeating the bits to the next station.

A token-ring architecture is not especially attractive for a high-speed optical network (where $S \sim B$ is above 1 Gb/s) because of the cost of high-speed repeater optoelectronics at each station and the packet-processing delay. In addition, at high bit rates, the packet time may be much shorter than the propagation time around the ring, unless a packet contains an impractically large number of bits. Efficient use of the ring with short packets may call for multiple tokens, which can lead to complex protocols. Increasing the number of bits per packet increases the packet time but places added burden on the high-speed buffer at each station.

Reliability—if one station is disabled or if the fiber breaks—is a problem in both fiber and wire rings. To address these reliability problems, a double-ring optical network can provide for bypass of defective stations and loop back around a fiber break. Each station has two inputs and two outputs connected to two rings that operate in opposite directions; this, of course, increases the cost.

The fiber distributed data interface (FDDI) [Ross, 1986, 1989] is a standard proposed by the American National Standard Institute for a 100-Mb/s double-ring, time-division multiplexed (TDM) LAN that uses 1.3-μm multimode fiber and LED (or single-mode fiber and laser diode) data links between stations. This LAN is designed to provide both backbone services that interconnect lower speed LANs and back-end services that interconnect mainframe computers, mass storage systems, and other high-speed peripherals. FDDI provides datagram packet service with up to about 4,500 data bytes per frame. It employs 4B/5B coding so that the clock rate is 125 MHz for maximum S of 100 Mb/s. The FDDI network is designed to operate with low-cost components that were commercially available in 1986. The standard can provide both packet-switched and circuit-switched services. As many as 1000 stations can be connected, with a maximum of 2 km between stations and a maximum perimeter of 200 km.

An FDDI follow on LAN (FFOL) will operate with laser and single-mode fiber data links at bit rates corresponding to the synchronous optical network (SONET) or synchronous digital hierarchy (SDH) standards

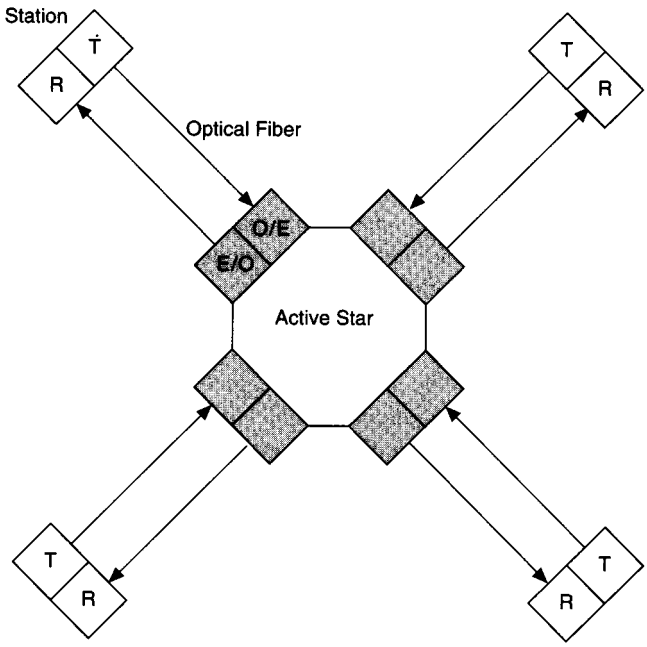


FIGURE 71.9 Active star network. Optical-to-electrical (O/E) and electrical-to-optical (E/O) converters must be provided at the star. R and T represent the receiver and transmitter functions, respectively.

of 622 Mb/s and 2.5 Gb/s, possibly with ATM (asynchronous transfer mode) cells. The geographical size will also be increased.

Active Star Networks: Ethernet, Datakit®

Carrier sense multiple access with collision detection (CSMA/CD) Ethernet networks operating at 10 Mb/s connect users on a copper bus. The length of the bus must be less than 1/2 the distance light propagates in the time required to transmit a packet frame. Thus, for speeds much greater than 10 Mb/s, where optics might be needed, the length of the bus will be limited unless the maximum frame contains an impractically large number of bits. Further, the number of stations that can be supported by an optical bus is limited by the nature of optical taps, as opposed to electrical taps [Kaminow, 1989]. Finally, the collision detection algorithm does not work well on an optical bus because the intensities of optical packets from two different stations may vary considerably along the bus. Thus, an active electronic star, as shown in Fig. 71.9, with optical data links from users is often employed.

The AT&T Datakit [Fraser, 1983] packet switch behaves as a virtual circuit switch (VCS) in that a reliable data path is set up for each session, and packet retransmission because of collisions is not required. Remote stations that may consist of mainframe computers, concentrators that bring together many terminals, or gateways to other networks are connected by 8-Mb/s fiber-optic data links to individual electronic modules at the central node as shown in Fig. 71.10 [Kaminow, 1988]. These modules plug into two electronic buses that are short (about 1 m) compared to a packet propagation length (16 bytes).

In the module, packets are formed and stored with a header that contains the source address. When the packet is complete (it has the full number of bytes, or a fixed waiting period for added bytes has passed), the module transmits its binary address on the contention bus while listening for bits transmitted by others. If the module transmits a 1 and hears a 1, it transmits the next address bit. But if it transmits a 0 and hears a 0, it transmits the next bit; and, if it transmits a 0 and hears a 1, it stops transmission, having lost the contention. This process is equivalent to a logical OR operation and assigns the contention to the highest address. The winner transmits the packet on the contention bus in the next time frame.

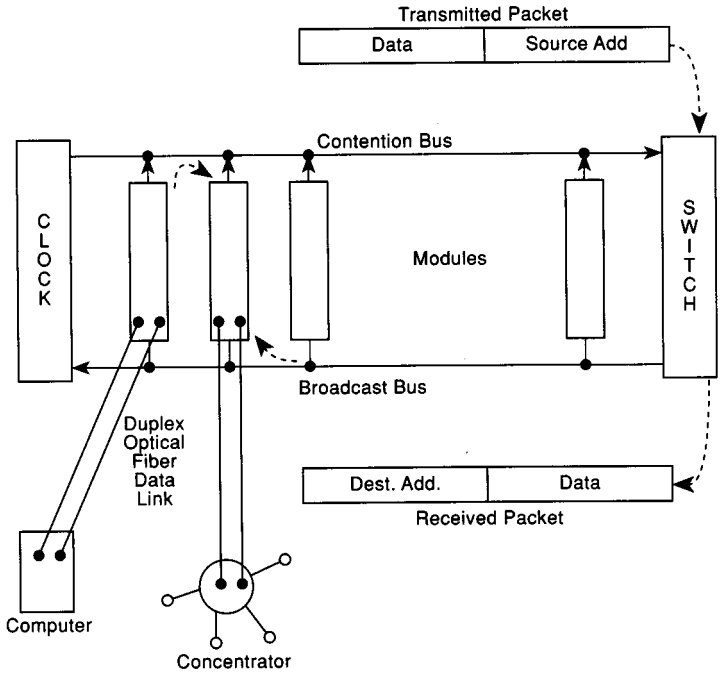


FIGURE 71.10 Datakit® VCS network. Remote stations are connected to the electrical node by 8-Mb/s data links. The length of the bus is much shorter than the propagation length of a packet.

The switch at the end of the bus replaces the source address with the destination address and transmits the packet on the broadcast bus, where the appropriate module records the destination address and sends the packet to the remote station over the fiber link. Because the switch establishes a correspondence between source and destination at the beginning of a session (as in a circuit switch), source modules need not know the bus position of destination modules. The switch has a directory of positions and terminal names.

If we were to go to very high bit rates, the physical bus length might no longer be short compared with a packet, and collisions caused by delays might upset the "perfect scheduling" of packets. Although methods have been proposed [Acampora and Hluchyj, 1984] for overcoming this limitation, the electronic circuit costs and electrical reflections on the bus may limit the effectiveness of a centralized bus at very high data rates.

New Approaches to Optical Networks

The preceding conventional networks with optical data links replacing copper can improve their throughputs thanks to the increased bandwidth of the transmission medium. However, a revolutionary improvement in throughput to terabit-per-second levels with gigabit-per-second access requires entirely new approaches for the physical connectivity, architecture, and access protocols. We can use much of the photonic technology employed in long-haul lightwave systems to provide physical connectivity, but we also need devices with new functionality to realize proposed architectures, and, conversely, with new component functionality we can dream of new architectures.

We can provide connectivity among users in three dimensions: space, time, and optical frequency or wavelength, employing space-division multiplexing (SDM), optical time-division multiplexing (OTDM), and optical frequency-division multiplexing (OFDM) or wavelength-division multiplexing (WDM), respectively. To control the path routing we need optical switches for OTDM and frequency routing technology for OFDM. At present, network architectures and protocols are at the research stage. We mention some of these components and switches in the following paragraphs. More details can be found in the References [Miller and Kaminow, 1988; Special Issue, 1990].

A star topology seems most attractive for gigabit-per-second multiple-access photonic networks [Kaminow, 1989], as shown in Fig. 71.11. Each station has its own transmitter and receiver. For optical TDM, the connectivity can be provided by an $N \times N$ electrooptic switch and suitable controller, and for optical FDM, the connectivity is provided by a passive $N \times N$ star coupler. Electrooptic $N \times N$ switches based on integrated titanium-diffused lithium niobate waveguide elements [Korotky and Alferness, 1988] have been demonstrated with $N = 16$ and operating at $B = 2.5$ Gb/s for each input. The switch connections can be rearranged in a few nanoseconds. It is estimated that such switches could be interconnected to provide $N = 256$. Unlike electronic switches, electrooptic switches are transparent to the bit rate, i.e., they can connect any bit stream independent of B. The problems of suitable multiple-access protocols and controls have not yet been fully addressed.

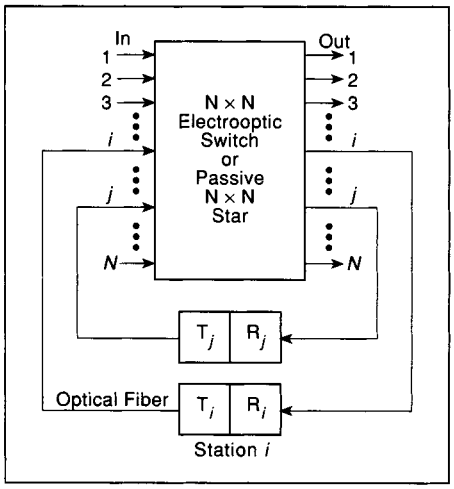


FIGURE 71.11 Electrooptic switch network for optical TDM or passive star network for optical FDM. The $N \times N$ switch or star have N single-mode optical fiber input ports and N optical fiber output ports. As in the active star (Fig. 71.9), two fibers connect a remote station with the star. R_x and T_x represent the receiver and transmitter functions, respectively, for station x.

The passive $N \times N$ star coupler [Kaminow, 1989] in Fig. 71.11 has N single-mode fiber inputs and N outputs. In an ideal passive star, a signal incident on any input is divided equally among all the outputs, i.e., the star broadcasts every input to every output. Unlike the OTDM case, each transmitter uses a different optical frequency and each receiver must tune to the frequency of the channel intended for it, as illustrated in Fig. 71.12. Alternatively, the receiver frequencies may be fixed and the transmitters tunable. Thus, the control can be distributed in the terminals. Calculations indicate that such a network can support throughputs of several terabits per second. Current research [Special Issue, 1990] is aimed at devising multiple-access protocols and demonstrating the novel devices needed for optical frequency routing. These include fast tunable lasers and receivers that can cover many channels (switching speeds of ~10 ns at 2.5 Gb/s with ~50 channels appear feasible), optical frequency translators for frequency reuse, and integrated star couplers and integrated optical frequency routers.

One challenge in photonic network design is to make them "all-optical," as nearly as possible, in order to avoid throughput bottlenecks by electronic components and the expense of O/E and E/O conversions. In principle, clear all-optical channels would offer connectivity independent of data-rate and format for a wide variety of applications. However, many physical technology problems remain and new concepts for multple access and congestion control [Special Issue, 1991] suited to large bit-rate-delay (M) systems must be found.

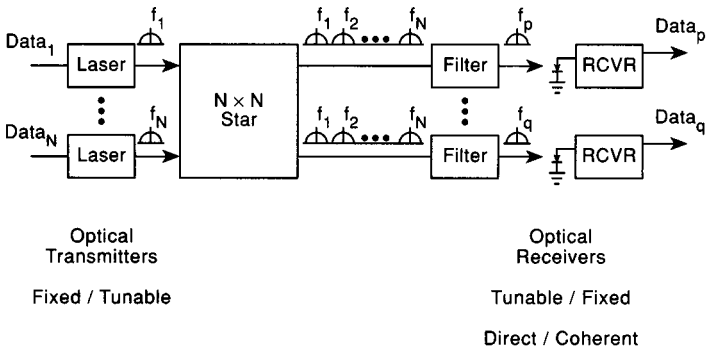


FIGURE 71.12 An optical FDM network with passive star distribution. Optical transmitter frequencies $f_{1 \ldots} f_N$ are modulated with data at the transmitter and selected by a filter at the receiver.

Related Topic

72.2 Computer Communication Networks

References

A. S. Acampora and M. G. Hluchyj, "A new local area network architecture using a centralized bus," *IEEE Communications Magazine,* vol. 22, no. 8, pp. 12–21, 1984.

A. G. Fraser, "Towards a universal data transport system," *IEEE J. Selected Areas in Communications,* vol. SAC-1, no. 5, pp. 803–816, 1983.

I. P. Kaminow, "Photonic multiple access networks," *AT&T Technical Journal,* vol. 68, no. 2, pp. 61–86, 1989.

I. P. Kaminow, "Photonic local networks," in *Optical Fiber Telecommunications, II,* New York: Academic Press, 1988, chap. 26.

S. K. Korotky and R. C. Alferness, "Waveguide electrooptic devices for optical fiber communication," in *Optical Fiber Telecommunications, II,* New York: Academic Press, 1988, chap. 11.

S. E. Miller and I. P. Kaminow, Eds., *Optical Fiber Telecommunications, II,* New York: Academic Press, 1988.

F. E. Ross, "FDDI—A tutorial," *IEEE Communications Magazine,* vol. 24, no. 5, pp. 10–17, 1986.

F. E. Ross, "An overview of FDDI—The fiber distributed data interface," *IEEE J. Selected Areas in Communications,* vol. 7, no. 7, pp. 1043–1051, 1989.

Special Issue, "Congestion control in high speed networks," *IEEE Communications Magazine,* vol. 29, no. 10, 1991.

Special Issue "Dense wavelength division multiplexing techniques for high capacity and multiple access communications systems," *IEEE J. Selected Areas in Communication,* vol. 8, no. 6, 1990.

Further Information

J. G. Nellist, *Understanding Telecommunications and Lightware Systems,* 2nd ed. IEEE Press, 1996.

J. Gibson, *The Mobile Communication Handbook,* Boca Raton, Fla.: CRC Press, 1996.

S. Betti, *Coherent Optical Communications Systems,* New York: Wiley, 1995.

B. Saleh, *Fundamentals of Photonics,* New York: Wiley, 1992.

I. P. Kaminow and T. L. Koch, *Optical Fiber Telecommunications, III*, New York: Academic Press, 1997.

Huber, M.N., Daigle, J.N., Bannister, J., Gerla, M., Robrock II, R.B. “Networks”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

72

Networks

Manfred N. Huber
Siemens

J. N. Daigle
University of Mississippi

Joseph Bannister
University of Southern California Information Sciences Institute

Mario Gerla
University of California, Los Angeles

Richard B. Robrock II
Bell Communications Research

72.1 B-ISDN
B-ISDN Services and Applications • Asynchronous Transfer Mode • Transmission of B-ISDN Signals • ATM Adaptation Layer • B-ISDN Signaling

72.2 Computer Communication Networks
General Networking Concepts • Computer Communication Network Architecture • Local-Area Networks and Internets • Some Additional Recent Developments

72.3 Local-Area Networks
The LAN Service Model • Other Features • The Importance of LAN Standards

72.4 The Intelligent Network
A History of Intelligence in the Network • The Intelligent Network • Intelligent Network Systems • The CCS7 Network • The Service Control Point • Data Base 800 Service • Alternate Billing Services • Other Services • The Advanced Intelligent Network • Back to the Future

72.1 B-ISDN

Manfred N. Huber

Since the mid-1980s the idea of the **integrated services digital network** (ISDN) has become reality. In ISDN voice services with supplementary features and data services with a bit rate of up to 64 kbit/s are integrated in one network. For voice communication and many text and data applications the 64-kbit/s ISDN will be sufficient. Although it is minor as yet, there exists already a growing demand for **broadband** communication with bit rates from some megabits per second up to approximately 130 Mbit/s [Wiest, 1990] (e.g., high-speed data communication, video communication, high-resolution graphics).

In order to provide the same advantages of ISDN to broadband communication users, network operators, and service providers, the development of an intelligent broadband-ISDN (B-ISDN) is necessary. The future B-ISDN will become the universal network integrating different kinds of services with their individual features and requirements. B-ISDN will support switched, semipermanent and permanent, point-to-point, and point-to-multipoint connections and provide on-demand, reserved, and permanent services. B-ISDN connections support packet mode and circuit mode services of mono- and/or multimedia type of a connection-oriented or connectionless nature in a unidirectional or bidirectional configuration [Händel and Huber, 1991b].

B-ISDN Services and Applications

As already mentioned, there exists some demand for broadband communication which originates from business customers as well as residential customers. In the residential area, on the one hand, people are interested in video distribution services for entertainment purposes, like television and high-definition TV; on the other

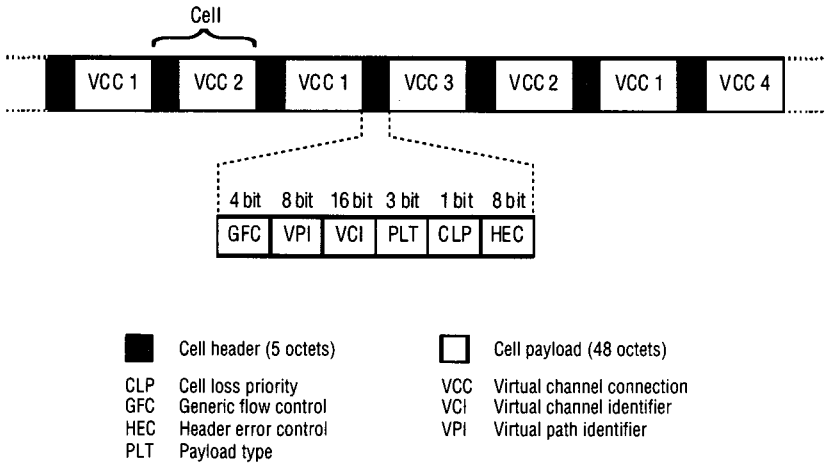


FIGURE 72.1 ATM principle.

hand, they will use video telephony with acceptable quality. Over the long term, video mail services and video retrieval services will become more important.

Voice and text are no longer sufficient for business customers. In the offices and factories of tomorrow, interactive broadband services will be required. Handling complex tasks in the future demands comprehensive support by services for voice, text, data, graphics, video, and documents. In addition to the individual services, the multimedia services and the simultaneous or alternating use of several services with multifunction workstations will gain importance [Armbrüster, 1990].

Interconnection of local-area networks (LANs) or large computers, computer-aided design, and computer-aided manufacturing will become important data applications. The first video services will be video telephony and video conferencing (studio-to-studio and workstation video conferencing). Initially these services may have diminished quality, but for the long term TV quality can be expected.

The bit rates of all services mentioned above are in the range of 2 to 130 Mbit/s (depending on the individual application). Taking into account that in the future more enhanced video coding mechanisms will be available, the required bit rates for video services will become lower without influencing quality significantly.

Asynchronous Transfer Mode

In today's public switched networks the synchronous transfer mode (STM) predominates. Applying STM technology, for the duration of a connection a synchronous channel with constant bit rate is allocated to that connection. STM does not fit very well for the integration of services with bit rates from some kilobits per second to 130 Mbit/s. Therefore, in B-ISDN a new transfer mode called **asynchronous transfer mode** (ATM) is used.

In ATM all kinds of information is transported in **cells**. A cell is a block of fixed length, which consists of a 5-octet cell header and a 48-octet cell payload (see Fig. 72.1). The cell header contains all necessary information for transferring the cell through the network and the cell payload includes the user information. The cell rate of a connection is proportional to the service bit rate. Only if information is available is a cell used by the connection. By having different routing labels, cells of different connections can be transported on the same transmission line (cell multiplexing). If no connection has information ready to transport, idle cells will be inserted. Idle cells do not belong to any connection; they are identified by a standardized cell header.

ATM uses only cells; multiplexing and switching of cells is independent of the applications and of the bit rates of the individual connections. Applying ATM technology, the idea of one universal integrated network becomes a reality. However, the ATM technology also causes some problems. Because of the asynchronous multiplexing buffers are necessary, which results in cell delay, cell delay variation, and cell loss. In order to compensate for these effects additional measures have to be provided.

Figure 72.1 also shows the individual subfields of the cell header. The first field, called generic flow control (GFC), is only available at the user-network interface (UNI). Its main purpose is media access control in shared medium configurations (LAN-like configurations) within the customer premises [Göldner and Huber, 1991]. The proposed GFC procedures are based either on the distributed queue algorithm or the reset timer control mechanism [Händel and Huber, 1991a]. At the network-node interface (NNI) these bits are part of the virtual path identifier (VPI).

The VPI together with the virtual channel identifier (VCI) form the routing label (identifier of the connection). The VPI itself marks only the virtual path (VP). The VP concept allows the flexible configuration of individual subnetworks (e.g., **signaling** network or virtual private network), which can be independent of the underlying transmission network. VP networks are under the control of network management. The bandwidth of a VP will be allocated according to its requirements. Within the VP network the individual connections are established and cleared down dynamically (by signaling).

The payload type field in the cell header differentiates the information in the cell payload of one connection (e.g., user information, operation and maintenance information for ATM). The value of the cell loss priority bit distinguishes cells that can be discarded under some exceptional network conditions without disturbing the quality significantly from those cells that may not be discarded. The last field of the cell header forms the header error control field. The cell header is protected against errors with a mechanism that allows the correction of a single bit error and the detection of multibit errors.

The high transmission speeds for ATM cell transfer require very high-performance switching nodes. Therefore, the switching networks (SNs) have to be implemented in fast hardware. Within the SN the self-routing principle will be applied [Schaffer, 1990]. At the inlet of the SN the cell is extended by an SN-internal header. It is evident that the SN-internal operational speed has to be increased. When passing the individual switching elements, for the processing of the SN-internal header only simple hard-wired logic is necessary. This reduces the control complexity and provides a better failure behavior. When starting several years ago with the implementation of the ATM technology, only the emitter coupled logic (ECL) was available. Nowadays, the complementary metal-oxide semiconductor (CMOS) technology with its low power consumption is used [Fischer et al., 1991].

Transmission of B-ISDN Signals

Transmission systems at the UNI provide bit rates of around 150 and 622 Mbit/s. In addition to these rates, at the NNI around 2.5 Gbit/s and up to 10 Gbit/s will be used in the future [Baur, 1991]. In addition to the high-capacity switching and multiplexing technology, high-speed transmission systems are required. Optical fibers are especially suitable for this purpose; however, for the lower bit rates coaxial cables can be used. Optical transmission uses optical fibers as the transmission medium in low-diameter and low-weight cables to provide large transmission capacities over long distances without the need for repeaters. Optical transmission equipment currently tends to mono-mode fiber and laser diodes with wavelengths of around 1310 nm. For both directions in a transmission system either two separate fibers or one common fiber with wavelength division multiplexing can be used. The second solution may be a good alternative for subscriber lines and short trunk lines [Bauch, 1991].

For ATM cell transmission, two possibilities exist, which are shown in Fig. 72.2: synchronous pulse frame or continuous cell stream (cell-based). The basis for the pulse frame concept is the existing **synchronous digital hierarchy** (SDH). In SDH the cells are transported within the SDH payload; the frame overhead includes operation and maintenance (OAM) of the transmission system. In the cell-based system the OAM for the transmission system is transported within cells. The SDH solution is already defined, whereas for cell-based transmission some problems remain to be solved (e.g., OAM is not yet fully defined).

ATM Adaptation Layer

The **ATM adaptation layer** (AAL) is between the ATM layer and higher layers. Its basic function is the enhanced adaptation of the services provided by ATM to the requirements of the layers above. In order to minimize the number of AAL protocols, the service classification shown in Fig. 72.3 was defined. This classification was made with respect to timing relation, bit rate, and connection mode.

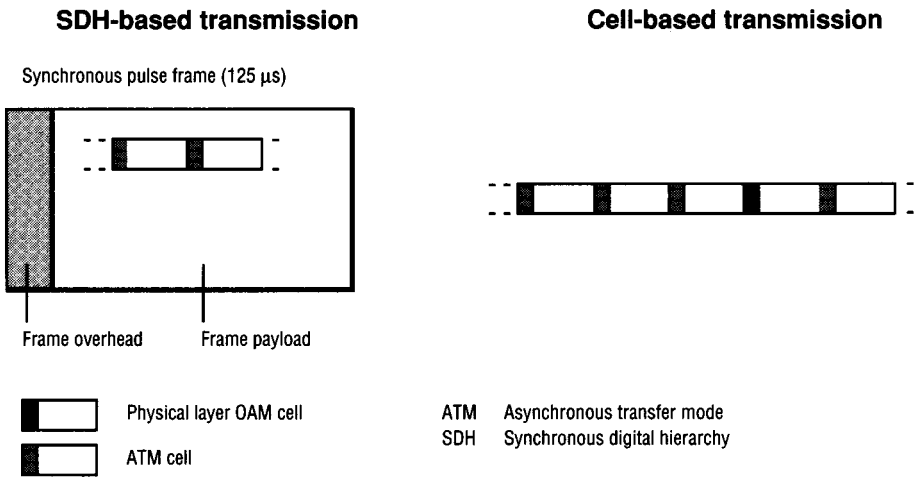


FIGURE 72.2 Transmission principles for B-ISDN.

	Class A	Class B	Class C	Class D
Timing relation between source and destination	Required		Not required	
Bitrate	Constant	Variable		
Connection mode	Connection oriented			Connection-less

FIGURE 72.3 AAL service classification.

The AAL protocols are subdivided into two parts. The lower part performs, at the sending side, the segmentation of long messages into the cell payload and, at the receiving side, reassembly into long messages. The upper part is service dependent and provides the AAL service to the higher layer.

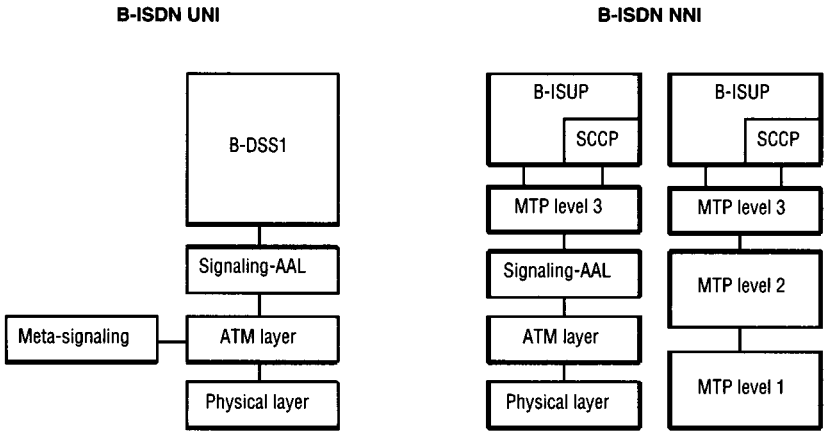
B-ISDN Signaling

For signaling in B-ISDN, existing protocols and infrastructure will be reused as much as possible. Figure 72.4 shows the protocol stacks for UNI and NNI. The upper part concerns signaling applications and the lower part signaling transfer.

For the introduction of simple switched services in B-ISDN, at UNI and NNI, existing signaling application protocols will be reused. The 64-kbit/s ISDN-specific information elements will be removed and new B-ISDN-specific information elements will be added. Right from the beginning these protocols will provide means that allow smooth migration toward future applications, which will include highly sophisticated features like multimedia services [Huber et al., 1992]. This approach guarantees compatibility for future protocol versions.

At the NNI the existing signaling system no. 7 (SS7) can be reused (see right part of the NNI protocol stack in Fig. 72.4). SS7 is a powerful and widespread network that will continue to be applied for rather a long period until ATM penetration has been reached. For the middle term, however, a fully ATM-based network will be available which also carries signaling messages (see left part of the NNI protocol stack in Fig. 72.4). ATM-based signaling at the NNI needs a suitable AAL which provides the services of the existing message transfer part level 2.

At the UNI, right from the beginning, all kinds of traffic (including signaling) is carried within cells. An AAL for signaling at the UNI is also required. This AAL has to provide the services of the existing layer 2 UNI



AAL	ATM adaptation layer
ATM	Asynchronous transfer mode
B-DSS1	Broadband digital subscriber signaling system no. 1
B-ISUP	Broadband-ISDN user part
MTP	Message transfer part
NNI	Network-node interface
SCCP	Signaling connection control part
UNI	User-network interface

FIGURE 72.4 Protocol stacks for B-ISDN signaling.

protocol. The AAL for signaling at UNI and NNI will be common as much as possible. In contrast to the NNI, at the UNI meta-signaling is necessary. Meta-signaling establishes, checks, and removes the signaling channels between customer equipment and the central office in a dynamic way. The signaling channels at the NNI are semipermanent and, therefore, meta-signaling is not required.

Defining Terms

Asynchronous transfer mode: A transfer mode in which the information is organized into cells; it is asynchronous in the sense that the recurrence of cells containing information from an individual user is not necessarily periodic.

ATM adaptation layer: A layer which provides the adaptation of higher layers to ATM.

Broadband: A service or system requiring transmission channels capable of supporting bit rates greater than 2 Mbit/s.

Cell: A block of fixed length which is subdivided into a cell header and an information field. The cell header contains a label which allows the clear allocation of a cell to a connection.

Integrated services digital network: A network which provides end-to-end digital connectivity to support a wide range of services, including voice and nonvoice services, to which users have access by a limited set of standard multipurpose user-network interfaces.

Signaling: Procedures which are used to control (set up and clear down) calls and connections within a telecommunication network.

Synchronous digital hierarchy: A standard for optical transmission which provides transmission facilities with flexible add/drop capabilities to allow simple multiplexing and demultiplexing of signals.

Related Topic

72.2 Computer Communications Networks

References

H. Armbrüster, "Blueprint for future telecommunications," *Telcom Report International*, vol. 13, no. 1, pp. 5–8, 1990.

H. Bauch, "Transmission systems for B-ISDN," *IEEE LTS, Magazine of Lightwave Telecommunication,* vol. 2, no. 3, pp. 31–36, 1991.

H. Baur, "Technological perspective of telecommunications for the nineties," *Integration, Interoperation and Interconnection: This Way to Global Services, Proceedings of the Technical Symposium*, Geneva, part 2, vol. 1 paper 1.1, 1991.

W. Fischer, O. Fundneider, E.-H. Goeldner, and K.A. Lutz, "A scalable ATM switching system architecture," *IEEE Journal on Selected Areas in Communication,* vol. 9, no. 8, pp. 1299–1307, 1991.

E.-H. Göldner and M.N. Huber, "Multiple access for B-ISDN," *IEEE LTS, Magazine of Lightwave Telecommunication,* vol. 2, no. 3, pp. 37–43, 1991.

R. Händel and M.N. Huber, "Customer network configurations and generic flow control," *International Journal of Digital and Analog Communication Systems,* vol. 4, no. 2, pp. 117–122, 1991a.

R. Händel and M.N. Huber, *Integrated Broadband Networks — An Introduction to ATM-Based Networks*, Reading, Mass.: Addison-Wesley, 1991b.

M.N. Huber, V. Frantzen, and G. Maegerl, "Proposed evolutionary paths for B-ISDN signalling," *Proceedings of the XIV International Switching Symposium,* Yokohama, vol. 1, pp. 334–338, 1992.

B. Schaffer, "ATM switching in the developing telecommunication network," *Proceedings of the XIII International Switching Symposium,* vol. 1, pp. 105–110, 1990.

G. Wiest, "More intelligence and flexibility for communication network—Challenges for tomorrow's switching systems," *Proceedings of the XIII International Switching Symposium,* vol. 5, pp. 201–204, 1990.

Further Information

CCITT Recommendations and CCITT Draft Recommendations concerning B-ISDN (parts of F, G, I and Q series), which are published by the International Telecommunication Union.

Journals of the IEEE Communication Society (*Communications Magazine, Journal on Selected Areas in Communications, LTS: Magazine of Lightwave Telecommunication, Networks, Transactions on Communications*), which are published by the Institute of Electrical and Electronics Engineers, Inc.

International Journal of Digital and Analog Communication System, which is published by John Wiley & Sons, Ltd.

Proceedings of international conferences such as GLOBECOM, INFOCOM, International Conference on Communications, International Conference on Computer Communication, International Switching Symposium, International Symposium on Subscriber Loops and Services, and International Teletraffic Congress.

A detailed description of ISDN is given in *ISDN—The Integrated Services Digital Network— Concepts, Methods, Systems*, by P. Bocker, published by Springer-Verlag.

72.2 Computer Communication Networks

J. N. Daigle

A **computer communication network** is a collection of applications hosted on different machines and interconnected by an infrastructure that provides communications among the communicating entities. While the applications are generally understood to be computer programs, the generic model includes the human being as an application. In fact, one or all of the "applications" that are communicating may be human beings.

This section summarizes the major characteristics of computer communication networks. The objective is to provide a concise introduction that will allow the reader to gain an understanding of the key distinguishing characteristics of the major classes of networks that exist today and some of the issues involved in the introduction of emerging technologies.

There are a significant number of well-recognized books in this area. Among these are the excellent texts by Schwartz [1987], Tanenbaum [1988], and Spragins [1991], which have enjoyed wide acceptance by both students and practicing engineers and cover most of the general aspects of computer communication networks. Stallings

[1990a, 1990b, 1990c] covers a broad array of standards in this area. Other books that have been found to be especially useful by practitioners are those by Rose [1990] and Black [1992].

The latest developments are, of course, covered in the current literature, conference proceedings, and the notes of standards meetings. A pedagogically oriented magazine that specializes in computer communications networks is *IEEE Network,* but *IEEE Communications* and *IEEE Computer* often also contain interesting articles in this area. *ACM Communications Review,* in addition to presenting pedagogically oriented articles, often presents very useful summaries of the latest standards activities. Major conferences that specialize in computer communications include the IEEE INFOCOM and ACM SIGCOMM series, which are held annually.

We will begin our discussion with a brief statement of how computer networking came about and a capsule description of the networks that resulted from the early efforts. Networks of this generic class, called **wide-area networks (WANs),** are broadly deployed today, and there are still a large number of unanswered questions with respect to their design. The issues involved in the design of those networks are basic to the design of most networks, whether wide area or otherwise. In the process of introducing these early systems, we will describe and contrast three basic types of communication switching: circuit, message, and packet.

We will next turn to a discussion of computer communication **architecture,** which describes the structure of communication-oriented processing software within a communication processing system. Our discussion is limited to the **International Standards Organization/Open Systems Interconnection** (ISO/OSI) **reference model** (ISORM) because it provides a framework for discussion of some of the modern developments in communications in general and communication networking in particular. This discussion is necessarily simplified in the extreme, thorough coverage requiring on the order of several hundred pages, but we hope our brief description will enable the reader to appreciate some of the issues.

Having introduced the basic architectural structure of communication networks, we will next turn to a discussion of an important variation on this architectural scheme: the **local-area network** (LAN). Discussion of this topic is important because it helps to illustrate what the reference model is and what it is not. In particular, the architecture of LANs illustrates how the ISORM can be adapted for specialized purposes. Specifically, early network architectures anticipate networks in which individual node pairs are interconnected via a single link, and connections through the network are formed by concatenating node-to-node connections.

LAN architectures, on the other hand, anticipate all nodes being interconnected in some fashion over the same communication link (or medium). This, then, introduces the concept of adaption layers in a natural way. It also illustrates that if the services provided by an architectural layer are carefully defined, then the services can be used to implement virtually any service desired by the user, possibly at the price of some inefficiency.

After discussing LANs, we will conclude our article with a discussion of two of the variants in packet switching transmission technology: frame relay and a recent development in basic transmission technology called the **asynchronous transfer mode**, which is a part of the larger **broadband integrated services digital network** effort. These technologies are likely to be important building blocks for the computer communication networks of the future.

General Networking Concepts

Data communication networks have existed since about 1950. The early networks existed primarily for the purpose of connecting users of a large computer to the computer itself, with additional capability to provide communications between computers of the same variety and having the same operating software. The lessons learned during the first twenty or so years of operation of these types of networks have been valuable in preparing the way for modern networks. For the purposes of our current discussion, however, we will think of communication networks as being networks whose purpose is to interconnect a set of applications that are implemented on hosts manufactured by possibly different vendors and managed by a variety of operating systems. Networking capability is provided by software systems that implement standardized interfaces specifically designed for the exchange of information among heterogeneous computers.

During the late 1960s, many forward-looking thinkers began to recognize that significant computing resources (that is, supercomputers) would be expensive and unlikely to be affordable by many of the researchers needing this kind of computer power. In addition, they realized that significant computing resources would not be needed all of the time by those having local access. If the computing resource could be shared by a number of research sites, then the cost of the resource could be shared by its users.

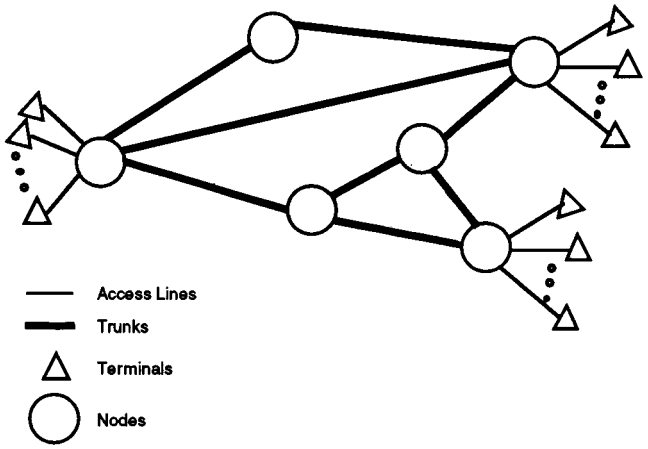


FIGURE 72.5 Generic computer communication network.

Many researchers at this time had computing resources available under the scenario described in the first paragraph above. The idea of interconnecting the computers to extend the reach of these researchers to other computers developed. In addition, the interconnection of the computers would provide for communication among the researchers themselves. In order to investigate the feasibility of providing the interconnectivity anticipated for the future using a new technology called **packet switching,** the Advanced Research Projects Agency (ARPA) of the Department of the Army sponsored a networking effort, which resulted in the computer communication network called the ARPANET.

The end results of the ARPA networking effort, its derivatives, and the early initiatives of many companies such as AT&T, DATAPOINT, DEC, IBM, and NCR have been far-reaching in the extreme. Any finitely delimited discussion of the accomplishments of those efforts would appear to underestimate their impact on our lives. We will concentrate on the most visible product of these efforts, which is a collection of programs that allows applications running in different computers to intercommunicate. Before turning to our discussion of the software, however, we will provide a brief description of a generic computer communication network.

Figure 72.5 shows a diagram of a generic computer communication network. The most visible components of the network are the terminals, the **access lines,** the **trunks,** and the **switching nodes.** Work is accomplished when the users of the network, the terminals, exchange messages over the network.

The terminals represent the set of communication terminating equipment communicating over the network. Equipment in this class includes, but is not limited to, user terminals, general-purpose computers, and database systems. This equipment, either through software or through human interaction, provides the functions required for information exchange between pairs of application programs or between application programs and people. The functions include, but are not limited to, call set-up, session management, and message transmission control. Examples of applications include electronic mail transfer, terminal-to-computer connection for time sharing or other purposes, and terminal-to-database connections.

Access lines provide for data transmission between the terminals and the network switching nodes. These connections may be set up on a permanent basis or they may be switched connections, and there are numerous transmission schemes and protocols available to manage these connections. The essence of these connections, however, from our point of view is a channel that provides data transmission at some number of bits per second (bps), called the channel capacity, *C*. The access line capacities may range from a few hundred bits per second to in excess of millions of bits per second, and they are usually not the same for all terminating equipments of a given network. The actual information-carrying capacity of the link depends upon the protocols employed to effect the transfer; the interested reader is referred to Bertsekas and Gallagher [1987], especially Chapter 2, for a general discussion of the issues involved in transmission of data over communication links.

Trunks, or internodal trunks, are the transmission facilities that provide for transmission of data between pairs of communication switches. These are analogous to access lines, and, from our point of view, they simply provide a communication path at some capacity, specified in bits per second.

There are three basic switching paradigms: circuit, message, and packet switching. **Circuit switching** and packet switching are transmission technologies while message switching is a service technology. In circuit switching, a call connection between two terminating equipments corresponds to the allocation of a prescribed set of physical facilities that provide a transmission path of a certain bandwidth or transmission capacity. These facilities are dedicated to the users for the duration of the call. The primary performance issues, other than those related to quality of transmission, are related to whether or not a transmission path is available at call set-up time and how calls are handled if facilities are not available.

Message switching is similar in concept to the postal system. When a user wants to send a message to one or more recipients, the user forms the message and addresses it. The message switching system reads the address and forwards the complete message to the next switch in the path. The message moves asynchronously through the network on a message switch-to-message switch basis until it reaches its destination. Message switching systems offer services such as mail boxes, multiple destination delivery, automatic verification of message delivery, and bulletin board. Communication links between the message switches may be established using circuit or packet switching networks as is the case with most other networking applications.

Examples of message switching protocols that have been used to build message switching systems are Simple Mail Transfer Protocol (SMTP) and the International Telegraph and Telephone Consultative Committee (CCITT) X.400 series. The former is much more widely deployed, while the latter has significantly broader capabilities, but its deployment is plagued by having two incompatible versions (1984 and 1988) and other problems. Many commercial vendors offer message switching services based on either one of the above protocols or a proprietary protocol.

In the circuit switching case, there is a one-to-one correspondence between the number of trunks between nodes and the number of simultaneous calls that can be carried. That is, a trunk is a facility between two switches that can service exactly one call, and it does not matter how this transmission facility is derived. Major design issues include the specification of the number of trunks between node pairs and the routing strategy used to determine the path through a network in order to achieve a given call blocking probability. When blocked calls are queued, the number of calls that may be queued is also a design question.

A packet-switched communication system exchanges messages among users by transmitting sequences of packets which comprise the messages. That is, the sending terminal equipment partitions a message into a sequence of packets, the packets are transmitted across the network, and the receiving terminal equipment reassembles the packets into messages. The transmission facility interconnecting a given node pair is viewed as a single trunk, and the transmission capacity of this trunk is shared among all users whose packets traverse both nodes. While the trunk capacity is specified in bits per second, the packet handling capacity of a node pair depends both upon the trunk capacity and the nodal processing power.

In many packet-switched networks, the path traversed by a packet through the network is established during a call set-up procedure, and the network is referred to as a virtual circuit packet switching network. Other networks provide datagram service, a service that allows users to transmit individually addressed packets without the need for call set-up. Datagram networks have the advantage of not having to establish connections before communication takes place, but they have the disadvantage that every packet must contain complete addressing information. Virtual circuit networks have the advantage that addressing information is not required in each packet, but have the disadvantage that a call set-up must take place before communication can occur. Datagram is an example of **connectionless service** while virtual circuit is an example of **connection-oriented service.**

Prior to the late 1970s, signaling for circuit establishment was in-band. That is, in order to set up a call through the network, the call set-up information was sent sequentially from switch to switch using the actual circuit that would eventually become the circuit used to connect the end users. In an extreme case, this amounted to trying to find a path through a maze, sometimes having to retrace one's steps before finally emerging at the destination or just simply giving up when no path could be found. This had two negative characteristics: first, the rate of signaling information transfer was limited to the circuit speed, and second, the circuits that could have been used for accomplishing the end objective were being consumed simply to find a path between the end-points. This resulted in tremendous bottlenecks on major holidays, which were solved by virtually disallowing alternate routes through the toll switching network.

An alternate out-of-band signaling system, usually called **common-channel interoffice signaling** (CCIS), was developed primarily to solve this problem. Signaling now takes place over a signaling network that is

partitioned from the network that carries the user traffic. This principle is incorporated into the concept of integrated services digital networks (ISDNs), which is described thoroughly in Helgert [1991]. The basic idea of ISDN is to offer to the user some number of 64-kbps access lines plus a 16-kbps access line through which the user can describe to an ISDN how the user wishes to use each of the 64-kbps circuits at any given time. The channels formed by concatenating the access lines with the network interswitch trunks having the requested characteristics are established using an out-of-band signaling system, the most modern of which is signaling system #7 (SS#7).

In either virtual circuit or datagram networks, packets from a large number of users may simultaneously need transmission services between nodes. Packets arrive at a given node at random times. The switching node determines the next node in the transmission path, and then places the packet in a queue for transmission over a trunk facility to the next node. Packet arrival processes tend to be bursty, that is, the number of packet arrivals over fixed-length intervals of time has a large variance. Because of the burstiness of the arrival process, packets may experience significant delays at the trunks. Queues may also build due to the difference in transmission capacities of the various trunks and access lines. Combining of packets that arrive at random times from different users onto the same line, in this case a trunk, is called statistical multiplexing.

In addition to the delays experienced at the input to trunks, packets may also experience queueing delays within the switching nodes. In particular, the functions required for packet switching are effected by executing various software processes within the nodes, and packets must queue while awaiting execution of the various processes on their behalf.

Both transmission capacities and nodal processing capabilities are available over a wide range of values. If the trunk capacities are relatively low compared to nodal processing capability, then delays at switching nodes may be relatively small. If line capacities are large compared to nodal processing capabilities, however, delays due to nodal processing may be significant. In the general case, all possible sources of delay should be examined to determine where bottlenecks, and consequently delay, occur.

It is often the case that a particular point in the communication network, either a processing node or a trunk, is the primary source of delay. In this case, this point is usually singled out for analysis, and a simple model is invoked to analyze the performance at that point. The results of this analysis, combined with results of other analyses, result in a profile of overall system performance. In this case, the key aspect of the analysis is to choose an appropriate model for the isolated analysis. In this way, a simplified analysis leading to useful results can be performed, and this can lead to an improved network design.

Protocol design and performance issues are frequent topics of discussion at both general conferences in communications and those specialized to networking. The reader is encouraged to consult the proceedings of the conferences mentioned earlier for a better appreciation of the range of issues and the diversity of the proposed solutions to the issues.

Computer Communication Network Architecture

In this section, we will begin with a brief, high-level definition of the ISORM. The reference model has seven layers, none of which can be bypassed conceptually. In general, a layer is defined by the types of services it provides to its users and the quality of those services. For each layer in the ISO/OSI architecture, the user of a layer is the next layer up in the hierarchy, except for the highest layer for which the user is an application. Clearly, when a layered architecture is implemented under this philosophy, then the quality of service obtained by the end user, the application, is a function of the quality of service provided by all of the layers. In order to clarify the communications strategy of the ISO/OSI architecture, we will provide a discussion of the layer 2 services in some detail.

There is significant debate over whether the efforts of the ISO/OSI community are leading to the best standards (or even standards that have any merit whatever!). Limiting our discussion to the ISORM is, by no means, an endorsement of the actual protocols that have been developed in the ISO arena; there are actually more widely deployed and successful standards in other arenas. On the other hand, the ISORM is very useful for discussing network architecture principles, and these principles apply across the board. Thus, we choose to base our discussion on the ISORM.

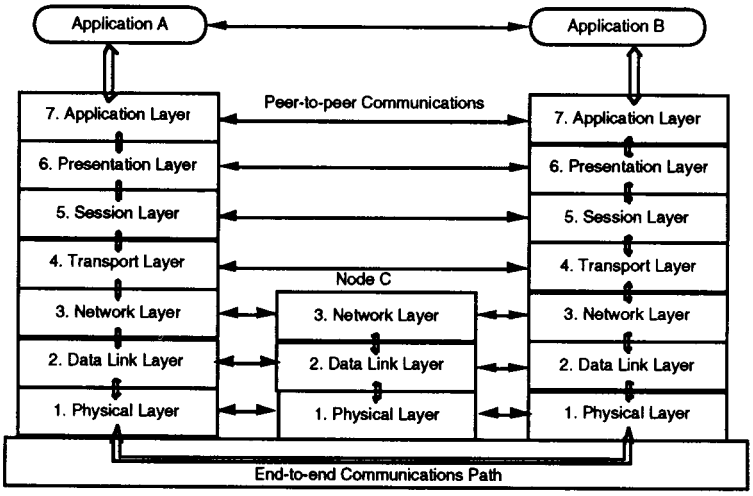


FIGURE 72.6 Layered architecture for ISO/OSI reference model.

Figure 72.6, adopted from Spragins [1991], shows the basic structure of the OSI architecture and how this architecture is envisaged to provide for exchange of information between applications. As shown in the figure, there are seven layers: application, presentation, session, transport, network, data link, and physical. Brief definitions of the layers follow, but the reader should bear in mind that substantial further study will be required to develop an understanding of the practical implications of the definitions:

- *Physical layer:* Provides electrical, functional, and procedural characteristics to activate, maintain, and deactivate physical data links that transparently pass the bit stream for communication between data link **entities.**
- *Data link layer:* Provides functional and procedural means to transfer data between network entities; provides for activation, maintenance, and deactivation of data link connections, character and frame synchronization, grouping of bits into characters and frames, error control, media access control, and flow control.
- *Network layer:* Provides switching and routing functions to establish, maintain, and terminate network layer connections, and transfer data between transport layers.
- *Transport layer:* Provides host-to-host, cost-effective, transparent transfer of data, end-to-end flow control, and end-to-end quality of service as required by applications.
- *Session layer:* Provides mechanisms for organizing and structuring dialogues between application processes.
- *Presentation layer:* Provides for independent data representation and syntax selection by each communicating application and conversion between selected contexts and the internal architecture standard.
- *Application layer:* Provides applications with access to the ISO/OSI communication stack and certain distributed information services.

As we have mentioned previously, a layer is defined by the types of services it provides to its users. In the case of a request or a response, these services are provided via invocation of **service primitives** of the layer in question by the layer that wants the service performed. In the case of an indication or a confirm, these services are provided via invocation of service primitives of the layer in question by the same layer that wants the service performed.

This process is not unlike a user of a programming system calling a subroutine from a scientific subroutine package in order to obtain a service, say, matrix inversion or memory allocation. For example, a request is analogous to a CALL statement in a FORTRAN program, and a response is analogous to the RETURN statement in the subroutine that has been CALLed. The requests for services are generated asynchronously by all of the

users of all of the services and these join (typically prioritized) queues along with other requests and responses while awaiting servicing by the processor or other resource such as a transmission line.

The service primitives fall into four basic types: request, indication, response, and confirm. These types are defined as follows:

- *Request:* A primitive sent by layer $(N + 1)$ to layer N to request a service.
- *Indication:* A primitive sent by layer N to layer $(N + 1)$ to indicate that a service has been requested of layer N by a different layer $(N + 1)$ entity.
- *Response:* A primitive sent by layer $(N + 1)$ to layer N in response to an *indication* primitive.
- *Confirm:* A primitive sent by layer N to layer $(N + 1)$ to indicate that a response to an earlier *request* primitive has been received.

In order to be more specific about how communication takes place, we will now turn to a brief discussion of layer 2, the data link layer. The primitives provided by the ISO data link (DL) layer are as follows [Stallings, 1990a]:

```
DL_CONNECT.request          DL_RESET.request
DL_CONNECT.indication       DL_RESET.indication
DL_CONNECT.response         DL_RESET.response
DL_CONNECT.confirm          DL_RESET.confirm
DL_DATA.request             DL_DISCONNECT.request
DL_DATA.indication          DL_DISCONNECT.indication
DL_DATA.response            DL_UNITDATA.request
DL_DATA.confirm             DL_UNITDATA.indication
```

Each primitive has a set of **formal parameters,** which are analogous to the formal parameters of a procedure in a programming language. For example, the parameters for the DL_CONNECT.request primitive are the Called Address, the Calling Address, and the Quality of Service Parameter Set. The four primitives are used in the establishment of data link connections. The called address and the calling address are analogous to the telephone numbers of two parties of a telephone call, while the quality of service parameter set allows for the negotiation of various agreements such as throughput measured in bits per second.

All four DL_CONNECT primitives are used to establish a data link. An analogy to an ordinary phone call can better illustrate the basic idea of the primitives. The DL_CONNECT.request is equivalent to picking up the phone and dialing. The phone ringing at the called party's end is represented by the DL_CONNECT.indication. DL_CONNECT.response is equivalent to the called party lifting the receiver and answering, and DL_CONNECT.confirm is equivalent to the calling party hearing the response of the called party.

In general, communication takes place between peer layer protocols by the exchange of **protocol data units (PDUs)**, which contain all of the information required for the receiving protocol entity to provide the required service. In order to exchange PDUs, entities at a given layer use the services of the next lower layer. The data link primitives listed above include both connection-mode primitives and connectionless-mode primitives. For connection-mode communications, a connection must be established between two peer entities before they can exchange PDUs.

For example, suppose a network layer entity in Host A wishes to be connected to a network layer entity in Host B, as shown in Fig. 72.6. Then the connection would be accomplished by the concatenation of two data link connections: one between A and C, and one between C and B. In order to establish the connection, the network layer entity in Host A would issue a DL_CONNECT.request to its associated data link entity, providing the required parameters. This data link entity would then transmit this request to a data link entity in C, which would issue a DL_CONNECT.indication to a network entity in C. The network entity in C would then analyze the parameters of the DL_CONNECT.indication and realize that the target destination is B. This network layer entity would then reissue the DL_CONNECT.request to its data link entity, which would transmit the request to a data link entity in B. The data link entity in B would send a DL_CONNECT.indication to a network layer entity in B, and this entity would issue a DL_CONNECT.response back to the data link entity in B. This DL_CONNECT.response would be relayed back to the data link entity in A following the same sequence of events as in the forward path. Eventually, this DL_CONNECT.response would be converted to a

DL_CONNECT.confirm by the data link entity in A and passed to the network entity in A, thus completing the connection.

Once the connection is established, data exchange between the two network layer entities can take place; that is, the entities can exchange PDUs. For example, if a network layer entity in Host A wishes to send a PDU to a network layer entity in Host B, the network layer entity in Host A would issue a DL_DATA.request to the appropriate data link layer entity in Host A. This entity would package the PDU together with appropriate control information into a data link service data unit (DLSDU) and send it to its peer at C. The peer at C would deliver it to the network entity at C, which would forward it to the data link entity in C providing the connection to Host B. This entity would then send the DLSDU to its peer in Host B, and this data link entity would pass the PDU to Host B network entity via a DL_DATA.indication.

Network layer PDUs are called packets and data link layer PDUs are called frames. The data link layer does not know that the information it is transmitting is a packet; to the data link layer entity, the packet is simply user information. From the perspective of a data link entity, it is not necessary to have a network layer. The network layer exists to add value for the user of the network layer to the services provided by the data link layer. In the example above, value was added by the network layer by providing a relaying capability since Hosts A and C were not directly connected. Similarly, the data link layer functions on a hop-by-hop basis, each hop being completely unaware that there are any other hops involved in the communication. We will see later that the data link need not be limited to a single physical connection.

The philosophy of the ISO/OSI architecture is that in addition to the software being layered, implementations are not allowed to bypass entire layers; that is, every layer must appear in the implementation. This approach was developed after the approach defined for the ARPANET project, which is hierarchical, was fully developed. In the hierarchical approach, the layer interfaces are carefully designed, but any number of layers of software can be bypassed by any application (or other higher-layer protocol) that provides the appropriate functionality. These two approaches have been hotly debated for a number of years, but as the years pass, the approaches are actually beginning to look more and more alike for a variety of reasons that will not be discussed here.

The ISO/OSI layered architecture described above would appear to be very rigid, not allowing for any variations in underlying topology or variations in link reliability. However, as we shall see, this is not necessarily the case. As an example, ISO 8348, which developed as a result of the X.25 project, provides only connection-oriented service, and it was originally intended as the only network layer standard for ISO/OSI. However, ISO 8473, or ISO-IP, which is virtually identical to the Department of Defense (DoD) internet protocol (DoD-IP) developed in the ARPANET project, has since been added to the protocol suite to provide connectionless service as well as internet service. An interesting aside is that because of the addressing limitations of DoD-IP, the Internet Administrative Board (IAB) has recently recommended replacement of the DoD-IP protocol by the ISO-IP protocol, thus bringing the process full circle.

The ISO/OSI protocol suite is in a constant state of revision as new experience reveals the need for additional capabilities and flexibility. Some of this additional flexibility and functionality is being provided through the use of so-called **adaption sublayers**, which enhance the capabilities of a given layer so that it can use the services of a lower layer with which it was not specifically designed for compatibility.

Interestingly, the use of adaption sublayers is only a short step away from using adaption layers that would allow applications to directly interface with any ISO layer. This would result in a hierarchical rather than layered architecture; to wit: ISORM becomes DoDRM. Indeed, fundamental changes in the national (and worldwide) communications infrastructure appear to be leading naturally in the hierarchical direction. Of course, the indiscriminate use of such adaptions would lead back to the proliferation of incompatible protocols and interfaces, the frustration that led to the current twenty-year standardization crusade! It is refreshing to note that a return to our former state does not appear to be around the corner; most standardization work is actually headed in the direction of allowing open systems to intercommunicate.

Local-Area Networks and Internets

We will now turn to a discussion of LANs, which have inherent properties that make the use of sublayers particularly attractive. In this section, we will discuss the organization of communications software for LANs. In addition, we will introduce the idea of **internets**, which were brought about to a large extent by the advent

of LANs. We will discuss the types of networks only briefly and refer the reader to the many excellent texts on the subject. Layers 4 and above for local-area communications networks are identical to those of wide-area networks. However, because the hosts communicating over a LAN share a single physical transmission facility, the routing functions provided by the network layer, layer 3, are not necessary. Thus, the functionality of a layer 3 in a LAN can be substantially simplified without loss of utility. On the other hand, a data link layer entity must now manage many simultaneous data link layer connections because all connections entering and leaving a host on a single LAN do so over a single physical link. Thus, in the case of connection-oriented communications, the software must manage several virtual connections over a single physical link.

There were several basic types of transmission schemes in use in early LANs. Three of these received serious consideration for standardization: the **token ring, token bus,** and **carrier-sense multiple access** (CSMA). In a token ring network, the stations are configured on a physical ring around the medium. A token rotates around this physical ring, visiting each host (or station) in turn. A station wishing to transmit data must wait until the token is available to that station. In a token bus LAN, the situation is the same, except that the stations share a common bus and the ring is logical rather than physical. In a CSMA network, the stations are bus connected, and a station may transmit whenever other stations are not currently transmitting. That is, a station wishing to transmit senses the channel, and if there is no activity, the station may transmit. Of course, the actual access protocol is significantly more complicated than this.

In the early 1980s, there was significant debate over which LAN connection arrangement was superior, a single choice being viewed as necessary. This debate centered on such issues as cost, network throughput, network delay, and growth potential. Performance evaluation based on queueing theory played a major role in putting these issues in perspective. For thorough descriptions of LAN protocols and queueing models used to evaluate their performance, the interested reader is referred to Hammond and O'Reilly [1986].

All three of the access methods mentioned above became IEEE standards (IEEE 802) and eventually became ISO standards (ISO 8802 series) because all merited standardization. On the other hand, all existed for the express purpose of exchanging information among peers, and it was recognized at the outset that the upper end of the data link layer could be shared by all three access techniques. On the other hand, the lower-level functions of the layer deal with interfacing to the physical media. Here, drastic differences in the way the protocol had to interface with the media were recognized. Thus, a different **media-access control sublayer** (MAC) was needed for each of the access techniques.

The decision to use a common logical link control (LLC) sublayer for all of the LAN protocols apparently ushered in the idea of adaption sublayers. The reason for splitting the layer is simple: a user of the data link control (DLC) layer need not know what kind of medium provides the communications; all that is necessary is that the user understand the interface to the DLC layer.

On the other hand, the media of the three types of access protocols provide transmission service in different ways, so software is needed to bridge the gap between what the user of the service needs, which is provided by the LLC, and how the LLC uses the media to provide the required service. Thus, the MAC sublayer was born.

This idea has proven to be valuable as new types of technologies have become available. For example, the new fiber-distributed digital interface (FDDI) uses the LLC of all other LAN protocols, but its MAC is completely different from the token ring MAC even though FDDI is a token ring protocol. Reasons for needing a new MAC for LLC are provided in Stallings [1990b].

One of the more interesting consequences of the advent of local-area networking is that many traditional computer communication networks became internets overnight. LAN technology was used to connect stations to a host computer, and these host computers were already on a WAN. It was then a simple matter to provide a relaying, or bridging, service at the host in order to provide wide-area interconnection of stations to LANs to each other. In short, the previously established WANs became networks for interconnection of LANs; that is, they were interconnecting networks rather than stations. Internet performance suddenly became a primary concern in the design of networks.

More recently, FDDI is being thought of as a mechanism to provide LAN interconnection on a site basis, and a new type of network, the **metropolitan-area network** (MAN) has been under study for the interconnection of LANs within a metropolitan area. The primary media configuration for MANs is a dual bus configuration and it is implemented via the distributed queue, dual bus (DQDB) protocol, also known as IEEE 802.6. The net effect of this protocol is to use the dual bus configuration to provide service approaching the

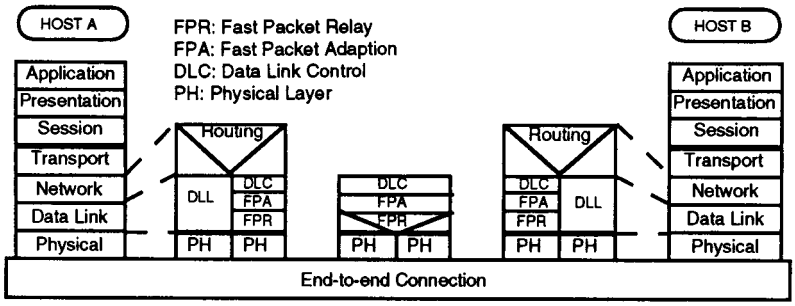


FIGURE 72.7 Fast packet switched layered architecture.

first-come–first-served service discipline to the traffic entering the FDDI network, which is remarkable considering that the LANs being interconnected are geographically dispersed. Interestingly, DQDB concepts have recently been adapted to provide wide-area communications. Specifically, structures have been defined for transmitting DQDB frames over standard DS-1 (1.544 megabits per second [Mb/s]) and DS-3 (6.312 Mb/s) facilities, and these have been used as the basis for a new service offering called switched multimegabit data services (SMDS).

As of this writing, advances in LANs design and new forms of LANs are emerging. One example is wireless LANs, which are LANs in which radio or photonic links serve as cable replacements. Wireless LAN technology is viewed by many as crucial to the evolution of personal communication networks. Another example is the asynchronous transfer mode-based LAN, which is mentioned in the next section following a general discussion of asynchronous transfer mode.

Some Additional Recent Developments

In this subsection, we will describe two recent developments of significant interest in communication networking: **fast packet networks** and transmission using the *asynchronous transfer mode* (ATM), which is a part of the larger *broadband integrated services digital network* (B-ISDN) effort.

As we have mentioned previously, there is really no requirement that the physical media between two adjacent data link layers be composed of a single link. In fact, if a path through the network is initially established between two data link entities, there is no reason that DLC protocols need to be executed at intermediate nodes. Figure 72.7, adapted from Bhargava and Hluchyj [1990] shows how the end-to-end connection might be implemented. A network implemented in the fashion indicated in Fig. 72.7 is called a fast packet network (FPN).

From Fig. 72.7, it is seen that the data link layer is partitioned into three sublayers: the data link control sublayer (which parallels the LLC layer of LANs), the fast packet adaption (FPA) sublayer, and the fast packet relay (FPR) sublayer. The function of the fast packet adaption sublayer is to segment the layer-2 PDU, the frame, into smaller units, called fast packets, for transmission over the FPN. These fast packets contain information that identifies the source and destination node names and the frame to which they belong so that they can be routed through the network and reassembled at the destination.

The fast packets are statistically multiplexed onto a common physical link by the FPR sublayer for transmission. At intermediate nodes, minor error checking, fast packet framing, fast packet switching, and queueing takes place. If errors are found, then the fast packet is dropped. When the fast packets reach their destination, they are reassembled into a frame by the FPA sublayer and passed on to the DLC sublayer where normal DLC functions are performed.

The motivation for FPNs is that since link transmission is becoming more reliable, extensive error checking and flow control are not needed across individual links; an end-to-end check should be sufficient. Meanwhile, the savings in processing due to not processing at the network layer can be applied to frame processing, which allows interconnection of the switches at higher line speeds.

Since bits-per-second costs decrease with increased line speed, service providers can offer savings to their customers through FPNs. Significant issues are fast packet loss probability and retransmission delay. Such factors will determine the retransmission strategy deployed in the network. Of course, the goal is to improve

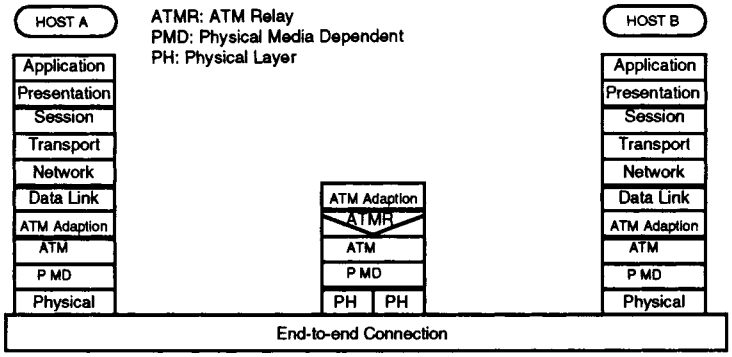


FIGURE 72.8 Asynchronous transfer mode layered architecture.

network efficiency, so a significant issue is whether FPNs are better than ordinary packet networks and, if so, by how much.

Another recent innovation is the ATM, usually associated with B-ISDN. The idea of ATM is to partition a user's data into many small segments, called cells, for transmission over the network. Independent of the data's origin, the cell size is 53 octets, of which 5 octets are for use by the network itself for routing and error control. Users of the ATM are responsible for segmentation and reassembly of their data. Any control information required for this purpose must be included in the 48 octets of user information in each cell. In the usual case, these cells would be transmitted over networks that would provide users with 135 Mb/s and above data transmission capacity (with user overhead included in the capacity).

The segmentation of units of data into cells introduces tremendous flexibility for handling different types of information, such as voice, data, image, and video, over a single transmission facility. As a result, LANs, WANs, and MANs based on the ATM paradigm are being designed, and indeed deployed. A significant portion of the deployment activity is a national testbed program, which involves industrial/academia cooperation, under joint sponsorship of the National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA). There is also significant private investment in developing this technology; for example, experimental ATM-based LANs are already in place at the Digital Equipment Corporation (DEC) research facility in Palo Alto, California. There is some possibility that LANs of this type, rather than of the FDDI type, will be the dominant means of providing high-speed LAN and LAN-interconnect services.

There are numerous possibilities for connection of hosts to ATM networks, but they all share a common architecture, which consists of three sublayers: the ATM adaption layer (AAL), the ATM layer, and the physical media-dependent (PMD) layer. Connection of hosts to ATM at a given layer is achieved through developing an AAL for the layer in question. For example, one might decide to adapt to ATM at the network layer. In that case, the transport layer would operate as usual, and the AAL would be designed to process data structures from the transport layer to produce data structures for use by the ATM layer and vice versa. Of course, all hosts communicating with each other in this way would use the same AAL.

Figure 72.8 shows an example of how an ISO/OSI host might connect to an ATM network. Below the data link layer is the ATM adaption layer (AAL), which provides for call control across the ATM network and for segmentation and reassembly of frames from the data link layer. The current estimate for the amount of overhead needed per cell for AAL purposes is 4 octets, leaving 44 octets for user information.

At the present time, end-to-end connections at the ATM level are expected to be connection oriented. As cells traverse the network, they are switched on a one-by-one basis, using information contained in the five ATM overhead octets to follow the virtual path established during the ATM call set-up. Typically, cells outbound on a common link are statistically multiplexed, and if buffers are full, cells are dropped. In addition, if one or more errors are found in a cell, then the cell is dropped.

In the case of data transmission, a lost cell will result in an unusable frame unless the data is encoded to guard against cell loss prior to transmission. Coding might be provided by the AAL, for example. The trade-offs involved in coding and retransmission and their impact upon network throughput, delay and complexity are not well understood at the time of this writing. Part of the reason for this is that cell loss probability and

the types of traffic that are likely to use the network are not thoroughly understood. Resolution of these issues accounts for a significant portion of the research activity in computer communication networking at this time. The relevant American National Standards Institute and CCITT documents are frequently updated to include the results.

Defining Terms

Access line: A communication line that connects a user's terminal equipment to a switching node.

Adaption sublayer: Software that is added between two protocol layers to allow the upper layer to take advantage of the services offered by the lower layer in situations where the upper layer is not specifically designed to interface directly to the lower layer.

Architecture: The set of protocols defining a computer communication network.

Asynchronous transfer mode (ATM): A mode of communication in which communication takes place through the exchange of tiny units of information called cells.

Broadband integrated services digital network (B-ISDN): A generic term that generally refers to the future network infrastructure that will provide ubiquitous availability of integrated voice, data, imagery, and video services.

Carrier-sense multiple access: A random-access method of sharing a bus-type communications medium in which a potential user of the medium listens before beginning to transmit.

Circuit switching: A method of communication in which a physical circuit is established between two terminating equipments before communication begins to take place. This is analogous to an ordinary phone call.

Common-channel interoffice signaling: The use of a special network, dedicated to signaling, to establish a path through a communication network, which is dedicated to the transfer of user information.

Computer communication network: Collection of applications hosted on different machines and interconnected by an infrastructure that provides intercommunications.

Connection-oriented service: A mode of packet switching in which a call is established prior to any information exchange taking place. This is analogous to an ordinary phone call, except that no physical resources need to be allocated.

Connectionless service: A mode of packet switching in which packets are exchanged without first establishing a connection. Conceptually, this is very close to message switching, except that if the destination node is not active, then the packet is lost.

Entity: A software process that implements a part of a protocol in a computer communication network.

Fast packet networks: Networks in which packets are transferred by switching at the frame layer rather than the packet layer. Such networks are sometimes called frame relay networks. At this time, it is becoming in vogue to think of frame relay as a service, rather than transmission, technology.

Formal parameters: The parameters passed during the invocation of a service primitive; similar to the arguments passed in a subroutine call in a computer program.

International Standards Organization reference model: A model, established by ISO, that organizes the functions required by a complete communication network into seven layers.

Internet: A network formed by the interconnection of networks.

Local-area network: A computer communication network spanning a limited geographic area, such as a building or college campus.

Media-access control: A sublayer of the link layer protocol whose implementation is specific to the type of physical medium over which communication takes place and which controls access to that medium.

Message switching: A service-oriented class of communication in which messages are exchanged among terminating equipments by traversing a set of switching nodes in a store-and-forward manner. This is analogous to an ordinary postal system. The destination terminal need not be active at the same time as the originator in order that the message exchange take place.

Metropolitan-area network: A computer communication network spanning a limited geographic area, such as a city; sometimes features interconnection of LANs.

Packet switching: A method of communication in which messages are exchanged between terminating equipments via the exchange of a sequence of fragments of the message called packets.

Protocol data unit (PDU): The unit of exchange of protocol information between entities. Typically, a PDU is analogous to a structure in C or a record in Pascal; the protocol is executed by processing a sequence of PDUs.

Service primitive: The name of a procedure that provides a service; similar to the name of a subroutine or procedure in a scientific subroutine library.

Switching node: A computer or computing equipment that provides access to networking services.

Token bus: A method of sharing a bus-type communications medium that uses a token to schedule access to the medium. When a particular station has completed its use of the token, it broadcasts the token on the bus, and the station to which it is addressed takes control of the medium.

Token ring: A method of sharing a ring-type communications medium that uses a token to schedule access to the medium. When a particular station has completed its use of the token, it transmits the token on the ring, and the station that is physically next on the ring takes control.

Trunk: A communication line between two switching nodes.

Wide-area network: A computer communication network spanning a broad geographic area, such as a state or country.

Related Topics

71.3 Protonic Networks • 72.1 B-ISDN

References

D. Bertsekas and R. Gallagher, *Data Networks,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

A. Bhargava and M. G. Hluchyj, "Frame losses due to buffer overflow in fast packet networks," *Proc. IEEE INFOCOM '90*, San Francisco, 1990, pp. 132–139.

U. Black, *Computer Networks, Protocols and Standards,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1994.

U. Black, *TCP/IP and Retated Protocols*, New York: McGraw-Hill, 1992.

J. L. Hammond and P. J. P. O'Reilly, *Performance Analysis of Local Computer Networks,* Reading, Mass.: Addison-Wesley, 1986.

S. Haykin, *Communication Systems,* 3rd ed., New York: Wiley, 1994.

H. J. Helgert, *Integrated Services Digital Networks.* Reading, Mass.: Addison-Wesley, 1991.

M. Rose, *The Open Book: A Practical Perspective on OSF;* Englewood Cliffs; N.J.: Prentice-Hall, 1990.

M. Schwartz, *Telecommunications Networks: Protocols, Modeling and Analysis,* Reading, Mass.: Addison-Wesley, 1987.

J. D. Spragins, *Telecommunications: Protocols and Design,* Reading, Mass.: Addison-Wesley, 1991.

W. Stallings, *Handbook of Computer-Communications Standards: The Open Systems Interconnection (OSI) Model and OSI-Related Standards,* New York: Macmillan, 1990a.

W. Stallings, *Handbook of Computer-Communications Standards: Local Network Standards,* New York: Macmillan, 1990b.

W. Stallings, *Handbook of Computer-Communications Standards: Department of Defense (DOD) Protocol Standards,* New York: Macmillan, 1990c.

A. S. Tanenbaum, *Computer Networks,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1988.

M. E. Woodword, *Communication and Computer Networks,* IEEE Press, 1993.

Further Information

There are many conferences and workshops that provide up-to-date coverage in the computer communications area. Among these are the IEEE INFOCOM and ACM SIGCOMM conferences and the IEEE Computer Communications Workshop, which specialize in computer communications and are held annually. In addition, IEEE GLOBCOM (annual), IEEE ICC (annual), IFIPS ICCC (biannual), and the International Telecommunications Congress (biannual) regularly feature a substantial number of paper and panel sessions in networking.

The *ACM Communications Review,* a quarterly, specializes in computer communications and often presents summaries of the latest standards activities. *IEEE Network,* a bimonthly, contains tutorial articles on all aspects

of computer communications and includes a regular column on books related to the discipline. Additionally, *IEEE Communications* and *IEEE Computer,* monthly magazines, frequently have articles on specific aspects of networking.

For those who wish to be involved in the most up-to-date activities, there are many interest groups on the Internet, a worldwide TCP/IP-based network, that specialize in some aspect of networking. *The User's Directory of Computer Networks* (Digital Press, T. L. LaQuey, Ed.) provides an excellent introduction to the activities surrounding internetworking and how to obtain timely information.

72.3 Local-Area Networks

Joseph Bannister and Mario Gerla

The local-area network (LAN) is a communication network that interconnects computers and computer-based devices, such as file servers, printers, and graphics terminals. The LAN is characterized as being contained completely within the premises of a single business entity—which almost always owns and operates the network—and this distinguishes the LAN from public-domain networks such as metropolitan- or wide-area networks. The LAN, then, is normally restricted to a few hundred stations (i.e., devices that attach directly to the LAN) that span a limited geographical area, so that no two connected stations are separated by a distance of more than a few kilometers. Moreover, the LAN can be distinguished from the computer or backplane bus, which interconnects components, boards, or devices that comprise a single computer. The LAN uses serial—rather than parallel—transmission, which also differentiates it from the computer bus. In contrast to today's wide-area networks, information is transmitted over LANs at high speeds and with very low error rates.

The LAN often employs fully broadcast media, or **physical media** that allow each station's transmissions to be received by all other stations. Thus, a broadcast capability is often an integral feature of the LAN. Frequently, the LAN also provides for multicasting, a generalization of broadcasting in which a specified subset of stations receives a transmission.

LANs are based on a variety of technologies that include twisted copper-wire pairs, coaxial cable, optical fibers, wireless infrared and radio for signal transport, as well as several integrated circuit families for transmitters, receivers, and the implementation of low-level protocols.

The **topology** of a LAN refers to the physical layout of the transmission media and the logical arrangement of the stations on those media. Four topologies are commonly used in LANs: the bus, ring, star, and tree topologies, which are illustrated in Fig. 72.9.

The LAN Service Model

Within the scope of the well-known Open Systems Interconnection seven-layer reference model, the LAN occupies the two bottom layers, namely, the physical and data link layers, as shown in Fig. 72.10. The physical layer specifies the most primitive services of the LAN, e.g., media characteristics, signal formats, waveforms, signaling rates, timing, and mechanical aspects of connectors, etc. The data link layer uses the services of the physical layer to provide multiple access for stations sharing the media. Station or network management, which is shown as a vertical "layer" in Fig. 72.10, is responsible for maintaining a necessary level of performance, fault detection and recovery, and configuration and security functions.

The Physical Layer

Since the physical layer provides the most primitive services to LAN users, this layer is most closely associated with the implementation technology of the LAN. At a fundamental level the transmission media can be either electrical or optical waveguides. The physical media can be laid out as one of those topologies illustrated in Fig. 72.9. However, certain media are better suited to some topologies, e.g., the bus is frequently used with electrical but not with optical media, because there is high insertion loss associated with taps in the latter.

Signaling is also a critical element of the physical layer. Baseband modulation, in which digital signals (0s and 1s) are transmitted as electrical or optical pulses, is common in LANs because of its simplicity. Modulation of carriers is less common but especially important when several independent channels are employed. Amplitude,

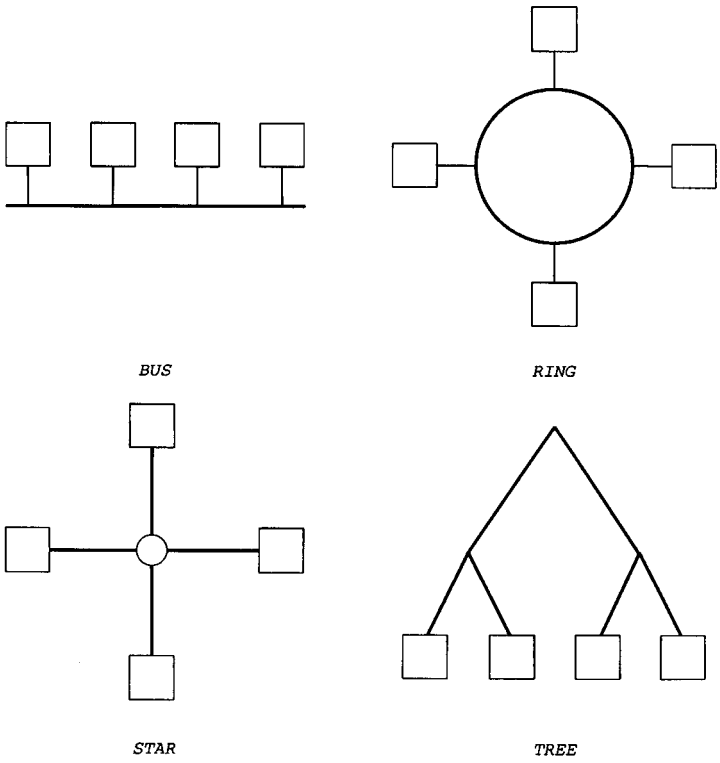


FIGURE 72.9 LAN topologies.

frequency, and phase modulation have been used in community antenna television (CATV) systems to support multichannel LANs. Coherent lightwave systems, although still largely experimental, are expected to increase in importance because they permit multiplexing a large number of channels over a single-mode optical fiber. Also of increasing importance is atmospheric propagation of electromagnetic signals. The growing demand for mobile communication and ubiquitous computing is driving the development of the personal communication network, which is to be based on code-division multiple access.

The signaling rates and formats are also part of the physical layer specification. Electrical media generally use Manchester baseband encoding, which has a 50% duty cycle and operates at rates below 100 Mb/s. Optical media often use the so-called 4B/5B intensity-modulation encoding, which achieves 80% efficiency by representing 27 distinct symbols (of which 16 are data and 11 are control symbols) by five bits in such a manner that four consecutive 0s (i.e., low-light power levels) should never occur. Similarly, 8B/6T encoding is sometimes used with electrical media to encode an octet of data as six ternary digits, achieving an efficiency of 75%.

Connector and cable-plant technology is another critical element of the physical layer. Thorough characterization of the transmission media is required if users are to interoperate with each other. The type of cable—e.g., shielded or unshielded twisted copper-wire pairs, coaxial cable, and single-mode or multimode optical fiber—must be specified. Furthermore, the connectors between stations and the cable plant are defined as part of the physical layer. Stations can attach via passive taps or can actively repeat signals; in the latter case a bypass switch is usually provided as the station's interface to the cable plant.

The Data Link Layer

The data link layer is often divided into two sublayers, i.e., the media-access and logical-link control (MAC and LLC) sublayers, as shown in Fig. 72.10. The LLC sublayer [see *Logical Link Control*] uses the services of the MAC sublayer to provide to its user a connection-oriented service between stations that includes flow and error control or a connectionless service that does little more than multiplex upper-layer connections. The

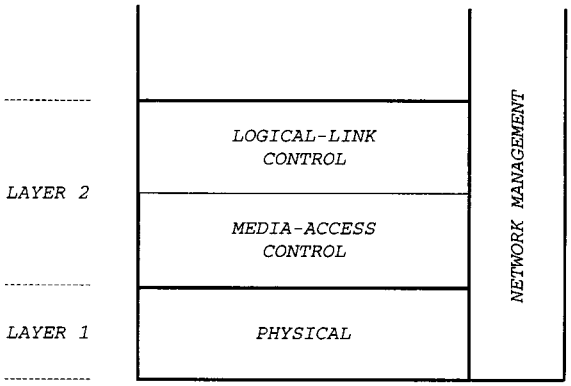


FIGURE 72.10 LANservice model.

connection-oriented LLC protocol gives the service user the illusion of having a dedicated point-to-point link between a pair of communicating stations.

The MAC sublayer specifies the **media access protocol** that stations use to share the media. In fully broadcast media no more than one station may transmit at a time, so the MAC sublayer manages exclusive access to the broadcast media. The ring topology is well suited to a token-passing MAC protocol, which gives transmission rights to the station holding the token. The token is represented by a special packet that is passed sequentially from station to station. When a station recognizes the token, it seizes it and begins transmitting buffered packets, or passes the token to the next station if it has no packet to transmit. To limit the time that a station can hold the token, the MAC protocol can implement one of several disciplines:

- One-shot service, in which the station releases the token when it has transmitted one packet
- Exhaustive service, in which the station releases the token when it has no more packets to transmit
- Gated service, in which the station releases the token when all packets that were buffered at token-acquisition time have been transmitted
- Token-timing service, in which the station releases the token at the expiration of a timer

The IEEE 802.5 token ring standard specifies a token-timing service discipline that requires transmissions to be completed within a fixed time after the token is seized, but implementations sometimes use the simpler one-shot service discipline. The ANSI X3T9.5 fiber distributed data interface (FDDI) standard uses an adaptive token-timing service discipline that is intended to guarantee a minimum amount of (synchronous) bandwidth to each station.

A variation of the token ring protocol is the token bus protocol, which allows the token to be passed in a specified order. In the token bus protocol, which is often used with the bus or tree topologies, a station broadcasts the token, specifying the successor station in an address field of the token packet. Although all stations receive the token, only the addressed successor station can seize it.

A MAC scheme that is widely used with the bus topology is carrier-sense multiple access (CSMA). A contention protocol, CSMA operates by allowing any station to transmit a buffered packet if it senses that the bus is idle. If two stations are ready to transmit their packets at nearly the same time, they will both sense that the bus is idle and their transmissions will collide, i.e., the superimposed bits of the packets will be garbled. The propagation delay—or time it takes for the packet to travel from one station to the other—dictates the window of vulnerability for CSMA; the larger the window, the more collisions are likely. To overcome the problem of collisions, CSMA is often enhanced with collision detection (CSMA/CD) by enabling stations to monitor their transmissions for the garbled bits associated with collisions. When a collision is detected, the station aborts its transmission and reschedules it by backing off for a period of time. The binary exponential backoff algorithm specifies that the random backoff time is drawn uniformly from the interval between 0 and $2^n - 1$ time units, where n is the number of times the packet has collided.

A time-slotted bus maintains on the bus a continuous stream of short, fixed-length frames that are initially empty but can be filled with data as they pass stations with waiting packets. The distributed queue, dual bus (DQDB) local- and metropolitan-area network uses two-directional buses so that a station can reserve on its downstream bus a frame for its upstream-destined packets.

In the star topology stations are homed into a central hub which can manage their access to the media. Active hubs physically control media access, while passive hubs merely broadcast incoming packets to specific output ports. Linear combiner/dividers based on lithium niobate technology allow incoming optical signals to be combined and distributed to output ports according to electronically programmed combining and dividing ratios. A common scheme is to use time-division multiplexing with the star topology. The hub can serve as the central controller, allocating time slots to individual stations, or reservations can be used in the manner of a satellite-based network.

The Management Layer

LAN-specific network management functions are referred to as station management. Station management covers five areas—configuration, performance, fault, accounting, and security management.

Monitoring and controlling the LAN are essential elements of station management. By monitoring the media, stations maintain a record of important measurements, such as the number of a specific type of packet transmitted or received, the number of different kinds of errors, and the source addresses of received packets. Such measurements are made available to an application in the station or to a management center. Thus are applications able to monitor and collect, correlate, and act upon key LAN statistics. Likewise, designated applications are able to effect changes in the LAN by writing to specific variables within stations, which collectively comprise the so-called management information base. For instance, station management informs the MAC sublayer of its unique LAN address by writing the value to a special MAC register.

Some management functions are distributed across the LAN and are implemented at a low level. To recover after the failure of a dual-fiber cable, stations automatically enter into a procedure to reconfigure around the failure and reestablish connectivity. Although such procedures can be viewed as station management, they are sometimes specified as part of the MAC sublayer, because they are so tightly integrated with media access.

Other Features

The basic features of media access are often augmented to provide specialized services and features.

Specialized LAN Services

LAN users have special communication requirements that must be supported by the physical and data link layers. In particular, the MAC sublayer is responsible for providing specialized services. Although all MAC sublayers support asynchronous traffic by providing for the simple, best-effort delivery of packets, some MAC sublayers also support other classes of traffic. To synchronous traffic, which requires a set amount of preallocated bandwidth, the properly designed MAC sublayer guarantees a maximum packet response time. The adaptively timed token-passing protocol of FDDI is capable of supporting synchronous traffic, i.e., at token-capture time the station has a fixed amount of time during which to transmit synchronous packets, and the token is guaranteed to return within a certain amount of time. Isochronous traffic, which requires a fixed amount of traffic to be periodically delivered, is also accommodated by some MAC sublayers. DQDB uses preallocated time slots to provide isochronous service.

Priorities are also important in LANs. Therefore many LANs transmit queued packets in accordance with priorities assigned to the packets. Prioritization can be on a LAN-wide basis or merely within the station. Most LANs offer some method for prioritizing the transmission of packets.

Reliability and Availability

Being a shared resource, the LAN should have a high degree of reliability and availability. The media should not be a single point of failure, and no individual station should be able to prevent—maliciously or otherwise—the delivery of service to other stations. LANs are designed to withstand both transient and permanent failures of the media and stations.

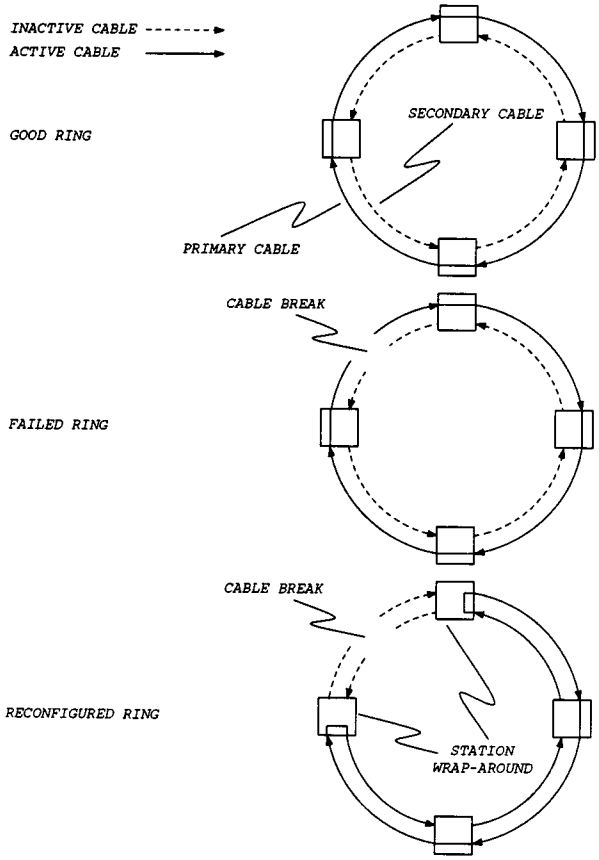


FIGURE 72.11 Reconfiguration of dual counter-rotating rings.

Transmitted information is subject to short bursts of errors and must also be protected. The connection-oriented service at the LLC sublayer is intended to recover from errors—such as garbled, dropped, or out-of-sequence packets—by positively acknowledging packets and retransmitting packets not acknowledged within the timeout window. The MAC sublayer usually provides error-detecting codes that can recognize an error burst of several consecutive bits (a favorite is the 32-bit cyclic redundancy code, which is easily implemented as a linear feedback shift register). Errors can also be recognized at the physical layer when they cause code violations, e.g., the absence of transitions in the Manchester or 4B/5B codes. Some LANs even use error-correcting codes for protecting time-sensitive information.

Other protection mechanisms are used to tolerate cable breaks and station malfunctions. The use of fully broadcast media makes a LAN vulnerable to media failure, since this effectively partitions the stations into noncommunicating groups. To cope with this problem, redundant cables are provided and a mechanism for reconfiguring from the bad to the good cable is built into the LAN protocols. A popular approach for the token ring can be seen in the counter-rotating dual-ring scheme, which is illustrated in Fig. 72.11. If a cable segment or an active station fails, the stations adjacent to the failure can reconfigure the ring by executing "wrap-around" operations. The new configuration uses the spare cable in conjunction with the original cable to form a new ring. Given the complexity of such a reconfiguration procedure, it is usually necessary for station management to coordinate the actions of the stations.

Special mechanisms for adding and removing stations to and from the LAN might also be required. Since the physical addition or removal of a station can disrupt the transmission of data, protocols for reestablishing a lost token could also be necessary.

TABLE 72.1 Characteristics of Standard LANs

	CSMA/CD	Token Ring	Token Bus	FDDI	DQDB
Standard	IEEE 802.3	IEEE 802.5	IEEE 802.4	ANS X3T9.5	IEEE 802.6
Topology	Bus, tree, star	Ring	Tree	Ring	Pseudobus
Media	Coax, UTP, MMF	STP	Coax	MMF, SMF	SMF
Encoding	MC, FSK, AM/PSK 4B/5B, 8B/6T	DMC	FSK, AM/PSK	4B/5B	4B/5B
Data rate	10 Mb/s, 100 Mb/s	4 Mb/s 16 Mb/s	1 Mb/s 5 Mb/s 10 Mb/s	100 Mb/s	34 Mb/s 45 Mb/s 140 Mb/s 155 Mb/s
Features		Priorities	Priorities, ST, multichannel	Priorities, ST, dual ring	Priorities, IT, dual bus

AM/PSK = amplitude modulation/phase-shift keying
ANS = American National Standard
CSMA/CD = carrier-sense multiple-access with collision detection
DMC = differential Manchester coding
DQDB = distributed queue, dual bus
FDDI = fiber distributed data interface
FSK = frequency-shift keying
IEEE = Institute of Electrical and Electronics Engineers
IT = isochronous traffic
MC = Manchester coding
MMF = multimode fiber
SMF = single-mode fiber
ST = synchronous traffic
STP = shielded twisted pair
UTP = unshielded twisted pair
*n*B/*m*B = *n*-bit/*m*-bit
*n*B/*m*T = *n*-bit/*m*-trit

The Importance of LAN Standards

LAN standards play a central role in promoting the goal of universal connectivity among a community of users. The standardization of communication services and protocols allows all conforming implementations to exchange information. Consequently, the importance of LAN standards has grown steadily. Currently, several LAN standards have been established to support the different communication requirements of users.

The first LANs—developed in the early 1970s—were proprietary products meant to interconnect one vendor's computer products. By 1980, however, Project 802 of the Institute of Electrical and Electronics Engineers (IEEE) had recognized the need for publicly disseminated LAN standards and eventually published a specification of the CSMA/CD protocol that any vendor may implement. Furthermore, the definition of the standard was sanctioned by companies that participated in the IEEE Working Group's balloting process, so that the standard was viewed as an open, nonproprietary solution. The IEEE 802.3 Working Group chose a protocol that was based closely on the Ethernet LAN originally developed at Xerox by Robert Metcalfe and David Boggs.

The IEEE Project 802 has broadened its scope to encompass other LAN standards. These include the following:

- 802.3: The CSMA/CD protocols for baseband coaxial cable (10Base5 and 10Base2), unshielded twisted copper-wire pairs (10BaseT), broadband coaxial cable (10Broad36), and optical fiber (10BaseF)
- 802.4: The token bus protocol for multichannel broadband coaxial cable
- 802.5: The token ring protocol for shielded twisted copper-wire pairs
- 802.6: The DQDB protocol for redundant optical fibers

Other standards-making bodies, such as the American National Standards Institute (ANSI) and the International Organization for Standards/International Electrotechnical Committee (ISO/IEC) have developed or cross-adopted LAN standards. ANSI's X3T9.5 committee is responsible for the FDDI LAN standard, a high-speed token ring that uses redundant optical fibers. Some of the important LAN standards and their characteristics are shown in Table 72.1.

The trend is for vendors to market LAN products that conform to specific standards. However, proprietary networks have been successfully marketed and were instrumental in the development of LAN standards. Some of the better known proprietary-LAN product offerings were the Xerox Ethernet, Datapoint Arcnet, Network Systems Hyperchannel, Proteon Pronet, Sytek System 20, and AT&T DATAKIT.

Summary

The LAN is the preferred method for connecting computers within a customer's premises. A number of transmission media, topologies, data rates, and services are available to meet users' needs. The services offered by the LAN are used to implement higher-layer protocols that are required by distributed computing systems. LANs will continue to grow more capable in the data and bit-error rates they achieve, the functionality they provide, and the number and geographical span of the stations they support.

Defining Terms

Media-access protocol: The protocol that permits one of a group of contending stations to access the media exclusively. Media-access protocols are generally based on token passing or carrier sense.

Physical media: The communication channel over which signals are transmitted. Broadcast media, in which all stations receive each transmission, are primarily used in local-area networks. Common media are optical fibers, coaxial cable, twisted copper-wire pairs, and airwaves.

Topology: The paths and switches of a local-area network that provide the physical interconnection among stations. The most common topologies are the bus, ring, tree, and star.

Related Topics

72.2 Computer Communication Networks • 75.3 Wireless Local-Area Networks for the 1990s

References

American National Standard for Information Systems—Fiber Distributed Data Interface (FDDI), ANSI Standards X3.139, X3.148, X3.166, X3.184.

Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications, ANSI/IEEE Standard 802.3, ISO/IEC Standard 8802/3.

Distributed Queue Dual Bus (DQDB) Metropolitan Area Network (MAN), Proposed IEEE Standard 802.6.

Logical Link Control, ANSI/IEEE Standard 802.2, ISO/IEC Standard 8802/2.

Token-Passing Bus Access Method and Physical Layer Specifications, ANSI/IEEE Standard 802.4, ISO/IEC Standard 8802/4.

Token Ring Access Method and Physical Layer Specifications, ANSI/IEEE Standard 802.5, ISO/IEC Standard 8802/5.

Further Information

A popular, frequently updated textbook on LANs is W. Stallings, *Local Networks,* 3rd ed., New York: Macmillan, 1990.

Leading journals that publish research articles on LANs include:

- IEEE Transactions on Communications
- IEEE Transactions on Networking
- Computer Networks and ISDN Systems
- IEEE Network

Four annual conferences that cover the topic of LANs are:

- The IEEE INFOCOM Conference on Computer Communications
- The IEEE Conference on Local Computer Networks
- The ACM SIGCOMM Conference on Communications Architectures and Protocols
- The EFOC/LAN European Fibre Optic Communications and Local Area Networks Conference

72.4 The Intelligent Network

Richard B. Robrock II

The term *intelligent network* refers to the concept of deploying centralized databases in the telecommunications network and querying those databases to provide a wide variety of network services such as 800 Service (toll-free service) and credit card calling. The first use of these centralized databases was in AT&T's network in 1981 where they were used to facilitate the setup of telephone calls charged to a Calling Card. Today such databases are widely deployed throughout North America and support the handling of close to 100 billion telephone calls per year.

The words *intelligent network,* when first used, had a relatively narrow definition, but that definition has broadened considerably with the introduction of the advanced intelligent network, the wireless intelligent network, and soon, the broadband intelligent network. The advanced intelligent network has introduced powerful service creation tools which have empowered network providers to create their own network services. The network providers, in turn, are beginning to broaden the participation in service creation by allowing their customers or third parties to use these tools to create services. The result has been a rapid growth in new network services.

A History of Intelligence in the Network

The first "intelligence" in the telephone network took the form of rows of human telephone operators, sitting side by side, plugging cords into jacks to facilitate the handling of calls. These operators established calls to far-away points, selected the best routes and provided billing information. They were also an information source—providing time or weather or perhaps disseminating the local news. Moreover, they had the opportunity to demonstrate a kind of heroism—gathering volunteers to save a house from fire, helping to catch a prowler, locating a lost child, and on and on. In the early years of telephony, the feats of the telephone operator were indeed legendary.

In the 1920s, however, technology became available that allowed automatic switching of telephone calls through the use of sophisticated electromechanical switching systems. Initially, these switches served as an aid to operators; ultimately, they led to the replacement of operators. The combination of the rotary telephone dial and the electromechanical switch allowed customers to directly dial calls without the assistance of operators. This led to a reduction of human intelligence in the network.

Another dramatic change took place in the telephone network in 1965; it was called software. It came with the marriage of the computer and the telephone switching system in the first stored-program control switch. With the introduction of switching software came a family of Custom Calling services (speed calling, call waiting, call forwarding, and three-way calling) for residential customers, and a robust set of Centrex features (station attendant, call transfer, abbreviated dialing, etc.) for business customers. The first software programs for these stored-program control switches contained approximately 100,000 lines of code; by 1990 some of these switching systems became enormously complex, containing 10 million lines of code and offering hundreds of different services to telephone users.

During the 1980s, a new architectural concept was introduced; it came to be called the intelligent network. It allowed new telecommunications services to be introduced rapidly and in a ubiquitous and uniform fashion. Feature and service availability in the network ceased to be solely dependent upon the hardware and software in stored-program control switches. Rather some new intelligence was centralized in databases which were accessed using packet switching techniques. Most significantly, the intelligent network started to provide some of the capabilities that operators had made available in the early years of telephony. The remaining sections of this chapter describe the intelligent network, its characteristics, and its services. They also provide a description of the advanced intelligent network, which dramatically broadens the participation in the creation of new services.

The Intelligent Network

The intelligent network architecture is illustrated in Fig. 72.12; its primary elements are a switching system, a signaling network, a centralized database, and an operations support system which supports the database. The

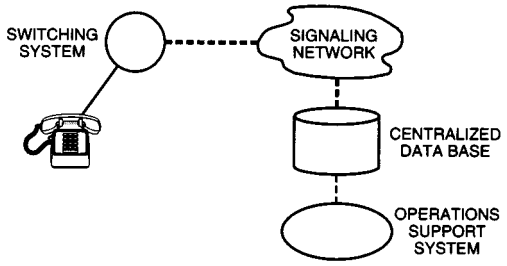


FIGURE 72.12 Intelligent network architecture—telephone calls which require special handling are intercepted in a switching system which launches queries through a signaling network to a centralized database. (*Source:* R. B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1, pp. 7–20, January 1991. © 1991 IEEE.)

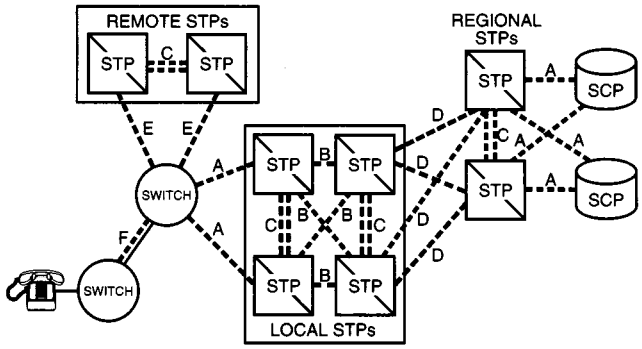


FIGURE 72.13 Link arrangements in a CCS7 signaling network. (*Source:* R. B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1, pp. 7–20, January 1991. © 1991 IEEE.)

architectural concept is a simple one. When a customer places a telephone call which requires special handling, such as a toll-free call (800 Service) or credit card call, that call is intercepted by the switching system which suspends call processing while it launches a query through a signaling network to a centralized database. The database, in turn, retrieves the necessary information to handle the call and returns that information through the signaling network to the switch so that the call can be completed. The role of the operations support system is to administer the appropriate network and customer information that resides in the database.

It is conceivable that the database in this architecture could reside in the switching system, and the signaling network in this instance would not be required. However, that would magnify the task of administering the customer information, since that information would be contained in thousands of switches instead of dozens of centralized databases. In addition, even more importantly, there are two shortcomings associated with basing many of the potential new services in switches, rather than utilizing centralized databases to provide information for the switches. The first is a deployment problem. As of 1990 there were more than 15,000 switches in the United States, and a single switch can cost millions of dollars. To introduce a new service in local switches and to make it widely available generally requires some not-so-simple changes in those switches or, in some cases, replacement of certain switch types altogether. These switch modifications typically take years to implement and require a tremendous capital investment. As a result, ten years after introduction, Custom Calling services were available to fewer than 1% of the residential customers in the United States.

A second problem with switch-based services has been that a single service sometimes behaves differently in different switch types. For example, the speed calling access patterns are different in various stored-program control switches. The public is not particularly sensitive to this fact, because speed calling is not associated with an individual but rather an individual's station set. People live in a mobile society, however, and they want to have their services available from any station set and have them behave the same from any station set.

The intelligent network architecture has been the key to solving both the deployment problem and service uniformity problem associated with switch-based services. Services deployed using an intelligent network centralized database are immediately ubiquitous and uniform throughout a company's serving area.

Intelligent Network Systems

In 1981, AT&T introduced into the Bell System a set of centralized databases called network control points; they supported two applications—the Billing Validation Application for Calling Card Service (credit card calling) and the INWATS database used to support 800 Service. Queries were launched to these databases through AT&T's common-channel interoffice signaling (CCIS) network.

In 1984, following the divestiture of the Regional Bell Operating Companies from AT&T, the regional companies began planning to deploy their own **common-channel signaling (CCS)** networks and their own centralized databases. They selected the **signaling system 7** protocol for use in their signaling networks, called CCS7 networks, and they named their databases **service control points (SCPs).**

The CCS7 Network

A general architecture for a regional signaling network is shown in Fig. 72.13. The network is made up of **signal transfer points (STPs),** which are very reliable, high-capacity packet switches that route signaling messages between network access nodes such as switches and SCPs. To perform these routing functions, the STPs each possess a large routing database containing translation data.

The CCS7 network in Fig. 72.13 contains both local STPs and regional STPs. The STPs are typically deployed in geographically separated pairs so that in the event of a natural disaster at one site, such as an earthquake, flood, or fire, the total traffic volume can be handled by the second site. Indeed, redundancy is provided at all key points so that no single failure can isolate a node.

As illustrated in Fig. 72.13, the following link types have been designated:

- A-links connect an access node, such as a switching system or SCP, to both members of an STP pair.
- B-links interconnect two STP pairs forming a "quad" of four signaling links where each STP independently connects to each member of the other pair.
- C-links are the high-capacity connections between the geographically separated members of an STP pair.
- D-links connect one STP pair to a second STP pair at another level in the signaling hierarchy or to another carrier.
- E-links connect an access node to a remote STP pair in the signaling network and are rarely used.
- F-links directly interconnect two access nodes without the use of an STP; they are nonredundant.

The CCS7 links normally function at 56 kb/s in North America while links operating at 64 kb/s are common in Europe.

The CCS7 signaling network provides the underlying foundation for the intelligent network, and the regional telephone companies in the United States began wide-scale deployment of these networks in 1986; several large independent telephone companies and interexchange carriers (ICs) soon followed. They used these networks for both trunk signaling between switches as well as for direct signaling from a switch to a database.

The Service Control Point

The "brains" of the intelligent network is the SCP. It is an on-line, fault-tolerant, transaction-processing database which provides call handling information in response to network queries. The SCP deployed for 800 Service is a high-capacity system capable of handling more than 900 queries per second or 3 million per hour. It is a real-time system with a response time of less than one half second, and it is a high-availability system with a downtime of less than 3 minutes per year for a mated SCP pair. The SCP is also designed to accommodate growth, which means that processing power or memory can be added to an in-service system without interrupting service. In addition, it is designed to accommodate graceful retrofit, which means that a new software program can be loaded into an in-service SCP without disrupting service.

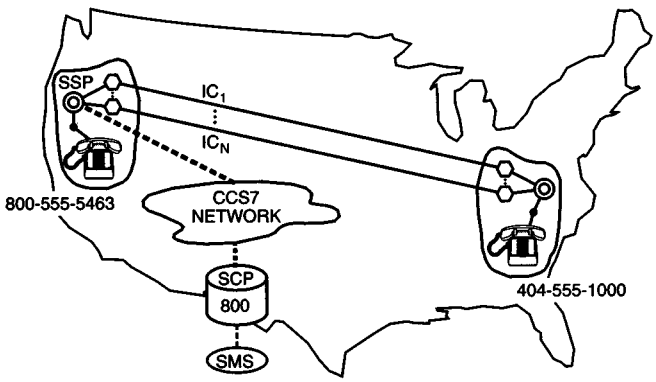


FIGURE 72.14 Data Base 800 Service—800-number calls are routed to an SSP which launches queries through a CCS7 network to an SCP containing the 800 database. In this example, the SCP translates the 800 Service number of 800-555-5463 into the POTS number of 404-555-1000. (*Source:* R. B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE*, vol. 79, no. 1. pp. 7–20, January 1991. © 1991 IEEE.)

Data Base 800 Service

SCPs have been deployed throughout the United States in support of the Data Base 800 Service mandated by the Federal Communications Commission. This service provides its subscribers with number portability so that a single 800 number can be used with different carriers. The Data Base 800 Service architecture is shown in Fig. 72.14. With this architecture, 800-number calls are routed from an end office to a service switching point (SSP) which launches queries through a CCS7 signaling network to the SCP. The SCP identifies the appropriate carrier, as specified by the 800 Service subscriber, and then, if appropriate, translates the 800 number to a plain old telephone (POTS) number. This information is subsequently returned to the SSP so that the call can be routed through the network by handing the call off to the appropriate carrier. This technology allows subscribers to select the carrier and the POTS number as a function of criteria such as time of day, day of week, percent allocation, and the location of the calling station. Thus the SCP provides two customer-specified routing information functions: a carrier identification function and an address translation function.

The SCP 800 Service database is administered by a single national **service management system (SMS).** The SMS is an interactive operations support system that is used to process and update customer records. It is the interface between the customer and the SCP. It translates a language which is friendly to a customer into a language which is friendly to on-line, real-time databases. Along the way, it validates the customer input.

Alternate Billing Services

Alternate billing services (ABS) have also been implemented using the intelligent network architecture. Alternate billing is an umbrella title which includes Calling Card Service, collect calling, and bill-to-third-number calling. The network configuration supporting ABS is shown in Fig. 72.15.

With this architecture, when a customer places a Calling Card call, the call is routed to an operator services system (OSS) which suspends call processing and launches a query through a CCS7 signaling network. The query is delivered to an SCP which contains the **line information database (LIDB)** application software. The LIDB application can provide routing information, such as identifying the customer-specified carrier which is to handle the call, as well as provide screening functions, such as the Calling Card validation used to authorize a call. The LIDB then returns the appropriate information to the OSS so that the call can be completed. The LIDBs are supported by the **database administration system (DBAS),** which is an operations support system that processes updates for Calling Card Service as well as other services. Multiple DBAS systems typically support each LIDB.

During 1991, each of the Regional Bell Operating Companies and a number of large independent telephone companies interconnected their CCS7 networks, mostly through STP hubs, to create a national signaling network; it was a process called LIDB interconnect. When it was finished, it meant that a person carrying a

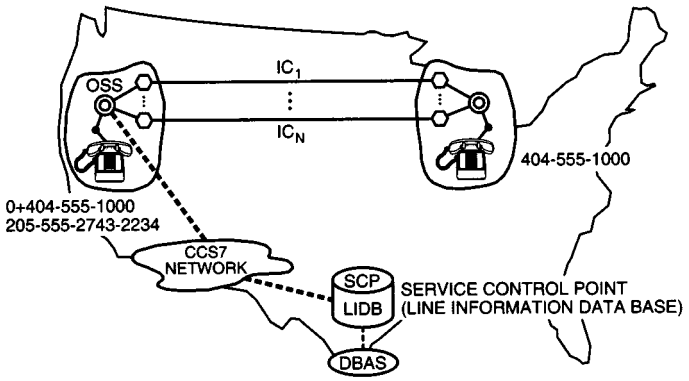


FIGURE 72.15 Alternate billing services—calls are routed to an OSS which launches queries through the CCS7 network to SCPs containing the LIDB application. (*Source:* R. B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991. © 1991 IEEE.)

particular company's Calling Card could, from anywhere in the United States, query the LIDB containing the associated Calling Card number.

Although the LIDB was originally developed to support Calling Card Service, it has since found wide application in the telecommunications industry. For example, the LIDB is used to translate the telephone number of a calling party to a name as part of Calling Name Delivery service, or to convert that number to a nine-digit ZIP code as part of Single Number Service. The LIDB databases now contain more than a quarter of a billion customer records which are updated at a rate of more than a million changes per day. Although physically distributed, the LIDBs appear logically as a single database. They represent a national resource.

Other Services

For alternate billing services, the SCP is essentially designed to perform two functions: carrier identification and billing authorization. For 800 Service, the SCP provides carrier identification and address translation. These basic functions of authorization, address translation and carrier identification can be used again and again in many different ways. For example, the intelligent network has been used to support private virtual networks (PVNs). PVNs make use of the public telephone network but, by means of software control, appear to have the characteristics of private networks. A PVN serves a closed-user group, and a caller requires authorization to gain access to the network. This screening function on originating calls uses an authorization function. Second, a PVN may offer an abbreviated dialing plan, for example, four-digit dialing. In this instance, the SCP performs an address translation function, converting a four-digit number to a ten-digit POTS number. There may also be a customer-specified routing information function which involves selecting from a hierarchy of facilities; this can be accomplished through use of the SCP carrier identification function.

The SCP in the intelligent network can support a vast number of services ranging from Calling Name Delivery service to messaging service. With Calling Name Delivery, a switch sends a query to the SCP with the ten-digit calling party number; the response is the calling party name which is then forwarded by the switch to a display unit attached to the called party station set. In support of messaging services, the address translation capability of the SCP can be used to translate a person's telephone number to an electronic-mail address. As a result, the sender of electronic mail need only know a person's telephone number.

The Advanced Intelligent Network

The intelligent network architecture discussed thus far is often referred to in the literature as Intelligent Network/1; this architecture has addressed the deployment problem and the service uniformity problem. The next phase in the evolution of this network has come to be called the advanced intelligent network (AIN), with the AIN standards defined by Bellcore.

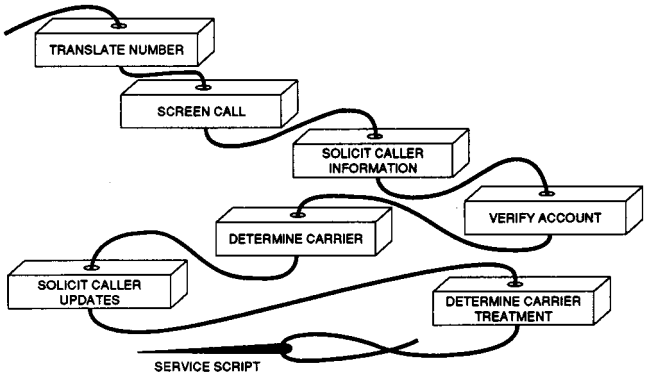


FIGURE 72.16 Creating the service script or scenario for a call by stitching together functional blocks. (*Source:* R. B. Robrock II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991. © 1991 IEEE.)

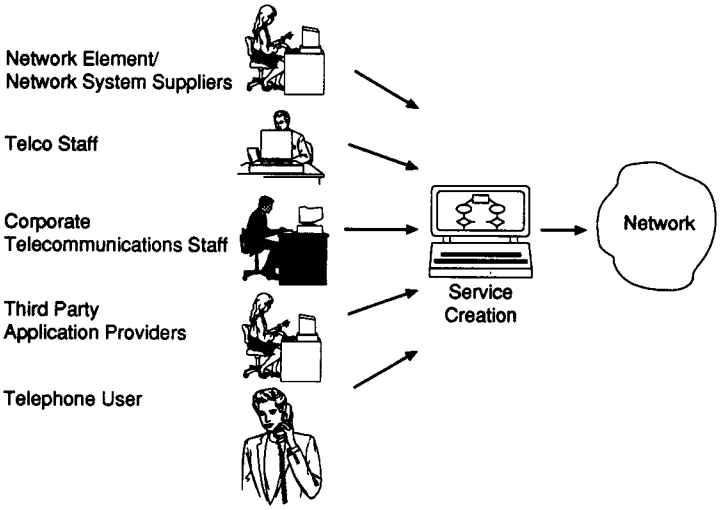


FIGURE 72.17 The advanced intelligent network—a business perspective. (*Source:* R. B. Robrock II, "Putting the Telephone User in the Driver's Seat," International Council for Computer Communication Conference on Intelligent Networks, pp. 144–150, May 1992.)

The concept of AIN is that new services can be developed and introduced into the network without requiring carriers to wait for switch generics to be upgraded. Some AIN applications introduce powerful service-creation capabilities which allow nonprogrammers to invoke basic functions offered in the network and stitch together those functions, as illustrated in Fig. 72.16, to constitute a new service. As a result, AIN promises to dramatically shorten the interval required to develop new services. Perhaps of greater significance, it promises to broaden the participation in service creation. In addition, it offers the opportunity to personalize or customize services. The silicon revolution has driven the cost of memory down to the point where it is economically viable to have enough memory in the network to store the service scripts or call processing scenarios that are unique to individuals.

Many people think of the AIN as a collection of network elements, network systems and operations systems; this view might be called a technologist's view. Perhaps a better representation is shown in Fig. 72.17; it shows a collection of people—people empowered to create services.

Historically, the creation of new services provided by the telephone network has been the sole domain of the network element and network system suppliers. There is perhaps a good analogy with the automobile

industry. A market study in the early 1900s predicted that 200,000 was the maximum number of cars that could ever be sold in a single year in the United States; the reasoning was that 200,000 was the maximum number of chauffeurs that could enter the workforce in a single year. In the telecommunications business, the network element and network system suppliers have been the chauffeurs of the network services business.

The service-creation tools offered by the AIN, however, empower telephone company staff to create new services. Moreover, similar tools may well be used by the telecommunications staff of large corporations, or by third-party application providers or even by some segment of the telephone user population. As a result, we may see an explosion in the number of network services.

The AIN introduces very powerful service-creation tools which are used to produce service-logic scripts (programs). In one arrangement, the service creation is done by assembling service-logic graphs from graphical icons that represent functional components of services. The completed graph is then validated with an expert system and tested off-line by executing every leg of the service-logic graph. At this point the service-logic program can be downloaded into the service control point so that it is ready for execution.

To make use of the new service, it is then necessary to set "triggers" in the appropriate service switching point. These triggers can be set for both originating and terminating calls, and they represent events which, should they occur, indicate the need for the switch to launch a query to the SCP for information the switch needs to process the call. The AIN switch generics, which are presently deployed, support several triggers such as "immediate off hook" or "called address." Future AIN switch generics are expected to support several dozen triggers. The first phase of the AIN, called AIN 0, became reality in late 1991 when friendly user trials began in two of the Regional Bell Operating Companies.

AIN 0 evolved to AIN 0.1 and then AIN 0.2, with each new version of AIN containing additional triggers. Today over 100 AIN services are deployed in North America and the number is growing rapidly.

The European Telecommunications Standards Institute (ETSI) has defined a European AIN standard referred to as Core INAP, and deployment of Core INAP systems in Europe began in 1996.

The architectural concepts of AIN are now beginning to carry over into wireless networks as well as broadband networks. Although the standards in these domains are just being developed, the value added by the wireless intelligent network (WIN) and the broadband intelligent network (BIN) promises to surpass the value seen in the narrowband wireline world.

Back to the Future

The intelligent network, with its centralized databases, has offered a means to rapidly introduce new services in a ubiquitous fashion and with operational uniformity as seen by the end user. The advanced intelligent network has gone on to provide a service-independent architecture, and, with its powerful service-creation capabilities, has empowered nonprogrammers to participate in the development of new services. In many ways, as we go into the future, we are going back to a time when operators were the "human intelligence" in the network. The human intelligence was all but eliminated with the introduction of switching systems, but now the intelligent network is working to put the intelligence of the human operator back into the network.

Defining Terms

Common-channel signaling (CCS): A technique for routing signaling information through a packet-switched network.

Database administration systems (DBAS): An operations support system that administers updates for the line information database.

Line information database (LIDB): An application running on the service control point that contains information on telephone lines and Calling Cards.

Service control point (SCP): An on-line, real-time, fault-tolerant, transaction-processing database which provides call-handling information in response to network queries.

Service management system (SMS): An operations support system which administers customer records for the service control point.

Signal transfer point (STP): A packet switch found in the common-channel signaling network; it is used to route signaling messages between network access nodes such as switches and SCPs.

Signaling system 7 (SS7): A communications protocol used in common-channel signaling networks.

Related Topic

72.2 Computer Communication Networks

References

AT&T Bell Laboratories, "Common channel signaling," *The Bell System Tech. J.,* vol. 57, no. 2, pp. 221–477, February 1978.

AT&T Bell Laboratories, "Stored program controlled network," *The Bell System Tech. J.,* vol. 61, no. 7, part 3, pp. 1573–1815, September 1982.

Bell Communications Research, "Advanced intelligent network (AIN) 0.1 switch-service control point (SCP) application protocol interface generic requirements," *Bell Commun. Res. Technical Ref.,* TR-NWT-001285, Issue 1, August 1992.

Bell Communications Research, "Advanced intelligent network (AIN) switch-service control point (SCP)/Adjunct interface generic requirements," *Bell Commun. Res.,* Generic Requirements, GR-1299-CORE, Issue 2, December 1994.

European Telecommunications Standards Institute, "Intelligent network (IN): Intelligent network capability set 1 (CS1) core intelligent network applications protocol (INAP) part 1: Protocol specification," *Eur. Telecom. Stds. Inst.,* ETS 300 374-1, draft, May 1994.

Globecom '86: The Global Telecommunications Conference, *Conference Record,* vol. 3, pp. 1311– 1335, December 1986.

R.J. Hass and R.W. Humes, "Intelligent network/2: A network architecture concept for the 1990s," International Switching Symposium, *Conference Record,* vol. 4, pp. 944–951, March 1987.

R.B. Robrock, II, "The intelligent network—Changing the face of telecommunications," *Proc. IEEE,* vol. 79, no. 1. pp. 7–20, January 1991.

R.B. Robrock, II, "Putting the telephone user in the driver's seat," International Council for Computer Communication Intelligent Networks Conference, pp. 144–150, May 1992.

R.B. Robrock, II, "The many faces of the LIDB data base," International Conference on Communications, *Conference Record,* June 1992.

Further Information

The bimonthly magazine *Bellcore Exchange* has numerous articles on the intelligent network, particularly in the following issues: July/August 1986, November/December 1987, July/August 1988, and March/April 1989. Articles on AIN service creation appear in the January/February 1992 issue. Subscriptions or single copies are available from the Bellcore Exchange Circulation Manager, 60 New England Avenue, Piscataway, NJ 08854-4196.

The monthly publication *IEEE Communications Magazine* contains numerous articles on the intelligent network. A special issue on the subject was published in January 1992. Copies are available from the IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854-4150.

The monthly publication *The Bellcore Digest* lists recent Bellcore publications. There are a series of technical advisories, technical requirements, and special reports that have been issued on the intelligent network. Copies are available by contacting Bellcore Customer Service Toll-Free 1-800-521-CORE (2673).

The bimonthly publication *The AT&T Technical Journal* contains numerous articles on the intelligent network. The advanced intelligent network is the subject of a special issue: Summer 1991, vol. 70, nos. 3–4. Current or recent issues may be obtained from the AT&T Customer Information Center, P.O. Box 19901, Indianapolis, IN 46219.

Poor, H.V., Looney, C.G., Marks II, R.J., Verdú, S., Thomas, J.A., Cover, T.M.
"Information Theory"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

73

Information Theory

73.1 Signal Detection
General Considerations • Detection of Known Signals • Detection of Parametrized Signals • Detection of Random Signals • Deciding Among Multiple Signals • Detection of Signals in More General Noise Processes • Robust and Nonparametric Detection • Distributed and Sequential Detection • Detection with Continuous-Time Measurements

73.2 Noise
Statistics of Noise • Noise Power • Effect of Linear Transformations on Autocorrelation and Power Spectral Density • White, Gaussian, and Pink Noise Models • Thermal Noise as Gaussian White Noise • Some Examples • Measuring Thermal Noise • Effective Noise and Antenna Noise • Noise Factor and Noise Ratio • Equivalent Input Noise • Other Electrical Noise • Measurement and Quantization Noise • Coping with Noise

73.3 Stochastic Processes
Introduction to Random Variables • Stochastic Processes • Classifications of Stochastic Processes • Stationarity of Processes • Gaussian and Markov Processes • Examples of Stochastic Processes • Linear Filtering of Weakly Stationary Processes • Cross-Correlation of Processes • Coherence • Ergodicity

73.4 The Sampling Theorem
The Cardinal Series • Proof of the Sampling Theorem • The Time-Bandwidth Product • Sources of Error • Generalizations of the Sampling Theorem

73.5 Channel Capacity
Information Rates • Communication Channels • Reliable Information Transmission: Shannon's Theorem • Bandwidth and Capacity • Channel Coding Theorems

73.6 Data Compression
Entropy • The Huffman Algorithm • Entropy Rate • Arithmetic Coding • Lempel–Ziv Coding • Rate Distortion Theory • Quantization and Vector Quantization • Kolmogorov Complexity • Data Compression in Practice

H. Vincent Poor
Princeton University

Carl G. Looney
University of Nevada

R. J. Marks II
University of Washington

Sergio Verdú
Princeton University

Joy A. Thomas
IBM

Thomas M. Cover
Stanford University

73.1 Signal Detection

H. Vincent Poor

The field of signal detection and estimation is concerned with the processing of information-bearing signals for the purpose of extracting the information they contain. The applications of this methodology are quite broad, ranging from areas of electrical engineering such as automatic control, digital communications, image processing, and remote sensing, into other engineering disciplines and the physical, biological, and social sciences.

There are two basic types of problems of interest in this context. *Signal detection* problems are concerned primarily with situations in which the information to be extracted from a signal is discrete in nature. That is, signal detection procedures are techniques for deciding among a discrete (usually finite) number of possible alternatives. An example of such a problem is the demodulation of a digital communication signal, in which the task of interest is to decide which of several possible transmitted symbols has elicited a given received signal. *Estimation* problems, on the other hand, deal with the determination of some numerical quantity taking values in a continuum. An example of an estimation problem is that of determining the phase or frequency of the carrier underlying a communication signal.

Although signal detection and estimation is an area of considerable current research activity, the fundamental principles are quite well developed. These principles, which are based on the theory of statistical inference, explain and motivate most of the basic signal detection and estimation procedures used in practice. In this section, we will give a brief overview of the basic principles underlying the field of signal detection. Estimation is treated elsewhere this volume, notably in Section 16.2. A more complete introduction to these subjects is found in Poor [1994].

General Considerations

The basic principles of signal detection can be conveniently discussed in the context of decision-making between two possible statistical models for a set of real-valued measurements, $Y_1, Y_2, \ldots, Y_n$. In particular, on observing $Y_1, Y_2, \ldots, Y_n$, we wish to decide whether these measurements are most consistent with the model

$$Y_k = N_k, \qquad k = 1, 2, \ldots, n \tag{73.1}$$

or with the model

$$Y_k = N_k + S_k, \quad k = 1, 2, \ldots, n \tag{73.2}$$

where $N_1, N_2, \ldots, N_n$ is a random sequence representing noise, and where $S_1, S_2, \ldots, S_n$ is a sequence representing a (possibly random) signal.

In deciding between Eqs. (73.1) and (73.2), there are two types of errors possible: a *false alarm,* in which (73.2) is falsely chosen, and a *miss,* in which (73.1) is falsely chosen. The probabilities of these two types of errors can be used as performance indices in the optimization of rules for deciding between (73.1) and (73.2). Obviously, it is desirable to minimize both of these probabilities to the extent possible. However, the minimization of the **false-alarm probability** and the minimization of the **miss probability** are opposing criteria. So, it is necessary to effect a trade-off between them in order to design a signal detection procedure. There are several ways of trading off the probabilities of miss and false alarm: the **Bayesian detector** minimizes an average of the two probabilities taken with respect to prior probabilities of the two conditions (73.1) and (73.2), the *minimax* detector minimizes the maximum of the two error probabilities, and the **Neyman-Pearson detector** minimizes the miss probability under an upper-bound constraint on the false-alarm probability.

If the statistics of noise and signal are known, the Bayesian, minimax, and Neyman-Pearson detectors are all of the same form. Namely, they reduce the measurements to a single number by computing the **likelihood ratio**

$$L(Y_1, Y_2, \ldots, Y_n) \triangleq \frac{p_{S+N}(Y_1, Y_2, \ldots, Y_n)}{p_N(Y_1, Y_2, \ldots, Y_n)} \tag{73.3}$$

where p_{S+N} and p_N denote the probability density functions of the measurements under signal-plus-noise (73.2) and noise-only (73.1) conditions, respectively. The likelihood ratio is then compared to a *decision threshold,* with the signal-present model (73.2) being chosen if the threshold is exceeded, and the signal-absent model (73.1) being chosen otherwise. Choice of the decision threshold determines a trade-off of the two error probabilities, and the optimum procedures for the three criteria mentioned above differ only in this choice.

There are several basic signal detection structures that can be derived from Eqs. (73.1) to (73.3) under the assumption that the noise sequence consists of a set of independent and identically distributed (i.i.d.) Gaussian random variables with zero means. Such a sequence is known as **discrete-time white Gaussian noise.** Thus, until further notice, we will make this assumption about the noise. It should be noted that this assumption is physically justifiable in many applications.

Detection of Known Signals

If the signal sequence $S_1, S_2, \ldots, S_n$ is known to be given by a specific sequence, say $s_1, s_2, \ldots, s_n$ (a situation known as *coherent detection*), then the likelihood ratio (73.3) is given in the white Gaussian noise case by

$$\exp\left\{\left(\sum_{k=1}^{n} s_k Y_k - \frac{1}{2}\sum_{k=1}^{n} s_k^2\right) / \sigma^2\right\} \tag{73.4}$$

where σ^2 is the variance of the noise samples. The only part of (73.4) that depends on the measurements is the term $\Sigma_{k=1}^{n} s_k Y_k$ and the likelihood ratio is a monotonically increasing function of this quantity. Thus, optimum detection of a coherent signal can be accomplished via a correlation detector, which operates by comparing the quantity

$$\sum_{k=1}^{n} s_k Y_k \tag{73.5}$$

to a threshold, announcing signal presence when the threshold is exceeded.

Note that this detector works on the principle that the signal will correlate well with itself, yielding a large value of (73.5) when present, whereas the random noise will tend to average out in the sum (73.5), yielding a relatively small value when the signal is absent. This detector is illustrated in Fig. 73.1.

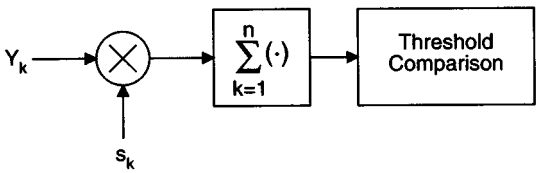


FIGURE 73.1 Correlation detector for a coherent signal in additive white Gaussian noise.

Detection of Parametrized Signals

The correlation detector cannot usually be used directly unless the signal is known exactly. If, alternatively, the signal is known up to a short vector $\boldsymbol{\theta}$ of random parameters (such as frequencies or phases) that are independent of the noise, then an optimum test can be implemented by threshold comparison of the quantity

$$\int_{\Lambda} \exp\left\{\left(\sum_{k=1}^{n} s_k(\boldsymbol{\theta}) Y_k - \frac{1}{2}\sum_{k=1}^{n} \left[s_k(\boldsymbol{\theta})\right]^2\right) / \sigma^2\right\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{73.6}$$

where we have written $S_k = s_k(\boldsymbol{\theta})$ to indicate the functional dependence of the signal on the parameters, and where Λ and $p(\boldsymbol{\theta})$ denote the range and probability density function, respectively, of the parameters.

The most important example of such a parametrized signal is that in which the signal is a modulated sinusoid with random phase; i.e.,

$$S_k = a_k \cos(\omega_c k + \theta), \qquad k = 1, 2, \ldots, n \tag{73.7}$$

where $a_1, a_2, \ldots, a_n$ is a known amplitude modulation sequence, ω_c is a known (discrete-time) carrier frequency, and the random phase θ is uniformly distributed in the interval $[-\pi,\pi]$. In this case, the likelihood ratio is a monotonically increasing function of the quantity

$$\left[\sum_{k=1}^{n} a_k \cos(\omega_c k) Y_k\right]^2 + \left[\sum_{k=1}^{n} a_k \sin(\omega_c k) Y_k\right]^2 \tag{73.8}$$

Thus, optimum detection can be implemented via comparison of (73.8) with a threshold, a structure known as an **envelope detector.** Note that this detector correlates the measurements with two orthogonal components of the signal, $a_k \cos(\omega_c k)$ and $a_k \sin(\omega_c k)$. These two correlations, known as the in-phase and quadrature components of the measurements, respectively, capture all of the energy in the signal, regardless of the value of θ. Since θ is unknown, however, these two correlations cannot be combined coherently, and thus they are combined noncoherently via (73.8) before the result is compared with a threshold. This detector is illustrated in Fig. 73.2.

Parametrized signals also arise in situations in which it is not appropriate to model the unknown parameters as random variables with a known distribution. In such cases, it is not possible to compute the likelihood ratio (73.6) so an alternative to the likelihood ratio detector must then be used. (An exception is that in which the likelihood ratio detector is invariant to the unknown parameters—a case known as *uniformly most powerful detection.*) Several alternatives to the likelihood ratio detector exist for these cases.

One useful such procedure is to test for the signal's presence by threshold comparison of the *generalized likelihood ratio,* given by

$$\max_{\boldsymbol{\theta} \in \Lambda} L_{\boldsymbol{\theta}}(Y_1, Y_2, \ldots, Y_n) \tag{73.9}$$

where $L_{\boldsymbol{\theta}}$ denotes the likelihood ratio for Eqs. (73.1) and (73.2) for the known-signal problem with the parameter vector fixed at $\boldsymbol{\theta}$. In the case of white Gaussian noise, we have

$$L_{\boldsymbol{\theta}}(Y_1, Y_2, \ldots, Y_n) = \exp\left\{\left(\sum_{k=1}^{n} s_k(\boldsymbol{\theta}) Y_k - \frac{1}{2}\sum_{k=1}^{n}\left[s_k(\boldsymbol{\theta})\right]^2\right) / \sigma^2\right\} \tag{73.10}$$

It should be noted that this formulation is also valid if the statistics of the noise have unknown parameters, e.g., the noise variance in the white Gaussian case.

One common application in which the generalized likelihood ratio detector is useful is that of detecting a signal that is known except for its time of arrival. That is, we are often interested in signals parametrized as

$$s_k(\theta) = a_{k-\theta} \tag{73.11}$$

where $\{a_k\}$ is a known finite-duration signal sequence and where θ ranges over the integers. Assuming white Gaussian noise and an observation interval much longer than the duration of $\{a_k\}$, the generalized likelihood ratio detector in this case announces the presence of the signal if the quantity

$$\max_{\theta} \sum_{k} a_{k-\theta} Y_k \tag{73.12}$$

exceeds a fixed threshold. This type of detector is known as a *matched filter,* since it can be implemented by filtering the measurements with a digital filter whose pulse response is a time-reversed version of the known

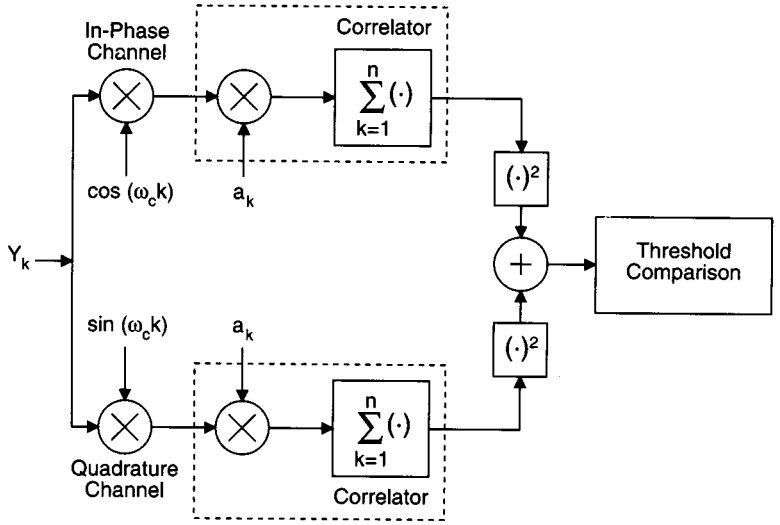


FIGURE 73.2 Envelope detector for a noncoherent signal in additive white Gaussian noise.

signal $\{a_k\}$ (hence it is "matched" to the signal), and announcing the signal's presence if the filter output exceeds the decision threshold at any time.

Detection of Random Signals

In some applications, particularly in remote sensing applications such as sonar and radio astronomy, it is appropriate to consider the signal sequence $S_1, S_2, \ldots, S_n$ itself to be a random sequence, statistically independent of the noise. In such cases, the likelihood ratio formula of (73.6) is still valid with the parameter vector $\boldsymbol{\theta}$ simply taken to be the signal itself. However, for long measurement records (i.e., large n), (73.6) is not a very practical formula except in some specific cases, the most important of which is the case in which the signal is Gaussian.

In particular, if the signal is Gaussian with zero-mean and autocorrelation sequence $r_{k,l} \triangleq E\{S_k S_l\}$, then the likelihood ratio is a monotonically increasing function of the quantity

$$\sum_{k=1}^{n} \sum_{l=1}^{n} q_{k,l} Y_k Y_l \tag{73.13}$$

with $q_{k,l}$ the element in the kth row and lth column of the positive-definite matrix

$$\mathbf{Q} \triangleq \mathbf{I} - (\mathbf{I} + \mathbf{R}/\sigma^2)^{-1} \tag{73.14}$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix, and R is the covariance matrix of the signal, i.e., it is the $n \times n$ matrix with elements $r_{k,l}$.

Note that (73.13) is a quadratic function of the measurements; thus, a detector based on the comparison of this quantity to a threshold is known as a **quadratic detector.** The simplest form of this detector results from the situation in which the signal samples are, like the noise samples, i.i.d. In this case, the quadratic function (73.13) reduces to a positive constant multiple of the quantity

$$\sum_{k=1}^{n} Y_k^2 \tag{73.15}$$

A detector based on (73.15) simply measures the energy in the measurements and then announces the presence of the signal if this energy is large enough. This type of detector is known as a *radiometer.*

Thus, radiometry is optimum in the case in which both signal and noise are i.i.d. Gaussian sequences with zero means. Since in this case the presence of the signal is manifested only by an increase in energy level, it is intuitively obvious that radiometry is the only way of detecting the signal's presence. More generally, when the signal is correlated, the quadratic function (73.13) exploits both the increased energy level and the correlation structure introduced by the presence of the signal. For example, if the signal is a narrowband Gaussian process, then the quadratic function (73.13) acts as a narrowband radiometer with bandpass characteristic that approximately matches that of the signal. In general, the quadratic detector will make use of whatever spectral properties the signal exhibits.

If the signal is random but not Gaussian, then its optimum detection [described by (73.6)] typically requires more complicated nonlinear processing than the quadratic processing of (73.13) in order to exploit the distributional differences between signal and noise. This type of processing is often not practical for implementation, and thus approximations to the optimum detector are typically used. An interesting family of such detectors uses cubic or quartic functions of the measurements, which exploit the higher-order spectral properties of the signal [Mendel, 1991]. As with deterministic signals, random signals can be parametrized. In this case, however, it is the distribution of the signal that is parametrized. For example, the power spectrum of the signal of interest may be known only up to a set of unknown parameters. Generalized likelihood ratio detectors (73.9) are often used to detect such signals.

Deciding Among Multiple Signals

The preceding results have been developed under the model (73.1)–(73.2) that there is a single signal that is either present or absent. In digital communications applications, it is more common to have the situation in which we wish to decide between the presence of two (or more) possible signals in a given set of measurements. The foregoing results can be adapted straightforwardly to such problems. This can be seen most easily in the case of deciding among known signals. In particular, consider the problem of deciding between two alternatives:

$$Y_k = N_k + s_k^{(0)}, \qquad k = 1, 2, \ldots, n \tag{73.16}$$

and

$$Y_k = N_k + s_k^{(1)}, \qquad k = 1, 2, \ldots, n \tag{73.17}$$

where $s_1^{(0)}, s_2^{(0)}, \ldots, s_n^{(0)}$ and $s_1^{(1)}, s_2^{(1)}, \ldots, s_n^{(1)}$ are two known signals. Such problems arise in data transmission problems, in which the two signals $s^{(0)}$ and $s^{(1)}$ correspond to the waveforms received after transmission of a logical "zero" and "one," respectively. In such problems, we are generally interested in minimizing the *average probability of error,* which is the average of the two error probabilities weighted by the prior probabilities of occurrence of the two signals. This is a Bayesian performance criterion, and the optimum decision rule is a straightforward extension of the correlation detector based on (73.5). In particular, under the assumptions that the two signals are equally likely to occur prior to measurement, and that the noise is white and Gaussian, the optimum decision between (73.16) and (73.17) is to choose the model (73.16) if $\Sigma_{k=1}^{n} s_k^{(0)} Y_k$ is larger than $\Sigma_{k=1}^{n} s_k^{(1)} Y_k$, and to choose the model (73.17) otherwise.

More generally, many problems in digital communications involve deciding among M equally likely signals with $M > 2$. In this case, again assuming white Gaussian noise, the decision rule that minimizes the error probability is to choose the signal $s_1^{(j)}, s_2^{(j)}, \ldots, s_n^{(j)}$, where j is a solution of the maximization problem

$$\sum_{k=1}^{n} s_k^{(j)} Y_k = \max_{0 \le m \le M-1} \sum_{k=1}^{n} s_k^{(m)} Y_k \tag{73.18}$$

There are two basic types of digital communications applications in which the problem (73.18) arises. One is in *M-ary data transmission,* in which a symbol alphabet with M elements is used to transmit data, and a decision among these M symbols must be made in each symbol interval [Proakis, 1983]. The other type of application in which (73.18) arises is that in which data symbols are correlated in some way because of intersymbol interference, coding, or multiuser transmission. In such cases, each of the M possible signals represents a frame of data symbols, and a joint decision must be made about the entire frame since individual symbol decisions cannot be decoupled. Within this latter framework, the problem (73.18) is known as *sequence detection.* The basic distinction between M-ary transmission and sequence detection is one of degree. In typical M-ary transmission, the number of elements in the signaling alphabet is typically a small power of 2 (say 8 or 32), whereas the number of symbols in a frame of data could be on the order of thousands. Thus, solution of (73.18) by exhaustive search is prohibitive for sequence detection, and less complex algorithms must be used. Typical digital communications applications in which sequence detection is necessary admit dynamic programming solutions to (73.18) (see, e.g., Verdú [1993]).

Detection of Signals in More General Noise Processes

In the foregoing paragraphs, we have described three basic detection procedures: correlation detection of signals that are completely known, envelope detection of signals that are known except for a random phase, and quadratic detection for Gaussian random signals. These detectors were all derived under an assumption of white Gaussian noise. This assumption provides an accurate model for the dominant noise arising in many communication channels. For example, the thermal noise generated in signal processing electronics is adequately described as being white and Gaussian. However, there are also many channels in which the statistical behavior of the noise is not well described in this way, particularly when the dominant noise is produced in the physical channel rather than in the receiver electronics.

One type of noise that often arises is noise that is Gaussian but not white. In this case, the detection problem (73.1)–(73.2) can be converted to an equivalent problem with white noise by applying a linear filtering process known as *prewhitening* to the measurements. In particular, on denoting the noise covariance matrix by Σ, we can write

$$\Sigma = \mathbf{C}\mathbf{C}^T \tag{73.19}$$

where $\mathbf{C}$ is an $n \times n$ invertible, lower-triangular matrix and where the superscript T denotes matrix transposition. The representation (73.19) is known as the *Cholesky decomposition.* On multiplying the measurement vector $\mathbf{Y} \triangleq (Y_1, Y_2, \ldots, Y_n)^T$ satisfying (73.1)–(73.2) with noise covariance Σ, by $\mathbf{C}^{-1}$, we produce an equivalent (in terms of information content) measurement vector that satistifies the model (73.1)–(73.2) with white Gaussian noise and with the signal conformally transformed. This model can then be treated using the methods described previously.

In other channels, the noise can be modeled as being i.i.d. but with an amplitude distribution that is not Gaussian. This type of model arises, for example, in channels dominated by impulsive phenomena, such as urban radio channels. In the non-Gaussian case the procedures discussed previously lose their optimality as defined in terms of the error probabilities. These procedures can still be used, and they will work well under many conditions; however, there will be a resulting performance penalty with respect to optimum procedures based on the likelihood ratio. Generally speaking, likelihood-ratio-based procedures for non-Gaussian noise channels involve more complex nonlinear processing of the measurements than is required in the standard detectors, although the retention of the i.i.d. assumption greatly simplifies this problem. A treatment of methods for such channels can be found in Kassam [1988].

When the noise is both non-Gaussian and dependent, the methodology is less well developed, although some techniques are available in these cases. An overview can be found in Poor and Thomas [1993].

Robust and Nonparametric Detection

All of the procedures outlined in the preceding subsection are based on the assumption of a known (possibly up to a set of unknown parameters) statistical model for signals and noise. In many practical situations it is

not possible to specify accurate statistical models for signals or noise, and so it is of interest to design detection procedures that do not rely heavily on such models. Of course, the parametrized models described in the foregoing paragraphs allow for uncertainty in the statistics of the observations. Such models are known as *parametric* models, because the set of possible distributions can be parametrized by a finite set of real parameters.

While parametric models can be used to describe many types of modeling uncertainty, composite models in which the set of possible distributions is much broader than a parametric model would allow are sometimed more realistic in practice. Such models are termed *nonparametric models.* For example, one might be able to assume only some very coarse model for the noise, such as that it is symmetrically distributed. A wide variety of useful and powerful detectors have been developed for signal-detection problems that cannot be parametrized. These are basically of two types: *robust* and *nonparametric.* Robust detectors are those designed to perform well despite small, but potentially damaging, nonparametric deviations from a nominal parametric model, whereas nonparametric detectors are designed to achieve constant false-alarm probability over very wide classes of noise statistics.

Robustness problems are usually treated analytically via minimax formulations that seek best worst-case performance as the design objective. This formulation has proven to be very useful in the design and characterization of robust detectors for a wide variety of detection problems. Solutions typically call for the introduction of light limiting to prevent extremes of gain dictated by an (unrealistic) nominal model. For example, the correlation detector of Fig. 73.1 can be made robust against deviations from the Gaussian noise model by introducing a soft-limiter between the multiplier and the accumulator.

Nonparametric detection is usually based on relatively coarse information about the observations, such as the algebraic signs or the ranks of the observations. One such test is the *sign test,* which bases its decisions on the number of positive observations obtained. This test is nonparametric for the model in which the noise samples are i.i.d. with zero median and is reasonably powerful against alternatives such as the presence of a positive constant signal in such noise. More powerful tests for such problems can be achieved at the expense of complexity by incorporating rank information into the test statistic.

Distributed and Sequential Detection

The detection procedures discussed in the preceding paragraphs are based on the assumption that all measurements can and should be used in the detection of the signal, and moreover that no constraints exist on how measurements can be combined. There are a number of applications, however, in which constraints apply to the information pattern of the measurements.

One type of constrained information pattern that is of interest in a number of applications is a network consisting of a number of distributed or local decision makers, each of which processes a subset of the measurements, and a *fusion center,* which combines the outputs of the distributed decision makers to produce a global detection decision. The communication between the distributed decision makers and the fusion center is constrained, so that each local decision maker must reduce its subset of measurements to a summarizing local decision to be transmitted to the fusion center. Such structures arise in applications such as the testing of large-scale integrated circuits, in which data collection is decentralized, or in detection problems involving very large data sets, in which it is desirable to distribute the computational work of the detection algorithm. Such problems lie in the field of *distributed detection.* Except in some trivial special cases, the constraints imposed by distributing the detection algorithm introduce a further level of difficulty into the design of optimum detection systems. Nevertheless, considerable progress has been made on this problem, a survey of which can be found in Tsitsiklis [1993].

Another type of nonstandard information pattern that arises is that in which the number of measurements is potentially infinite, but in which there is a cost associated with taking each measurement. This type of model arises in applications such as the synchronization of wideband communication signals. In such situations, the error probabilities alone do not completely characterize the performance of a detection system, since consideration must also be given to the cost of sampling. The field of *sequential detection* deals with the optimization of detection systems within such constraints. In sequential detectors, the number of measurements taken becomes a random variable depending on the measurements themselves. A typical performance criterion for optimizing such a system is to seek a detector that minimizes the expected number of measurements for given levels of miss and false-alarm probabilities.

The most commonly used sequential detection procedure is the *sequential probability ratio test,* which operates by recursive comparison of the likelihood ratio (73.3) to two thresholds. In this detector, if the likelihood ratio for a given number of samples exceeds the larger of the two thresholds, then the signal's presence is announced and the test terminates. Alternatively, if the likelihood ratio falls below the smaller of the two thresholds, the signal's absence is announced and the test terminates. However, if neither of the two thresholds is crossed, then another measurement is taken and the test is repeated.

Detection with Continuous-Time Measurements

Note that all of the preceding formulations have involved the assumption of discrete-time (i.e., sampled-data) measurements. From a practical point of view, this is the most natural framework within which to consider these problems, since implementations most often involve digital hardware. However, the procedures discussed in this section all have continuous-time counterparts, which are of both theoretical and practical interest. Mathematically, continuous-time detection problems are more difficult than discrete-time ones, because they involve probabilistic analysis on function spaces. The theory of such problems is quite elegant, and the interested reader is referred to Poor [1994] or Grenander [1981] for more detailed exposition.

Continuous-time models are of primary interest in the front-end stages of radio frequency or optical communication receivers. At radio frequencies, continuous-time versions of the models described in the preceding paragraphs can be used. For example, one may consider the detection of signals in continuous-time Gaussian white noise. At optical wavelengths, one may consider either continuous models (such as Gaussian processes) or point-process models (such as Poisson counting processes), depending on the type of detection used (see, e.g., Snyder and Miller [1991]). In the most fundamental analyses of optical detection problems, it is sometimes desirable to consider the quantum mechanical nature of the measurements [Helstrom, 1976].

Defining Terms

Bayesian detector: A detector that minimizes the average of the false-alarm and miss probabilities, weighted with respect to prior probabilities of signal-absent and signal-present conditions.

Correlation detector: The optimum structure for detecting coherent signals in the presence of additive white Gaussian noise.

Discrete-time white Gaussian noise: Noise samples modeled as independent and identically distributed Gaussian random variables.

Envelope detector: The optimum structure for detecting a modulated sinusoid with random phase in the presence of additive white Gaussian noise.

False-alarm probability: The probability of falsely announcing the presence of a signal.

Likelihood ratio: The optimum processor for reducing a set of signal-detection measurements to a single number for subsequent threshold comparison.

Miss probability: The probability of falsely announcing the absence of a signal.

Neyman-Pearson detector: A detector that minimizes the miss probability within an upper-bound constraint on the false-alarm probability.

Quadratic detector: A detector that makes use of the second-order statistical structure (e.g., the spectral characteristics) of the measurements. The optimum structure for detecting a zero-mean Gaussian signal in the presence of additive Gaussian noise is of this form.

Related Topics

16.2 Parameter Estimation • 70.3 Spread Spectrum Communications

References

U. Grenander, *Abstract Inference*, New York: Wiley, 1981.

C.W. Helstrom, *Quantum Detection and Estimation Theory*, New York: Academic Press, 1976.

S.A. Kassam, *Signal Detection in Non-Gaussian Noise*, New York: Springer-Verlag, 1988.

J.M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and systems theory: Theoretical results and some applications," *Proc. IEEE,* vol. 79, pp. 278–305, 1991.
H.V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed., New York: Springer-Verlag, 1994.
H.V. Poor and J. B. Thomas, "Signal detection in dependent non-Gaussian noise," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.
J.G. Proakis, *Digital Communications*, New York: McGraw-Hill, 1983.
D.L. Snyder and M.I. Miller, *Random Point Processes in Time and Space*, New York: Springer-Verlag, 1991.
J. Tsitsiklis, "Distributed detection," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.
S. Verdú, "Multiuser detection," in *Advances in Statistical Signal Processing*, vol. 2, Signal Detection, H.V. Poor and J.B. Thomas, Eds., Greenwich, Conn.: JAI Press, 1993.

Further Information

Except as otherwise noted in the accompanying text, further details on the topics introduced in this section can be found in the textbook:
Poor, H.V. *An Introduction to Signal Detection and Estimation,* 2nd ed., New York: Springer-Verlag, 1994.

The bimonthly journal, *IEEE Transactions on Information Theory,* publishes recent advances in the theory of signal detection. It is available from the Institute of Electrical and Electronics Engineers, Inc., 345 East 47th Street, New York, NY 10017.

Papers describing applications of signal detection are published in a number of journals, including the monthly journals *IEEE Transactions on Communications, IEEE Transactions on Signal Processing,* and the *Journal of the Acoustical Society of America.* The IEEE journals are available from the IEEE, as above. The *Journal of the Acoustical Society of America* is available from the American Institute of Physics, 335 East 45th Street, New York, NY 10017.

73.2 Noise

Carl G. Looney

Every information signal $s(t)$ is corrupted to some extent by the superimposition of extra-signal fluctuations that assume unpredictable values at each time instant t. Such undesirable signals were called **noise** due to early measurements with sensitive audio amplifiers.

Noise sources are (1) *intrinsic*, (2) *external*, or (3) *process induced*. Intrinsic noise in conductors comes from thermal agitation of molecularly bound ions and electrons, from microboundaries of impurities and grains with varying potential, and from transistor junction areas that become temporarily depleted of electrons/holes. External electromagnetic interference sources include airport radar, x-rays, power and telephone lines, communications transmissions, gasoline engines and electric motors, computers and other electronic devices; and also include lightning, cosmic rays, plasmas (charged particles) in space, and solar/stellar radiation (conductors act as antennas). Reflective objects and other macroboundaries cause multiple paths of signals. Process-induced errors include measurement, quantization, truncation, and signal generation errors. These also corrupt the signal with noise power and loss of resolution.

Statistics of Noise

Statistics allow us to analyze the spectra of noise. We model a noise signal by a **random** (or *stochastic*) **process** $N(t)$, a function whose realized value $N(t) = x_t$ at any time instant t is chosen by the outcome of the random variable $N_t = N(t)$. $N(t)$ has a probability distribution for the values x it can assume. Any particular trajectory $\{(t, x_t)\}$ of outcomes is called a **realization** of the noise process. The *first-order statistic* of $N(t)$ is the *expected value* $\mu_t = E[N(t)]$. The *second-order statistic* is the *autocorrelation function* $R_{NN}(t, t + \tau) = E[N(t)N(t + \tau)]$, where $E[-]$ is the expected value operator. **Autocorrelation** measures the extent to which noise random variables $N_1 = N(t_1)$ and $N_2 = N(t_2)$ at times t_1 and t_2 depend on each other in an average sense.

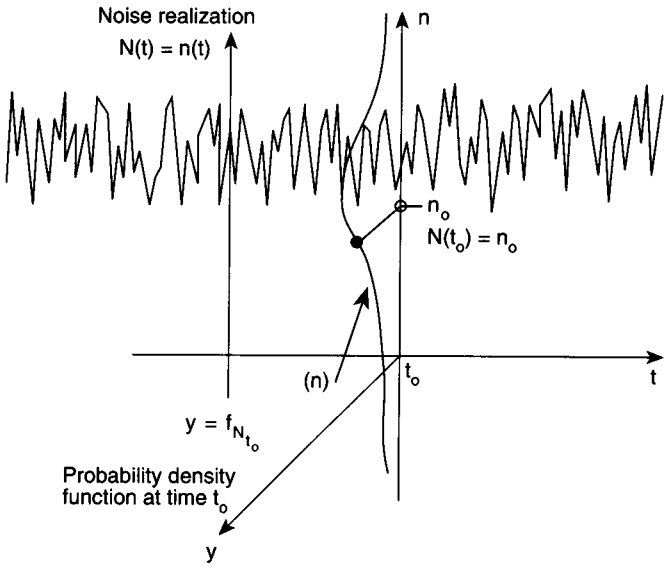


FIGURE 73.3 A noise process.

When the first- and second-order statistics do not change over time, we call the noise a **weakly** (or *wide-sense*) **stationary process**. This means that: (1) $E[N(t)] = \mu_t = \mu$ is constant for all t, and (2) $R_{NN}(t, t+\tau) = E[N(t)N(t+\tau)] = E[N(0)N(\tau)] = R_{NN}(\tau)$ for all t [see Brown, 1983, p. 82; Gardner, 1990, p. 108; or Peebles, 1987, p. 153 for properties of $R_{NN}(\tau)$]. In this case the autocorrelation function depends only on the *offset* τ. We assume hereafter that $\mu = 0$ (we can subtract μ, which does not change the autocorrelation). When $\tau = 0$, $R_{NN}(0) = E[N(t)N(t+0)] = E[(N(t))^2] = \sigma_N^2$, which is the fixed variance of each random variable N_t for all t. Weakly stationary (ws) processes are the most commonly encountered cases and are the ones considered here. *Evolutionary* processes have statistics that change over time and are difficult to analyze.

Figure 73.3 shows a realization of a noise process $N(t)$, where at any particular time t, the probability density function is shown coming out of the page in a third dimension. For a ws noise, the distributions are the same for each t. The most mathematically tractable noises are *Gaussian* ws processes, where at each time t the probability distribution for the random variable $N_t = N(t)$ is Gaussian (also called *normal*). The first- and second-order statistics completely determine Gaussian distributions, and so ws makes their statistics of all orders stationary over time also. It is well known [see Brown, 1983, p. 39] that linear transformations of Gaussian random variables are also Gaussian random variables. The probability density function for a Gaussian random variable N_t is $f_N(x) = \{1/[2\pi\sigma_N{}^2]^{1/2}\exp[-(x-\mu_N)^2/2\sigma_N^2]$, which is the familiar bell-shaped curve centered on $x = \mu_N$. The standard Gaussian probability table [Peebles, 1987, p. 314] is useful, e.g., $\Pr[-\sigma_N < N_t < \sigma_N) = 2\Pr[0 < N_t < \sigma_N) = 0.8413$ from the table.

Noise Power

The noise signal $N(t)$ represents voltage, so the autocorrelation function at offset 0, $R_{NN}(0) = E[N(t)N(t)]$ represents expected power in volts squared, or watts per ohm. When $R = 1\ \Omega$, then $N(t)N(t) = N(t)[N(t)/R] = N(t)I(t)$ volt-amperes = watts (where $I(t)$ is the current in a 1-Ω resistor). The Fourier transform $\mathbf{F}[R_{NN}(\tau)]$ of the autocorrelation function $R_{NN}(\tau)$ is the power spectrum, called the **power spectral density function** (psdf), $S_{NN}(w)$ in W/(rad/s). Then

$$S_{NN}(w) = \int_{-\infty}^{\infty} R_{NN}(\tau)e^{-jws}d\tau = \mathbf{F}[R_{NN}(\tau)]$$
$$R_{NN}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{NN}(w)e^{jws}dw = \mathbf{F}^{-1}[S_{NN}(w)] \tag{73.20}$$

The psdf at frequency f is defined to be the expected power that the voltage $N(t)$, bandlimited to an incremental band df centered at f, would dissipate in a 1-Ω resistance, divided by df.

Equations (73.20) are known as the *Wiener-Khinchin* relations that establish that $S_{NN}(w)$ and $R_{NN}(\tau)$ are a Fourier transform pair for ws random processes [Brown, 1983; Gardner, 1990, p. 230; Peebles, 1987]. The psdf $S_{NN}(w)$ has units of W/(rad/s), whereas the autocorrelation function $R_{NN}(\tau)$ has units of watts. When $\tau = 0$ in the second integral of Eq. (73.20), the exponential becomes $e^0 = 1$, so that $R_{NN}(0)$ $(= E[N(t)^2] = \sigma_N^2)$ is the integral of the psdf $S_{NN}(w)$ over all radian frequencies, $-\infty < w < \infty$. The rms (root-mean-square) voltage is $N_{\text{rms}} = \sigma_N$ (the *standard deviation*). The power spectrum in W/(rad/s) is a density that is summed up via an integral over the radian frequency band w_1 to w_2 to obtain the total power over that band.

$$P_{NN}(w_1, w_2) = \frac{1}{2\pi}\int_{w1}^{w2} S_{NN}(w)\cdot dw \quad \text{watts}$$
$$P_{NN} = \sigma_N^2 = E[N(t)^2] = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{NN}(w)\cdot dw \quad \text{watts} \tag{73.21}$$

The variance $\sigma_N^2 = R_{NN}(0)$ is the mean instantaneous power P_{NN} over all frequencies at any time t.

Effect of Linear Transformations on Autocorrelation and Power Spectral Density

Let $h(t)$ be the impulse response function of a time-invariant linear system L and $H(w) = \mathbf{F}[h(t)]$ be its transfer function. Let an input noise signal $N(t)$ have autocorrelation function $R_{NN}(\tau)$ and psdf $S_{NN}(w)$. We denote the output noise signal by $Y(t) = L[N(t)]$. The Fourier transforms $Y(w) \equiv \mathbf{F}[Y(t)]$ and $N(w) \equiv \mathbf{F}[N(t)]$ do not exist, but they are not needed. The output $Y(t)$ of a linear system is ws whenever the input $N(t)$ is ws [see Gardner, 1990, p. 195; or Peebles, 1987, p. 215]. The output psdf $S_{YY}(w)$ and autocorrelation function $R_{YY}(\tau)$ are given by, respectively,

$$S_{YY}(w) = |H(w)|^2 S_{NN}(w),\ R_{YY}(\tau) = \mathrm{F}^{-1}[S_{YY}(w)] \tag{73.22}$$

[see Gardner, 1990, p. 223]. The output noise power is

$$\sigma_Y^2 = P_{YY} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{YY}(w)dw = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(w)|^2 S_{NN}(w)dw \tag{73.23}$$

White, Gaussian, and Pink Noise Models

White noise [see Brown, 1983; Gardner, 1990, p. 234; or Peebles, 1987] is a theoretical model $W(t)$ of noise that is ws with zero mean. It has a constant power level n_o over all frequencies (analogous to white light), so its psdf is $S_{WW}(w) = n_o$ W/(rad/s), $-\infty < w < \infty$. The inverse Fourier transform of this is the impulse function $R_{WW}(\tau) = (n_o)\delta(\tau)$, which is zero for all offsets except $\tau = 0$. Therefore, white noise $W(t)$ is a process that is uncorrelated over time, i.e., $E[W(t_1)W(t_2)] = 0$ for t_1 not equal to t_2. Figure 73.4(a) shows the autocorrelation and psdf for white noise where the offset is $s = \tau$. A *Gaussian white noise* is white noise such that the probability distribution of each random variable $W_t = W(t)$ is Gaussian. When two Gaussian random variables W_1 and W_2 are *uncorrelated*, i.e., $E[W_1W_2] = 0$, they are independent [see Gardner, 1990, p. 37]. We use Gaussian models because of the *central limit theorem* that states that the sum of a number of random variables is approximately Gaussian.

Actual circuits attenuate signals above cut-off frequencies, and also the power must be finite. However, for white noise, $P_{WW} = R_{NN}(0) = \infty$, so we often truncate the white noise spectral density (psdf) at cut-offs $-w_c$ to w_c. The result is known as *pink noise*, $P(t)$, and is usually taken to be Gaussian because linear filtering of any white noise (through the effect of the central limit theorem) tends to make the noise Gaussian [see Gardner,

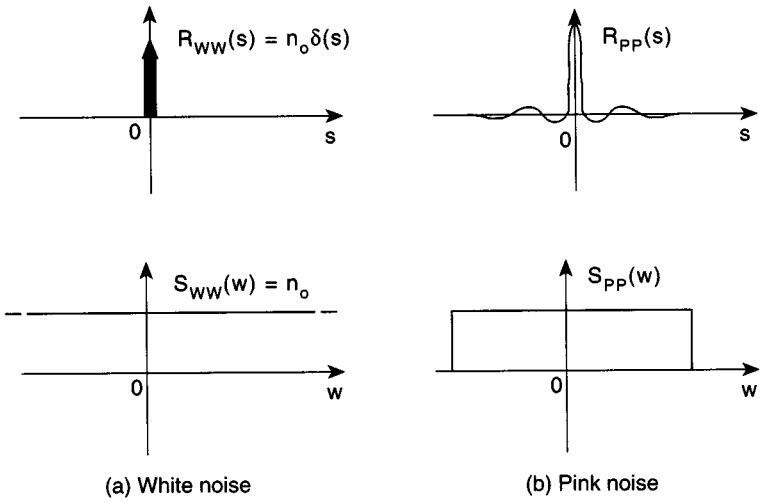


FIGURE 73.4 Power transform pairs for white and pink noise.

Figure 73.5 not available

FIGURE 73.5 Thermal noise in a resistor.

1990, p. 241]. Figure 73.4(b) shows the sinc function $R_{PP}(s) = \mathrm{F}^{-1}[S_{PP}(w)]$ for pink noise. Random variables P_1 and P_2 at times t_1 and t_2 are correlated only for t_1 and t_2 close.

Thermal Noise as Gaussian White Noise

Brown observed in 1828 that pollen and dust particles moved randomly when suspended in liquid. In 1906, Einstein analyzed such motion based on the random walk model. Perrin confirmed in 1908 that the thermal activity of molecules in a liquid caused irregular bombardment of the much larger particles. It was predicted that charges bound to thermally vibrating molecules would generate electromotive force (emf) at the open terminals of a conductor, and that this placed a limit on the sensitivity of galvanometers. Thermal noise (also called *Johnson noise*) was first observed by J. B. Johnson at Bell Laboratories in 1927. Figure 73.5 displays white noise as seen in the laboratory on an oscilloscope.

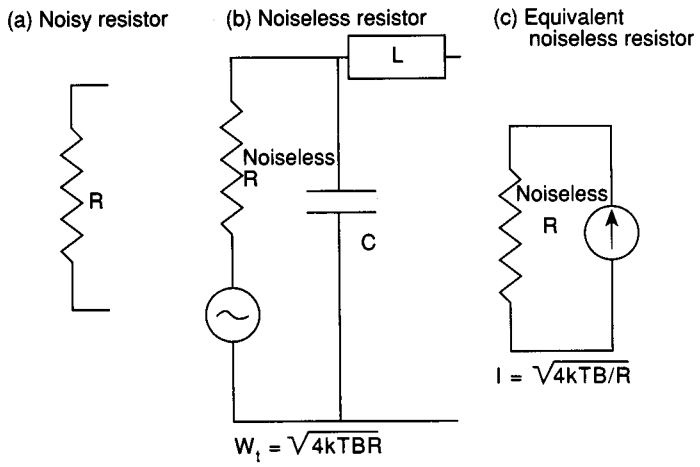


FIGURE 73.6 Thermal noise in a resistor.

The voltage $N(t)$ generated thermally between two points in an open circuit conductor is the sum of an extremely large number of superimposed, independent electronically and ionically induced microvoltages at all frequencies up to f_c = 6,000 GHz at room temperature [see Gardner 1990, p. 235], near infrared. The mean relaxation time of free electrons is $1/f_c = 0.5 \times 10^{-10}/T$ s, so at room temperature of T = 290K, it is 0.17 ps (1 picosecond = 10^{-12} s). The values of $N(t)$ at different times are uncorrelated for time differences (offsets) greater than $\tau_c = 1/f_c$. The expected value of $N(t)$ is zero. The power is fairly constant across a broad spectrum, and we cannot sample signals at picosecond periods, so we model Johnson noise $N(t)$ with Gaussian white noise $W(t)$. Although $\mu = E[W(t)] = 0$, the average power is positive at temperatures above 0K, and is $\sigma_W^2 = R_{WW}(0)$ [see the right side of Eq. (73.21)]. A disadvantage of the white noise model is its infinite power, i.e., $R_{WW}(0) = \sigma_W^2 = \infty$, but it is valid over a limited bandwidth of B Hz, in which case its power is finite.

In 1927, Nyquist [1928] theoretically derived thermal noise power in a resistor to be

$$P_{WW}(B) = 4kTRB \text{ (watts)} \tag{73.24}$$

where R is resistance (ohms), B is the frequency bandwidth of measurement in Hz (all emf fluctuations outside of B are ignored), $P_{WW}(B)$ is the mean power over B (see Eq. 73.21), and Boltzmann's constant is $k = 1.38 \times 10^{-23}$ J/K [see Ott, 1988; Gardner, 1990, p. 288; or Peebles, 1987, p. 227]. Under external emf, the thermally induced collisions are the main source of resistance in conductors (electrons pulled into motion by an external emf at 0K meet no resistance). The rms voltage is $W_{\text{rms}} = \sigma_W = [(4kTRB)]^{1/2}$ V over a bandwidth of B Hz.

Planck's radiation law is $S_{NN}(w) = (2h|f|)/[\exp(h|f|/kT) - 1]$, where $h = 6.63 \times 10^{-34}$ J/s is Planck's constant, and f is the frequency [see Gardner, 1990, p. 234]. For $|f|$ much smaller than $kT/h = 6.04 \times 10^{12}$ Hz $\approx$ 6,000 GHz, the exponential above can be approximated by $\exp(\text{h}|f|/\text{kT}) = 1 + \text{h}|f|/\text{kT}$. The denominator of $S_{NN}(w)$ becomes $h|f|/kT$, so $S_{NN}(w) = (2h|f|)/(h|f|/kT) = 2kT$ W/Hz in a 1-Ω resistor. Over a resistance of R Ω and a bandwidth of B Hz (positive frequencies), this yields the total power $P_{WW}(B) = 2BRS_{NN}(w) = 4kTRB$ W over the two-sided frequency spectrum. This is Nyquist's result.

Thermal noise is the same in a 1000-Ω carbon resistor as it is in a 1000-Ω tantalum thin-film resistor [see Ott, 1988]. While the intrinsic noise may never be less, it may be higher because of other superimposed noise (described in later sections). We model the thermal noise in a resistor by an internal source (generator), as shown in Fig. 73.6. Capacitance cannot be ignored at high f, but pure reactance (C or L) cannot dissipate energy, and so cannot generate thermal noise. The white noise model $W(t)$ for thermal noise $N(t)$ has a constant psdf $S_{WW}(w) = n_o$ W/(rad/s) for $-\infty < w < \infty$. By Eq. 73.21, the white noise mean power over the frequency bandwidth B is

$$P_{WW}(B) = \frac{1}{2\pi}\int_{-2\pi B}^{2\pi B} S_{WW}(w)dw = n_o(4\pi B/2\pi) = 2n_oB \tag{73.25}$$

Solving for the constant n_o, we obtain $n_o = P_{WW}(B)/2B$, which we put into Eq. (73.20) to get the spectral density as a function of temperature and resistance using Nyquist's result above.

$$S_{WW}(w) = n_o = P_{WW}(B)/4\pi B = 4kTR2\pi B/4\pi B = 2kTR \text{ watts/(rad/s)} \qquad (73.26)$$

Some Examples

The parasitic capacitance in the terminals of a resistor may cause a roll-off of about 20 dB/octave in actual resistors [Brown, 1983, p. 139]. At 290K (room temperature), we have $2kT = 2 \times 1.38 \times 10^{-23} \times 290 = 0.8 \times 10^{-20}$ W/Hz due to each ohm [see Ott, 1988]. For $R = 1$ MΩ (10^6 Ω), $S_{WW}(w) = 0.8 \times 10^{-14}$. Over a band of 10^8 Hz, we have $P_{WW}(B) = S_{WW}(w)B = 0.8 \times 10^{-14} \times 10^8 = 0.8 \times 10^{-6}$ W = 0.8 μW by Eqs. (73.24) and (73.26). In practice, parasitic capacitance causes thermal noise to be bandlimited (pink noise). Now consider Fig. 73.6(b) and let the temperature be 300K, $R = 10^6$ Ω, $C = 1$ pf (1 picofarad = 10^{-12} farads), and assume L is 0H. By Eq. (73.26), the thermal noise power is

$$S_{WW}(w) = 2kTR = 2 \times 1.38 \times 10^{-23} \times 300 \times 10^6 = 828 \times 10^{-17} \text{ W/Hz}$$

The power across a bandwidth $B = 10^6$ is $P_{WW}(B) = S_{WW}(w)B = 8280 \times 10^{-12}$ W, so the rms voltage is $W_{\text{rms}} = [P_{WW}(B)]^{1/2} = 91$ μV.

Now let $Y(t)$ be the output voltage across the capacitor. The transfer function can be seen to be $H(w) = \{I(w)(1/jwC)\}/\{I(w)[R + (1/jwC)]\} = (1/jwC)/[R + 1/jwC] = 1/[1 + jwRC]$ (where $I(w)$ is the Fourier transform of the current). The output psdf [see Eq. (73.22)] is

$$S_{YY}(w) = |H(w)|^2 S_{WW}(w) = (1/[1 + w^2R^2C^2])S_{WW}(w)$$

Integrating $S_{YY}(w) = (1/[1 + w^2R^2C^2])S_{WW}(w)$ over all radian frequencies $w = 2\pi f$ [see Eq. (73.21)], we obtain the antiderivative $(828 \times 10^{-17})(1/RC)\text{atan}(RCw)/2\pi$. Upon substituting the limits $w = \pm\infty$, this becomes $828 \times 10^{-17}[\pi/2 + \pi/2]/2\pi RC = 414 \times 10^{-17}(1/2\text{RC}) = 207 \times 10^{-17} \times 10^6 = 2070 \times 10^{-12}$ W/Hz. Then $\sigma_Y^2 = E[Y(t)^2] = P_{YY}(-\infty,\infty) = 2070 \times 10^{-12}$ W, so $Y_{\text{rms}}(\text{t}) = \sigma_Y = [P_{YY}(-\infty,\infty)]^{1/2} = 45.5$ μV. The half-power (cut-off) radian frequency is $w_c = 1/RC = 10^6$ rad/s, or $f_c = w_c/2\pi = 159.2$ kHz. Approximating $S_{YY}(w)$ by the rectangular spectrum $S_{YY}(w) = n_o, -10^6 < w < 10^6$ rad/s (0 elsewhere), we have that $R_{YY}(\tau) = (w_c/\pi)\text{sinc}(w_c\tau)$, which has the first zeros at $|w_c\tau| = \pi$, that is $|\tau| = 1/(2f_c)$ [see Fig. 73.4(b)]. We approximate the autocorrelation by $R_{YY}(\tau) = 0$ for $|s| \geq 1/2f_c$.

Measuring Thermal Noise

In Fig. 73.7, the thermal noise from a noisy resistor R is to be measured, where R_L is the measurement load. The incremental noise power in R over an incremental frequency band of width df is $P_{WW}(df) = 4kTRdf$ W, by Eq. (73.24). $P_{YY}(df)$ is the integral of $S_{YY}(w)$ over df by Eqs. (73.21), where $S_{YY}(w) = |H(w)|^2 S_{WW}(w)$, by Eq. (73.22). In this case, the transfer function $H(w)$ is nonreactive and does not depend upon the radian frequency (we can factor it out of the integral). Thus,

$$P_{YY}(df) = \int_{-df}^{df} |H(f)|^2 \, (2kTR)df = \{R_L/(R + R_L)^2\}(4kTRdf)$$

To maximize the power measured, let $R_L = R$. The *incremental available power* measured is then $P_{YY}(df) = 4kTR^2df/(4R^2) = kTdf$ [see Ott, 1988, p. 201; Gardner, 1990, p. 288; or Peebles, 1987, p. 227]. Thus, we have the result that incremental available power over bandwidth df depends only on the temperature T.

$$P_{YY}(df) = kTdf \qquad (\text{output power over } df) \qquad (73.27)$$

Albert Einstein used statistical mechanics in 1906 to postulate that the mean kinetic energy per degree of freedom of a particle, $(1/2)mE[v^2(t)]$, is equal to $(1/2)kT$, where m is the mass of the particle, $v(t)$ is its

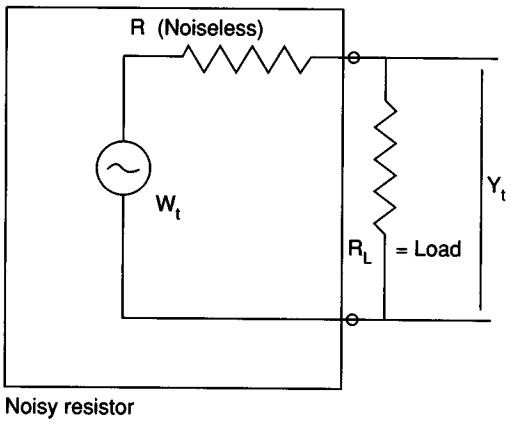


FIGURE 73.7 Measuring thermal noise voltage.

instantaneous velocity in a single dimension, k is Boltzmann's constant, and T is the temperature in kelvin. A shunt capacitor C is charged by the thermal noise in the resistor [see Fig. 73.6(b), where L is taken to be zero]. The average potential energy stored is $(1/2)CE[W(t)^2]$. Equating this to $1/2kT$ and solving, we obtain the mean square power

$$E[W(t)^2] = kT/C \tag{73.28}$$

For example, let $T = 300$K and $C = 50$ pf, and recall that $k = 1.38 \times 10^{-23}$ J/K. Then $E[W(t)^2] = kT/C = 82.8 \times 10^{-12}$, so that the input rms voltage is $\{E[W(t)^2]\}^{1/2} = 9.09$ μV.

Effective Noise and Antenna Noise

Let two series resistors R_1 and R_2 have respective temperatures of T_1 and T_2. The total noise power over an incremental frequency band df is $P_{\text{Total}}(df) = P_{11}(df) + P_{22}(df) = 4kT_1R_1\,df + 4kT_2R_2\,df = 4k(T_1R_1 + T_2R_2)df$. By putting

$$T_E = (T_1R_1 + T_2R_2)/(R_1 + R_2) \tag{73.29}$$

we can write $P_{\text{Total}}(df) = 4kT_E(R_1 + R_2)df$. T_E is called the *effective noise temperature* [see Gardner, 1990, p. 289; or Peebles, 1987, p. 228]. An antenna receives noise from various sources of electromagnetic radiation, such as radio transmissions and harmonics, switching equipment (such as computers, electrical motor controllers), thermal (blackbody) radiation of the atmosphere and other matter, solar radiation, stellar radiation, and galaxial radiation (the ambient noise of the universe). To account for noise at the antenna output, we model the noise with an equivalent thermal noise using an effective noise temperature T_E. The incremental available power (output) over an incremental frequency band df is $P_{YY}(df) = kT_E\,df$, from Eq. (73.27). T_E is often called *antenna temperature*, denoted by T_A. Although it varies with the frequency band, it is usually virtually constant over a small bandwidth.

Noise Factor and Noise Ratio

In reference to Fig. 73.8(a), we define the *noise factor* F = (*noise power output of actual device*)/(*noise power output of ideal device*), where (noise power output of ideal device) = (power output due to thermal noise source). The noise source is taken to be a noisy resistor R at a temperature T, and all output noise measurements must be taken over a resistive load R_L (reactance is ignored). Letting $P_{WW}(B) = 4kTRB$ be the open circuit thermal noise power of the source resistor over a frequency bandwidth B, and noting that the gain of the device is G, the output power due to the resistive noise source becomes $G^2P_{WW}(B) = 4kTRBG^2/R_L$. Now let $Y(t)$ be the output voltage measured at the output across R_L. Then the noise factor is

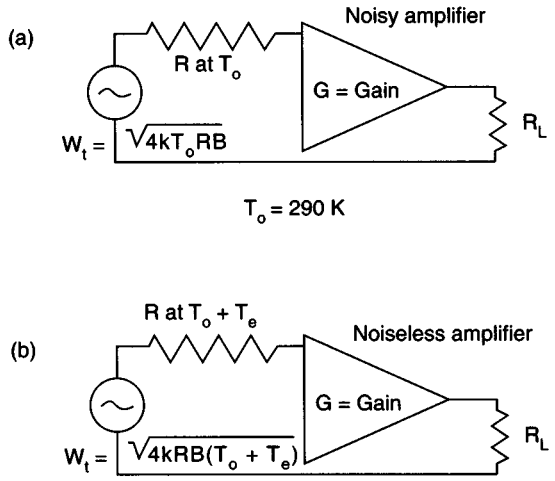


FIGURE 73.8 Equivalent input noise and noise factor.

$$F = (P_{YY}(B)/R_L)/(G^2P_{WW}(B)/R_L) = (P_{YY}(B))/(4kTRBG^2) \tag{73.30}$$

F is seen to be independent of R_L, but not R. To compare two noise factors, the same source must be used. In the ideal noiseless case, $F = 1$, but as the noise level in the device increases, F increases. Because this is a power ratio, we may take the logarithm, called the *noise ratio*, which is

$$N_F = 10\log_{10}(F) = 10\log_{10}(P_{YY}(B)) - 10\log_{10}(4kTRBG^2) \tag{73.31}$$

The noise power output $P_{YY}(B)$ of an actual device is a superposition of the amplified source thermal noise $G^2P_{WW}(B)$ and the device noise, i.e., $P_{YY}(B) = G^2P_{WW}(B)$ + (device noise). The output noise across R_L can be measured by putting a single frequency (in the passband) source generator $S(t)$ as input. First, $S(t)$ is turned off, and the output rms voltage $Y(t)$ is measured and the output power $P_{Y(W)}(B)$ is recorded. This is the sum of the thermal available power and the device noise. Next, $S(t)$ is turned on and adjusted until the output power doubles, i.e., until the output power $P_{Y(W)}(B) + P_{Y(S)}(B) = 2P_{Y(W)}(B)$. This $P_{SS}(B)$ is recorded. Solving for $P_{Y(S)}(B) = P_{Y(W)}(B)$, we substitute this in $F = P_{Y(W)}(B)/(G^2P_{WW}(B))$ to obtain

$$F = P_{Y(S)}(B)/(G^2 \cdot P_{WW}(B)) = (G^2P_{SS}(B))/(G^24kTRB) = P_{SS}(B)/4kTRB \tag{73.32}$$

A better way is to input white noise $W(t)$ in place of $S(t)$ (a noise diode may be used). The disadvantages of noise factors are (1) when the device has low noise relative to thermal noise, the noise factor has value close to 1; (2) a low resistance causes high values; and (3) increasing the source resistance decreases the noise factor while increasing the total noise in the circuit [Ott, 1988, p. 216]. Thus, accuracy is not good. For cascaded devices, the noise factors can be conveniently computed [see Buckingham, 1985, p. 67; or Ott, 1988, p. 228].

Equivalent Input Noise

Shot noise (see below) and other noise can be modeled by equivalent thermal noise that would be generated in an input resistor by increased temperature. Recall that the (maximum) incremental available power (output) in a frequency bandwidth df is $P_{WW}(df) = kTdf$ from Eq. (73.27). Figure 73.8(b) presents the situation. Let the resistor be the noise source at temperature T_o with thermal noise $W(t)$. Then $E[W(t)^2] = 4kT_oRdf$, by Eq. (73.24) (Nyquist's result). Let the open circuit output noise power at R_L be $E[Y(t)^2]$. The incremental available noise power $P_{YY}(df)$ at the output ($R_L = R$) can be considered to be due to the resistor R having a higher temperature and an ideal (noiseless) device, usually an amplifier. We must find a temperature T_e at which a pseudothermal

noise power $E[W_e(t)^2] = 4kT_eRdf$ yields the extra "input" noise power. Let $V(t) = W(t) + W_e(t)$. Then $P_{VV}(df) = 4kT_oRdf + 4kT_eRdf = 4k(T_o + T_e)Rdf$ W, from Eq. (73.24). T_e is called the *equivalent input noise temperature*. It is related to the noise factor F by $T_e = 290(F - 1)$. In cascaded amplifiers with gains $G_1, G_2, \ldots$ and equivalent input noise temperatures $T_{e1}, T_{e2}, \ldots$, the total equivalent input noise temperature is

$$T_{e(\text{Total})} = T_{e1} + T_{e2}/G_1 + T_{e3}/G_1G_2 + \ldots \tag{73.33}$$

[see Gardner, 1990, p. 289].

Other Electrical Noise

Thermal noise and shot noise (which can be modeled by thermal noise with equivalent input noise) are the main noise sources. Other noises are discussed in the following paragraphs.

Shot Noise

In a conductor under an external emf, there is an average flow of electrons, holes, photons, etc. In addition to this induced net flow and thermal noise, there is another effect. The potential differs across the boundaries of metallic grains and particles of impurities, and when the kinetic energy of electrons exceeds this potential, electrons jump across the barrier. This summed random flow is known as *shot noise* [see Gardner, 1990, p. 239; Ott, 1988, p. 208]. The shot effect was analyzed by Schottky in 1918 as $I_{\text{sh}} = (2qI_{\text{dc}}B)^{1/2}$, where $q = 1.6 \times 10^{-19}$ coulombs per electron, I_{dc} = average dc current in amperes, and B = noise bandwidth (Hz).

Partition Noise

Partition noise is caused by a parting of the flow of electrons to different electrodes into streams of randomly varying density. Suppose that electrons from some source S flow to destination electrodes A and B. Let $n(A)$ and $n(B)$ be the average numbers of electrons per second that go to nodes A and B respectively, so that $n(S) = n(A) + n(B)$ is the average total number of electrons emitted per second. It is a success when an electron goes to A, and the probability of success on a single trial is p, where

$$p = n(A)/n(S),\ 1 - p = n(B)/n(S) \tag{73.34}$$

The current to the respective destinations is $I(A) = n(A)q$, $I(B) = n(B)q$, where q is the charge of an electron, so that $I(A)/I(S) = p$ and $I(B)/I(S) = 1 - p$. Using the binomial model, the average numbers of successes are $E[n(A)] = n(S)p$ and $E[n(B)] = n(S)(1 - p)$. The variance is $\text{Var}(n(A)) = n(S)p(1 - p) = \text{Var}(n(B))$ (from the binomial formula for variance). Therefore, substitution yields

$$\text{Var}(I(A)) = q^2[n(S)p(1 - p)] = q^2n(S)\{I(A)I(B)/[I(A) + I(B)]\} \tag{73.35}$$

Partition noise applies to pentodes, where the source is the cathode, A is the anode (success), and B is the grid. For transistors, the source is the emitter, A is the collector, and B represents recombination in the base. In photo devices, a photoelectron is absorbed, and either an electron is emitted (a success) or not. Even a partially silvered mirror can be considered to be a partitioner: the passing of a photon is a success and reflection is a failure. While the binomial model applies to partitions with destinations A and B, multinomial models are analogous for more than two destinations.

Flicker, Contact, and Burst Noise

J.B. Johnson first noticed in 1925 that noise across thermionic gates exceeded the expected shot noise at lower frequencies. It is most noticeable up to about 2 kHz. The psdf of the extra noise, called *flicker noise*, is

$$S(f) = I^2/\alpha f,\ f > 0 \tag{73.36}$$

where I is the dc current flowing through the device and f is the positive frequency. Empirical values of α are about 1 to 1.6 for different sources. These sources vary but include the irregularity of the size of macro regions

of the cathode surface, impurities in the conducting channel, and generation and recombination noise in transistors. In the early days of transistors, this generation-recombination was of great concern because the materials were not of high purity. Flicker noise occurs in thin layers of metallic or semiconducting material, solid state devices, carbon resistors, and vacuum tubes [see Buckingham, 1985, p. 143]. It includes *contact noise* because it is caused by fluctuating conductivity due to imperfect contact between two surfaces, especially in switches and relays. Flicker noise may be high at low frequencies.

Burst noise is also called *popcorn noise:* audio amplifiers sound like popcorn popping in a frying pan background (thermal noise). Its characteristic is $1/f^n$ (usually $n = 2$), so its power density falls off rapidly, where f is frequency. It may be problematic at low frequencies. The cause is manufacturing defects in the junction of transistors (usually a metallic impurity).

Barkhousen and Other Noise

Barkhousen noise is due to the variations in size and orientation of small regions of ferromagnetic material and is especially noticeable in the steeply rising region of the hysteresis loop. There is also secondary emission, photo and collision ionization, etc.

Measurement and Quantization Noise

Measurement Error

The measurement X_t of a signal $X(t)$ at any t results in a measured value $X_t = x$ that contains error, and so is not equal to the true value $X_t = x_T$. The probability is higher that the magnitude of $e = (x - x_T)$ is closer to zero. The bell-shaped Gaussian probability density $f(e) = [1/(2\pi\sigma^2]^{1/2}\exp(-e^2/2\pi\sigma)$ fits the error well. This noise process is stationary over time. The expected value is $\mu_e = 0$, the mean-square error is σ_e^2, and the rms error is σ_e. Its instantaneous power at time t is σ_e^2. To see this, the error signal $e(t) = (x - x_T)$ has instantaneous power per Ω of

$$P_i = e(t)\,i(t) = e(t)[\,e(t)/R] = e^2(t) \tag{73.37}$$

where $R = 1\ \Omega$ and $i(t)$ is the current. The average power is the summed instantaneous power over a period of time T, divided by the time, taken in the limit as $T \to \infty$, i.e.,

$$P_{\text{ave}} = \lim_{T\to\infty}(1/T)\int_0^T e^2(t)dt$$

This average power can be determined by sampling on known signal values and then computing the sample variance (assuming ergodicity: see Gardner [1990, p. 163]). The error and signal are probabilistically independent (unless the error depends on the values of X). The signal-to-noise power ratio is computed by $S/N = P_{\text{signal}}/P_{\text{ave}}$.

Quantization Noise

Quantization noise is due to the digitization of an exact signal value $v_t = v(t)$ captured at sampling time t by an A/D converter. The binary representation is $b_{n-1}b_{n-2}\ldots b_1b_0$ (an n-bit word). The n-bit digitization has 2^n different values possible, from 0 to 2^n–1. Let the voltage range be R. The *resolution* is $dv = R/2^n$. Any voltage v_t is coded into the nearest lower binary value x_b, where the error $e = x_t - x_b$ satisfies $0 \le e \le dv$. Thus, the errors e are distributed over the interval $[0, dv]$ in an equally likely fashion that implies the uniform distribution on $[0, dv]$. The expected value of $e = e_t = e(t)$ at any time is $\mu_e = dv/2$, and the variance is $\mu_e^2 = dv^2/12$ (the variance of a uniform distribution on an interval $[a,b]$ is $\sigma = (b - a)^2/12$). Thus the noise is ws and the power of quantization noise is

$$\begin{aligned}\sigma_e^2 &= \int_0^{dv}(e - dv/2)^2(1/dv)de \\ &= (e - dv/2)^3/3dv\,\Big|_0^{dv} = [(dv)^3 + (dv)^3]/24dv = dv^2/12\end{aligned} \tag{73.38}$$

We can find the signal-to-noise voltage ratio for the total range R via $R/(dv/(12)^{1/2}) = 2^n dv/(dv/(12)^{1/2}) = 2^n(12)^{1/2}$. The power ratio is the square of this, which is $(2^{2n})(12)$. In decibels this becomes $(S/N)_{\text{dB}} = 10\log_{10}(2^{2n}\cdot 12) = 10\log_{10}(12) + 20n\log_{10}(2) = 10.8 + 6.02n$. Thus, quantization S/N power ratio depends directly upon the number of bits n in that the higher S/N power ratio is better, just as we would have expected.

Coping with Noise

External interference is ubiquitous. Intrinsic noise is present up to the incremental available power at temperatures above absolute zero, and other intrinsic noises depend on material purity and connection integrity. Processing error is always introduced in some form.

External Sources

Standard defenses are (1) shielding of lines and circuits, (2) twisted wire pairs or coaxial cables, (3) short lines and leads, (4) digital regeneration at waypoints of digital signals, (5) narrowband signals, (6) correlation of received signals with multipaths, and (7) adaptive notch filtering to eliminate interference at known frequencies; e.g., the second harmonic of 60-Hz ac power lines may interfere with biological microvoltage measurements but could be eliminated via adaptive notch filtering. Ferrite beads can dampen interference [Barnes, 1987]. Digital signal processing, spectral shaping filters [see Brown, 1983], and frequency-shift filters [see Gardner, 1990, p. 400] can be used to lower noise power. Kalman filtering is a powerful estimation method, and frequency-shift filtering is a newer technique for discriminating against both measurement error (e.g., in system identification applications) and extrinsic sources of both noise and interference [Gardner, 1990, p. 400].

Intrinsic Sources

Strategies for minimizing intrinsic noise are (a) small bandwidth B, (b) small resistances R, (c) low temperature T (higher temperatures can be devastating), (d) low voltage and currents (CMOS transistors), (e) modern materials of high purity, (f) wrapped wire resistors (thermal noise is the same, but other noise will be less), (g) fewer and better connections (of gold), (h) smaller circuits of lower power, and (i) shunt capacitors to reduce noise bandwidth. Greater purity of integrated circuit materials nowadays essentially reduces intrinsic noise to thermal noise. Better design and materials are the keys to lower noise.

Processing Sources

Processing errors can be reduced by using higher resolution of analog-to-digital converters, i.e., more bits to represent each value. This lowers the quantization error power. Measurement error can be reduced while using the same instruments by taking multiple measurements and averaging. Other estimation/correlation can yield better values (e.g., the Global Positioning System location determination can be reduced from meters to a few centimeters by multiple measurement estimation).

Defining Terms

Autocorrelation: A function associated with a random signal $X(t)$ that is defined on pairs of time instants t_1 and t_2 and whose value is the expected value of the product of the random variables $X(t_1)$ and $X(t_2)$, i.e., $R_{XX}(t_1,t_2) = E[X(t_1)X(t_2)]$. For weakly stationary random signals, it depends only on the offset $\tau = t_2 - t_1$, so we write $R_{XX}(\tau) = E[X(t)X(t+\tau)]$.

Noise: A signal $N(t)$ whose value at any time t is randomly selected by events beyond our control. At any time instant t, $N(t)$ is a random variable N_t with a probability distribution that determines the relative frequencies at which N_t assumes values. The statistics of the family of random variables $\{N_t\}$ may be constant (stationary) over time (the usual case) or may vary.

Power spectral density: The Fourier transform of the power $X^2(t)$ does not necessarily exist, but it does for $X_T^2(t)/2T$ ($X_T(t) = 0$ for $|t| > T$, $= X(t)$ elsewhere), for any $T > 0$. Letting $T \to \infty$, the expected value of the Fourier transforms $E[\mathrm{F}[X_T^2(t)/2T] = \mathrm{F}[E[X_T^2(t)]/2T$ goes to the limit of the average power in $X(t)$ over $-T$ to T, known as the power spectral density function $S_{xx\,(w)}$. Summed up over all frequencies, it gives the total power in the signal $X(t)$.

Random process: (signal): A signal that is either a noise, an interfering signal $s(t)$, or a sum of these such as $X(t) = s_1(t) + \ldots + s_m(t) + N_1(t) + \ldots + N_n(t)$.

Realization: A trajectory $\{(t, x_t)\colon X(t) = x_t\}$ determined by the actual outcomes $\{x_t\}$ of values from a random signal $X(t)$, where $X(t) = x_t$ at each instant t. A trajectory is also called a *sample function* of $X(t)$.

Weakly stationary (ws) random process (signal): A random signal whose first- and second-order statistics remain stationary (fixed) over time.

Related Topic

15.2 Speech Enhancement and Noise Reduction

References

J. R. Barnes, *Electronic System Design: Interference and Noise Control,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering,* New York: Wiley, 1983.

M. J. Buckingham, *Noise in Electronic Devices and Systems,* New York: Halstead Press, 1985.

W. A. Gardner, *Introduction to Random Processes,* 2nd ed., New York: McGraw-Hill, 1990.

J. B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.,* vol. 29, pp. 367–368, 1927.

J. B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.,* vol. 32, pp. 97–109, 1928.

H. Nyquist, "Thermal agitation of electric charge in conductors," *Phys. Rev.,* vol. 32, pp. 110–113, 1928.

H. W. Ott, *Noise Reduction Techniques in Electronic Systems,* 2nd ed., New York: Wiley-Interscience, 1988.

P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles,* 2nd ed., New York: McGraw-Hill, 1987.

Further Information

The IEEE Individual Learning Program, *Random Signal Analysis with Random Processes and Kalman Filtering,* prepared by Carl G. Looney (IEEE Educational Activities Board, PO Box 1331, Piscataway, NJ 08855-1331, 1989) contains a gentle introduction to estimation and Kalman filtering.

Also see H. M. Denny, *Getting Rid of Interference,* IEEE Video Conference, Educational Activities Dept., Piscataway, NJ, 08855-1331, 1992.

73.3 Stochastic Processes

Carl G. Looney

Introduction to Random Variables

A *random variable* (rv) A is specified by its *probability density function* (pdf)

$$f_A(a) = \lim_{\varepsilon\to 0}\ (1/\varepsilon)P[a - (\varepsilon/2) < A \le a + (\varepsilon/2)]$$

In other words, the rectangular area $\varepsilon \cdot f_A(a)$ approximates the probability $P[(A \le a + (\varepsilon/2)] - P[a - (\varepsilon/2) < A]$. The joint pdf of two rv's A and B is specified by

$$f_{AB}(a,b) = \lim_{\varepsilon\to 0}\ (1/\varepsilon^2)P[\,a - \varepsilon < A \le a + (\varepsilon/2) \text{ and } b - \varepsilon < B \le b + (\varepsilon/2)]$$

A similar definition holds for any finite number of rv's.

The *expected value* $E[A]$, or *mean* μ_A, of a rv A is the first moment of the pdf, and the *variance* of A is the second centralized moment, defined respectively by

$$\mu_A = E[A] \equiv \int_{-\infty}^{\infty} a f_A(a)da \tag{73.39a}$$

$$\sigma_A^2 = E[(A-\mu_A)^2] \equiv \int_{-\infty}^{\infty} (a-\mu_A)^2 f_A(a)da \tag{73.39b}$$

The square root of the variance is the *standard deviation*, which is also called the *root mean square* (rms) error. The *covariance* of two rv's A and B is the second-order centralized joint moment

$$\sigma_{AB} = E[(A-\mu_A)(B-\mu_B)] \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a-\mu_A)(b-\mu_B) f_{AB}(a,b)dadb \tag{73.40}$$

The noncentralized second moments are the *mean-square value* and the *correlation*, respectively,

$$E[A^2] = \int_{-\infty}^{\infty} a^2 f_A(a)da = \sigma_A^2 + \mu_A^2, \quad E[AB] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} abf_{AB}(a,b)dadb = \sigma_{AB} + \mu_A\mu_B$$

A set of rv's A, B, and C is defined to be *independent* whenever their joint pdf factors as

$$f_{ABC}(a,b,c) = f_A(a)f_B(b)f_C(c) \tag{73.41}$$

for all a, b, and c, and similarly for any finite set of rv's. A *weak independence* holds when the second moment of the joint pdf, the correlation, factors as $E[AB] = E[A]E[B]$, so that $\sigma_{AB} = 0$, in which case the rv's are said to be *uncorrelated*. The covariance of A and B is a measure of how often A and B vary together (have the same sign), how often they vary oppositely (different signs), and by how much, on the average over trials of outcomes. To standardize so that units do not influence the measure of dependence, we use the *correlation coefficient*

$$\rho_{AB} \equiv \sigma_{AB}/\sigma_A\sigma_B$$

The accuracy of approximating a rv A as a linear function of another rv B, $A \approx cB + d$, for real coefficients c and d, is found by minimizing the mean-square error $\varepsilon = E\{[A - (cB + d)]^2\}$. Upon squaring and taking the expected values, we can obtain $\varepsilon_{\min} = \sigma_A^2(1 - |\rho_{AB}|^2)$, which shows $|\rho_{AB}|$ to be a measure of the degree of linear relationship between A and B. Because $\varepsilon_{\min} \geq 0$, this shows that $|\rho_{AB}| \leq 1$, which demonstrates the Cauchy-Schwarz inequality

$$|E[AB]| \leq \{E[A^2]E[B^2]\}^{1/2} \tag{73.42}$$

When $|\rho_{AB}| = 1$, then knowledge of one of A or B completely determines the other ($c \neq 0$), and so A and B are completely dependent, while $|\rho_{AB}| = 0$ indicates there is no linear relationship, i.e., that A and B are uncorrelated.

An important result is the *fundamental theorem of expectation*: if $g(\cdot)$ is any real function, then the expected value of the rv $B = g(A)$ is given by

$$E[B] = E[g(A)] = \int_{-\infty}^{\infty} g(a)f_A(a)da \tag{73.43}$$

Stochastic Processes

A **stochastic** (or *random*) **process** is a collection of random variables $\{X_t : t \in T\}$, indexed on an ordered set T that is usually a subset of the real numbers or integers. Examples are the Dow-Jones averages $D(t)$ at each time t, the pressure $R(x)$ in a pipe at distance x, or a noise voltage $N(t)$ at time t. A process is thus a *random function* $X(t)$ of t whose value at each t is drawn randomly from a range of outcomes for the rv $X_t = X(t)$ according to a probability distribution for X_t. A trajectory $\{x_t : t \in T\}$ of outcomes over all $t \in T$, where $X_t = x_t$ is the realized value at each t, is called a **sample function** (or *realization*) of the process. A stochastic process $X(t)$ has mean

TABLE 73.1 Continuous/Discrete Classification of Stochastic Processes

	X Values	
T Values	Continuous	Discrete
Continuous	Continuous stochastic processes	Discrete valued stochastic processes
Discrete	Continuous random sequences	Discrete valued random sequences

value $E[X(t)] = \mu(t)$ at time t, and **autocorrelation function** $R_{XX}(t, t+\tau) = E[X(t)X(t+\tau)]$ at times t and $t+\tau$, the correlation of two rv's at two times offset by τ. When $\mu(t) = 0$ for all t, the autocorrelation function equals the *autocovariance function* $C_{XX}(t, t+\tau) = E[(X(t) - \mu(t))(X(t+\tau) - \mu(t+\tau))]$.

A process $X(t)$ is completely determined by its joint pdf's $f_{X(t(1)) \ldots X(t(n))}(x(t_1), \ldots, x(t_n))$ for all time combinations $t_1, \ldots, t_n$ and all positive integers n (where $t(j) = t_j$). When the rv's $X(t)$ are *iid* (independent, identically distributed), then knowledge of one pdf yields the knowledge of all joint pdf's. This is because we can construct the joint pdf by factorization, per Eq. (73.41).

Classifications of Stochastic Processes

The ordered set T can be continuous or discrete, and the values that $X(t)$ assumes at each t may also be continuous or discrete, as shown in Table 73.1.

In another classification, a stochastic process $X(t)$ is *deterministic* whenever an entire sample function can be determined from an initial segment $\{x;\ t \le t_1\}$ of $X(t)$. Otherwise, it is *nondeterministic* [see Brown, 1983, p. 79; or Gardner, 1990, p. 304].

Stationarity of Processes

A stochastic process is *nth order* (*strongly*) *stationary* whenever all joint pdf's of n and fewer rv's are independent of all translations of times $t_1, \ldots, t_n$ to times $\tau + t_1, \ldots, \tau + t_n$. The case of $n = 2$ is very useful. Another type of process is called **weakly stationary** (ws), or *wide-sense stationary*, and is defined to have first- and second-order moments that are independent of time (see Section 73.2 on noise). These satisfy (1) $\mu(t) = \mu$ (constant) for all t, and (2) $R_{XX}(t, t+\tau) = R_{XX}(t+s, t+s+\tau)$ for all values of s. For $s = -t$, this yields $R_{XX}(t, t+\tau) = R_{XX}(0, 0+\tau)$, which is abbreviated to $R_{XX}(\tau)$. $X(t)$ is *uncorrelated* whenever $C_{XX}(\tau) = 0$ for τ not zero [we say $X(t)$ has *no memory*]. If $X(t)$ is correlated, then $X(t_1)$ depends on values $X(t)$ for $t \ne t_1$ [$X(t)$ has *memory*].

Some properties of autocorrelation functions for ws processes follow. First, $|R_{XX}(\tau)| \le R_{XX}(0)$, $-\infty < \tau < \infty$, as can be seen from Eq. (73.42) with $|R_{XX}(\tau)|^2 = E[X(0)X(\tau)] \le E[X(0)^2]E[X(\tau)^2] = R_{XX}(0)R_{XX}(\tau)$. Next, $R_{XX}(\tau)$ is real and even, i.e., $R_{XX}(-\tau) = R_{XX}(\tau)$, which is evident from substituting $s = t - \tau$ in $E[X(s)X(s+\tau)]$and using time independence. If $X(t)$ has a periodic component, then $R_{XX}(\tau)$ will have that same periodic component, which follows from the definition. Finally, if $X(t)$ has a nonzero mean μ and no periodic components, then the variance goes to zero (the memory fades) and so $\lim_{\tau\to\infty} R_{XX}(\tau) \to 0 + \mu^2 = \mu^2$.

Gaussian and Markov Processes

A process $X(t)$ is defined to be *Gaussian* if for every possible finite set $\{t_1, \ldots, t_n\}$ of times, the rv's $X(t_1), \ldots, X(t_n)$ are *jointly Gaussian*, which means that every linear combination $Z = a_1X(t_1) + \ldots + a_nX(t_n)$ is a Gaussian rv, defined by the Gaussian pdf

$$f_Z(z) = \left[1/(\sigma_Z\sqrt{2\pi})\right] \exp\{-(z - \mu_Z)^2/2\sigma_Z^2\} \tag{73.44}$$

In case the n rv's are *linearly independent*, i.e., $Z = 0$ only if $a_1 = \cdots = a_n = 0$, the joint pdf has the Gaussian form [see Gardner, 1990, pp. 39–40]

$$f_{X(t(1))\ldots X(t(n))}(x_1, \ldots, x_n) = [1/(2\pi)^{n/2}|C|^{1/2}] \cdot \exp\{-(\mathrm{x} - \boldsymbol{\mu})^t C^{-1}(\mathrm{x} - \boldsymbol{\mu})\} \quad (73.45)$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is a column vector, $\mathbf{x}^t$ is its transpose, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is the mean vector, C is the *covariance matrix*

$$C = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \vdots & \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{pmatrix} \quad (73.46)$$

and $|C|$ is the determinant of C. If $X(t_1), \ldots, X(t_n)$ are linearly dependent, then the joint pdf takes on a form similar to Eq. (73.45), but contains impulses [see Gardner, 1990, p. 40].

A weakly stationary Gaussian process is strongly stationary to all orders *n:* all Gaussian joint pdf's are completely determined by their first and second moments by Eq. (73.45), and those moments are time independent by weak stationarity, and so all joint pdf's are also. Every second-order strongly stationary stochastic process $X(t)$ is also weakly stationary because the time translation independence of the joint pdf's determines the first and second moments to have the same property. However, non-Gaussian weakly stationary processes need not be strongly second-order stationary.

Rather than with pdf's, a process $X(t)$ may be specified in terms of conditional pdf's

$$f_{X(t(1))\ldots X(t(n))}(x_1, \ldots, x_n) = f_{X(t(n))|X(t(n-1))}(x_n|x_{n-1}) \cdot \ldots \cdot f_{X(t(2))|X(t(1))}(x_2|x_1) f_{X(t(1))}(x_1)$$

by successive applications of Bayes' law, for $t_1 < t_2 < \cdots < t_n$. The conditional pdf's satisfy

$$f_{A|B}(a|b) = f_{AB}(a,b)/f_B(b) \quad (73.47)$$

The conditional factorization property may satisfy

$$f_{X(t(n))|X(t(n-1))\ldots X(t(1))}(x_n|x_{n-1}, \ldots, x_1) = f_{X(t(n))|X(t(n-1))}(x_n|x_{n-1}) \quad (73.48)$$

which indicates that the pdf of the process at any time t_n, given values of the process at any number of previous times $t_{n-1}, \ldots, t_1$, is the same as the pdf at t_n given the value of the process at the most recent time t_{n-1}. Such an $X(t)$ is called a *first-order Markov process*, in which case we say the process remembers only the previous value (the previous value has influence). In general, an *nth-order Markov* process remembers only the n most recent previous values. A first-order Markov process can be fully specified in terms of its first-order conditional pdf's $f_{X(t)|X(s)}(x_t, x_s)$ and its unconditional first-order pdf at some initial time t_0, $f_{X(t(0))}(x_0)$.

Examples of Stochastic Processes

Figure 73.9 shows two sample functions of nonstationary processes. Now consider the discrete time process $X(k) = A$, for all $k \geq 0$, where A is a rv (a *random initial condition*) that assumes a value 1 or –1 with respective probabilities p and $1 - p$ at $k = 0$. This value does not change, once the initial random draw is done at $k = 0$. This stochastic sequence has two sample functions only, the constant sequences $\{-1\}$ and $\{1\}$. The expected value of $X(k)$ at any time k is $E[X(k)] = E[A] = p \cdot 1 + (1 - p) \cdot (-1) = 2p - 1$, which is independent of k. The autocorrelation function is, by definition, $E[X(k)X(k + m)] = E[A \cdot A] = E[A^2] = p \cdot 1^2 + (1 - p) \cdot (-1)^2 = 1$ which is also independent of time k. Thus $X(k)$ is perfectly correlated for all time (the process has *infinite memory*). This process is deterministic.

For another example, put $X(t) = (c) \cdot \cos(wt + \Phi)$, where Φ is the uniform rv on $(-\pi, \pi)$. Then $X(t)$ is a function of the rv Φ (as well as t), so by use of Eq. (73.39a), we obtain

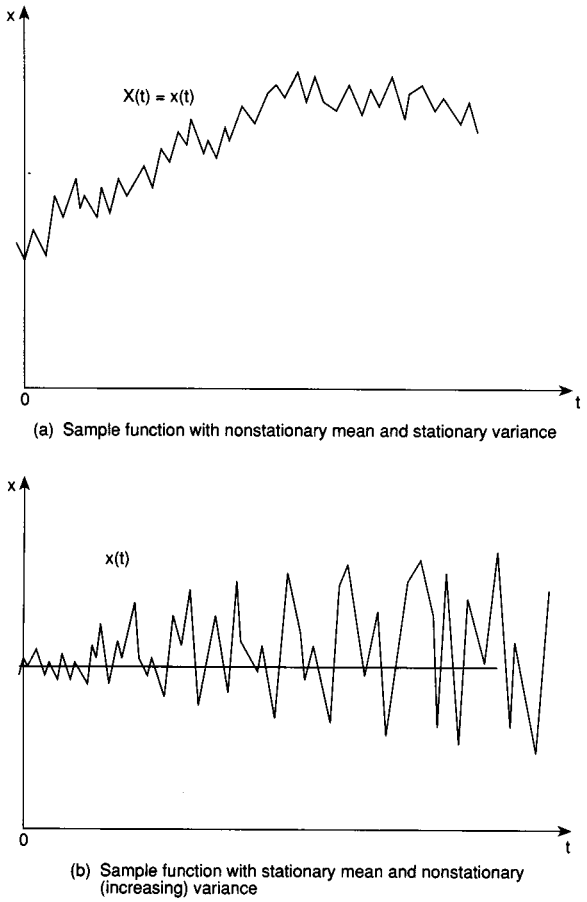


FIGURE 73.9 Examples of nonstationary processes.

$$E[X(t) = c \cdot \int_{-\pi}^{\pi} \cos(wt + \phi) f_{\Phi}(\phi) d\phi = (c/2\pi) \sin(wt + \phi)\Big|_{-\pi}^{\pi} = 0$$

Therefore, the mean does not vary with time t. The autocorrelation is

$$\begin{aligned} R_{XX}(t, t + \tau) &= E[(c) \cdot \cos(wt + \Phi)(c) \cdot \cos(wt + w\tau + \Phi)] \\ &= c^2 E[\cos(wt + \Phi) \cos(wt + w\tau + \Phi)] \\ &= c^2 \int_{-\pi}^{\pi} \cos(wt + \phi) \cos(wt + w\tau + \Phi) f_{\Phi}(\phi) d\phi \\ &= (c^2/2) \int_{-\pi}^{\pi} \{\cos(2wt + 2\phi + w\tau) + \cos(w\tau)\}(1/2\pi) d\phi \\ &= (c^2/4\pi) \cdot \{\sin(\Theta + 2\pi) - \sin(\Theta - 2\pi) + \cos(w\tau) \cdot 2\pi\} \\ &= (c^2/4\pi) \cdot \{\cos(w\tau) \cdot 2\pi\} = (c^2/2) \cos(w\tau) \end{aligned}$$

[using $\cos(x)\cos(y) = \frac{1}{2}\{\cos(x + y) + \cos(x - y)\}$ and letting $\Theta = 2wt + 2\Phi + w\tau$]. Therefore, $X(t)$ is ws. The autocorrelation is periodic in the offset variable τ.

Now consider the example $X(t) = A\cos(2\pi f_0 t)$ for each t, where f_0 is a constant frequency, and the amplitude A is a random initial condition as given above. There are only two sample functions here: (1) x(t) = $\cos(2\pi f_0 t)$ and (2) $x(t) = -\cos(2\pi f_0 t)$. A related example is $X(t) = A\cos(2\pi f_0 t + \Phi)$, where A is given above, the phase Φ is the uniform random variable on $[0,\pi]$, and A and Φ are independent. Again, Φ and A do not depend on time (initial random conditions). Thus, the sample functions for $X(t)$ are $x(t) = \pm\cos(2\pi ft + \phi)$, where $\Phi = \phi$ is the value assumed initially. There are infinitely many sample functions because of the phase. Equation (73.39b) and the independence of A and Φ yield

$$\begin{aligned} E[X(t)] &= E[A\cos(2\pi ft + \Phi)] = E[A]E[g(\Phi)] = \mu_A \int_0^{\pi} \cos(2\pi ft + \phi)(1/\pi)d\phi \\ &= (\mu_A/\pi)\sin(2\pi t + \phi)\Big|_{\phi=0}^{\pi} = (\mu_A/\pi)[\sin(2\pi t + \pi) - \sin(2\pi t)] \\ &= (\mu_A/\pi)[\sin(-2\pi t) - \sin(2\pi t)] = (-2\mu_A/\pi)\sin(2\pi t) \end{aligned}$$

which is dependent upon time. Thus, $X(t)$ is not ws.

Next, let $X(t) = [a + S(t)]\cos[2\pi ft + \Phi]$, where the signal $S(t)$ is a nondeterministic stochastic process. This is an amplitude-modulated sine wave carrier. The carrier $\cos[2\pi ft + \Phi]$ has random initial condition Φ and is deterministic. Because $S(t)$ is nondeterministic, $X(t)$ is also. The expected value $E[X(t)] = E[a + S(t)]E[\cos(2\pi ft + \Phi)]$ can be found as above by independence of $S(t)$ and Φ.

Finally, let $X(t)$ be uncorrelated ($E[X(t)X(t + \tau)] = 0$ for τ not zero) such that each rv $X(t) = X_t$ is Gaussian with zero mean and variance $\sigma^2(t) = t$, for all $t > 0$. Any realized sample function $x(t)$ of $X(t)$ cannot be predicted in any average sense based on past values (uncorrelated Gaussian random variables are independent). The variance grows in an unbounded manner over time, so $X(t)$ is neither stationary nor deterministic. This is called the *Wiener* process.

A useful model of a ws process is that for which $\mu = 0$ and $R_{XX}(\tau) = \sigma_X^2\exp(-\alpha|\tau|)$. If this process is also Gaussian, then it is strongly stationary and all of its joint pdf's are fully specified by $R_{XX}(\tau)$. In this case it is also a first-order Markov process and is called the *Ornstein-Uhlenbeck* process [see Gardner, 1990, p. 102]. Unlike white noise, many real-world ws stochastic processes are correlated ($|R_{XX}(t, t + \tau)| > 0$) for $|\tau| > 0$. The autocorrelation either goes to zero as τ goes to infinity, or else it has periodic or other nondecaying memory. We consider ws processes henceforth [for nonstationary processes, see Gardner, 1990]. We will also assume without loss of generality that $\mu = 0$.

Linear Filtering of Weakly Stationary Processes

Let the ws stochastic process $X(t)$ be the input to a linear time-invariant stable filter with impulse response function $h(t)$. The output of the filter is also a ws stochastic process and is given by the convolution

$$Y(t) = h(t) * X(t) = \int_{-\infty}^{\infty} h(s)X(t-s)ds \tag{73.49}$$

The mean of the output process is obtained by using the linearity of the expectation operator [see Gardner, 1990, p. 32]

$$\begin{aligned} \mu_Y &= E[Y(t)] = E\left[\int_{-\infty}^{\infty} h(s)X(t - s)ds\right] = \int_{-\infty}^{\infty} h(s)E[X(t - s)]ds = \int_{-\infty}^{\infty} h(s)\mu_X ds \\ &= \mu_X \int_{-\infty}^{\infty} h(s)ds = \mu_X \cdot H(0) \end{aligned} \tag{73.50}$$

where $H(f) = \int_{-\infty}^{\infty} h(t)\,e^{-j2\pi ft}dt$ is the filter transfer function and $H(0)$ is the dc gain of the filter.

The autocorrelation of the output process, obtained by using the linearity of $E[\cdot]$, is

$$\begin{aligned} R_{YY}(\tau) &= E[Y(t)Y(t+\tau)] = E\left[\int_{-\infty}^{\infty} h(v)X(t-v)dv \int_{-\infty}^{\infty} h(u)X(t+\tau-u)du\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[X(t-v)X(t+\tau-u)]h(v)h(u)dvdu \\ &= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} R_{XX}(\tau-u+v)h(u)du\right\}h(v)dv \\ &= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} R_{XX}([\tau-(-v)]-u)h(u)du\right\}h(-v)(dv) \\ &= \int_{-\infty}^{\infty}\left\{R_{XX}(\tau+v) * h(\tau+v)\right\}h(-v)dv \\ &= [R_{XX}(\tau) * h(\tau)] * h(-\tau) = R_{XX}(\tau) * [h(\tau) * h(-\tau)] = R_{XX}(\tau) * r_h(\tau) \end{aligned} \tag{73.51}$$

where $r_h(\tau) = \int_{-\infty}^{\infty} h(\tau + u)h(u)du$. However, $r_h(\tau)$ has Fourier transform $H(f)H^*(f) = |H(f)|^2$, because the Fourier transform of the convolution of two functions is the product of their Fourier transforms, and the Fourier transform of $h(-\tau)$ is the complex conjugate $H^*(f)$ of the Fourier transform $H(f)$ of $h(\tau)$. Thus, the Fourier transform of $R_{YY}(\tau)$, denoted by $\mathrm{F}\{R_{YY}(\tau)\}$, is

$$\mathbf{F}\{R_{YY}(\tau)\} = \mathbf{F}\{R_{XX}(\tau) * h(\tau) * h(-\tau)\} = \mathbf{F}\{R_{XX}(\tau)\} \cdot H(f)H^*(f) = \mathbf{F}\{R_{XX}(\tau)\} \cdot |H(f)|^2$$

Upon defining the functions

$$S_{XX}(f) \equiv \mathbf{F}\{R_{XX}(\tau)\},\ S_{YY}(f) \equiv \mathbf{F}\{R_{YY}(\tau)\} \tag{73.52}$$

we can also determine $R_{YY}(\tau)$ via the two steps

$$S_{YY}(f) = S_{XX}(f) \cdot |H(f)|^2 \tag{73.53}$$

$$R_{YY}(\tau) = \mathbf{F}^{-1}\{S_{YY}(f)\} = \int_{-\infty}^{\infty} S_{YY}(f)e^{j2\pi f\tau}df \tag{73.54}$$

Equations (73.52) define the *power spectral density functions* (psdf's) $S_{XX}(f)$ for $X(t)$ and $S_{YY}(f)$ for $Y(t)$. Thus, $R_{XX}(\tau)$ and $S_{XX}(f)$ are Fourier transform pairs, as are $R_{YY}(\tau)$ and $S_{YY}(f)$ (see Eq. 73.20). Further, the psdf $S_{XX}(f)$ of $X(t)$ is a power spectrum (in an average sense). If $X(t)$ is the voltage dropped across a 1-Ω resistor, then $X^2(t)$ is the instantaneous power dissipation in the resistance. Consequently, $R_{XX}(0) = E[X^2(t)]$ is the expected power dissipation over all frequencies, i.e., by Eq. (73.54) with $\tau = 0$, we have

$$R_{XX}(0) = \int_{-\infty} S_{XX}(f)df$$

We want to show that when we pass $X(t)$ through a narrow bandpass filter with a bandwidth δ centered at the frequency $\pm f_0$, the expected power at the output terminals, divided by the bandwidth δ, is $S_{XX}(f_0)$ in the limit as $\delta \to 0$. This shows that $S_{XX}(f)$ is a density function (whose area is the total expected power over all frequencies, just as the area under a pdf is the total probability). This result that $R_{XX}(\tau)$ and $S_{XX}(f)$ are a Fourier transform pair is known as the *Wiener-Khinchin* relation [see Gardner, 1990, p. 230].

To verify this relation, let $H(f)$ be the transfer function of an ideal bandpass filter, where

$$H(f) = 1, |f - f_0| < \delta/2; \qquad H(f) = 0, \text{ otherwise}$$

Let Y(t) be the output of the filter. Then Eqs. (73.54) and (73.53) provide

$$E[Y^2(t)] = R_{YY}(0) = \int_{-\infty}^{\infty} S_{YY}(f)df = \int_{-\infty}^{\infty} S_{XX}(f) \mid H(f) \mid^2 df$$

$$= \int_{f_0-\delta/2}^{f_0+\delta/2} S_{XX}(f)df + \int_{-f_0-\delta/2}^{-f_0+\delta/2} S_{XX}(f)df$$

Dividing by 2δ and taking the limit as $\delta \to 0$ yields $(1/2)S_{XX}(f_0) + (1/2)S_{XX}(-f_0)$, which becomes $S_{XX}(f_0)$ when we use the fact that psdf's are even and real functions (because they are the Fourier transforms of autocorrelation functions, which are even and real).

For example, let $X(t)$ be white noise, with $S_{XX}(f) = N_0$, being put through a first-order linear time-invariant system with respective impulse response and transfer functions

$$h(t) = \exp\{-\alpha t\},\ t \geq 0;\ h(t) = 0,\ t < 0 \qquad H(f) = 1/[\alpha + j2\pi f],\ \text{all } f$$

The temporal correlation of $h(t)$ with itself is $r_h(\tau) = (1/2\alpha)\exp\{-\alpha|\tau|\}$, so the power transfer function is $|H(f)|^2 = 1/[\alpha^2 + (2\pi f)^2]$. The autocorrelation for the input $X(t)$ is

$$R_{XX}(\tau) = \int_{-\infty}^{\infty} N_0 e^{j2\pi f\tau} df = N_0\delta(\tau)$$

which is an impulse. It follows (see Eq. 73.22) that the output $Y(t)$ has respective autocorrelation and psdf

$$R_{YY}(\tau) = [N_0\delta(\tau)] * [(1/2\alpha)\, e^{-\alpha|\tau|}] = (N_0/2\alpha)\, e^{-\alpha|\tau|},\ S_{YY}(f) = N_0/[\alpha^2 + (2\pi f)^2]$$

The output expected power $E[Y^2(t)]$ can be found from either one of

$$E[Y^2(t)] = R_{YY}(0) = N_0/2\alpha \qquad \text{or} \qquad E[Y^2(t)] = \int_{-\infty}^{\infty} S_{YY}(f)df = N_0/2\alpha$$

If the input $X(t)$ to a linear system is Gaussian, then the output will also be Gaussian [see Brown, 1983; or Gardner, 1990]. Thus, the output of a first-order linear time-invariant system driven by Gaussian white noise is the Ornstein–Uhlenbeck process, which is also a first-order Markov process.

For another example, let $X(t) = A\cos(w_o t + \Theta)$, where the random amplitude A has zero mean, the random phase Θ is uniform on $[-\pi, \pi]$, and A and Θ are independent. As before, we obtain $R_{XX}(\tau) = \sigma_A{}^2\cos(w_o\tau)$, from which it follows that $S_{XX}(f) = (\sigma_A{}^2/2)[\delta(f - w_o/2\pi) + \delta(f + w_o/2\pi)]$. These impulses in the psdf, called *spectral lines*, represent positive amounts of power at discrete frequencies.

Cross-Correlation of Processes

The *cross-correlation function* for two random processes $X(t)$ and $Y(t)$ is defined via

$$R_{XY}(t,\ t+\tau) \equiv E[X(t)Y(t+\tau)] \tag{73.55}$$

Let both processes be ws with zero means, so the covariance coincides with the correlation function. We say that two ws processes $X(t)$ and $Y(t)$ are *jointly ws* whenever $R_{XY}(t,t+\tau) = R_{XY}(\tau)$. In case $Y(t)$ is the output of a filter with impulse response $h(t)$,we can find the cross-correlation $R_{XY}(\tau)$ between the input and output via

$$R_{XY}(\tau) = E[X(t)Y(t+\tau)] = E[X(t)\int_{-\infty}^{\infty} h(u)X(t+\tau-u)du]$$
$$= \int_{-\infty}^{\infty} h(u)E[X(t)X(t+\tau-u)]\,du \tag{73.56}$$
$$= \int_{-\infty}^{\infty} h(u)R_{XX}(\tau-u)du = R_{XX}(\tau) * h(\tau)$$

Cross-correlation functions of ws processes satisfy (1) $R_{XY}(-\tau) = R_{YX}(\tau)$, (2) $|R_{XY}(\tau)|^2 \le R_{XX}(0)R_{YY}(0)$, and (3) $|R_{XY}(\tau)| \le (1/2)[R_{XX}(0) + R_{YY}(0)]$. The first follows from the definition, while the second comes from expanding $E[\{Y(t+\tau) - \alpha X(t)\}^2] \ge 0$. The third comes from the fact that the geometric mean cannot exceed the arithmetic mean [see Peebles, 1987, p. 154].

Taking the Fourier transform of the leftmost and rightmost sides of Eqs. (73.56) yields

$$S_{XY}(f) = S_{XX}(f)H(f) \tag{73.57}$$

The Fourier transform of the cross-correlation function is the *cross-spectral density function*

$$S_{XY}(f) = \int_{-\infty}^{\infty} R_{XY}(\tau)e^{-j2\pi f\tau}d\tau \tag{73.58}$$

According to Gardner [1990, p. 228], this is a *spectral correlation density function* that does not represent power in any sense.

Equation (73.57) suggests a method for identifying a linear time-invariant system. If the system is subjected to a ws input $X(t)$ and the power spectral density of $X(t)$ and the cross-spectral density of $X(t)$ and the output $Y(t)$ are measured, then the ratio yields the system transfer function

$$H(f) = S_{XY}(f)/S_{XX}(f) \tag{73.59}$$

In fact, it can be shown that this method gives the best linear time-invariant model of the (possibly time varying and nonlinear) system in the sense that the time-averaged mean-square error between the outputs of the actual system and of the model, when both are subjected to the same input, is minimized [see Gardner, 1990, pp. 282–286].

As an application, suppose that an undersea sonar-based device is to find the range to a target, as shown in Fig. 73.10, by transmitting a sonar signal $X(t)$ and receiving the reflected signal $Y(t)$. If v is the velocity of the sonar signal, and τ_o is the offset that maximizes the cross-correlation $R_{XY}(\tau)$, then the range (distance) d can be determined from $d = v\tau_o/2$ (note that the signal travels twice the range d).

Coherence

When $X(t)$ and $Y(t)$ have no spectral lines at f, the finite spectral correlation $S_{XY}(f)$ is actually a spectral covariance and the two associated variances are $S_{XX}(f)$ and $S_{YY}(f)$. We can normalize $S_{XY}(f)$ to obtain a *spectral correlation coefficient* $\Upsilon_{XY}(f)$ defined by

$$\Upsilon_{XY}(f)^2 = |S_{XY}(f)|^2/S_{XX}(f)S_{YY}(f) \tag{73.60}$$

We call $\Upsilon_{XY}(f)$ the *coherence function*. It is a measure of the power correlation of $X(t)$ and $Y(t)$ at each frequency f. When $Y(t) = X(t)*h(t)$, it has a maximum: by Eqs. (73.53), (73.59), and (73.60), $|\Upsilon_{XY}(f)|^2 = |S_{XX}(f) \cdot H(f)|^2/[S_{XX}(f) \cdot S_{XX}(f) \cdot |H(f)|^2] = 1$. In the general case we have

$$|S_{XY}(f)| \le [S_{XX}(f)S_{YY}(f)]^{1/2} \tag{73.61}$$

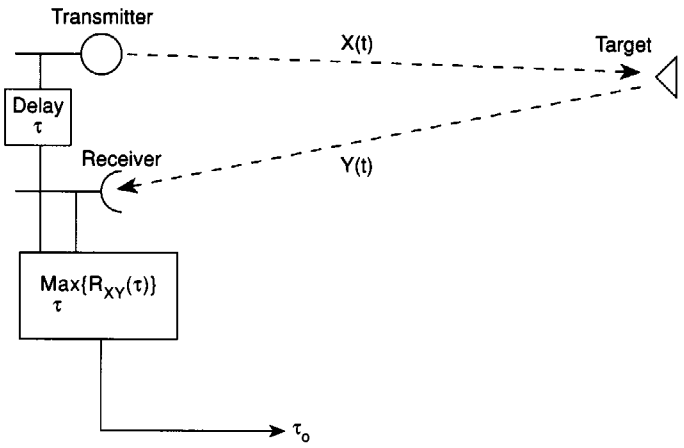


FIGURE 73.10 A sonar range finder.

Upon minimizing the mean-square error $\varepsilon = E[(Y(t) - X(t)*h(t))^2]$ over all possible impulse response functions $h(t)$, the optimal one, $h_o(t)$, has transfer function

$$H_o(f) = S_{XY}(f)/S_{XX}(f) \tag{73.62}$$

Further, the resultant minimum value is given by

$$\varepsilon_{\min} = \int_{-\infty}^{\infty} S_{YY}(f)[1 - |\Upsilon_{XY}(f)|^2]\,df$$

[see Gardner, 1990, pp. 434–436; or Bendat and Piersol, 1986]. At frequencies f where $|\Upsilon_{XY}(f)| \approx 1$, $\varepsilon_{\min} \approx 0$. Thus $1 - |\Upsilon_{XY}(f)|^2$ is the mean-square proportion of $Y(t)$ not accounted for by $X(t)$, while $|\Upsilon_{XY}(f)|^2$ is the proportion due to $X(t)$. When $Y(t) = X(t)*h(t)$, $\varepsilon_{\min} = 0$.

The optimum system $H_o(f)$ of Eq. (73.62) is known as the *Wiener filter* for minimum mean-square error estimation of one process $Y(t)$ using a filtered version of another process $X(t)$ [see Gardner, 1990; or Peebles, 1987, p. 262].

Ergodicity

When the **time average**

$$\lim_{T\to\infty}(1/T)\int_{-T/2}^{T/2} X(t)dt$$

exists and equals the corresponding expected value $E[X(t)]$, then the process $X(t)$ is said to possess an *ergodic property associated with the mean*. There are ergodic properties associated with the mean, autocorrelation (and power spectral density), and all finite-order joint moments, as well as finite-order joint pdf's. If a process has all possible ergodic properties, it is said to be an *ergodic process*.

Let $Y(t) = g[X(t + t_1), \ldots, X(t + t_n)]$, where $g[\cdot]$ is any nonrandom real function, so that $Y(t)$ is a function of a finite number of time samples of a strongly stationary process. For example, let (1) $Y(t) = X(t + t_1)X(t + t_2)$, $E[Y(t)] = R_{XX}(t_1 - t_2)$ and (2) $Y(t) = 1$ if $X(t) < x$, $Y(t) = 0$, otherwise, so that

$$E[Y(t)] = 1\cdot P(X(t) < x) + 0\cdot P(X(t) \geq x) = P(X(t) < x) = \int_{-\infty}^{1} f_{X(t)}(z)dz$$

We want to know under what conditions the mean-square error between the time average

$$\langle Y(t) \rangle_T \equiv (1/T)\int_{-T/2}^{T/2} Y(t)dt$$

and the expected value $E[Y(t)]$ will converge to zero. It can be shown that a necessary and sufficient condition for the mean-square ergodic property

$$\lim_{T\to\infty} E[\{\langle Y(t) \rangle_T - E[Y(t)]\}^2] = 0 \tag{73.63}$$

to hold is that

$$\lim_{T\to\infty} (1/T)\int_0^T C_{YY}(\tau)d\tau = 0 \tag{73.64}$$

For example, if $C_{YY}(\tau) \to 0$ as $\tau \to \infty$, then Eq. (73.64) will hold, and thus Eq. (73.63) will also, where $C_{YY}(\tau)$ is the covariance function of $Y(t)$. As long as the two sets of rv's $\{X(t + t_1), \ldots, X(t + t_n)\}$ and $\{X(t + t_1 + \tau), \ldots, X(t + t_n + \tau)\}$ become independent of each other as $\tau \to \infty$, the above condition holds, so Eq. (73.63) holds [see Gardner, 1990, pp. 163–174].

In practice, if $X(t)$ exhibits ergodicity associated with the autocorrelation, then we can estimate $R_{XX}(\tau)$ using the time average

$$\langle X(t)X(t+\tau) \rangle_T \equiv (1/T)\int_{-T/2}^{T/2} X(t)X(t+\tau)dt \tag{73.65}$$

In this case the mean-square estimation error $E[\{\langle X(t)X(t+\tau)\rangle_T - R_{XX}(\tau)\}^2]$ will converge to zero as T increases to infinity, and the power spectral density $S_{XX}(f)$ can also be estimated via time averaging [see Gardner, 1990, pp. 230–231].

Defining Terms

Autocorrelation function: A function $R_{XX}(t, t + \tau) = E[X(t)X(t + \tau)]$ that measures the degree to which any two rv's $X(t)$ and $X(t + \tau)$, at times t and $t + \tau$, are correlated.

Coherence function: A function of frequency f that provides the degree of correlation of two stochastic processes at each f by the ratio of their cross-spectral density function to the product of their power spectral density functions.

Power spectral density function: The Fourier transform of the autocorrelation function of a stochastic process $X(t)$, denoted by $S_{XX}(f)$. The area under its curve between f_1 and f_2 represents the total power over all t in $X(t)$ in the band of frequencies f_1 to f_2. Its dimension is watts per Hz.

Sample function: A real-valued function $x(t)$ of t where at each time t the value $x(t)$ at the argument t was determined by the outcome of a rv $X_t = x(t)$.

Stochastic process: A collection of rv's $\{X_t; t \in T\}$, where T is an ordered set such as the real numbers or integers [$X(t)$ is also called a random function, on the domain T].

Time average: Any real function $g(t)$ of time has average value g_{ave} on the interval $[a,b]$ such that the rectangular area $g_{ave}(b - a)$ is equal to the area under the curve between a and b, i.e., $g_{ave} = [1/(b - a)] \int_a^b g(t)dt$. The time average of a sample function $x(t)$ is the limit of its average value over $[0, T]$ as T goes to infinity.

Weakly stationary: The property of a stochastic process $X(t)$ whose mean $E[X(t)] = \mu(t)$ is a fixed constant μ over all time t, and whose autocorrelation is also independent of time in that $R_{XX}(t, t + \tau) = R_{XX}(s + t, s + t + \tau)$ for any s. Thus, $R_{XX}(t, t + \tau) = R_{XX}(0, \tau) = R_{XX}(\tau)$.

Related Topic

16.1 Spectral Analysis

References

The author is grateful to William Gardner of the University of California, Davis for making substantial suggestions.

J.S. Bendat and A.G. Piersol, *Random Data: Analysis and Measurement,* 2nd ed., New York: Wiley-Interscience, 1986.

R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering,* New York: Wiley, 1983.

W. A. Gardner, *Introduction to Random Processes,* 2nd ed., New York: McGraw-Hill, 1990.

P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Principles,* 2nd ed., New York: McGraw-Hill, 1987.

Further Information

The IEEE Individual Learning Package, *Random Signal Analysis with Random Processes and Kalman Filtering,* prepared for the IEEE in 1989 by Carl G. Looney, IEEE Educational Activities Board, PO Box 1331, Piscataway, NJ 08855-1331.

R. Iranpour and P. Chacon, *Basic Stochastic Processes: The Mark Kac Lectures,* New York: Macmillan, 1988.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* 3rd ed., New York: Macmillan, 1991.

73.4 The Sampling Theorem

R. J. Marks II

Most signals originating from physical phenomena are analog. Most computational engines, on the other hand, are digital. Transforming from analog to digital is straightforward: we simply sample. Regaining the original signal from these samples and assessing the information lost in the sampling process are the fundamental questions addressed by the **sampling theorem**.

The fundamental result of the sampling theorem is, remarkably, that a bandlimited signal is uniquely specified by its sufficiently close equally spaced samples. Indeed, the sampling theorem illustrates how the original signal can be regained from knowledge of the samples and the sampling rate at which they were taken.

Popularization of the sampling theorem is credited to Shannon [1948] who, in 1948, used it to show the equivalence of the information content of a bandlimited signal and a sequence of discrete numbers. Shannon was aware of the pioneering work of Whittaker [1915] and Whittaker's son [1929] in formulating the sampling theorem. Kotel'nikov's [1933] independent discovery in the then Soviet Union deserves mention. Higgins [1985] credits Borel [1897] with first recognizing that a signal could be recovered from its samples.

Surveys of sampling theory are in the widely cited paper of Jerri [1977] and in two books by the author [1991, 1993]. Marvasti [1987] has written a book devoted to nonuniform sampling.

The Cardinal Series

If a signal has finite energy, the minimum **sampling rate** is equal to two samples per period of the highest frequency component of the signal. Specifically, if the highest frequency component of the signal is B Hz, then the signal, $x(t)$, can be recovered from the samples by

$$x(t) = \frac{1}{\pi} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \frac{\sin[\pi(2Bt - n)]}{2Bt - n} \tag{73.66}$$

The frequency B is also referred to as the signal's bandwidth and, if B is finite, $x(t)$ is said to be bandlimited. The signal, $x(t)$, is here being sampled at a rate of 2 B samples per second. If sampling were done at a lower

rate, the replications would overlap and the information about $X(\omega)$ [and thus $x(t)$] is irretrievably lost. Undersampling results in *aliased* data. The minimum sampling rate at which **aliasing** does not occur is referred to as the **Nyquist rate** which, in our example, is $2B$. Eq. (73.66) was dubbed the **cardinal series** by the junior Whittaker [1929].

A signal is bandlimited in the low-pass sense if there is a $B > 0$ such that

$$X(\omega) = X(\omega)\Pi\left(\frac{\omega}{4\pi B}\right) \tag{73.67}$$

where the gate function $\Pi(\xi)$ is one for $\xi \le 1/2$ and is otherwise zero, and

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{73.68}$$

is the **Fourier transform** of $x(t)$. That is, the spectrum is identically zero for $|\omega| > 2\pi B$. The B parameter is referred to as the signal's bandwidth. The inverse Fourier transform is

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)\, e^{j\omega t}d\omega \tag{73.69}$$

The sampling theorem reduces the normally continuum infinity of ordered pairs required to specify a function to a countable—although still infinite—set. Remarkably, these elements are obtained directly by sampling.

How can the cardinal series interpolate uniquely the bandlimited signal from which the samples were taken? Could not the same samples be generated from another bandlimited signal? The answer is no. Bandlimited functions are smooth. Any behavior deviating from smooth would result in high-frequency components which in turn invalidates the required property of being bandlimited. The smoothness of the signal between samples precludes arbitrary variation of the signal there.

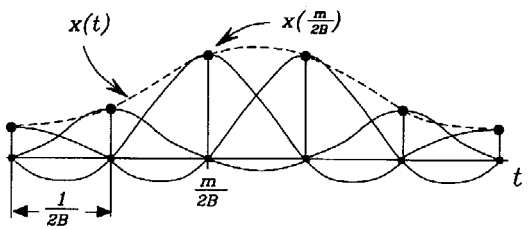


FIGURE 73.11 Illustration of the interpolation that results from the cardinal series. A sinc function, weighted by the sample, is placed at each sample bottom. The sum of the sincs exactly generates the original bandlimited function from which the samples were taken.

Let's examine the cardinal series more closely. Evaluate Eq. (73.74) at $t = m/2B$. Since $\text{sinc}(n)$ is one for $n = 0$ and is otherwise zero, only the sample at $t = m/2B$ contributes to the interpolation at that point. This is illustrated in Fig. 73.11, where the reconstruction of a signal from its samples using the cardinal series is shown. The value of $x(t)$ at a point other than a sample location [e.g., $t = (m + 1/2)/2B$] is determined by all of the sample values.

Proof of the Sampling Theorem

Borel [1897] and Shannon [1948] both discussed the sampling theorem as the Fourier transform dual of the Fourier series. Let $x(t)$ have a bandwidth of B. Consider the periodic signal

$$Y(\omega) = \sum_{n=-\infty}^{\infty} X(\omega - 4\pi nB) \tag{73.70}$$

The function $Y(\omega)$ is a periodic function with period $4\pi B$. From Eq. (73.67) $X(\omega)$ is zero for $\omega > 2\pi B$ and is thus finite in extent. The terms in Eq. (73.70) therefore do not overlap. Periodic functions can be expressed as a Fourier series.

$$Y(\omega) = \sum_{n=-\infty}^{\infty} \alpha_n \exp\left(\frac{-jn\omega}{2B}\right) \tag{73.71}$$

where the Fourier series coefficients are

$$\alpha_n = \frac{1}{4\pi B} \int_{-2\pi B}^{2\pi B} Y(\omega) \exp\left(\frac{jn\omega}{2B}\right) d\omega$$

or

$$\alpha_n = \frac{1}{2B} x\left(\frac{n}{2B}\right) \tag{73.72}$$

where we have used the inverse Fourier transform in Eq. (73.69). Substituting into the Fourier series in Eq. (73.71) gives

$$Y(\omega) = \frac{1}{2B} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \exp\left(\frac{-jn\omega}{2B}\right) \tag{73.73}$$

Since a period of $Y(\omega)$ is $X(\omega)$, we can get back the original spectrum by

$$X(\omega) = Y(\omega) \prod\left(\frac{\omega}{4\pi B}\right)$$

Substitute Eq. (73.73) and inverse transforming gives, using Eq. (73.69),

$$x(t) = \frac{1}{4\pi B} \int_{-2\pi B}^{2\pi B} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \exp\left(\frac{-jn\omega}{2B}\right) e^{j\omega t} d\omega$$

or

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \text{sinc}\,(2Bt - n) \tag{73.74}$$

where

$$\text{sinc}\,(t) = \frac{\sin \pi t}{\pi t}$$

is the inverse Fourier transform of $\Pi\,(\omega/2\pi)$. Eq. (73.74) is, of course, the cardinal series.

The sampling theorem generally converges uniformly, in the sense that

$$\lim_{N\to\infty} \left| x(t) - x_N(t) \right|^2 = 0$$

where the truncated cardinal series is

$$x_N(t) = \sum_{n=-N}^{N} x\left(\frac{n}{2B}\right) \operatorname{sinc}(2Bt - n) \tag{73.75}$$

Sufficient conditions for uniform convergence are [Marks, 1991]

1. the signal, $x(t)$, has finite energy, E,

$$E = \int_{-\infty}^{\infty} \left| x(t) \right|^2 dt < \infty$$

2. or $X(\omega)$ has finite area,

$$A = \int_{-\infty}^{\infty} \left| X(\omega) \right| d\omega < \infty$$

Care must be taken in the second case, though, when singularities exist at $\omega = \pm 2\pi B$. Here, sampling may be required to be strictly greater than $2B$. Such is the case, for example, for the signal, $x(t) = \sin(2\pi Bt)$. Although the signal is bandlimited, and although its Fourier transform has finite area, all of the samples of $x(t)$ taken at $t = n/2B$ are zero. The cardinal series in Eq. (73.74) will thus interpolate to zero everywhere. If the sampling rate is a bit greater than $2B$, however, the samples are not zero and the cardinal series will uniformly converge to the proper answer.

The Time-Bandwidth Product

The cardinal series requires knowledge of an infinite number of samples. In practice, only a finite number of samples can be used. If most of the energy of a signal exists in the interval $0 \le t \le T$, and we sample at the Nyquist rate of $2B$ samples per second, then a total of $S = \langle 2BT \rangle$ samples are taken. ($\langle\theta\rangle$ denotes the largest number not exceeding θ.) The number S is a measure of the degrees of freedom of the signal and is referred to as its **time-bandwidth product.** A 5-min single-track audio recording requiring fidelity up to 20,000 Hz, for example, requires a minimum of $S = 2 \times 20{,}000 \times 5 \times 60 = 12$ million samples. In practice, audio sampling is performed well above the Nyquist rate.

Sources of Error

Exact interpolation using the cardinal series assumes that (1) the values of the samples are known exactly, (2) the sample locations are known exactly, and (3) an infinite number of terms are used in the series. Deviation from these requirements results in interpolation error due to (1) data noise, (2) jitter, and (3) truncation, respectively. The effect of data error on the restoration can be significant. Some innocently appearing sampling theorem generalizations, when subjected to performance analysis in the presence of data error, are revealed as ill-posed. In other words, a bounded error on the data can result in unbounded error on the restoration [Marks, 1991].

Data Noise

The source of data noise can be the signal from which samples are taken, or from round-off error due to finite sampling precision. If the noise is additive and random, instead of the samples

$$x\left(\frac{n}{2B}\right)$$

we must deal with the samples

$$x\left(\frac{n}{2B}\right)+\xi\left(\frac{n}{2B}\right)$$

where $\xi(t)$ is a stochastic process. If these noisy samples are used in the cardinal series, the interpolation, instead of simple $x(t)$, is

$$x(t)\ +\ \eta(t)$$

where the interpolation noise is

$$\eta(t)=\sum_{n=-\infty}^{\infty}\xi\left(\frac{n}{2B}\right)\mathrm{sinc}\,(2Bt-n)$$

If $\xi(t)$ is a zero mean process, then so is the interpolation noise. Thus, the noisy interpolation is an unbiased version of $x(t)$. More remarkably, if $\xi(t)$ is a zero-mean (wide-sense) stationary process with uncertainty (variance) σ^2, then so is $\eta(t)$. In other words, *the uncertainty at the sample point locations is the same as at all points of interpolation* [Marks, 1991].

Truncation

The truncated cardinal series is in Eq. (73.75). A signal cannot be both bandlimited and of finite duration. Indeed, a bandlimited function cannot be identically zero over any finite interval. Thus, other than the rare case where an infinite number of the signal's zero crossings coincide with the sample locations, truncation will result in an error.

The magnitude of this **truncation error** can be estimated through the use of Parseval's theorem for the cardinal series that states

$$\begin{aligned}E\ &=\ \int_{-\infty}^{\infty}\left|\,x(t)\,\right|^2dt\\ &=\ \frac{1}{2B}\sum_{-\infty}^{\infty}\left|x\left(\frac{n}{2B}\right)\right|^2\end{aligned}$$

The energy of a signal can thus be determined directly from either the signals or the samples. The energy associated with the truncated signal is

$$E_N\ =\ \frac{1}{2B}\sum_{-N}^{N}\left|x\left(\frac{n}{2B}\right)\right|^2$$

If $E - E_N << E$, then the truncation error is small.

Jitter

Jitter occurs when samples are taken near to but not exactly at the desired sample locations. Instead of the samples $x\,(n/2W)$, we have the samples

$$x\left(\frac{n}{2W} - \sigma_n\right)$$

where σ_n is the jitter offset of the nth sample. For jitter, the σ_n's are not known. If they were, an appropriate nonuniform sampling theorem [Marks, 1993; Marvasti, 1987] could be used to interpolate the signal.

Using the jittered samples in the cardinal series results in an interpolation that is not an unbiased estimate of $x(t)$. Indeed, if the probability density function of the jitter is the same at all sample locations, the expected value of the jittered interpolation is the convolution of $x(t)$ with the probability density function of the jitter. This bias can be removed by inverse filtering at a cost of decreasing the signal-to-noise ratio of the interpolation [Marks, 1993].

Generalizations of the Sampling Theorem

There exist numerous generalizations of the sampling theorem [Marks, 1991; Marks, 1993].

1. **Stochastic processes.** A wide-sense stationary stochastic process, $\chi(t)$, is said to be bandlimited if its autocorrelation, $R_\chi(t)$, is a bandlimited function. The cardinal series

$$\hat{\chi}(t) = \sum_{n=-\infty}^{\infty} \chi\left(\frac{n}{2B}\right) \operatorname{sinc}(2Bt - n)$$

converges to $\chi(t)$ in the sense that

$$E\left[\left|\hat{\chi}(t) - \chi(t)\right|^2\right] = 0$$

where E denotes expectation.
2. **Nonuniform sampling.** There exist numerous scenarios wherein interpolation can be performed from samples that are not spaced uniformly. Marvasti [1987] devotes a book to the topic.
3. **Kramer's generalization.** Kramer generalized the sampling theorem to integral transforms other than Fourier, for example, to Legendre and Laguerre transforms.
4. **Papoulis' generalization.** Shannon noted that a bandlimited signal could be restored when sampling was performed at half the Nyquist rate if, at every sample location, a sample of the signal's derivative were also taken. Recurrent nonuniform sampling is where P samples are spaced the same in every P Nyquist intervals. Another sampling scenario is when a signal and its Hilbert transform are both sampled at half their respective Nyquist rates. Restoration of the signal from these and numerous other sampling scenarios are subsumed in an eloquent generalization of the sampling theorem by Papoulis.
5. **Lagrangian interpolation.** Lagrangian interpolation is a topic familiar in numerical analysis. An Nth order polynomial is fit to $N + 1$ arbitrarily spaced sample points. If an infinite number of samples are equally spaced, then Lagrangian interpolation is equivalent to the cardinal series.
6. **Trigonometric polynomials.** All periodic bandlimited signals can be expressed as trigonometric polynomials (i.e., a Fourier series with a finite number of terms). If the series has M terms, then the signal has M degrees of freedom which can be determined from M samples taken within a single period.
7. **Multidimensional sampling theorems.** Multidimensional signals, such as images, require dimensional extensions of the sampling theorem. The sampling of the signal now requires geometrical interpretation. Uniform sampling of an image, for example, can either be done on a rectangular or hexagonal grid. The minimum sampling density for one geometry may differ from that of another. The smallest sampling density that does not result in aliasing can be achieved, in many cases, with a number of different uniform sampling geometries and is referred to as the Nyquist density. Interestingly, sampling can sometimes be performed below the Nyquist density with nonuniform sampling geometries such that the multidimensional signal can be restored. Such is not the case for one dimension.

8. **Continuous sampling.** When a signal is known on one or more disjoint intervals, it is said to have been continuously sampled. Divide the time line into intervals of T. Periodic continuous sampling assumes that the signal is known on each interval over an interval of αT where α is the duty cycle. Continuously sampled signals can be accurately interpolated even in the presence of aliasing. Other continuously sampled cases, each of which can be considered as a limiting case of continuously periodically sampled restoration, include
   (a) **Interpolation.** The tails of a signal are known and we wish to restore the middle.
   (b) **Extrapolation.** We wish to generate the tails of a function with knowledge of the middle.
   (c) **Prediction.** A signal for $t > 0$ is to be estimated from knowledge of the signal for $t < 0$.

Final Remarks

Since its popularization in the late 1940s, the sampling theorem has been studied in depth. More than 1000 papers have been generated on the topic [Marks, 1993]. Its understanding is fundamental in matching the largely continuous world to digital computation engines.

Defining Terms

Aliasing: A phenomenon that occurs when a signal is undersampled. High-frequency information about the signal is lost.

Cardinal series: The formula by which samples of a bandlimited signal are interpolated to form a continuous time signal.

Fourier transform: The mathematical operation that converts a time-domain signal into the frequency domain.

Jitter: A sample is temporally displaced by an unknown, usually small, interval.

Kramer's generalization: A sampling theory based on other than Fourier transforms and frequency.

Lagrangian interpolation: A classic interpolation procedure used in numerical analysis. The sampling theorem is a special case.

Nyquist rate: The minimum sampling rate that does not result in aliasing.

Papoulis' generalization: A sampling theory applicable to many cases wherein signal samples are obtained either nonuniformly and/or indirectly.

Sampling rate: The number of samples per second.

Sampling theorem: Samples of a bandlimited signal, if taken close enough together, exactly specify the continuous time signal from which the samples were taken.

Signal bandwidth: The maximum frequency component of a signal.

Time bandwidth product: The product of a signal's duration and bandwidth approximates the number of samples required to characterize the signal.

Truncation error: The error that occurs when a finite number of samples are used to interpolate a continuous time signal.

Related Topic

8.5 Sampled Data

References

E. Borel, "Sur l'interpolation," *C.R. Acad. Sci. Paris,* vol. 124, pp. 673–676, 1897.

J. R. Higgins, "Five short stories about the cardinal series," *Bull. Am. Math. Soc.,* vol. 12, pp. 45–89, 1985.

A. J. Jerri, "The Shannon sampling theorem—its various extension and applications: a tutorial review," *Proc. IEEE,* vol. 65, pp. 1565–1596, 1977.

V. A. Kotel'nikov, "On the transmission capacity of 'ether' and wire in electrocommunications," *Izd. Red. Upr. Svyazi RKKA* (Moscow), 1933.

R. J. Marks II, *Introduction to Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1991.

R. J. Marks II, Ed., *Advanced Topics in Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1993.
F. A. Marvasti, *A Unified Approach to Zero-Crossing and Nonuniform Sampling,* Oak Park, Ill.: Nonuniform, 1987.
C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal,* vol. 27, pp. 379, 623, 1948.
E. T. Whittaker, "On the functions which are represented by the expansions of the interpolation theory," *Proc. Royal Society of Edinburgh,* vol. 35, pp. 181–194, 1915.
J. M. Whittaker, "The Fourier theory of the cardinal functions," *Proc. Math. Soc. Edinburgh,* vol. 1, pp. 169–176, 1929.
A. I. Zayed, Advances in Shannon's Sampling Theory, Boca Raton, Fla.: CRC Press, 1993.

Further Information

An in-depth study of the sample theorem and its numerous variations is provided in R. J. Marks II, Ed., *Introduction to Shannon Sampling and Interpolation Theory,* New York:Springer-Verlag, 1991.

In-depth studies of modern sampling theory with over 1000 references are available in R. J. Marks II, Ed., *Advanced Topics in Shannon Sampling and Interpolation Theory,* New York: Springer-Verlag, 1993.

The specific case of nonuniform sampling is treated in the monograph by F. A. Marvasti, *A Unified Approach to Zero-Crossing and Nonuniform Sampling*, Oak Park, Ill:Nonuniform, 1987.

The sampling theorem is treated generically in the *IEEE Transactions on Signal Processing*. For applications, topical journals are the best source of current literature.

73.5 Channel Capacity

Sergio Verdú

Information Rates

Tens of millions of users access the Internet daily via standard telephone lines. **Modems** operating at data rates of up to 28,800 bits per second enable the transmission of text, audio, color images, and even low-resolution video. The progression in modern technology for the standard telephone channel shown in Fig. 73.12 exhibits, if not the exponential increases ubiquitous in computer engineering, then a steady slope of abut 825 bits per second per year.

Few technological advances can result in as many time-savings for worldwide daily life as advances in modem information rates. However, modem designers are faced with a fundamental limitation in the maximum transmissible information rate. Every communication channel has a number associated with it called **channel capacity**, which determines the maximum information rate that can flow through the channel regardless of the complexity of the transmitting and receiving devices. Thus, the progression of modem rates shown in Fig. 73.12 is sure to come to a halt. But, at what rate? Answering this question for any communication channel model is one of the major goals of information theory—a discipline founded in 1948 by Claude E. Shannon [Shannon, 1948].

Communication Channels

The communication channel is the set of devices and systems that connects the transmitter to the receiver. The transmitter and receiver consist of an **encoder** and **decoder**, respectively, which translate the information stream produced by the source into a signal suitable for channel transmission and vice versa (Fig. 73.13). For example, in the case of the telephone line, two communication channels (one in each direction) share the same physical channel that connects the two modems. That physical channel usually consists of twisted copper wires at both ends and a variety of switching and signal processing operations that occur at the telephone exchanges. The modems themselves are not included in the communication channel. A microwave radio link is another example of a communication channel that consists of an amplifier and an antenna (at both ends) and a certain portion of the radio spectrum. In this case, the communication channel model does not fully correspond with the physical channel. Why not, for example, view the antenna as part of the transmitter rather than the channel

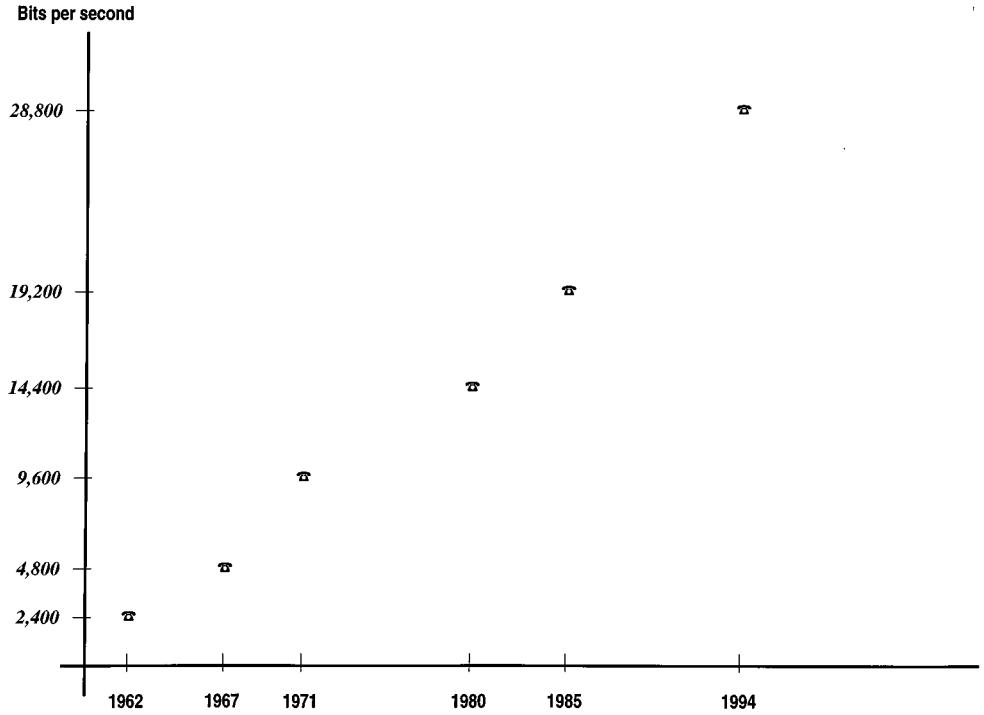


FIGURE 73.12 Information rates of modems for telephone channels.

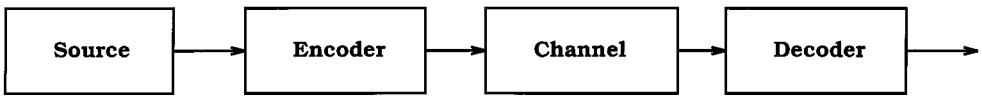


FIGURE 73.13 Elements of a communication system.

(Fig. 73.13)? Because considerations other than the optimization of the efficiency of the link are likely to dictate the choice of antenna size. This illustrates that the boundaries encoder–channel and channel–decoder in Fig. 73.13 are not always uniquely defined. This suggests an alternative definition of a channel as that part of the communication system that the designer is unable or unwilling to change.

A channel is characterized by the probability distributions of the output signals given every possible input signal. Channels are divided into (1) **discrete-time channels** and (2) **continuous-time channels** depending on whether the input/output signals are sequences or functions of a real variable. Some examples are as follows.

Example 1: Binary Symmetric Channel

A discrete-time **memoryless channel** with binary inputs and outputs (Fig. 73.14) where the probabilities that 0 and 1 are received erroneously are equal.

Example 2: Z-Channel

A discrete-time memoryless channel with binary inputs and outputs (Fig. 73.15) where 0 is received error-free.

Example 3: Erasure Channel

A discrete-time memoryless channel with binary inputs and ternary outputs (Fig. 73.16). The symbols 0 and 1 cannot be mistaken for each other but they can be "erased".

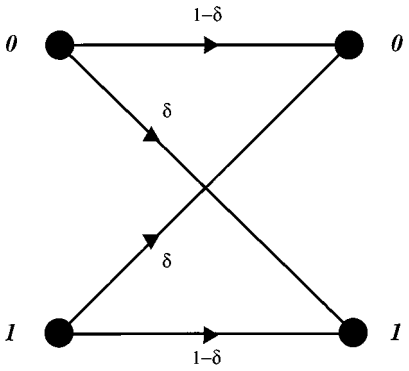


FIGURE 73.14 Binary symmetric channel.

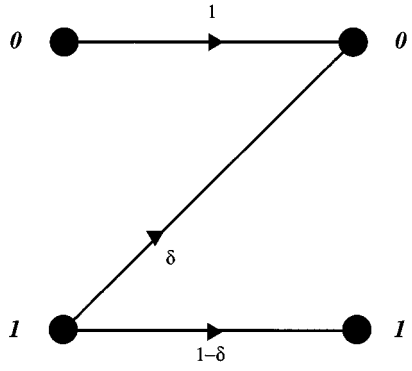


FIGURE 73.15 Z-channel.

Example 4: White Gaussian Discrete-Time Channel

A discrete-time channel whose output sequence is given by

$$y_i = x_i + n_i \tag{73.76}$$

where x_i is the input sequence and n_i is a sequence of independent Gaussian random variables with equal variance.

Example 5: Linear Continuous-Time Gaussian Channel

A continuous-time channel whose output signal is given by (Fig. 73.17)

$$y(t) = h(t) * x(t) + n(t) \tag{73.77}$$

where $x(t)$ is the input signal, $n(t)$ is a stationary Gaussian process, and $h(t)$ is the impulse response of a linear time-invariant system. The telephone channel is typically modeled by Eq. (73.77).

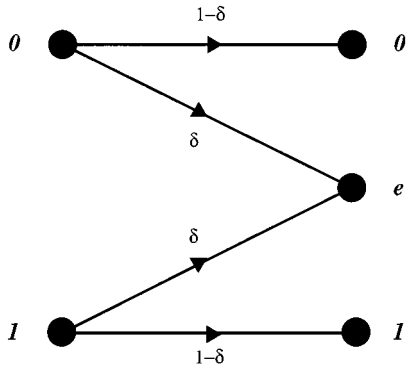


FIGURE 73.16 Erasure channel.

The goal of the encoder (Fig. 73.13) is to convert strings of binary data (messages) into channel-input signals. Source strings of m bits are translated into channel input strings of n symbols (with $m \leq n$) for discrete channels, and into continuous-time signals of duration T for continuous-time channels. The channel code (or more precisely the codebook) is the list of 2^m **codewords** (channel input signals) that may be sent by the encoder. The **rate** of the code is equal to the logarithm of its size divided by the duration of the codewords. Thus, the rate is equal to

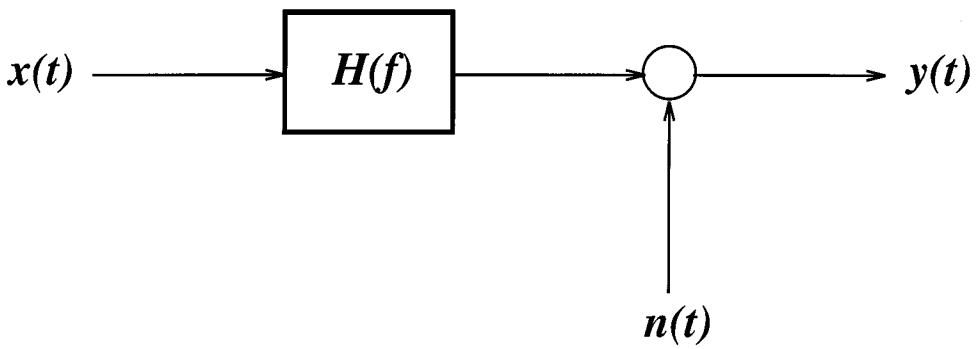


FIGURE 73.17 Linear continuous-time gaussian channel.

$$\frac{m}{n} \quad \text{bits per channel use}$$

for a discrete-time channel, whereas it is equal to

$$\frac{m}{T} \quad \text{bits per second}$$

for a continuous-time channel.

Once a codeword has been chosen by the encoder, the channel probabilistic mechanisms govern the distortion suffered by the transmitted signal. The role of the decoder is to recover the transmitted binary string (message) upon reception of the channel-distorted version of the transmitted codeword. To that end, the decoder knows the codebook used by the encoder. For most channels (including those above) there is a nonzero probability that the best decoder (maximum likelihood decoder) selects the wrong message. Thus, for a given channel the two figures of merit and of interest are the rate and the probability of error. The higher the tolerated probability of error, the higher the allowed rate; however, computing the exact tradeoff is a formidable task unless the code size either is very small or tends to infinity. The latter case was the one considered by Shannon and treated in the following section.

Reliable Information Transmission: Shannon's Theorem

Shannon [1948] considered the situation in which the codeword duration grows without bound. Channel capacity is the maximum rate for which encoders and decoders exist whose probability of error vanishes as the codewords get longer and longer.

Shannon's Theorem [Shannon, 1948] The capacity of a discrete memoryless channel is equal to

$$C = \max_X I(X; Y), \tag{73.78}$$

where $I(X; Y)$ stands for the input-output mutual information, which is a measure of the dependence of the input and the output defined as the divergence between the joint input/output distribution and the product of its marginals, $D(P_{XY}||P_X P_Y)$. For any pair of probability mass functions P and Q defined on the same space, divergence is an asymmetric measure of their similarity:

$$D(P||Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}. \tag{73.79}$$

Divergence is zero if both distributions are equal; otherwise it is strictly positive. The maximization in Eq. (73.78) is over the set of input distributions. Although, in general, there is no closed-form solution for that optimization problem, an efficient algorithm was obtained by Blahut and Arimoto in 1972 [Blahut, 1987]. The distribution that attains the maximum in Eq. (73.78) determines the statistical behavior of optimal codes and, thus, is of interest to the designer of the encoder. For the discrete memoryless channels mentioned above, the capacity is given by the following examples.

Example 1: Binary Symmetric Channel

$$C = 1 - \delta \log \frac{1}{\delta} - (1 - \delta) \log \frac{1}{1 - \delta}$$

attained by an equiprobable distribution and shown in Fig. 73.18.

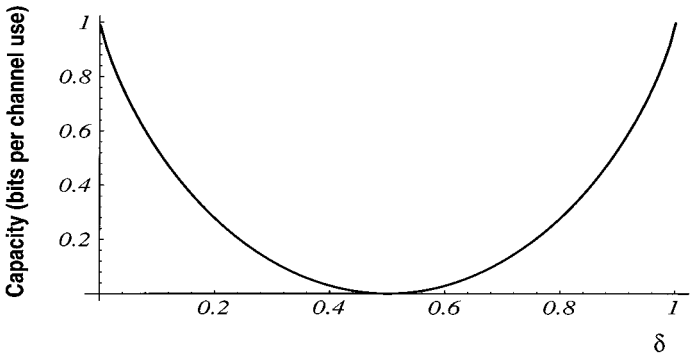


FIGURE 73.18 Capacity of the binary symmetric channel as a function of crossover probability.

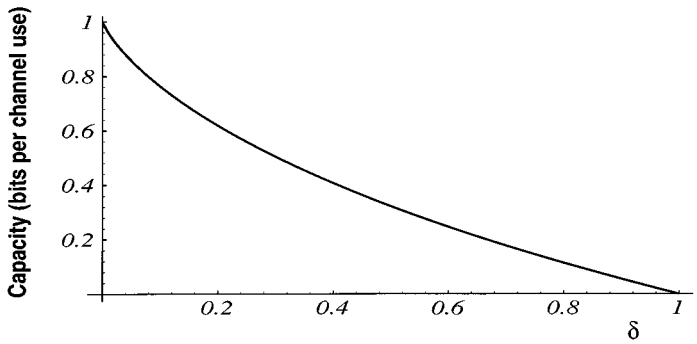


FIGURE 73.19 Capacity of the Z-channel.

Example 2: Z-Channel

$$C = \log\left(1 - \delta^{\frac{1}{1-\delta}} + \delta^{\frac{\delta}{1-\delta}}\right)$$

attained for a distribution whose probability mass at 0 ranges from 1/2 ($\delta = 0$) to $1/e$ ($\delta \to 1$) (Fig. 73.19).

Example 3: Erasure Channel

$$C = 1 - \delta$$

attained for equiprobable inputs.

Oftentimes the designer is satisfied with not exceeding a certain fixed level of bit error rate, ϵ, rather than the more stringent criterion of vanishing probability of selecting the wrong block of data. In such a case, it is possible to transmit information at a rate equal to capacity times

$$\left(1 - \epsilon \log \frac{1}{\epsilon} - (1 - \epsilon) \log \frac{1}{1 - \epsilon}\right)^{-1}$$

which is shown in Fig. 73.20.

If, contrary to what we have assumed thus far, the message source in Fig. 73.13 is not a source of pure bits, the significance of capacity can be extended to show that as long as the source entropy (see Chapter 73.6 on

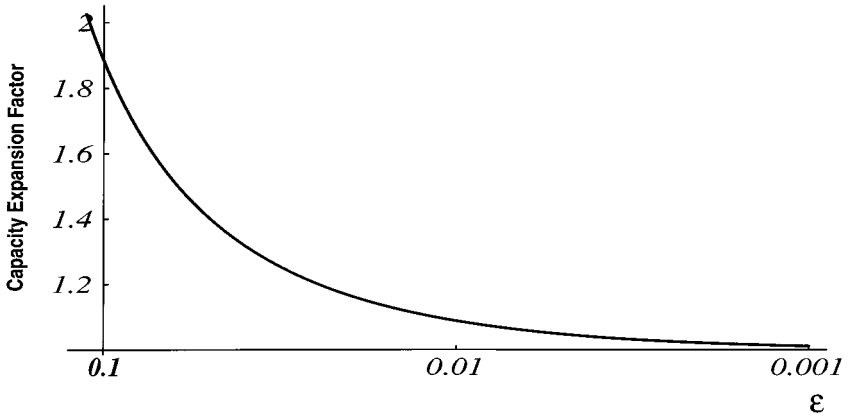


FIGURE 73.20 Capacity expansion factor as a function of bit-error-rate.

Data Compression) is below the channel capacity, an encoder/decoder pair exists that enables arbitrarily reliable communication. Conversely, if the source entropy is above capacity, then no such encoder/decoder pair exists.

Bandwidth and Capacity

The foregoing formulas for discrete channels do not lead to the capacity of continuous-time channels such as Example 5. We have seen that in the case of the telephone channel whose bandwidth is approximately equal to 3 kHz, capacity is lower bounded by 28,800 bits per second. How does bandwidth translate into capacity? The answer depends on the noise level and distribution. For example, if in the channel of Example 5, the noise is absent, capacity is infinite regardless of bandwidth. We can encode any amount of information as the binary expansion of a single scalar, which can be sent over the channel as the amplitude or phase of a single sinusoid; knowing the channel transfer function, the decoder can recover the transmitted scalar error-free. Clearly, such a transmission method is not recommended in practice because it hinges on the non-physical scenario of noiseless transmission.

In the simplest special case of Example 5, the noise is white, the channel has an ideal flat transfer function with bandwidth B (in Hz), and the input power is limited. Then, the channel capacity is equal to

$$C = B\ SNR^{dB} \log_2 10^{0.1} \quad \text{bits per second} \tag{73.80}$$

where $\log_2 10^{0.1} = 0.33$ and SNR^{dB} is equal to the optimum signal-to-noise ratio (in dB) of a linear estimate of a flat input signal given the channel output signal. Such an optimum signal-to-noise ratio is equal to one plus the power alloted to the input divided by the noise power in the channel band, i.e.,

$$SNR^{dB} = 10 \log_{10} \left(1 + \frac{P}{BN_0}\right).$$

It is interesting to notice that as the bandwidth grows, the channel capacity does not grow without bound. It tends to

$$\frac{P}{N_0} \log_2 e \qquad \text{bits per second}$$

where $\log_2 e = 1.44$. This means that the energy per bit necessary for reliable communication is equal to 0.69 times the noise power spectral density level. When the channel bandwidth is finite, the energy necessary to send one bit of information is strictly larger. The energy required to send one bit of information reliably can

be computed for other (non-Gaussian) channels even in cases where expressions for channel capacity are not known [Verdú, 1990].

When the channel transfer function $H(f)$ and/or noise spectral density $N(f)$ are not flat, the constant in Eq. (73.80) no longer applies. The so-called water-filling formula [Shannon, 1949] gives the channel capacity as

$$C = \frac{1}{2}\int \log\left(1 + \frac{\max\{0,\ w - M(f)\}}{M(f)}\right) df$$

where w is chosen so that

$$\int \max\{0,\ w - M(f)\}\, df = P,$$

and

$$M(f) = \frac{N(f)}{|H(f)|^2}.$$

The linear Gaussian-noise channel is a widely used model for space communication (in the power limited region) and for the telephone channel (in the bandwidth limited region). Thanks to the prevalence of digital switching and digital transmission in modern telephone systems, not only signal-to-noise ratios have improved over time but the Gaussian-noise model in Example 5 becomes increasingly questionable because quantization is responsible for a major component of the channel distortion. Therefore, future improvements in modem speeds are expected to arise mainly from finer modeling of the channel.

Due to the effect of time-varying received power (fading), several important channels fall outside the scope of Example 5 such as high-frequency radio links, tropospheric scatter links, and mobile radio channels.

Channel Coding Theorems

In information theory, the results that give a formula for channel capacity in terms of the probabilistic description of the channel are known as channel coding theorems. They typically involve two parts: an achievability part, which shows that codes with vanishing error probability exist for any rate below capacity; and a converse part, which shows that if the code rate exceeds capacity, then the probability of error is necessarily bounded away from zero. Shannon gave the first achievability results in [Shannon, 1948] for discrete memory channels. His method of proof, later formalized as the method of "typical sequences" (e.g., [Cover and Thomas, 1991]), is based on showing that the average probability of error of a code chosen at random vanishes with blocklength. Other known achievability proofs such as Feinstein's [1954], Gallager's [1968], and the method of types [Csiszar and Korner, 1981] are similarly non-constructive. The discipline of coding theory deals with constructive methods to design codes that approach the Shannon limit (see Chapter 71.1). The first converse channel coding theorem was not given by Shannon, but by Fano in 1952. A decade after Shannon's pioneering paper, several authors obtained the first channel coding theorems for channels with memory [Dobrushin, 1963]. The most general formula for channel capacity known to date can be found in [Verdú and Han, 1994]. The capacity of channels with feedback was first considered by Shannon in 1961 [Shannon, 1961], with later developments for Gaussian channels summarized in [Cover and Thomas, 1991]. In his 1961 paper [Shannon, 1961], Shannon founded the discipline of multiuser information theory by posing several challenging channels with more than one transmitter and/or receiver. In contrast to the **multiaccess channel** (one receiver) which has been solved in considerable generality, the capacities of channels involving more than one receiver, such as **broadcast channels** [Cover, 1972] and **interference channels** remain unsolved except in special cases.

Channel capacity has been shown to have a meaning outside the domain of information transmission [Han and Verdú, 1993]: it is the minimum rate of random bits required to generate any input random process so that the output process is simulated with arbitrary accuracy.

Defining Terms

Blocklength: The duration of a codeword, usually in the context of discrete-time channels.

Channel capacity: The maximum rate for which encoders and decoders exist whose probability of error vanishes as the codewords get longer and longer.

Codeword: Channel-input signal chosen by the encoder to represent the message.

Communication channel: Set of devices and systems that connect the transmitter to the receiver, not subject to optimization.

Broadcast channel: A communication channel with one input and several outputs each connected to a different receiver such that possibly different messages are to be conveyed to each receiver.

Discrete memoryless channel: A discrete-time memoryless channel where each channel input and output takes a finite number of values.

Discrete-time channel: A communication channel whose input/output signals are sequences of values. Its capacity is given in terms of bits per "channel use".

Continuous-time channel: A communication channel whose input/output signals are functions of a real variable (time). Its capacity is given in terms of bits per second.

Interference channel: A channel with several inputs/outputs such that autonomous transmitters are connected to each input and such that each receiver is interested in decoding the message sent by one and only one transmitter.

Memoryless channel: A channel where the conditional probability of the output given the current input is independent of all other inputs or outputs.

Multiaccess channel: A channel with several inputs and one output such that autonomous transmitters are connected to each input and such that the receiver is interested in decoding the messages sent by all the transmitters.

Decoder: Mapping from the set of channel-output signals to the set of messages.

Maximum-likelihood decoder: A decoder which selects the message that best explains the received signal, assuming all messages are equally likely.

Encoder: Mapping from the set of messages to the set of input codewords.

Modem: Device that converts binary information streams into electrical signals (and vice-versa) for transmission through the voiceband telephone channel.

Rate: The rate of a code is the number of bits transmitted (logarithm of code size) per second for a continuous-time channel or per channel use for a discrete-time channel.

Related Topics

70.1 Coding • 73.4 The Sampling Theorem

References

C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.,* 27, 379–423, 623–656, July–Oct. 1948.

R. E. Blahut, *Principles of Information Theory.* Reading, Mass.: Addison-Wesley, 1987.

S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Information Theory,* IT-36(5), 1019–1030, Sept. 1990.

C. E. Shannon, "Communication in the presence of noise," *Proc. Institute of Radio Engineers,* 37, 10–21, 1949.

T. M. Cover and J. A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

A. Feinstein, "A new basic theorem of information theory," *IRE Trans. PGIT,* pp. 2–22, 1954.

R. G. Gallager, *Information Theory and Reliable Communication,* New York: Wiley, 1968.

I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* New York: Academic Press, 1981.
R. L. Dobrushin, *General Formulation of Shannon's Main Theorem in Information Theory,* American Mathematical Society Translations, pp. 323–438, 1963.
S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. on Information Theory,* 40(4), 1147–1157, July 1994.
C. E. Shannon, "Two-way communication channels," *Proc. 4th. Berkeley Symp. Math. Statistics and Prob.,* pp. 611–644, 1961.
T. M. Cover, "Broadcast channels," *IEEE Trans. on Information Theory,* pp. 2–14, Jan. 1972.
T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. on Information Theory,* IT-39, 752–772, May 1993.

Further Information

The premier journal and conference in the field of information theory are the *IEEE Trans. on Information Theory* and the *IEEE International Symposium on Information Theory,* respectively. *Problems of Information Transmission* is a translation of a Russian-language journal in information theory. The newsletter of the IEEE Information Theory Society regularly publishes expository articles.

73.6 Data Compression

Joy A. Thomas and Thomas M. Cover

Data compression is a process of finding the most efficient representation of an information source in order to minimize communication or storage. It often consists of two stages—the first is the choice of a (probabilistic) model for the source and the second is the design of an efficient coding system for the model. In this section, we will concentrate on the second aspect of the compression process, though we will touch on some common sources and models in the last subsection.

Thus, a data compressor (sometimes called a source coder) maps an information source into a sequence of bits, with a corresponding decompressor, that given these bits provides a reconstruction of the source. Data compression systems can be classified into two types: *lossless,* where the reconstruction is exactly equal to the original source, and *lossy,* where the reconstruction is a distorted version of the original source. For lossless data compression, the fundamental lower bound on the rate of the data compression system is given by the entropy rate of the source. For lossy data compression, we have a tradeoff between the rate of the compressor and the distortion we incur, and the fundamental limit is given by the rate distortion function, which is discussed later in this section.

Shannon [1948] was the first to distinguish the probabilistic model that underlies an information source from the semantics of the information. An information source produces one of many possible messages; the goal of communication is to transmit an unambiguous specification of the message so that the receiver can reconstruct the original message. For example, the information to be sent may be the result of a horse race. If the recipient is assumed to know the names and numbers of the horses, then all that must be transmitted is the number of the horse that won. In a different context, the same number might mean something quite different, e.g., the price of a barrel of oil. The significant fact is that the difficulty in communication depends only on the length of the representation. Thus, finding the best (shortest) representation of an information source is critical to efficient communication.

When the possible messages are all equally likely, then it makes sense to represent them by strings of equal length. For example, if there are 32 possible equally likely messages, then each message can be represented by a binary string of 5 bits. However, if the messages are not equally likely, then it is more efficient on the average to allot short strings to the frequently occurring messages and longer strings to the rare messages. Thus, the Morse code allots the shortest string (a dot) to the most frequent letter (E) and allots long strings to the infrequent letters (e.g., dash, dash, dot, dash for Q). The minimum average length of the representation is a fundamental quantity called the entropy of the source, which is defined in the next subsection.

Entropy

An information source will be represented by a random variable X, which takes on one of a finite number of possibilities $i \in \mathcal{X}$ with probability $p_i = \Pr(X = i)$. The entropy of the random variable X is defined as

$$H(X) = -\sum_{i \in \mathcal{X}} p_i \log p_i \tag{73.81}$$

where the log is to base 2 and the entropy is measured in *bits*. We will use logarithms to base 2 throughout this chapter.

Example 73.1 Let X be a random variable that takes on a value 1 with probability θ and takes on the value 0 with probability $1 - \theta$. Then $H(X) = -\theta \log \theta - (1 - \theta) \log (1 - \theta)$. In particular, the entropy of a fair coin toss with $\theta = \frac{1}{2}$ is 1 bit.

This definition of entropy is related to the definition of entropy in thermodynamics. It is the fundamental lower bound on the average length of a code for the random variable.

A **code** for a random variable X is a mapping from $\mathcal{X}$, the range of X, to the set of finite-length binary strings. We will denote the code word corresponding to i by $C(i)$, and the length of the code word by l_i. The average length of the code is then $L(C) = \Sigma_i p_i l_i$.

A code is said to be *instantaneous* or *prefix-free* if no code word is a prefix of any other code word. This condition is sufficient (but not necessary) to allow a sequence of received bits to be parsed unambiguously into a sequence of code words.

Example 73.2 Consider a random variable X taking on the values $\{1, 2, 3\}$ with probabilities (0.5, 0.25, 0.25). An instantaneous code for this random variable might be (0, 10, 11). Thus, a string 01001110 can be uniquely parsed into 0, 10, 0, 11, 10, which decodes to the string $x = (1, 2, 1, 3, 2)$. Note that the average length of the code is 1.5 bits, which is the same as the entropy of the source.

For any instantaneous code, the following property of binary trees called the *Kraft inequality* [Cover and Thomas, 1991].

$$\sum_i 2^{-l_i} \leq 1 \tag{73.82}$$

must hold. Conversely, it can be shown that given a set of lengths that satisfies the Kraft inequality, we can find a set of prefix-free code words of those lengths.

The problem of finding the best source code then reduces to finding the optimal set of lengths that satisfies the Kraft inequality and minimizes the average length of the code. Simple calculus can then be used to show [Cover and Thomas, 1991] that the average length of any instantaneous code is larger than the entropy of the random variable, i.e., the minimum of $\Sigma p_i l_i$ over all l_i satisfying $\Sigma\, 2^{-l_i} \leq 1$ is $-\Sigma p_i \log p_i$. Also, if we take $l_i = \lceil \log 1/p_i \rceil$ (where $\lceil t \rceil$ denotes the smallest integer greater than or equal to t), we can verify that this choice of lengths satisfies the Kraft inequality and that

$$L(C) = \sum_i p_i \left\lceil \log \frac{1}{p_i} \right\rceil < \sum_i p_i \left(\log \frac{1}{p_i} + 1 \right) = H(X) + 1 \tag{73.83}$$

The optimal code can only have a shorter length, and therefore we have the following theorem:

Theorem 73.1 Let L^* be the average length of the optimal instantaneous code for a random variable X. Then

$$H(X) \leq L^* < H(X) + 1 \tag{73.84}$$

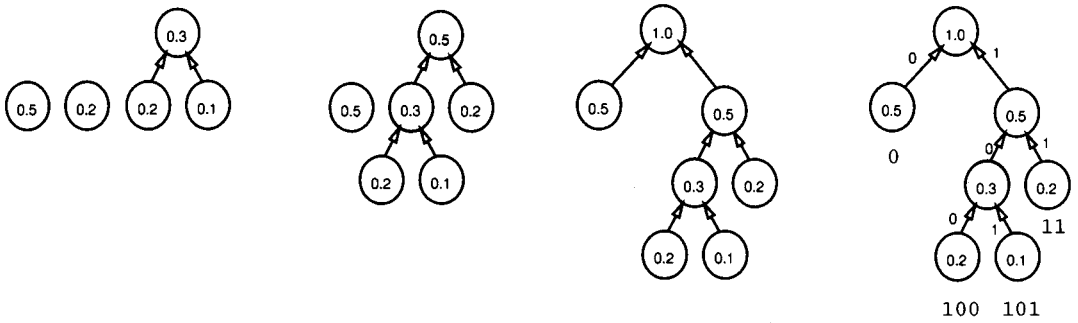


FIGURE 73.21 Example of the Huffman algorithm.

This theorem is one of the fundamental theorems of information theory. It identifies the entropy as the fundamental limit for the average length of the representation of a discrete information source and shows that we can find representations with average length within one bit of the entropy.

The Huffman Algorithm

The choice of code word lengths $l_i = \lceil \log 1/p_i \rceil$ (called the *Shannon code lengths*) is close to optimal, but not necessarily optimal, in terms of average code word length. We will now describe an algorithm (the **Huffman algorithm**) that produces an instantaneous code of minimal average length for a random variable with distribution $p_1, p_2, \ldots, p_m$. The algorithm is a greedy algorithm for building a tree from the bottom up.

Step 1. Arrange the probabilities in decreasing order so that $p_1 \geq p_2 \geq \ldots \geq p_m$.
Step 2. Form a subtree by combining the last two probabilities p_{m-1} and p_m to a single node of weight $p'_{m-1} = p_{m-1} + p_m$.
Step 3. Recursively execute Steps 1 and 2, decreasing the number of nodes each time, until a single node is obtained.
Step 4. Use the tree constructed above to allot code words.

The algorithm for tree construction is illustrated for a source with distribution (0.5, 0.2, 0.2, 0.1) in Fig. 73.21. After constructing the tree, the leaves of the tree (which correspond to the symbols of $\mathcal{X}$) can be assigned code words that correspond to the paths from the root to the leaf. We will not give a proof of the optimality of the Huffman algorithm; the reader is referred to Gallager [1968] or Cover and Thomas [1991] for details.

Entropy Rate

The entropy of a sequence of random variables $X_1, X_2, \ldots, X_n$ with joint distribution $p(x_1, x_2, \ldots, x_n)$ is defined analogously to the entropy of a single random variable as

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_1, x_2, \ldots, x_n)\log p(x_1, x_2, \ldots, x_n) \qquad (73.85)$$

For a stationary process $X_1, X_2, \ldots$, we define the *entropy rate* $\mathcal{H}(\mathcal{X})$ of the process as

$$\mathcal{H}(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \qquad (73.86)$$

It can be shown [Cover and Thomas, 1991] that the entropy rate is well defined for all stationary processes. In particular, if $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed (i.i.d.) random variables, then $H(X_1, X_2, \ldots, X_n) = nH(X_1)$, and $\mathcal{H}(\mathcal{X}) = H(X_1)$.

In the previous subsection, we showed the existence of a prefix-free code having an average length within one bit of the entropy. Now instead of trying to represent one occurrence of the random variable, we can form a code to represent a block of n random variables. In this case, the average code length is within one bit of $H(X_1, X_2, \ldots, X_n)$. Thus, the average length of the code per input symbol satisfies

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{L_n^*}{n} < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n} \tag{73.87}$$

Since $[H(X_1, X_2, \ldots, X_n)]/n \to \mathcal{H}(\mathcal{X})$, we can get arbitrarily close to the entropy rate by using longer and longer block lengths. Thus, the entropy rate is the fundamental limit for data compression for stationary sources, and we can achieve rates arbitrarily close to this limit by using long blocks.

All the above assumes that we know the probability distribution that underlies the information source. In many practical examples, however, the distribution is unknown or too complex to be used for coding. There are various ways to handle this situation:

- Assume a simple distribution and design an appropriate code for it. Use this code on the real source. If an estimated distribution $\hat{p}$ is used when in fact the true distribution is p, then the average length of the code is lower bounded by $H(X) + \Sigma_x p(x) \log [p(x)/\hat{p}(x)]$. The second term, which is denoted $D(p \| \hat{p})$ is called the *relative entropy* or the *Kullback Leibler distance* between the two distributions.
- Estimate the distribution empirically from the source and adapt the code to the distribution. For example, with *adaptive Huffman coding*, the empirical distribution of the source symbols is used to design the Huffman code used for the source.
- Use a *universal coding algorithm* like the **Lempel–Ziv algorithm** (see the subsection "Lempel–Ziv Coding").

Arithmetic Coding

In the previous subsections, it was shown how we could construct a code for a source that achieves an average length within one bit of the entropy. For small source alphabets, however, we have efficient coding only if we use long blocks of source symbols. For example, if the source is binary, and we code each symbol separately, we must use 1 bit per symbol, irrespective of the entropy of the source. If we use long blocks, we can achieve an expected length per symbol close to the entropy rate of the source.

It is therefore desirable to have an efficient coding procedure that works for long blocks of source symbols. Huffman coding is not ideal for this situation, since it is a bottom-up procedure with a complexity that grows rapidly with the block length. Arithmetic coding is an incremental coding algorithm that works efficiently for long block lengths and achieves an average length within one bit of the entropy for the block.

The essential idea of arithmetic coding is to represent a sequence $x^n = x_1, x_2, \ldots, x_n$ by the cumulative distribution function $F(x^n)$ (the sum of the probability of all sequences less than x^n) expressed to an appropriate accuracy. The cumulative distribution function for x^n is illustrated in Fig. 73.22. We can use any real number in the interval $[F(x^n) - p(x^n), F(x^n)]$ as the code for x^n. Expressing $F(x^n)$ to an accuracy of $\lceil \log 1/p(x^n) \rceil$ will give us a code for the source. The receiver can draw the cumulative distribution function, draw a horizontal line corresponding to the truncated value $\lfloor F(x^n) \rfloor$ that was sent, and read off the corresponding x^n. (This code is not prefix-free but can be easily modified to construct a prefix-free code [Cover and Thomas, 1991]). To implement arithmetic coding, however, we need efficient algorithms to calculate $p(x^n)$ and $F(x^n)$ to the appropriate accuracy based on a probabilistic model for the source. Details can be found in Langdon [1984] and Bell et al. [1990].

Lempel–Ziv Coding

The Lempel–Ziv algorithm [Ziv and Lempel, 1978] is a universal coding procedure that does not require knowledge of the source statistics and yet is asymptotically optimal. The basic idea of the algorithm is to construct a table or dictionary of frequently occurring strings and to represent new strings by pointing to their prefixes in the table. We first parse the string into sequences that have not appeared so far. For example, the

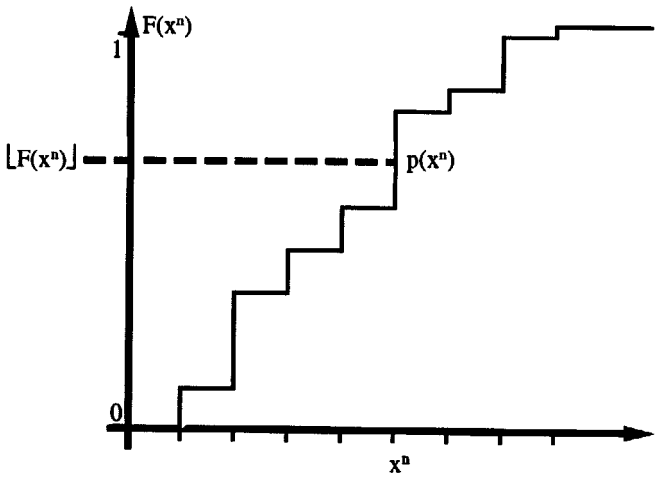


FIGURE 73.22 Cumulative distribution function for sequences x^n.

binary string 11010011011100 is parsed into 1,10,100,11,0,111,00. Then instead of sending the bits of each phrase, we send a pointer to its prefix and the value of the last bit. Thus, if we use three bits for the pointer, we will represent this string by (000,1), (001,0), (010,0), (001,1), (000,0), (100,1), (101,0), etc. For this short example, the algorithm has not compressed the string—it has in fact expanded it.

The very surprising fact is that, as Lempel and Ziv have shown, the algorithm is asymptotically optimal for any stationary ergodic source. This is expressed in the following theorem [Ziv and Lempel, 1978; Cover and Thomas, 1991]:

Theorem 73.2 Let L_n be the length of the Lempel–Ziv code for n symbols drawn from a stationary ergodic process $X_1, X_2, \ldots, X_n$ with entropy rate $\mathcal{H}(\mathcal{X})$. Then

$$\frac{L_n}{n} \to \mathcal{H}(\mathcal{X}) \quad \text{with probability 1} \tag{73.88}$$

Thus, for long enough block lengths, the Lempel–Ziv algorithm (which does not make any assumptions about the distribution of the source) does as well as if we knew the distribution in advance and designed the optimal code for this distribution.

The algorithm described above is only one of a large class of similar adaptive dictionary-based algorithms, which are all rather loosely called Lempel–Ziv. These algorithms are simple and fast and have been implemented in both software and hardware, e.g., in the *compress* command in UNIX and the *PKZIP* command on PCs. On ASCII text files, the Lempel–Ziv algorithm achieves compressions on the order of 50%. It has also been implemented in hardware and has been used to "double" the capacity of data storage media or to "double" the effective transmission rate of a modem. Many variations on the basic algorithm can be found in Bell et al. [1990].

Rate Distortion Theory

An infinite number of bits are required to describe an arbitrary real number, and therefore it is not possible to perfectly represent a continuous random variable with a finite number of bits. How "good" can the representation be? We first define a distortion measure, which is a measure of the distance between the random variable and its representation. We can then consider the trade-off between the number of bits used to represent a random variable and the distortion incurred. This trade-off is represented by the **rate distortion function** $R(D)$, which represents the minimum rate required to represent a random variable with a distortion D.

We will consider a discrete information source that produces random variables $X_1, X_2, \ldots, X_n$ that are drawn i.i.d. according to $p(x)$. (The results are also valid for continuous sources.) The encoder of the rate distortion

system of rate R will encode a block of n outputs X^n as an index $f(X^n) \in \{1, 2, \ldots, \lfloor 2^{nR} \rfloor\}$. (Thus, the index will require R bits/input symbol.) The decoder will calculate a representation $\hat{X}^n(f(X^n))$ of X^n. Normally, the representation alphabet $\hat{X}$ of the representation is the same as the source alphabet $\mathcal{X}$, but that need not be the case.

Definition: A *distortion function* or *distortion measure* is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow R^+ \tag{73.89}$$

from the set of source alphabet–reproduction alphabet pairs into the set of nonnegative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by the symbol $\hat{x}$.

Examples of common distortion functions are

- *Hamming (probability of error) distortion.* The Hamming distortion is given by

$$d(x, \hat{x}) = \begin{cases} 0 \text{ if } x = \hat{x} \\ 1 \text{ if } x \neq \hat{x} \end{cases} \tag{73.90}$$

  and thus $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$.
- *Squared error distortion.* The squared error distortion

$$d(x, \hat{x}) = (x - \hat{x})^2 \tag{73.91}$$

  is the most popular distortion measure used for continuous alphabets. Its advantages are its simplicity and its relationship to least squares prediction. However, for information sources such as images and speech, the squared error is not an appropriate measure for distortion as perceived by a human observer.

The distortion between sequences x^n and $\hat{x}^n$ of length n is defined by

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{73.92}$$

For a rate distortion system, the expected distortion D is defined as

$$D = Ed(X^n, \hat{X}^n(f(X^n))) = \sum_{x^n} p(x^n) d(x^n, \hat{X}^n(f(x^n))) \tag{73.93}$$

Definition: The rate distortion pair (R,D) is said to be achievable if there exists a rate distortion code of rate R with expected distortion D. The *rate distortion function* $R(D)$ is the infimum of rates R such that (R,D) is achievable for a given D.

Definition: The *mutual information* $I(X, \hat{X})$ between random variables X and $\hat{X}$, with joint probability mass function $p(x, \hat{x})$ and marginal probability mass functions $p(x)$ and $p(\hat{x})$ is defined as

$$I(X; \hat{X}) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} p(x, \hat{x}) \log \frac{p(x, \hat{x})}{p(x)p(\hat{x})} \tag{73.94}$$

The mutual information is a measure of the amount of information that one random variable carries about another.

The main result of rate distortion theory is contained in the following theorem, which provides a characterization of the rate distortion function in terms of the mutual information of joint distributions that satisfy the expected distortion constraint:

Theorem 73.3 The rate distortion function for an i.i.d. source X with distribution $p(x)$ and distortion function $d(x, \hat{x})$ is

$$R(D) = \min_{p(\hat{x}|x):\Sigma_{(x,\hat{x})}p(x)p(\hat{x}|x)d(x,\hat{x})\le D} I(X; \hat{X}) \tag{73.95}$$

We can construct rate distortion codes that can achieve distortion D at any rate greater than $R(D)$, and we cannot construct such codes at any rate below $R(D)$.

The proof of this theorem uses ideas of random coding and long block lengths as in the proof of the channel capacity theorem. The basic idea is to generate a code book of 2^{nR} reproduction code words $\hat{X}^n$ at random and show that for long block lengths, for any source sequence, it is very likely that there is at least one code word in this code book that is within distortion D of that source sequence. See Gallager [1968] or Cover and Thomas [1991] for details of the proof.

Example 73.3 (*Binary source*) The rate distortion function for a Bernoulli (p) source (a random variable that takes on values $\{0, 1\}$ with probabilities p, $1 - p$) with Hamming distortion is given by

$$R(D) = \begin{cases} H(p) - H(D), & 0 \le D \le \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\} \end{cases} \tag{73.96}$$

where $H(p) = -p \log p - (1 - p)\log (1 - p)$ is the binary entropy function.

Example 73.4 (*Gaussian source*) The rate distortion function for a Gaussian random variable with variance σ^2 and squared error distortion is

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases} \tag{73.97}$$

Thus, with nR bits, we can describe n i.i.d. Gaussian random variables $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ with a distortion of $\sigma^2 2^{-2R}$ per symbol.

Quantization and Vector Quantization

The rate distortion function represents the lower bound on the rate that is needed to represent a source with a particular distortion. We now consider simple algorithms that represent a continuous random variable with a few bits. Suppose we want to represent a single sample from a continuous source. Let the random variable to be represented be X and let the representation of X be denoted as $\hat{X}(X)$. If we are given R bits to represent X, then the function $\hat{X}$ can take on 2^R values. The problem of optimum **quantization** is to find the optimum set of values for $\hat{X}$ (called the reproduction points or code points) and the regions that are associated with each value $\hat{X}$ in order to minimize the expected distortion.

For example, let X be a Gaussian random variable with mean 0 and variance σ^2, and assume a squared error distortion measure. In this case, we wish to find the function $\hat{X}(X)$ such that $\hat{X}$ takes on at most 2^R values and minimizes $E(X - \hat{X}(X))^2$. If we are given 1 bit to represent X, it is clear that the bit should distinguish whether $X > 0$ or not. To minimize squared error, each reproduced symbol should be at the conditional mean of its region. If we are given 2 bits to represent the sample, the situation is not as simple. Clearly, we want to divide the real line into four regions and use a point within each region to represent the samples within that region.

We can state two simple properties of optimal regions and reconstruction points for the quantization of a single random variable:

- Given a set of reconstruction points, the distortion is minimized by mapping a source random variable X to the representation $\hat{X}(w)$ that is closest to it (in distortion). The set of regions defined by this mapping is called a Voronoi or Dirichlet partition defined by the reconstruction points.
- The reconstruction points should minimize the conditional expected distortion over their respective assignment regions.

These two properties enable us to construct a simple algorithm to find a "good" quantizer: we start with a set of reconstruction points, find the optimal set of reconstruction regions (which are the nearest neighbor regions with respect to the distortion measure), then find the optimal reconstruction points for these regions (the centroids of these regions if the distortion measure is squared error), and then repeat the iteration for this new set of reconstruction points. The expected distortion is decreased at each stage in the algorithm, so the algorithm will converge to a local minimum of the distortion. This algorithm is called the *Lloyd algorithm* [Gersho and Gray, 1992].

It follows from the arguments of rate distortion theory that we will do better if we encode long blocks of source symbols rather than encoding each symbol individually. In this case, we will consider a block of n symbols from the source as a vector-valued random variable, and we will represent these n-dimensional vectors by a set of 2^{nR} code words. This process is called **vector quantization** (VQ). We can apply the Lloyd algorithm to design a set of representation vectors (the code book) and the corresponding nearest neighbor regions. Instead of using the probability distribution for the source to calculate the centroids of the regions, we can use the empirical distribution from a training sequence. Many variations of the basic vector quantization algorithm are described in Gersho and Gray [1992].

Common information sources like speech produce continuous waveforms, not discrete sequences of random variables as in the models we have been considering so far. By sampling the signal at twice the maximum frequency present (the Nyquist rate), however, we convert the continuous time signal into a set of discrete samples from which the original signal can be recovered (the sampling theorem). We can then apply the theory of rate distortion and vector quantization to such waveform sources as well.

Kolmogorov Complexity

In the 1960s, the Russian mathematican Kolmogorov considered the question "What is the intrinsic descriptive complexity of a binary string?" From the preceding discussion, it follows that if the binary string were a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$, then on the average it would take $nH(X)$ bits to represent the sequence. But what if the bits were the first million bits of the binary expansion of π? In that case, the string appears random but can be generated by a simple computer program. So if we wanted to send these million bits to another location which has a computer, we could instead send the program and ask the computer to generate these million bits. Thus, the descriptive complexity of π is quite small.

Motivated by such considerations, Kolmogorov defined the complexity of a binary string to be the length of the shortest program for a universal computer that generates that string. (This concept was also proposed independently and at about the same time by Chaitin and Solomonoff.)

Definition: The ***Kolmogorov complexity*** $K_{\mathcal{U}}(x)$ of a string x with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p) \tag{73.98}$$

the minimum length over all programs that print x and halt. Thus $K_{\mathcal{U}}(x)$ is the shortest description length of x over all descriptions interpreted by computer $\mathcal{U}$.

A universal computer can be thought of as a Turing machine that can simulate any other universal computer. At first sight, the definition of Kolmogorov complexity seems to be useless, since it depends on the particular

computer that we are talking about. But using the fact that any universal computer can simulate any other universal computer, any program for one computer can be converted to a program for another computer by adding a constant length "simulation program" as a prefix. Thus, we can show [Cover and Thomas, 1991] that for any two universal computers, $\mathcal{U}$ and $\mathcal{A}$,

$$\left|K_{\mathcal{U}}(x) - K_{\mathcal{A}}(x)\right| < c \tag{73.99}$$

where the constant c, though large, does not depend on the string x under consideration. Thus, Kolmogorov complexity is universal in that it does not depend on the computer (up to a constant additive factor).

Kolmogorov complexity provides a unified way to think about problems of data compression. It is also the basis of principles of inference (Occam's razor: "The simplest explanation is the best") and is closely tied with the theory of computability.

Data Compression in Practice

The previous subsections discussed the fundamental limits to compression for a stochastic source. We will now consider the application of these algorithms to some practical sources, namely, text, speech, images, and video. In real applications, the sources may not be stationary or ergodic, and the distributions underlying the source are often unknown. Also, in addition to the efficiency of the algorithm, important considerations in practical applications include the computational speed and memory requirements of the algorithm, the perceptual quality of the reproductions to a human observer, etc. A considerable amount of research and engineering has gone into the development of these algorithms, and many issues are only now being explored. We will not go into the details but simply list some popular algorithms for the different sources.

Text

English text is normally represented in ASCII, which uses 8 bits/character. There is considerable redundancy in this representation (the entropy rate of English is about 1.3 bits/character). Popular compression algorithms include variants of the Lempel–Ziv algorithm, which compress text files by about 50% (to about 4 bits/character).

Speech

Telephone quality speech is normally sampled at 8 kHz and quantized at 8 bits/sample (a rate of 64 kbits/s) for uncompressed speech. Simple compression algorithms like adaptive differential pulse code modulation (ADPCM) [Jayant and Noll, 1984] use the correlation between adjacent samples to reduce the number of bits used by a factor of two to four or more with almost imperceptible distortion. Much higher compression ratios can be obtained with algorithms like linear predictive coding (LPC), which model speech as an autoregressive process, and send the parameters of the process as opposed to sending the speech itself. With LPC-based methods, it is possible to code speech at less than 4 kbits/s. At very low bit rates, however, the reproduced speech sounds synthetic.

Images

A single high-quality color image of 1024 by 1024 pixels with 24 bits per pixel represents about 3 MB of storage in an uncompressed form, which will take more than 14 minutes to transmit over a 28800-baud modem. It is therefore very important to use compression to save storage and communication capacity for images. Many different algorithms have been proposed for image compression, and standards are still being developed for compression of images. For example, the popular GIF standard uses a patented version of Lempel–Ziv coding, and the JPEG standard being developed by the Joint Photographic Experts Group uses an 8 by 8 discrete cosine transform (DCT) followed by quantization (the quality of which can be chosen by the user) and Huffman coding. Newer compression algorithms using wavelets or fractals offer higher compression than JPEG. The compression ratios achieved by these algorithms are very dependent on the image being coded. The lossless compression methods achieve compression ratios of up to about 3:1, whereas lossy compression methods achieve ratios up to 50:1 with very little perceptible loss of quality.

Video

Video compression methods exploit the correlation in both space and time of the sequence of images to improve compression. There is a very high correlation between successive frames of a video signal, and this can be exploited along with methods similar to those used for coding images to achieve compression ratios up to 200:1 for high-quality lossy compression. Standards for full-motion video and audio compression are being developed by the Moving Pictures Experts Group (MPEG). Applications of video compression techniques include video-conferencing, multimedia CD-ROMs, and high-definition TV.

A fascinating and very readable introduction to different sources of information, their entropy rates, and different compression algorithms can be found in the book by Lucky [1989]. Implementations of popular data compression algorithms including adaptive Huffman coding, arithmetic coding, Lempel–Ziv and the JPEG algorithm can be found in Nelson and Gailly [1995].

Defining Terms

Code: A mapping from a set of messages into binary strings.

Entropy: A measure of the average uncertainty of a random variable. For a random variable with probability distribution $p(x)$, the entropy $H(X)$ is defined as $\Sigma_x - p(x) \log p(x)$.

Huffman coding: A procedure that constructs the code of minimum average length for a random variable.

Kolmogorov complexity: The minimum length description of a binary string that would enable a universal computer to reconstruct the string.

Lempel-Ziv coding: A dictionary-based procedure for coding that does not use the probability distribution of the source and is nonetheless asymptotically optimal.

Quantization: A process by which the output of a continuous source is represented by one of a set of discrete points.

Rate distortion function: The minimum rate at which a source can be described to within a given average distortion.

Vector quantization: Quantization applied to vectors or blocks of outputs of a continuous source.

Related Topics

17.1 Digital Image Processing • 69.5 Digital Audio Broadcasting

References

T. Bell, J. Cleary, and I. Witten, *Text Compression*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

T. M. Cover and J. A. Thomas, *Elements of Information Theory,* New York: Wiley, 1991.

R. Gallager, *Information Theory and Reliable Communication,* New York: Wiley, 1968.

A. Gersho and R. Gray, *Vector Quantization and Source Coding,* Boston: Kluwer Academic, 1992.

N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video,* Englewood Cliffs, N.J.: Prentice-Hall, 1984.

G. Langdon, "An introduction to arithmetic coding," *IBM Journal of Research and Development,* vol. 28, pp. 135–149, 1984.

R. Lucky, *Silicon Dreams: Information, Man and Machine,* New York: St. Martin's Press, 1989.

M. Nelson and J. Gailly, *The Data Compression Book,* 2nd ed., San Mateo, Calif.: M & T Books, 1995.

C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal,* vol. 27, pp. 379–423, 623–656, 1948.

J. Ziv and A. Lempel, "Compression of individual sequences by variable rate coding," *IEEE Trans. Inform. Theory,* vol. IT–24, pp. 530–536, 1978.

Further Information

Discussion of various data compression algorithms for sources like speech and images can be found in the *IEEE Transactions on Communications* and the *IEEE Transactions on Signal Processing*, while the theoretical underpinnings of compression algorithms are discussed in the *IEEE Transactions on Information Theory*.

Some of the latest developments in the areas of speech and image coding are described in a special issue of the *IEEE Journal on Selected Areas in Communications*, June 1992. It includes an excellent survey by N.S. Jayant of current work on signal compression, including various data compression standards.

Special issues of the *IEEE Proceedings* in June 1994 and February 1995 also cover some of the recent developments in data compression and image and video coding.

A good starting point for current information on compression on the World Wide Web is the FAQ for the newsgroup comp.compression, which can be found at

http://www.cis.ohio-state.edu/hypertext/faq/usenet/compression-faq/top.html.

DiFonzo, D.F. “Satellites and Aerospace”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

74

Satellites and Aerospace

74.1 Introduction
74.2 Satellite Applications
74.3 Satellite Functions
74.4 Satellite Orbits and Pointing Angles
74.5 Communications Link
74.6 System Noise Temperature and *G/T*
74.7 Digital Links
74.8 Interference
74.9 Some Particular Orbits
74.10 Access and Modulation
74.11 Frequency Allocations
74.12 Satellite Subsystems
74.13 Trends

Daniel F. DiFonzo
Planar Communications Corporation

74.1 Introduction

The impact of satellites on world communications since commercial operations began in the mid-1960s is such that we now take for granted many services that were not available a few decades ago: worldwide TV, reliable communications with ships and aircraft, wide area data networks, communications to remote areas, direct TV broadcast to homes, position determination, and earth observation (weather and mapping). New and proposed satellite services include global personal communications to hand-held portable telephones, and broadband voice, video, and data to and from small user terminals at customer premises around the world.

Satellites function as line-of-sight microwave relays in orbits high above the earth which can see large areas of the earth's surface. Because of this unique feature, satellites are particularly well suited to communications over wide coverage areas such as for broadcasting, mobile communications, and point-to-multipoint communications. Satellite systems can also provide cost-effective access for many locations where the high investment cost of terrestrial facilities might not be warranted.

74.2 Satellite Applications

Figure 74.1 depicts several kinds of satellite links and orbits. The geostationary earth orbit (GEO) is in the equatorial plane at an altitude of 35,786 km with a period of one sidereal day (23h 56m 4.09s). This orbit is sometimes called the Clarke orbit in honor of Arthur C. Clarke who first described its usefulness for communications in 1945. GEO satellites appear to be almost stationary from the ground (subject to small perturbations) and the earth antennas pointing to these satellites may need only limited or no tracking capability.

An orbit for which the highest altitude (apogee) is greater than GEO is sometimes referred to as high earth orbit (HEO). Low earth orbits (LEO) typically range from a few hundred km to about 2000 km. Medium earth orbits (MEO) are at intermediate altitudes. Circular MEO orbits, also called Intermediate Circular Orbits (ICO)

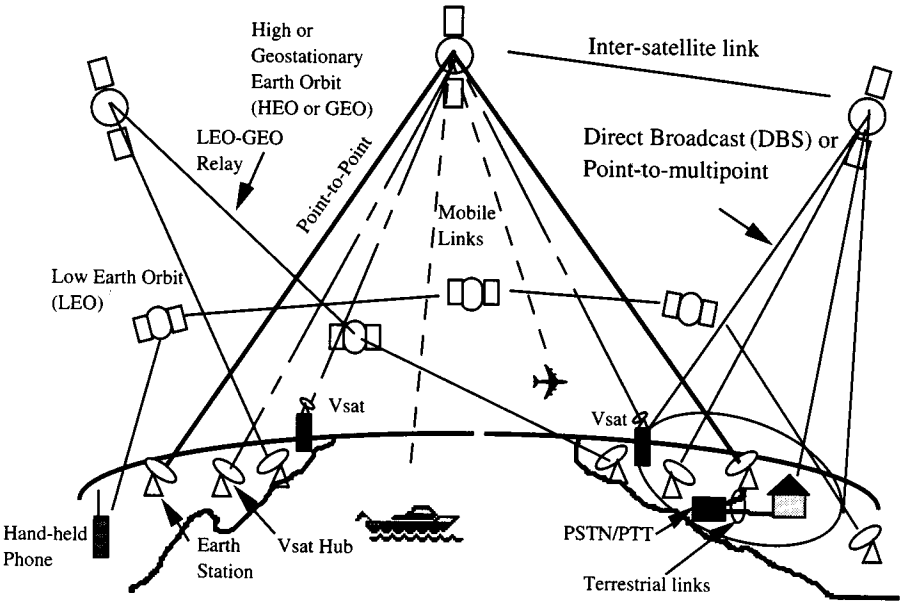


FIGURE 74.1 Several types of satellite links. Illustrated are point-to-point, point-to-multipoint, VSAT, direct broadcast, mobile, personal communications, and intersatellite links.

have been proposed at an altitude of about 10,400 km for global personal communications at frequencies designated for Mobile Satellite Services (MSS) [Johannsen, 1995].

LEO systems for voice communications are called *Big LEOs.* Constellations of so-called *Little LEOs* operating below 1 GHz and having only limited capacity have been proposed for low data rate non-voice services, such as paging and store and forward data for remote location and monitoring, for example, for freight containers and remote vehicles and personnel [Kiesling, 1996].

Initially, satellites were used primarily for point-to-point traffic in the GEO fixed satellite service (FSS), e.g., for telephony across the oceans and for point-to-multipoint TV distribution to cable *head end* stations. Large earth station antennas with high-gain narrow beams and high uplink powers were needed to compensate for limited satellite power. This type of system, exemplified by the early global network of the International Telecommunications Satellite Organization (INTELSAT) used Standard-A earth antennas with 30-m diameters. Since then, many other satellite organizations have been formed around the world to provide international, regional, and domestic services.

As satellites have grown in power and sophistication, the average size of the earth terminals has been reduced. High gain satellite antennas and relatively high power satellite transmitters have led to *very small aperture* earth *terminals* (VSAT) with diameters of less than 2 m, modest powers of less than 10 W [Gagliardi, 1991] and even smaller *ultra-small aperture terminals* (USAT) diameters typically less than 1 m. As depicted in Fig. 74.1, VSAT terminals may be placed atop urban office buildings, permitting private networks of hundreds or thousands of terminals, which bypass terrestrial lines. VSATs are usually incorporated into *star* networks where the small terminals communicate through the satellite with a larger *Hub* terminal. The hub retransmits through the satellite to another small terminal. Such links require two *hops* with attendant time delays. With high gain satellite antennas and relatively narrow-band digital signals, direct single-hop *mesh* interconnections of VSATs may be used.

74.3 Satellite Functions

The traditional function of a satellite is that of a bent pipe quasilinear repeater in space. As shown in Fig. 74.2, *uplink* signals from earth terminals directed at the satellite are received by the satellite's antennas, amplified, translated to a different *downlink* frequency band, channelized into *transponder channels,* further amplified to

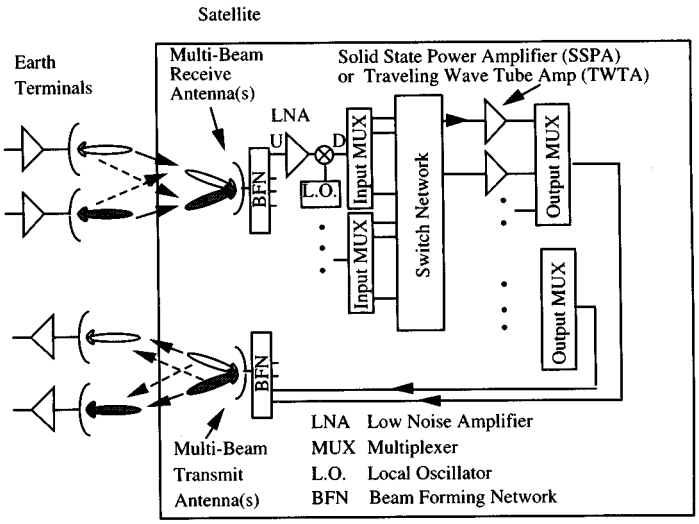


FIGURE 74.2 A satellite repeater receives uplink signals (U), translates them to a downlink frequency band (D), channelizes, amplifies to high power, and retransmits to earth. Multiple beams allow reuse of the available band. Interference (dashed lines) can limit performance. Downconversion may also occur after the input multiplexers. Several intermediate frequencies and downconversions may be used.

relatively high power, and retransmitted toward the earth. Transponder channels are generally rather broad (e.g., bandwidths from 24 MHz to more than 100 MHz) and each may contain many individual or user channels.

The functional diagram in Fig. 74.2 is appropriate to a satellite using frequency-division duplex (FDD), which refers to the fact that the satellites use separate frequency bands for the uplink and downlink and where both links operate simultaneously. This diagram also illustrates a particular *multiple access* technique, known as frequency-division multiple access (FDMA), which has been prevalent in mature satellite systems.

Multiple access, to be discussed later, allows many different user signals to utilize the satellite's resources of power and bandwidth without interfering with each other. Multiple access systems segregate users by frequency division (FDMA) where each user is assigned a specific frequency channel, space-division multiple access (SDMA) by *frequency reuse,* that is by reusing the same frequencies on multiple spatially isolated beams, time-division multiple access (TDMA) where each user signal occupies an entire allocated frequency band but for only part of the time, polarization-division (PD) where frequencies may be reused on spatially overlapping but orthogonally polarized beams, and code-division multiple access (CDMA) where different users occupy the same frequency band but use spread spectrum signals that contain orthogonal signaling codes [Sklar, 1988; Richharia, 1995].

Frequency modulation (FM) has been the most widely used modulation. However, advances in digital voice and video compression have led to the widespread use of digital modulation methods such as quadrature phase shift keying (QPSK) and quadrature amplitude modulation (QAM) [Sklar, 1988].

Newer satellite architectures incorporate digital modulations and on-board demodulation of the uplink signals to baseband bits, subsequent switching and assignment of the baseband signals to an appropriate downlink antenna beam, and re-modulation of the clean baseband signals prior to downlink transmission. These *regenerative repeaters* or *onboard processors* permit flexible routing of the user signals and can improve the overall communications link by separating the uplink noise from that of the downlink. The baseband signals may be those of individual users or they may represent frequency-division multiplexed (FDM) or time-division multiplexed (TDM) signals from many users.

Examples include the NASA Advanced Communications Technology Satellite (ACTS) and the Iridium® system built by Motorola for Iridium LLC. The ACTS is an FDD satellite system operating in the Ka-bands with uplink frequencies from 29.1 to 30.0 GHz and downlink frequencies from 19.2 to 20.1 GHz. It is intended to demonstrate technologies for future broadband voice, video, and data services applicable to the emerging concepts of the Global Information Infrastructure (GII) and National Information Infrastructure (NII) [Gedney, 1996].

Proposed Ka-band satellite systems that would operate at the 20- and 30-GHz bands may incorporate intersatellite links at Ka-band or even at 60 GHz. These systems are intended to provide broadband voice, video, and data services for the GII. Systems have been proposed for operation at GEO and LEO.

The Iridium satellites operate at LEO (altitude = 780 km) with time-division duplex (TDD), using the same 1.6-GHz L-band frequencies for transmission and reception but only receiving or transmitting for somewhat less than half the time each. Iridium uses 66 LEO satellites for personal communications systems (PCS) to enable communications directly to and from small handheld portable telephones at any time and anywhere in the world. Other PCS satellite systems will operate at 1.6 GHz for the uplink and 2.5 GHz for the downlink (e.g., FCC filings for Globalstar and Odyssey).

High-power *direct broadcast satellites* (DBS) or *direct-to-home* (DTH) satellites are operating at Ku-band. In the U.S., satellites operating in the broadcast satellite service (BSS) with downlink frequencies of 12.2 to 12.7 GHz, deliver TV directly to home receivers having parabolic dish antennas as small as 46 cm (18 in.) in diameter. DBS with digital modulation and compressed video is providing more than 150 National Television Systems Committee (NTSC) TV channels from a single orbital location having an allocation of 32 transponder channels, each with 24-MHz bandwidth. DBS is seen as an attractive medium for delivery of high-definition TV (HDTV) to a large number of homes. Other systems using analog FM are operational in Europe and Japan. In the U.S., DTH is also provided by satellites in the FSS frequency bands of 11.7 to 12.2 GHz. These are constrained by regulation to operate at lower downlink power and, therefore, require receiving dishes of about 1-m diameter.

Digital radio broadcast (DRB) from high power GEO satellites has been proposed for direct broadcast of digitally compressed near-CD quality audio to mobile and fixed users in the 2310-2360 MHz bands. [Briskman, 1996].

Mobile satellite services (MSS) operating at L-band around 1.6 GHz have revolutionized communications with ships and aircraft, which would normally be out of reliable communications range of terrestrial radio signals. The International Maritime Satellite Organization (INMARSAT) operates the dominant system of this type.

Links between LEO satellites (or the NASA Shuttle), and GEO satellites are used for data relay, for example, via the NASA tracking and data relay satellite system (TDRSS). Some systems will use intersatellite links (ISL) to improve the interconnectivity of a wide-area network. ISL systems would typically operate at frequencies such as 23 GHz, 60 GHz, or even use optical links.

74.4 Satellite Orbits and Pointing Angles

Reliable communication to and from a satellite requires a knowledge of its position and velocity relative to a location on the earth. Details of the relevant astrodynamic formulas for satellite orbits are given in Griffin and French [1991], Morgan and Gordon [1989], and Chobotov [1991]. Launch vehicles needed to deliver the satellites to their intended orbits are described in Isakowitz [1991].

A satellite, having mass m, in orbit around the earth, having mass M_e, traverses an elliptical path such that the centrifugal force due to its acceleration is balanced by the earth's gravitational attraction, leading to the equation of motion for two bodies:

$$\frac{d^2\mathbf{r}}{dt^2} + \frac{\mu}{r^3}\mathbf{r} = 0 \tag{74.1}$$

where r is the radius vector joining the earth's center and the satellite and $\mu = G(m + M_e) \approx GM_e = 398{,}600.5$ km^3/s^2 is the product of the gravitational constant and the mass of the earth. Because $m \ll M_e$, the center of rotation of the two bodies may be taken as the earth's center, which is at one of the focal points of the orbit ellipse.

Figure 74.3 depicts the orbital elements for a geocentric right-handed coordinate system where the x axis points to the first point of Aries, that is, the fixed position against the stars where the sun's apparent path around the earth crosses the earth's equatorial plane while traveling from the southern toward the northern

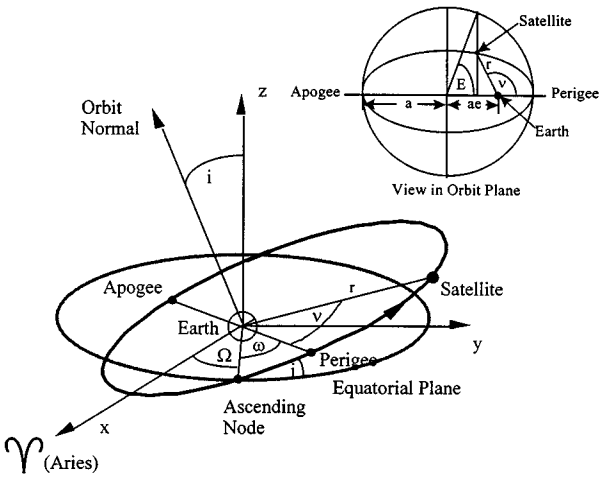


FIGURE 74.3 Orbital elements.

hemisphere at the vernal equinox. The z axis points to the north and the y axis is in the equatorial plane and points to the winter solstice. The elements shown are longitude or right ascension of the ascending node Ω measured in the equatorial plane, the orbit's inclination angle i relative to the equatorial plane; the ellipse semimajor axis length a, the ellipse eccentricity e, the argument (angle) of perigee ω, measured in the orbit plane from the ascending node to the satellite's closest approach to the earth; and the true anomaly (angle) in the orbit plane from the perigee to the satellite ν.

The mean anomaly M is the angle from perigee that would be traversed by a satellite moving at its mean angular velocity n. Given an initial value M_o, usually taken as 0 for a particular epoch (time) at perigee, the mean anomaly at time t is $M = M_o + n(t - t_o)$, where $n = \sqrt{\mu/a^3}$. The eccentric anomaly E may then be found from Kepler's transcendental equation $M = E - e\sin E$ which must be solved numerically by, for example, guessing an initial value for E and using a root finding method. For small eccentricities, the series approximation $E \approx M + e\sin M + (e^2/2)\sin 2M + (e^3/8)(3\sin 3M - \sin M)$ yields good accuracy [Morgan and Gordon, 1989, p. 806]. Other useful quantities include the orbit radius, r, the period, P, of the orbit, [i.e., for $n(t - t_o) = 2\pi$], the velocity, V, and the radial velocity, V_r:

$$r = a(1 = e\cos E) \tag{74.2}$$

$$P = 2\pi\sqrt{a^3/\mu} \tag{74.3}$$

$$V^2 = \mu\left(\frac{2}{r} - \frac{1}{a}\right) \tag{74.4}$$

$$V_r = \frac{e(\mu a)^{1/2}\sin E}{a(1 - e\cos E)} \tag{74.5}$$

Figure 74.4 depicts quantities useful for communications links in the plane formed by the satellite, a point on the earth's surface and the earth's center. Shown to approximate scale for comparison are satellites at altitudes representing LEO, MEO, and GEO orbits.

For a satellite at altitude h, and for the earth's radius at the equator $r_e = 6378.14$ km, the slant range r_s, elevation angle to the satellite from the local horizon el, and the satellite's nadir angle θ, are related by simple

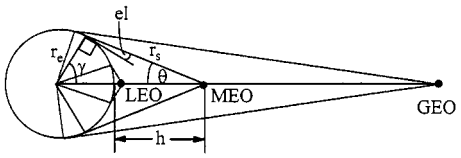


FIGURE 74.4 Geometry for a satellite in the plane defined by the satellite, the center of the earth, and a point on the earth's surface. The elevation angle, *el*, is the angle from the local horizon to the satellite. Shown to approximate scale are satellites at LEO, MEO (or ICO), and GEO.

trigonometry formulas. Note that $\theta + el + \gamma = 90°$, where γ is the earth's central angle and the ground range from the subsatellite point is γr_e. Then,

$$k = \frac{r_e + h}{r_e} = \frac{\cos(el)}{\sin\theta} \tag{74.6}$$

$$\tan(el) = \frac{(\cos\gamma - 1/k)}{\sin\gamma} \tag{74.7}$$

$$r_s = r_e\sqrt{1 + k^2 - 2k\cos\gamma} \tag{74.8}$$

The earth station azimuth angle to the satellite measured clockwise from north in the horizon plane is given in terms of the satellite's declination d, the observer's latitude, ϕ, and the difference of the east longitudes of observer and satellite, $\Delta\lambda$. Then:

$$\tan A = \frac{\sin\Delta\lambda}{(\cos\phi\tan\delta - \sin\phi\cos\Delta\lambda)} \tag{74.9}$$

taking due account of the sign of the denominator to ascertain the quadrant.

The fraction of the earth's surface area covered by the satellite within a circle for a given elevation angle, *el*, and the corresponding earth central angle, γ, is

$$\frac{a_c}{a_e} = \frac{1 - \cos\gamma}{2} \tag{74.10}$$

74.5 Communications Link

Figure 74.5 illustrates the elements of the radio frequency (RF) link between a satellite and earth terminals. The overall link performance is determined by computing the link equation for the uplink and downlink separately and then combining the results along with interference and intermodulation effects.

For a radio link with only thermal noise, the received carrier-to-noise power ratio is

$$\left(\frac{c}{n}\right) = (p_t g_t)\left(\frac{1}{4\pi r_s^2}\right)\left(\frac{g_r}{T}\right)\left(\frac{1}{k}\right)\left(\frac{\lambda^2}{4\pi}\right)\left(\frac{1}{a}\right)(\rho)\left(\frac{1}{b}\right) \tag{74.11a}$$

The same quantities expressed in dB are

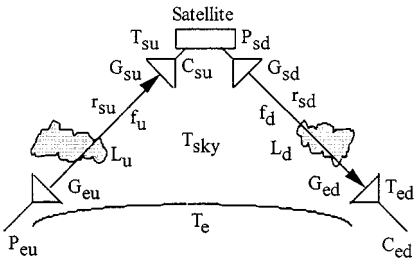


FIGURE 74.5 Quantities for a satellite RF link. P = transmit power (dBW). G = antenna gain (dBi.) C = received carrier power (dBW). T = noise temperature (K). L = dissipative loss (dB). r_s = slant range (m). f = frequency (Hz). u = uplink. d = downlink. e = earth. s = satellite.

$$\left(C/N\right) = EIRP - 10\log\left(4\pi r_s^2\right) + \left(G_r - 10\log T\right) + 228.6 - 10\log\left(4\pi/\lambda^2\right) - A + \Gamma - B \tag{74.11b}$$

where the subscripts in Eq. (74.11a) refer to transmit (t) and receive (r). Lower case terms are the actual quantities in watts, meters, etc. and the capitalized terms in Eq. (74.11b) correspond to the decibel (dB) versions of the parenthesized quantities in Eq. (74.11a). For example, $EIRP = P + G = 10\log p + 10\log g$ decibels relative to 1 W (dBW) and the expression (C/N) should be interpreted as $10\log c - 10\log n$. The uplink and downlink equations have identical form with the appropriate quantities substituted in Eq. (74.11). The relevant quantities are described below.

The ratio of received carrier power to noise power c/n, and its corresponding decibel value $(C/N) = 10\log(c/n)$ dB is the primary measure of link quality. The product of transmit power p_t (W) and the transmit antenna gain g_t, or equivalently, $P_t\,(dBW) + G_t$ [(dBi), that is, gain expressed in decibels relative to an isotropic antenna] is called the equivalent isotropically radiated power ($EIRP$) and its unit is dBW because the antenna gain is dimensionless. The antenna gain is that *in the direction of the link,* i.e., it is not necessarily the antenna's peak gain. The received thermal noise power is $n = kTB$ W where $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant and $10\log(k) = -228.6$ dBW/K/Hz. T is the system noise temperature in kelvins (K) and B is the bandwidth in dB Hz. Then, $G - 10\log T$ dB/K is a figure of merit for the receiving system. It is usually written as G/T and read as "gee over tee". The antenna gain and the noise temperature must be defined at the same reference point, e.g., at the receiver's input port or at the antenna terminals.

The spreading factor $4\pi r_s^2$ is independent of frequency and depends *only* on the slant range distance r_s. The gain of an antenna with an effective aperture area of 1 m^2 is $10\log(4\pi/\lambda^2)$, where the wavelength $\lambda = c/f$, f is the frequency in Hz, and $c = 2.9979 \times 10^8$ m/s is the velocity of light. The dB sum of the spreading factor and the gain of a 1-m^2 antenna is the frequency-dependent "path loss". "A" is the signal attenuation due to dissipative losses in the propagation medium. B is the bandwidth in dB Hz, i.e., $B = 10\log(b)$ where b is the bandwidth in Hz.

The polarization mismatch factor between the incident wave and the receive antenna, is given by $\Gamma = 10\log\rho$ where $0 \le \rho \le 1$. This factor may be obtained from the voltage axial ratio of the incident wave r_w, the voltage axial ratio of the receive antenna's polarization response r_a, and the difference in tilt angles of the wave and antenna polarization ellipses $\Delta\tau = \tau_w - \tau_a$, as follows

$$\rho = \frac{1}{2} + \frac{4r_w r_a + \left(r_w^2 - 1\right)\left(r_a^2 - 1\right)\cos\left(2\Delta\tau\right)}{2\left(r_w^2 + 1\right)\left(r_a^2 + 1\right)} \tag{74.12}$$

where the axial ratios are each signed quantities, having a positive sign for right-hand sense and a negative sign for left-hand sense. Therefore, if the wave and antenna are cross-polarized (have opposite senses), the sign of

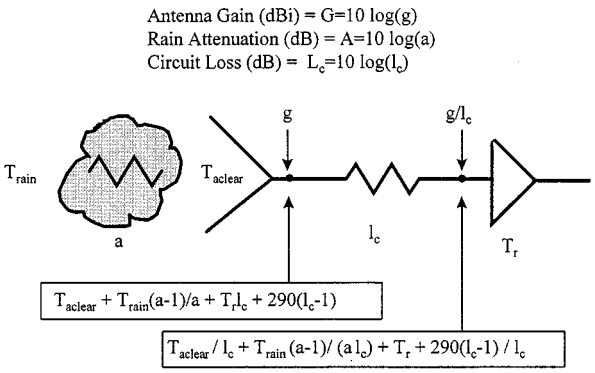


FIGURE 74.6 Tandem connection of antenna, loss elements such as waveguide, and receiver front end. The noise temperature depends on the reference plane but G/T is the same for both points shown.

$4r_w r_a$ is negative. The axial ratio in dB is given as $R = 10\log|r|$. The polarization coupling is maximum when the wave and antenna are copolarized, have identical axial ratios, and their polarization ellipses are aligned ($\Delta\tau = 0$). It is minimum when the axial ratios are identical, the senses are opposite, and the tilt angles differ by 90°.

74.6 System Noise Temperature and *G/T*

The system noise temperature, *T*, incorporates contributions to the noise power radiated into the receiving antenna from the sky, ground, and galaxy, as well as the noise temperature due to circuit and propagation losses, and the noise figure of the receiver. The clear sky antenna temperature for a directive earth antenna depends on the elevation angle since the antenna's sidelobes will receive a small fraction of the thermal noise power radiated by the earth which has a noise temperature $T_{earth} \approx 290$K. At 11 GHz, the clear sky *antenna* noise temperature, T_{aclear}, ranges from 5 to 10 K at zenith ($el = 90°$) to more than 50 K at $el = 5°$ [Pratt and Bostian, 1986].

As shown in Fig. 74.6, the system noise temperature is developed from the standard formula for the equivalent temperature of tandem elements including the antenna in clear sky, propagation (rain) loss of $A = 10\log(a)$ dB, circuit losses between the aperture and receiver of L_c dB, and receiver noise figure of F dB (corresponding to receiver noise temperature T_r K). The system noise temperature referred to the antenna aperture is approximated by the following equation where $T_{rain} \approx 280$ K is a reasonable approximation for the physical temperature of the rain [Pratt and Bostian, 1986, p. 342]:

$$T = T_{aclear} + T_{rain}(a-1)/a + T_r l_c + 290(l_c - 1) \qquad (74.13)$$

The system noise temperature is defined at a specific reference point such as the antenna aperture or the receiver input. However, G/T is independent of the reference point when G correctly accounts for circuit losses. The satellite's noise temperature is generally higher than an earth terminal's under clear sky conditions because the satellite antenna sees a warm earth temperature of ≈150–300 K, depending on the proportion of clouds, oceans, and land in the satellite antenna's beam, whereas a directive earth antenna generally sees cold sky and the sidelobes generally receive only a small fraction of noise power from the warm earth. Furthermore, a satellite receiving system generally has a higher noise temperature due to circuit losses in the beam forming networks, protection circuitry, and extra components for redundancy.

Figure 74.7 illustrates the link loss factors, maximum nadir angle, θ, earth central angle, γ, and earth-space time delay as a function of satellite altitude. The delay for a single *hop* between two earth locations includes the delays for the earth-space path, the space-earth path, and all circuit delays. The path losses are shown for several satellite frequencies in use. The variation in path loss and earth central angle is substantial. For example, L-band LEO personal communications systems to low-cost hand-held telephones with low gain (e.g., $G \approx$

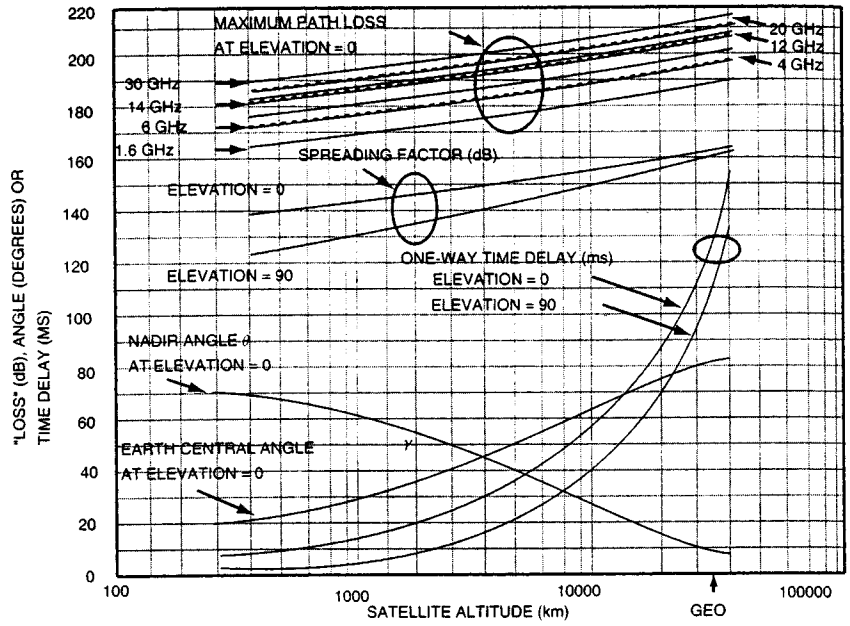


FIGURE 74.7 Satellite link losses, spreading factors, maximum nadir angle, θ–max, earth central angle, γ, and one-way time delay vs. satellite altitude, *h* km.

–2 τo +3 dBi) need less link power than for MEO or GEO. On the other hand, more satellites are needed from LEO constellations to provide full earth coverage since each satellite sees a much smaller fraction of the earth compared with higher orbits.

The design for a constellation of satellites to serve communications needs, such as the number of satellites, their orbital parameters, the satellite *G*/*T* and *EIRP*, etc. are topics related to mission analysis and design and involve trades of many factors such as total communications capacity, link margins, space and earth segment costs, reliability, interconnectivity, availability and cost of launch vehicles, mission lifetime, and system operations [Wertz and Larson, 1991].

74.7 Digital Links

For digital modulation systems, the bit error rate (BER) is related to the dimensionless ratio (dB difference) of energy per bit, E_b dB J and the total noise power density $N_o = 10\log(kT)$ dB J [Sklar, 1988]. For a system with only thermal noise N_o,

$$\left(E_b/N_o\right) = \left(C/N\right) + B - R = \left(C/N_o\right) - R \text{ dB} \tag{74.14}$$

where $R = 10\log$ (bit rate in bit/s), B is the bandwidth (dB Hz), and (C/N_o) is the *carrier-to-thermal noise density ratio,* that is, (C/N) normalized to unit bandwidth. Curves relating the communications performance measure of (BER) vs. (E_b/N_o) for different modulations may be found in [Sklar, 1988]. The link equation may then be expressed in terms of (E_b/N_o) and data rate, R, without explicit reference to the bandwidth:

$$\begin{aligned}\left(E_b/N_o\right) = {} & EIRP + \left(G/T\right) + 228.6 - 20\log\left(4\pi r_s/\lambda\right) \\ & - A + \Gamma - R \quad \text{dB}\end{aligned} \tag{74.15}$$

where the appropriate quantities are substituted depending on whether the uplink or downlink is being considered.

74.8 Interference

A complete transponder link analysis must include the contributions of the uplink, downlink, and also the power sum of all interference signals due, for example, to intermodulation products generated in the output stages of the amplifiers, external interference from other systems, and intra-system interference from reusing the same frequency band on spatially isolated or dual-polarized antenna beams to increase communications capacity. For most applications the total interference power may be taken as the power sum of interfering signals as long as they are not correlated with each other or the desired carrier. The values for the interfering signals due to, for example, frequency reuse cross-polarization, multiple beam interferers, and interference power received from other systems, must be obtained by carefully constructing the link equation for each case, taking into account the antenna gains for each polarization and beam direction of concern.

For an interference power i W, and carrier power, c W the interference ratio, c/i must be combined with the uplink and downlink c/n values to yield the total c/n. Here, the ratios are written in lower case to indicate they are *numerical power ratios.*

$$\left(\frac{c}{n}\right)_{total} = \frac{1}{\left(\frac{c}{n}\right)_u^{-1} + \left(\frac{c}{n}\right)_d^{-1} + \left(\frac{c}{i}\right)_{other}^{-1}} \tag{74.16}$$

Equation (74.16) applies to a "bent pipe" satellite. If on-board signal regeneration is used for digital transmission, the uplink signal is demodulated and a *clean* set of baseband bits is remodulated. This has the effect of separating the accumulation of uplink and downlink noise contributions by causing the uplink noise to be effectively modulated onto the downlink carrier with the desired signal [Gagliardi, 1991]. In that case, only the uplink or the downlink term in the denominator of Eq. (74.16) would be used as appropriate. Remodulation is also useful for intersatellite links. In each case, a savings in power or antenna size may be obtained at the expense of circuit and processing complexity.

The degradation to a digital link from interference follows a form similar to that of Eq. (74.16) in terms of e_b/n_o where the lower case quantities refer to numerical ratios. For a link that is subject to a *given* additive white noise-like interference power expressed as a ratio of desired signal power to interference power, c/i, and assuming digital modulation with m bits per symbol,

$$\frac{e_b}{i_o} = \frac{1}{m}\frac{c}{i} \tag{74.17}$$

The ratio of energy per bit to total thermal noise plus interference power density is

$$\left(\frac{e_b}{n_o + i_o}\right) = \frac{1}{\left(\frac{e_b}{n_o}\right)^{-1} + \left(\frac{e_b}{i_o}\right)^{-1}} \tag{74.18}$$

For a system employing frequency reuse via dual polarizations, the polarization coupling factor Γ between a wave and antenna determines the interference power. The (C/I) due to polarization is the ratio of desired (copolarized) receive power and undesired (cross polarized) receive powers as measured at the same receive port. This *polarization isolation* may be found by application of Eq. (74.12) to co-polarized and cross-polarized cases.

74.9 Some Particular Orbits

A *geosynchronous* orbit has a period that is a multiple of the earth's rotation period, but the orbit is not necessarily circular, and it may be inclined. A *geostationary* earth orbit (GEO) is a special case of a geosynchronous orbit

TABLE 74.1 Comparison of Orbit and Link Parameters for LEO, MEO, and GEO for the Particular Case of Circular Orbits (eccentricity, e, = 0) and for Elevation Angle (el = 10°)

Orbit	LEO	MEO/ICO	GEO
Example system	Iridium®	ICO-P	INTELSAT
Inclination, i (deg.)	86.4	±45	0
Altitude, h (km)	780	10,400	35,786
Semi-major axis radius, a (km)	7159	16,778	42,164
Orbit period (minutes)	100.5	360.5	1436.1
$(r_e + h)/r_e$	1.1222	2.6305	6.6107
Earth central angle, γ (deg.)	18.658	58.015	71.433
Nadir angle, θ (deg.)	61.3	22	8.6
Nadir spread factor			
$10 \log(4\pi h^2$ (dB m^2))	128.8	151.3	162.1
Slant range, r_s (km)	2325	14,450	40,586
One-way time delay (ms)	2.6	51.8	139.1
Maximum spread factor			
$10 \log(4\pi r_s^2$ (dB m^2))	138.3	154.2	163.2
$20 \log(r_s/h$ (dB))	9.5	2.9	1.1
Ground coverage area (km^2)	13.433×106	120.2×106	174.2×106
Fraction of earth area	0.026	0.235	0.34

Note: earth radius, r_e, (km) = 6378.14; earth surface area, a_e, (km^2) = 511.2×10^6; elevation angle, el (degrees) = 10.

where $e = 0$, $i = 0$, $k = (r_e + h)/r_e = 6.61$, and $h = 35{,}786$ km. When $el = 0$ the maximum nadir angle θ = 8.7°, the maximum slant range is 41,680 km, and, from Eq. (74.7), γ = 81.3°. Therefore, a GEO satellite cannot see the earth above 81.3° latitude [Gordon and Morgan, 1993].

Molniya and Tundra orbits have inclination $i = 63.4°$. This highly inclined elliptical orbit (HIEO) causes the satellite's subsatellite ground trace to dwell at apogee at the same place each day. One such orbit whose subsatellite path traces a repetitive loop (LOOPUS) allows several satellites to be phased to offer quasi-stationary satellite service at high latitudes. For full earth coverage from a constellation of LEO satellites, circular polar constellations [Adams and Rider, 1987] and constellations of orbit planes with different inclinations, e.g., Walker Orbits [Walker, 1977] have received attention.

The oblateness of the earth causes the right ascension of the ascending node Ω (Fig. 74.3) to move with time in the equatorial plane in a direction opposite to the satelliteís motion as seen from above the ascending node. This is called regression of the nodes. For inclination $i < 90°$ (prograde orbit) the ascending node rotates westward. For $i > 90°$ (retrograde orbit) the ascending node rotates eastward. For $i = 90°$ the regression is zero. The orbit parameters may be chosen such that the nodal regression is 360°/365.24 = 0.9856° eastward per day. In that case, the orbit plane will maintain a constant angle with the sun. The local solar time for the line of nodes is constant, that is, the satellite crosses a given latitude at the same solar time and same solar lighting conditions each day. This *sun-synchronous* orbit has advantages for certain applications such as weather and surveillance satellites [Roddy, 1996 p. 60].

Table 74.1 compares the geometry, coverage, and some parameters relevant to the communications links for typical LEO, MEO (or ICO), and GEO systems. Reference should be made to Fig. 74.4 for the geometry and to the given equations for geometrical and link parameters.

74.10 Access and Modulation

Satellites act as central relay nodes, which are visible to a large number of users who must efficiently use the limited power and bandwidth resources. For detailed discussions of access issues see Gagliardi [1991], Pritchard et al. [1993], Miya [1985], Roddy [1996], and Feher [1983]. A brief summary of issues specific to satellite systems is now given.

Frequency-division multiple access (FDMA) has been the most prevalent access for satellite systems until recently. Individual users assigned a particular frequency band may communicate at any time. Satellite filters

sub-divide a broad frequency band into a number of *transponder channels*. For example, the 500 MHz uplink FSS band from 5.925 to 6.425 GHz may be divided into 12 transponder channels of 36 MHz bandwidth plus guard bands. This limits the interference among adjacent channels in the corresponding downlink band of 3.7 to 4.2 GHz.

FDMA implies that several individual carriers co-exist in the transmit amplifiers. In order to limit intermodulation products caused by non-linearities, the amplifiers must be operated in a *backed off* condition relative to their saturated output power. For example, to limit third-order intermodulation power for two carriers in a conventional traveling wave tube (TWT) amplifier to Å –20 dB relative to the carrier, its input power must be reduced (*input backoff*) by about 10 dB relative to the power that would drive it to saturation. The output power of the carriers is reduced by about 4 to 5 dB (*output backoff*). Amplifiers with fixed bias levels will consume power even if no carrier is present. Therefore, DC-to-RF efficiency degrades as the operating point is backed off. For amplifiers with many carriers, the intermodulation products have a noise-like spectrum and the noise power ratio is a good measure of multi-carrier performance.

When reusing the available frequency spectrum by multiple spatially isolated beams (SDMA), interference can result if the sidelobes of one beam receives or transmits substantial energy in the direction of the other beams. Two beams that point in the same direction may reuse frequencies provided that they are orthogonally polarized, for example, vertical and horizontal linear polarizations or right- and left-hand circular polarizations. Typical values of sidelobe or polarization isolation among beams reusing the same frequency bands are from 27 to 35 dB.

Time-division multiple access (TDMA) users share a common frequency band and are each assigned a unique time slot for their digital transmissions. At any instant the DC-RF efficiency is high because there is only one carrier in the transmit amplifier, which may be operated near saturation. A drawback is the system complexity required to synchronize widely dispersed users in order to avoid intersymbol interference caused by more than one signal appearing in a given time slot. Also, the total transmission rate in a TDMA satellite channel must be essentially the sum of the users' rates, including overhead bits such as for framing, synchronization and clock recovery, and source coding. Earth terminal hardware costs for TDMA have been higher than for FDMA. Nevertheless, TDMA systems have gained acceptance for some applications as their costs decreased.

Code-division multiple access (CDMA) modulates each carrier with a unique pseudo-random code, usually by means of either a direct sequence or frequency hopping spread spectrum modulation. CDMA users occupy the same frequency band at the same time. The aggregate signal in the satellite amplifier is noise-like and individual signals are extracted at the receiver by correlation processes. CDMA tolerates noise-like interference but does not tolerate large deviations from average loading conditions. One or more very strong carriers could violate the noise-like interference condition and generate strong intermodulation signals. Careful power control of each user's signal is usually required in CDMA systems.

User access is via assignments of a frequency, time slot, or code. Fixed assigned channels allow a user unlimited access. However, this may result in poor utilization efficiency for the satellite resources and may imply higher user costs (analogous to a leased terrestrial line). Other assignment schemes include *demand assigned multiple access* (DAMA) and *random access* (e.g., for the Aloha concept). DAMA systems require the user to first send a channel request over a common control channel. The network controller (at another earth station) seeks an empty channel and instructs the sending unit to tune to it either in frequency or time slot. A link is maintained for the call duration and then released to the system for other users to request. Random access is economical for lightly used burst traffic such as data. It relies on random time of arrival of data packets and protocols are in place for repeat requests in the event of collisions [Gagliardi 1991].

In practice, combinations of multiplexing and access techniques may be used. A broad band may be channelized or *frequency-division multiplexed* (FDM) and FDMA may be used in each sub-band (FDM/FDMA).

74.11 Frequency Allocations

Table 74.2 contains a partial list of frequency allocations for satellite communications. The World Administrative Radio Conference, WARC-92, allocated L-band frequencies for LEO personal communications services and for LEO small satellite data relay. The World Radiocommunication Conference, WRC-95, allocated S-Band frequencies for Mobile Satellite Services (MSS). Most of the other bands have been in force for years.

TABLE 74.2 Partial List of Satellite Frequency Allocations

Band	Uplink	Downlink	Satellite Service
VHF		0.137–0.138	Mobile
VHF	0.3120–0.315	0.387–0.390	Mobile
L-Band		1.492–1.525	Mobile
	1.610–1.6138		Mobile, radio astronomy
	1.613.8–1.6265	1.6138–1.6265	Mobile LEO
	1.6265–1.6605	1.525–1.545	Mobile
		1.575	Global positioning system
		1.227	GPS
S-Band	1.980–2.010	2.170–2.200	MSS (available Jan. 1, 2000)
	1.980–1.990	2.165–2.200	(proposed for U.S. in 2000)
	2.110–2.120	2.290–2.300	Deep-space research
		2.4835–2.500	Mobile
C-Band	5.85–7.075	3.4–4.2	Fixed (FSS)
	7.250–7.300	4.5–4.8	FSS
X-Band	7.9–8.4	7.25–7.75	FSS
Ku-Band	12.75–13.25	10.7–12.2	FSS
	14.0–14.8	12.2–12.7	Direct Broadcast (BSS) (U.S.)
Ka-Band		17.3–17.7	FSS (BSS in U.S.)
			22.55–23.55 Intersatellite
			24.45–24.75 Intersatellite
			25.25–27.5 Intersatellite
	27–31	17–21	FSS
Q	42.5–43.5, 47.2–50.2	37.5–40.5	FSS, MSS
	50.4–51.4		Fixed
		40.5–42.5	Broadcast Satellite
V	54.24–58.2–		Intersatellite
	59–64		Intersatellite

Note: Frequencies in GHz. Allocations are not always global and may differ from region to region in all or subsets of the allocated bands.

Sources: Final Acts of the World Administrative Radio Conference (WARC-92), Malaga-Torremolinos, 1992; 1995 World Radiocommunication Conference (WRC-95). Also, see Gagliardi [1991].

74.12 Satellite Subsystems

The major satellite subsystems are described in, for example, Griffin and French [1991]. They are propulsion, power, antenna, communications repeater, structures, thermal, **attitude** determination and control, telemetry, tracking, and command. Thermal control is described in [Gilmore, 1994].

The satellite *antennas* typically are offset-fed paraboloids. Typical sizes are constrained by launch vehicles and have ranged from less than 1 m to more than 5 m for some applications. The INTELSAT VI satellite used a 3.2 m antenna at 4 GHz. Ku-band satellites may use a diameter $D > 2$ m (i.e., $D > 80\ \lambda$). Multiple feeds in the focal region each produce a narrow *component beam* whose beamwidth is $\approx 65\lambda/D$ and whose directions are established by the displacement of the feeds from the focal point. These beams are combined to produce a shaped beam with relatively high gain over a geographical region. Multiple beams are also used to reuse frequencies on the satellite. Figure 74.2 suggests that a satellite may have several beams for frequency reuse. In that case, the carriers occupying the same frequencies must be isolated from each other by either polarization orthogonality or antenna sidelobe suppression. As long as the sidelobes of one beam do not radiate strongly in the direction of another, both may use the same frequency band to increase the satellite's capacity.

The *repeaters* include the following main elements (see Fig. 74.2): a low noise amplifier (LNA) amplifies the received signal and establishes the uplink noise. The *G/T* of the satellite receiver includes the effect of losses in the satellite antenna, the noise figure of the LNA, and the noise temperature of the earth seen from space (from 150 to 290 K depending on the percentage of the beam area over oceans and clouds). In a conventional repeater, the overall frequency band is down-converted by a local oscillator (LO) and mixer from the uplink band to the downlink band. It is channelized by an input multiplexer into a number (e.g., 12) of transponder channels.

These channelized signals each are amplified by a separate high-power amplifier. Typically, a traveling wave tube amplifier (TWTA) is used with powers from a few watts to >200 W for a DBS. Solid-state amplifiers can provide more than 15 W at C- and Ku-Bands.

The *attitude determination and control system* (ADCS) must maintain the proper angular orientation of the satellite in its orbit in order to keep the antennas pointed to the earth and the solar arrays aimed toward the sun (for example). The two prevalent stabilization methods are spin stabilization and body stabilization. In the former, the satellite body spins and the angular momentum maintains gyroscopic stiffness. The latter uses momentum wheels to keep the spacecraft body orientation fixed. Components of this subsystem include the momentum wheels, torquers (which interact with the earth's magnetic field), gyros, sun and earth sensors, and thrusters to maintain orientation.

The *telemetry tracking and command* (TT&C) subsystem receives data from the ground and enables functions on the satellite to be activated by appropriate codes transmitted from the ground. This system operates with low data rates and requires omni-directional antennas to maintain ground contact in the event the satellite loses its orientation.

The *power* subsystem comprises batteries and a solar array. The solar array must provide enough power to drive the communications electronics as well as the housekeeping functions and it must also have enough capacity to charge the batteries that power the satellite during eclipse, that is, when it is shadowed and receives no power from the sun [Richharia, 1995, p. 39]. Typical battery technology uses nickel-hydrogen cells, which can provide a power density of more than 50 W-h/kg. Silicon solar cells can yield more than 170 W/m^2 at a satellite's beginning of life (BOL). Gallium arsenide solar cells (GaAs) yield more than 210 W/m^2. However, they are more expensive than silicon cells.

The space environment including radiation, thermal, and debris issues are described in Wertz and Larson [1991], Griffin and French [1991], and Committee on Space Debris [1995]. The structure must support all the functional components and withstand the rigors of the launch environment. The thermal subsystem must control the radiation of heat to maintain a required operating temperature for critical electronics [Gilmore, 1994].

74.13 Trends

Satellites continue to exploit their unique wide view of the earth for such applications as broadcast, mobile, and personal communications, and will find new niches for end-to-end broadband communications between customer premises by using the Ka-bands at 20 and 30 GHz and, perhaps, even higher frequencies. Historically, satellite construction has resembled a craft industry with extensive custom design, long lead times, long test programs, and high cost. New trends, pioneered by the lean production and design-to-cost concepts for the Iridium and Globalstar programs are leading to systems having lower cost per unit of capacity and higher reliability. Technology advances that are being pursued include development of light-weight small satellites for economical provision of data and communications services at low cost, more sophisticated on-board processing to improve interconnectivity, microwave and optical inter-satellite links, and improved components such as batteries and antennas with dynamically reconfigurable beams such as may be implemented by digital beam forming techniques [Bjornstrom, 1993].

Defining Terms

Attitude: The angular orientation of a satellite in its orbit, characterized by roll (R), pitch (P), and yaw (Y). The roll axis points in the direction of flight, the yaw axis points toward the earth's center, and the pitch axis is perpendicular to the orbit plane such that $R \times P \rightarrow Y$. For a GEO satellite, roll motion causes north-south beam pointing errors, pitch motion causes east-west pointing errors, and yaw causes a rotation about the subsatellite axis.

Backoff: Amplifiers are not linear devices when operated near saturation. To reduce intermodulation products for multiple carriers, the drive signal is reduced or backed off. Input backoff is the decibel difference between the input power required for saturation and that employed. Output backoff refers to the reduction in output power relative to saturation.

Beam and polarization isolation: Frequency reuse allocates the same bands to several independent satellite transponder channels. The only way these signals can be kept separate is to isolate the antenna response for one reuse channel in the direction or polarization of another. The beam isolation is the coupling factor for each interfering path and is always measured at the receiving site, that is, the satellite for the uplink and the earth terminal for the downlink.

Bus: The satellite bus is the ensemble of all the subsystems that support the antennas and payload electronics. It includes subsystems for electrical power, attitude control, thermal control, TT&C, and structures.

Frequency reuse: A way to increase the effective bandwidth of a satellite system when available spectrum is limited. Dual polarizations and multiple beams pointing to different earth regions may utilize the same frequencies as long as, for example, the gain of one beam or polarization in the directions of the other beams or polarization (and vice versa) is low enough. Isolations of 27 to 35 dB are typical for reuse systems.

Related Topics

69.1 Modulation and Demodulation • 73.2 Noise

References

W. S. Adams and L. Rider, "Circular polar constellations providing continuous single or multiple coverage above a specified latitude," *J. Astronautical Sci.,* 35(2), 155-192, April-June 1987.

G. Bjornstrom, "Digital payloads: enhanced performance through signal processing," *ESA Journal,* 17, 1-29, 1993.

R. D. Briskman, *Satellite Radio Technology,* Washington, D.C.: 16th International Communications Satellite Systems Conference, American Institute of Aeronautics and Astronautics, Feb. 25-29, 1996, pp. 821-825.

A. Chobotov, *Orbital Mechanics,* 2nd ed., Washington, D.C.: American Institute of Aeronautics and Astronautics, 1991.

S. De Gaudenzi, F. Gianetti, and M. Luise, " Advances in satellite CDMA transmission for mobile and personal communications," *Proc. IEEE,* 84 (1), 18-39, 1996.

Committee on Space Debris, National Research Council, *Orbital Debris,* Washington, D.C., National Academy Press, 1995.

K. Feher, *Digital Communications: Satellite/Earth Station Engineering,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.

M. Gagliardi, *Satellite Communications,* New York: Van Nostrand Reinhold, 1991.

R. Gedney, "Considerations for satellites providing NII/GII integrated services using ACTS results," Washington, D.C.: 16th International Communications Satellite Conference, paper AIAA-96-1027-CP, pp. 344-353, Feb. 25-29, 1996.

D. G. Gilmore, Ed., *Satellite Thermal Control Handbook,* El Segundo, Calif.: The Aerospace Corporation Press, 1994.

G. Gordon and W. Morgan, *Principles of Communications Satellites,* New York: John Wiley & Sons, 1993.

M. D. Griffin and J. R. French, *Space Vehicle Design,* Washington, D.C.: American Institute of Aeronautics and Astronautics, 1991.

J. Isakowitz, *International Reference Guide to Space Launch Systems,* 2nd ed., Washington, D.C.: American Institute of Aeronautics and Astronautics, 1991.

K. G. Johannsen, "Mobile P-service satellite system comparison," *Iny. J. Satellite Comm.,* 13, 453-471, 1995.

J. D. Kiesling, *"Little LEOs", an Important Satellite Service,* Washington, D.C.: American Institute of Aeronautics and Astronautics, 16th International Communications Satellite Systems Conference, Feb. 25-29, 1996, pp. 918-928.

K. Miya, Ed., *Satellite Communications Technology,* Tokyo: KDD Engineering and Consulting, Inc., 1985.

W. L Morgan and G. D. Gordon, *Communications Satellite Handbook,* New York: John Wiley & Sons, 1989.

T. Pratt and C. W. Bostian, *Satellite Communications,* New York: John Wiley & Sons, 1986.

W. L. Pritchard, H. G. Suyderhoud, and R. A. Nelson, *Satellite Communications Systems Engineering,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1993.

M. Richharia, *Satellite Systems, Design Principles,* New York: McGraw-Hill, 1995.
Roddy, *Satellite Communications,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1996.
A. Scott, *Understanding Microwaves,* New York: John Wiley & Sons, 1993.
B. Sklar, *Digital Communications,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.
J. G. Walker, "Continuous whole-earth coverage by circular orbit satellite patterns," Technical Report 77044, Farnborough, Hants, U.K.: Royal Aircraft Establishment, 1977.
J. R. Wertz, Ed., *Spacecraft Attitude Determination and Control,* Dordrecht, The Netherlands: D. Reidel Publishing Co., 1978.
J. R. Wertz and W. J. Larsen, Eds., *Space Mission Analysis and Design,* Dordrecht, The Netherlands: Kluwer Academic Publishers, 1991.

Further Information

For a brief history of satellite communications see *Satellite Communications: The First Quarter Century of Service,* by D. Reese, Wiley, 1990. Propagation issues are summarized in *Propagation Effects Handbook for Satellite Systems Design,* NASA Reference Publication 1082(04), 1989. Descriptions of the proposed LEO personal communications systems are in the FCC filings for *Iridium* (Motorola), *Globalstar* (SS/Loral), *Odyssey* (TRW), Ellipso (Ellipsat), and Aries (Constellation Communications), 1991 and 1992. Also, see the FCC filing of Teledesic for a Ka-band LEO broadband system employing 840 satellites. For a discussion of the trends in satellite communications see *An Assessment of the Status and Trends in Satellite Communications 1986-2000,* NASA Technical Memorandum 88867, NASA Lewis Research Center, Cleveland Ohio, November, 1986. For a broad collection of satellite papers, see the AIAA conference proceedings Feb. 25-29,1995, Washington, D.C.

Many of the organizations mentioned can be accessed via the Internet. Several examples include (with the usual http:// prefix): NASA (www.nasa.gov); International Telecommunications Union (ITU) (www.itu.ch); INTELSAT (www.intelsat.int:8080); Inmarsat (www.worldserver.pipex.com/inmarsat/index.htm); FCC (www.fcc.gov/); ICO Global Communications (www.i-co.co.uk); Motorola Satellite Communications (www.sat.mot.com); and Iridium LLC (www.iridium.com).

Lee, W.C.Y., Ziemer, R.E., Ovan, M. Mandyam, G.D. “Personal and Office”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

75

Personal and Office

75.1 Mobile Radio and Cellular Communications
The Difference between Fixed-to-Fixed Radio Communication and Mobile Communication • Natural Problems in Mobile Radio Communications • Description of Mobile Radio Systems • Mobile Data Systems • Personal Communication Service Systems

75.2 Facsimile
Scanning • Encoding • Modulation and Transmission • Demodulation and Decoding • Recording • Personal Computer Facsimile • Group 4 Facsimile

75.3 Wireless Local-Area Networks for the 1990s
The Wireless In-Building Vision • Market Research • LANMarket Factors • Cabling Problems • User Requirements Environment • Product Requirements: End User Reaction • Technology Alternatives in Meeting Customer Requirements

75.4 Wireless PCS
Cellular Band Systems • PCS Services • 3rd Generation Enhancements

William C. Y. Lee
AirTouch Communications, Inc.

Rodger E. Ziemer
University of Colorado at Colorado Springs

Mil Ovan
Motorola, Inc.

Giridhar D. Mandyam
Nokia Research Center

75.1 Mobile Radio and Cellular Communications

William C. Y. Lee

The Difference between Fixed-to-Fixed Radio Communication and Mobile Communication

In fixed-to-fixed radio communications, the transmitter power, antenna location, antenna height, and antenna gain can be determined after calculating the link budget. Also, depending on the frequency range of the carrier affected on the atmospheric variation, different "margin" values will be put in the budget calculation for different system applications. The fixed-to-fixed radio links are usually 10 miles or longer and high above the ground. The signal variation over the link is due mostly to atmospheric changes. Satellite communications, microwave links, troposcatter, etc. are fixed-to-fixed radio communications. In mobile radio communications, the parameters such as transmitter power, antenna location, antenna height, and antenna gain are determined by covering an area or cell. In mobile radio communications, at least one end is in motion. The sizes of cells in urban and suburban areas are less than 10 miles. In mobile radio communications, the design of cell coverage is based on the average power. No margin is applied in calculating the cell coverage.

Natural Problems in Mobile Radio Communications

In mobile radio communications, there are many problems which never occur in fixed-to-fixed radio communication system:

1. *Excessive pathloss:* Vehicles are referred to as mobile units. The antenna height of the mobile unit is very close to the ground. Therefore, the average signal strength received at the mobile unit has two components,

a direct wave and a ground-reflected wave. These two waves act in canceling their average signal strengths and result in excessive pathloss at the receiver.

2. *Multipath fading:* Due to the human-made environment in which mobile units travel, the instantaneous signal sent from the base station is reflected back and forth from buildings and other ground objects before arriving at the mobile unit and causes signal fading received in the time domain. This signal fading causes an increase in the bit error rate (BER) and in the degradation of voice quality.
3. *Human-made noise:* The antenna height of mobile units is usually low. Therefore, human-made industrial noise, automotive ignition noise, etc. are very easily received by the mobile unit. This noise will raise the noise floor and impact system performance.
4. *Dispersive medium:* Due to the human-made environment and the low antenna height of the mobile unit, the signal after bouncing back and forth from the human-made structures produces multiple reflected waves which arrive at the mobile unit at different times. One impulse sent from the base station propagating through the medium becomes multiple reflected impulses received at different times at the mobile unit. This medium is called a dispersive medium. First the dispersive medium does not affect the analog voice channel, but does affect the data channels. Second, the medium becomes effective depending on the transmission symbol rate of the system. The dispersive medium will impact the reception performance when the transmission rate is over 20 kbps. Third, the dispersive medium becomes more effective in urban areas than in suburban areas.

Description of Mobile Radio Systems

There are two basic systems: trunked systems and **cellular systems**.

Trunked Systems

A trunked system is assigned a channel from a number of available channels to a user. The user is never assigned to a fixed channel.

1. Specialized mobile radio (SMR) is a trunked system. The SMR operator is licensed by the FCC to a group of 10 or 20 channels within 14 MHz of the spectrum between 800 and 900 MHz.
   - Loading requirement: A minimum of 70 mobile units per channel is required. SMR can offer privacy, speedier channel access, and efficient services. It can serve up to 125–150 mobile units per channel.
   - Channel spacing: 25 kHz.
   - Channel allocation: The FCC allocates a spectrum of either 500 kHz or 1 MHz to a SMR operator who will serve 10 or 20 paired transmit-receiver voice channels.
   - Coverage: Coverage is about 25 miles in radius since SMR uses only one high-power transmitting tower covering a large area.
   - Telephone interconnect: The public service telephone network (PSTN) extends mobile telephone service to SMR users.
   - Roaming: Mobile units are equipped with software that allows the radio to roam to any SMR system in the network.
   - Handoff: No tower-to-tower **handoff** capability; the channel frequency does not change as the unit moves from one cell to another.
2. ESMR (enhanced SMR): A system used to enhance the SMR system. It was called MIRS (Mobile Integrated Radio Systems). Now it is called IDEN (Integrated Dispatch and Enhanced Network). Features are:
   - Uses the SMR band.
   - Uses **TDMA** (time division multiple access) digital technology, the same digital TDMA standard adopted by the cellular industry.
   - Applies network of low-power cells.
   - Provides cell-to-cell handoffs through a centralized switching facility.
   - A spectrum average of 7–8 MHz is used in each market. The spectrum is not contiguous.
   - A channel bandwidth of 25 kHz is specified with three time slots per channel.
   - Modulation 16 QAM is applied.
   - No equalizer is used.

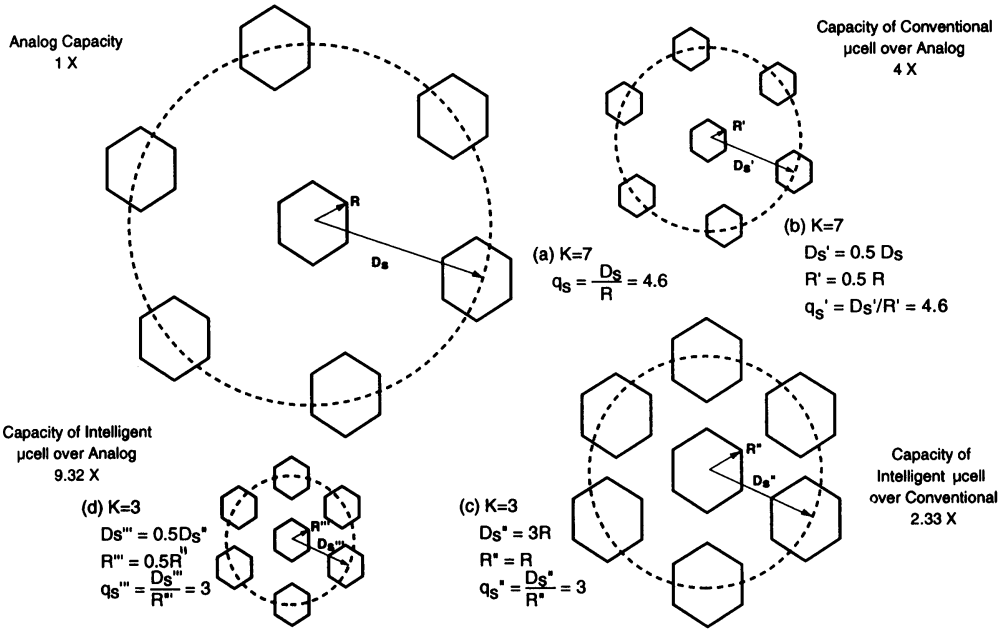


FIGURE 75.1 Four cases of expression of cochannel interference reduction factor.

Cellular Systems

The cellular system [Lee, 1989] is a high-capacity system that uses the frequency reuse concept. The same frequency is used over and over again in different geographical locations. In large cities, the same frequency can be reused over 30 times.

Key Elements: There are several key elements in the cellular system.

- **Cochannel interference reduction factor** q (see Fig. 75.1): Two cells using the same frequency channels are called cochannel cells. The required distance between two cochannel cells in order to receive the accepted voice quality is D_s, and the radius of the cell is R. Then the cochannel interference reduction factor q is

$$q = D_s/R$$

  There are six co-channel cells at the first tier seen from the center cell as shown in Fig. 75.1(a). For an analog cellular system $q_s = 4.6$, and the cell reuse factor K is $K = q_s^2/3 = 7$. The $K = 7$ means that a cluster of seven cells will reuse again and again in a serving area. The capacity increase in a cellular system can be achieved by reducing both the radius of cell R by one half and the separation D_s by one half such that the q_s remains constant and the capacity is increasing by four times. The reason is that a cell shown in Figure 75.1(a) can fit in four small cells shown in Fig. 75.1(b). In Fig. 75.1(c), the size of cells is the same as Fig. 75.1(a), but $q_s = 3$ is achieved by using an intelligent microcell system. The capacity of Fig. 75.1(c) is $7/3 = 2.33$ times over that of Fig. 75.1(a). In Fig. 75.1(d), the radius of the cell is reduced by one half, and K is reduced to 3. The capacity of Fig. 75.1(d) is $4 \times 2.33 = 9.32$ times over that of Fig. 75.1(a).

  The value of q is different in different kinds of cellular systems such as analog, TDMA, and **CDMA** (code division multiple access).

- **Handoff:** Handoff is a feature implemented in cellular systems to handoff a frequency of a cell while the mobile unit changes to another frequency of another cell while the vehicle is entering. The handoff is handled by the system and the user does not notice the handoff occurrences.

TABLE 75.1 Specifications of TDMA and CDMA Systems

TDMA		CDMA	
Bandwidth per channel	30 kHz	Bandwidth per channel	1.23 MHz
Time slots	3	Speech coder	8 kbps(max.)—a variable rate vocoder
Modulation	π/4-DQPSK	Forward radio channels	Pilot (1) sync (1), paging (7), traffic channels (55), total 64 channels
Speech coder	8 kbps—VSELP code (vector sum excited LPC*)		
Channel coding	Rate 1/2 convolutional (13 kbps)	Reverse radio channels	Access (9), traffic channels (55)
Total transmit rate	48 kbps per channel	Power control	Forward, reverse
Equalizer	Up to 40 μs	Diversity	Rake receiver

* LPC = linear predictive code.

- **Cell splitting:** When a cell provides a maximum of 60 radio channels and all are used during busy hours, the cell has to be split into smaller cells in order to provide more radio channels, normally reducing the cell by using a half radius. As a result a cell will be covered by four subcells. Each subcell provides 60 channels. The total area of an original cell will provide 240 radio channels which is four times higher in capacity as compared with the original cell capacity before splitting.

Spectrum Allocation in the United States, Europe, and Japan: In the United States there is 50 MHz of spectrum allocated to cellular radio within 800–900 MHz. Based on duopoly, each city has two licensed operators. Each one operates on a 25-MHz band. There are two bands, Band A and Band B. Each band consists of 416 channels. The channel bandwidth is 30 kHz. Among 416 channels, 21 channels are used for setting up and 395 are used for voice channels.

- *Analog:* The frequency management of both Band A and Band B is shown in Table 75.5.
- *Digital:* There are two potential systems, TDMA and CDMA shown in Table 75.1.

In Europe the spectrum allocation is as shown in Table 75.2 and 75.3.
In Japan the spectrum allocation is as shown in Table 75.4.

TABLE 75.2 Specification of Three European Systems

Analog	England	Scandinavia	West Germany
System	TACS*	NMT*	C450
Transmission frequency (kHz)			
Base station	935–960	463–467.5	461.3–465.74
Mobile station	890–915	453–457.5	451.3–455.74
Spacing between transmission and receiving frequencies (MHz)	45	10	10
Spacing between channels (kHz)	25	25	20
Number of channels	1000 (control channel 21 × 2); interleave used	180	222
Coverage radius (km)	2–20	1.8–40	5–30
Audio signal			
Type of modulation	FM	FM	FM
Frequency deviation (kHz)	±9.5	±5	±4
Control signal			
Type of modulation	FSK	FSK	FSK
Frequency deviation (kHz)	±6.4	±3.5	±2.5
Data transmission rate (kbps)	8	1.2	5.28
Message protection	Principle of majority decision is employed	Receiving steps are predetermined according to the content of the message	Message is sent again when an error is detected

* TACS = total access cellular system; NMT = nordic mobile telephone.

TABLE 75.3 GSM European Standard

GSM Characteristics
• TDMA: 8 slots/radio carrier
• 124 radio carriers (200 kHz/carrier) 935–960 MHz, 890–915 MHz
• GMSK modulation
• Slow frequency hopping (FH) (217 hops/s)
• Block and convolutional channel coding
• Synchronization (up to 233 μs absolute delay)
• Equalization (16 μs dispersion)
• TDMA structure: one frame (8 slots) 4.615 ms; each slot 0.557 ms
• Radio transmission rate: 270.833 kbps
GSM Physical Channels
• RACH: random-access control channel
• BCCH: broadcast common control channel (system parameters, sync.)
• PCH: paging channel
• SDCCH: stand-alone dedicated control channel (for transmit user's data)
• FACCH: fast associate control channel (for handoff)
• SACCH: slow associate control channel (for signaling)
• TCH: traffic channel Full rate: use full rate speech code Half rate

TABLE 75.4 Specification of the Japanese System

Analog	
System	NTT
Transmission frequency (kHz)	
Base station	870–885
Mobile station	925–940
Spacing between transmission and receiving frequencies (MHz)	55
Spacing between channels (kHz)	25
Number of channels	600
Coverage radius (km)	5 (urban area) 10 (suburbs)
Audio signal	
Type of modulation	FM
Frequency deviation (kHz)	±5
Control signal	
Type of modulation	FSK
Frequency deviation (kHz)	±4.5
Data transmission rate (kbps)	0.3
Message protection	Transmitted signal is checked when it is sent back to the sender by the receiver
Digital	
System	PHS* (Japan)
Frequency band	1.9 GHz
Access method	TDMA/TDD (MC)*
Traffic channels/RF carrier	1 (or 8 channels at half rate)
Modulation	π/4-QPSK
Voice codec	32 kbit/s ADPCM
Output power	10 mW
Radio transmission rate	384 kpbs
Carrier spacing	300 kHz

* PHS = personal handy phone system; TDD = time division duplexing; MC = multi-carrier; ADPCM = adaptive differential pulse code modulation.

TABLE 75.5 New Frequency Management (Full Spectrum)

Block A

1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B	1C	2C	3C	4C	5C	6C	7C
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84
85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126
127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147
148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168
169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189
190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210
211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231
232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252
253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273
274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294
295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	667	668	669
670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690
691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711
712	713	714	715	716	X	X	X	X	991	992	993	994	995	996	997	998	999	1000	1001	1002
1003	1004	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023
313*	**314**	**315**	**316**	**317**	**318**	**319**	**320**	**321**	**322**	**323**	**324**	**325**	**326**	**327**	**328**	**329**	**330**	**331**	**332**	**333**

Block B

1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B	1C	2C	3C	4C	5C	6C	7C
334*	**335**	**336**	**337**	**338**	**339**	**340**	**341**	**342**	**343**	**344**	**345**	**346**	**347**	**348**	**349**	**350**	**351**	**352**	**353**	**354**
355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375
376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396
397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417
418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438
439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459
460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480
481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501
502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522
523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543
544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564
565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585
586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606
607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627
628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648
649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	X	X	X
X	X	X	X	X	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732
733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753
754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774
775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795
796	797	798	799																	

*Boldface numbers indicate 21 control channels for Block A and Block B, respectively.

Mobile Data Systems

The design aspect of developing a mobile data system is different from that of developing a cellular voice system, although the mobile radio environment is the same. The quality of a voice channel has to be determined based on a subjective test. The quality of a data transmission is based on an objective test. In a data transmission, the bit error rate and the word error rate are the parameters to be used to measure the performance at any given carrier-to-interference ratio (C/I). The burst errors caused by the multipath fading and the intersymbol interference caused by the time delay spread are the major concerns in receiving the mobile data. The burst errors can be reduced by interleaving and coding. The intersymbol interference can be reduced by using equalizers or lowering the symbol rate or applying diversity.

The wireless data transmission can be sent via a circuit switched network or a packet switched network. Also, mobile data transmission can be implemented on cellular systems or on a stand-alone system.

ARDIS*	
Transmission rate	4.8 kbps and 19.2 kbps
Transmit power	1 W
Channel	Packet radio
Vendors	IBM/Motorola

RAM*	
Transmission rate	8 kbps
Transmit power	4 W
Channel	Packet radio
Vendor	Ericsson

Cellular Plan II	
Transmission rate	19.2 kbps
Transmit power	0.6–1.2 W
Channel	Packet cellular

Cellular Modems	
Transmission rate	38.4 kbps
Transmission	3 W
Channel	Circuit cellular, carry data over cellular voice channels
Modem vendor	AT&T, PowerTek, Vital

*ARDIS = advanced radio data

*RAM = mobile data service

Personal Communication Service Systems

In June 1990, the FCC started to ask the wireless communication industry to study the development of future personal communication service (PCS) systems. (In late 1994, the FCC started to auction off two of the six spectral bands for over 7 billion dollars. In 1996, Band C was auctioned off for over 4 billion dollars.) PCS systems need to have more capacity than cellular systems. The technologies of increasing the capacity not only apply to GSM, but also apply to CDMA (code division multiple access) and the new microcell system.

CDMA

A San Diego field test held in 1991 showed that a cellular CDMA scheme can provide higher capacity than cellular TDMA (time division multiple access). A cellular CDMA system [Lee, May 1991] does not require a frequency reuse scheme. All the CDMA cells share the same radio channel. Therefore, the capacity of a cellular CDMA system is higher than either cellular **FDMA** (frequency division multiple access) or cellular TDMA systems.

Assume that a spectral bandwidth of 1.2 MHz can be divided into 120 radio channels with a channel bandwidth of 10 kHz. This is an FDMA scheme. A spectral bandwidth of 1.2 MHz can also be divided into 40 radio channels with a radio channel bandwidth of 30 kHz but each radio channel carries three time slots. Therefore, a total of 120 time-slot channels is obtained. This is a TDMA scheme. A spectral bandwidth of 1.2 MHz can also be used as one radio channel but provide 40 code-sequence traffic channels for each sector of a cell. A cell of three sectors will have a total of 120 traffic channels. This is a CDMA scheme. Now we can visualize that as far as channel efficiency is concerned, TDMA, FDMA, and CDMA provide the same number of traffic channels. However, in FDMA or TDMA, frequency reuse has to be applied. Let the **frequency reuse factor** $K = 7$ maintain a required C/I $\geq$ 18 dB; then the total channels will be divided by 7 as:

$$\frac{120}{7} = 17 \text{ channels/cell (in TDMA or FDMA)}$$

In CDMA no frequency reuse is required. Therefore, every cell can have the same 120 channels: number of channels/cell (in CDMA). In cellular, because the frequency reuse factor is applied on FDMA and TDMA schemes but not on CDMA, therefore, cellular CDMA has a greater spectrum efficiency than cellular FDMA or TDMA [Lee, May 1991].

New Microcell System

The conventional microcell system [Lee, Nov. 1991, 1993] reduces the transmit power and makes a cell less than 1 km in radius. The concept of using cell splitting is to increase capacity. Furthermore, the new microcell system needs to find a way to make a conventional microcell to be intelligent. The conventional microcell does not have the intelligence to know where the mobile or portable units are located within the cell. Therefore, the cell site has to cover the signal strength over the whole cell or whole sector. The more unnecessary signal power transmitted, the more interference will be caused in the system and less capacity will be achieved. In this new intelligent microcell system, each cell is an intelligent cell. In a new microcell, there are three or more zones. The cell will know which zone a particular mobile unit is in. Then a small amount of power will be needed to deliver in that zone. The cochannel interference reduction factor (CIRF) now will be measured from two cochannel zones instead of two cochannel cells. Then the two cochannel cells can be located much closer. In this new microcell system, the frequency reuse factor K becomes $K = 3$. As compared to the conventional microcell $K = 7$, the new microcell system has a capacity increase of 2.33 (= 7/3) times. These two techniques can be used in buildings and outside buildings.

Defining Terms

CDMA: A multiple access scheme by using code sequences as traffic channels in a comom radio channel.
Cell splitting: A method of increasing capacity by reducing the size of the cell.
Cochannel interference reduction factor (CIRF): A key factor used to design a cellular system to avoid the cochannel interference.
FDMA: A multiple access scheme by dividing an allocated spectrum into different radio channels.
Frequency reuse factor (*K*): A number based on frequency reuse to determine how many channels per cell.
GSM (Global System Mobile): European digital cellular standard using TDMA.
Handoff: A frequency channel will be changed to a new frequency channel as the vehicle moves from one cell to another cell without the user's intervention.
IDEN (Integrated Dispatch and Enhanced Network): A cellular-like system.
Mobile cellular systems: A high-capacity system operating at 800–900 MHz using a frequency reuse scheme for vehicle and portable telephone communications.
PHS (Personal Handy Phone System): A TDD system deployed in Japan.
SMR (Specialized Mobile Radio): A trunked system for dispatch.
TDMA: A multiple access scheme by dividing a radio channel into many time slots where each slot carries a traffic channel.

Related Topic

69.2 Radio

References

W. C. Y. Lee, *Mobile Cellular Telecommunication Systems,* New York: McGraw Hill, 1989.
W. C. Y. Lee, "Overview of cellular CDMA," *IEEE Trans. on Veh. Tech.,* vol. 40, pp. 290–302, May 1991.
W. C. Y. Lee, "Microcell architecture—Smaller cells for greater performance," *IEEE Commun. Magazine,* vol. 29, pp. 19–23, Nov. 1991.
W. C. Y. Lee, *Mobile Communications Design Fundamentals,* 2nd ed., New York: Wiley, 1993.

Further Information

T. S. Rappaport, "The wireless revolution," *IEEE Commun. Magazine,* pp. 52–71, Nov. 1991.
Gilhousan et al., "On the capacity of cellular CDMA systems," *IEEE Trans. Vehicular Technol.,* vol. 40, no. 2, pp. 303–311, May 1991.
D. J. Goodman, "Trends in cellular and cordless communications, *IEEE Commun. Magazine,* pp. 31–39, June 1991.
Raith, K. and Uddenfeldt, J., "Capacity of digital cellular TDMA systems," *IEEE Trans. Vehicular Technol.,* vol. 40, no. 2, pp. 323–331, May 1991.

75.2 Facsimile

Rodger E. Ziemer

Facsimile combines copying with data transmission to produce an image of a **subject copy** at another location, either nearby or distant. Although the Latin phrase *fac simile* means to "make similar," the compressed phrase facsimile has been taken to mean "exact copy of a transmission" since 1815 [Quinn, 1989]. The image of the subject copy is referred to as a *facsimile copy*, or **record copy**. Often the abbreviated reference "fax" is used in place of the longer term *facsimile*.

Facsimile was invented by Alexander Bain in 1842; Bain's system used a synchronized pendulum arrangement to send a facsimile of dot patterns and record them on electrosensitive paper. Over the years, much technological development has taken place to make facsimile a practical and affordable document transmission process. An equally important role in the wide acceptance of facsimile for image transmission has been the adoption of standards by the Consultative Committee on International Telephone and Telegraph (CCITT). The advent of a nationwide dial telephone network in the 1960s provided impetus for the rebirth of facsimile after television put the damper on early facsimile use. Group 2 fax machines which appeared in the mid-1970s were capable of transmitting a page within a couple of minutes. These machines, based on analog transmission methods, were developed by Graphic Sciences and 3M. The Group 3 fax machines, developed in the mid-1970s by the Japanese, are based on digital transmission technology and are capable of transmitting a page in 20 seconds or less. They can automatically switch to an analog mode to communicate with the older Group 1 and 2 fax machines. Group 4 fax units offer the highest resolution at the fastest rates but rely on digital telephone lines which are just now becoming widely available [Quinn, 1989]. Group 3 facsimile will be featured in the remainder of this article. Group 3 facsimile refers to apparatus which is capable of transmitting an 8.5×11-inch page over telephone-type circuits in one minute or less. Detailed standards for Group 3 equipment may be found in Recommendation T.4 of CCITT, Vol. VII.

Facsimile transmission involves the separate processes of *scanning, encoding, modulation, transmission, demodulation, decoding,* and *recording*. Each of these will be described in greater detail below.

Scanning

Before transmission of the facsimile signal, the subject copy must be **scanned**. This involves the sensing of the diffuse reflectances of light from the elemental areas making up the subject copy. For CCITT Group 3 high-resolution facsimile, these elemental areas are rectangles 1/208 inch wide by 1/196 inch high. The signal corresponding to an elemental area is called a **pixel** which stands for picture element. For pixels that can assume only one of two possible states (i.e., white on black or vice versa), the term used is a **pel**. Various arrangements of illuminating sources, light-sensing transducers, and mechanical scanning methods can be employed. For more than six sweeps per second across the subject copy, electronic scanning utilizing a cathode-ray tube or photosensitive arrays or laser sources with polygon mirrors are utilized. A photosensitive array arrangement for scanning a flood-illuminated subject copy is illustrated in Fig. 75.2. This is the most often encountered scanning mechanism for modern facsimile scanners, and the sensors are typically silicon photosensitive devices. Two photosensor arrays in common use are photodiode arrays and charge-coupled device linear image sensors. For digital facsimile, the array is composed of 1728 sensors in a row 1.02 inches long with the optics designed so that an 8.5 inch subject copy can be scanned.

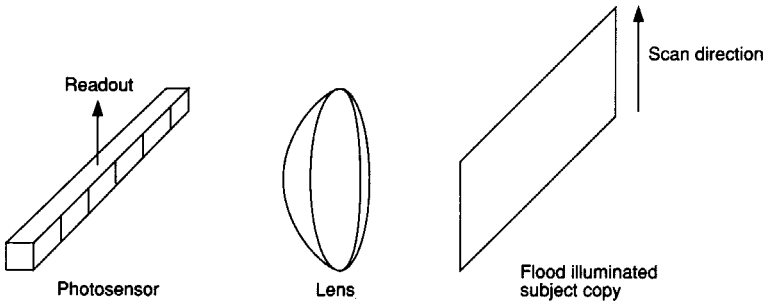


FIGURE 75.2 Arrangement for scanning by means of a linear photosensitive array.

Encoding

The output of the photosensor array for one scan or row of the subject copy consists of 1728 pels (1s or 0s) since Group 3 facsimile recognizes only black or white. Typically, facsimile subject copy is 85% white. The data from scanning the subject copy is reduced through run-length encoding. In the encoding process, it is assumed that a white pel (0) always occurs first. A white run is the number of 0s until the first 1 is encountered (the run length is 0 if the first pel is a 1); after a white run, a black run must follow with length equal to the number of 1s until the first 0 is encountered. All possible run lengths of white and black are then encoded into a binary code using a modified Huffman encoding technique [Jayant and Noll, 1984]. On the average, fewer binary symbols are needed to encode run lengths of the subject copy than if the binary values of the pels themselves were transmitted. For Group 3 facsimile, compression is optionally extended to the vertical dimension through employment of a READ (relative element address designate) code.

Run lengths from 0 to 63 are encoded by *terminating codes*, and run lengths in equal multiples of 64 from 64 to 1728 are encoded by *makeup codes*. Thus any run length up to 1728 can be described by a makeup code plus an appropriate terminating code. Additional makeup codewords are available for equipment that accommodates wider paper while maintaining the same resolution. Tables of modified Huffman run-length terminating and makeup codes are given in [Jayant and Noll, 1984].

Modulation and Transmission

Transmission of the encoded facsimile signal makes use of modem signaling techniques based on CCITT recommendations V.27 (standard) and V.29 (optional addition). The former utilizes 8-phase modulation at 4800 bits per second (bps), and the latter employs 16-QAM (quadrature amplitude modulation) at 9600 bps with adaptive, linear equalization. A facsimile telephone call consists of five phases, labeled A through E. In phase A, the telephone call is placed, with a training sequence sent consisting of signals to establish carrier detection, AGC, timing synchronization, and adjust equalizer tap settings. Phase B consists of the called station responding with a confirmation to receive (CFR) signal. The response is a 300 bps binary coded frequency-shift keyed signal (1 = 1650 ± 6 Hz and 0 = 1850 ± 6 Hz), except for the equalizer training sequence which is at the fast data rate of the digital modem. In phase C the encoded facsimile image is transmitted. Phase D consists of the end-of-transmission signal consisting of six consecutive end-of-lines (EOLs), with receipt required from the receiver. If no more images are to be sent or received, phase E (going on-hook) is effected at both terminals.

Demodulation and Decoding

Demodulation consists of the inverse of the modulation process. Standard techniques are used to demodulate the phase-modulated or QAM signals. Also included in the demodulation process is equalization. The decoding process converts the run-length encoded information to a series of 1s and 0s corresponding to the black and white pels of the image. The demodulated and decoded signal is then used to control the recording of the image.

Recording

Recording of the image at the receiver is effected by applying electricity, heat, light, ink jet, or pressure to a recording medium [Stamps, 1982]. Xerography or ink jet techniques can be used to record on plain paper. Other recording means using electricity, heat, or pressure require specially coated papers. Except for the ink jet, recording processes requiring only a one-step process utilize specially coated papers. Marking transducers are used to apply the image to the recording medium.

Personal Computer Facsimile

Whereas character-oriented text is readily transmitted between personal computers by means of teletex or computer mail, facsimile transmission in conjunction with personal computers extends this capability to images [Hayashi and Motegi, 1989].

Group 4 Facsimile

As mentioned previously, Group 4 facsimile apparatus is used mainly on public data networks of the circuit switched, packet switched, or integrated services digital network varieties (ISDN). Group 4 facsimile machines are subdivided into the following three classes [Yasuda, 1985]:

- Class 1 with the minimum requirement that such equipment can send and receive documents containing facsimile-encoded information
- Class 2, which in addition to the Class 1 capabilities, must be able to receive *teletex* and *mixed-mode documents*
- Class 3, which in addition to Class 1 and Class 2 capabilities, must be able to generate and send *teletex* and *mixed-mode documents.*

An additional feature of the specifications for Group 4 facsimile is that the resolution is equal in the horizontal and vertical directions. Standard resolution for Class 1 is 200 pels per 25.4 mm, and that for Classes 2 and 3 is 200 and 300 pels per 25.4 mm. When operating as a mixed-mode terminal, a receiving density of 240 pels per 25.4 mm is required, which is optional for all three classes. Bit rates range from 2.4 to 48 kbits/s with 64 kbits/s for ISDN. Compression techniques applicable to Group 4 facsimile are overviewed in Arps and Truong [1994].

Defining Terms

Facsimile: The process of making an exact copy of a document through scanning of the subject copy, electronic transmission of the resultant signals modulated by the subject copy, and making a record copy at a remote location.

Mixed-mode documents: Documents containing both character and facsimile information within a page. Such documents can be handled by Group 4 facsimile very efficiently. Group 3 facsimile treats each document as an image to be transmitted pixel by pixel. Mixed-mode documents are subdivided into increasingly smaller parts, such as pages, frames, and blocks. A block is a rectangular area that can contain only one category (character or facsimile information).

Pel: A picture element which has been encoded as black or white, with no gray scale in between.

Pixel: A picture element of a subject or record copy that is represented in shades of gray.

Record copy: The copy of the document made at the receiving end of a facsimile system.

Run-length encoding: The assignment of a codeword to each possible run of 0s (white pel sequence) or run of 1s (black pel sequence) in a scan of the subject copy.

Scanning: The process of scanning the subject copy in a facsimile transmission from left to right and from top to bottom.

Subject copy: The document that is scanned and transmitted in a facsimile system.

Teletex: Representation of character information by code words. Such a representation considerably improves the efficiency of the transmission process, but is not suitable for handwritten characters.

Related Topics

69.1 Modulation and Demodulation • 70.1 Coding

References

R. B. Arps and T. Truong, "Comparison of international standards for lossless still image compression," *IEEE Proc.,* vol. 82 (June), pp. 889–899, 1994.

K. Hayashi and C. Motegi, "Personal computer image communications using facsimile," *IEEE Journal on Selected Areas in Communications,* vol. 7, pp. 276–282, Feb. 1989.

N. S. Jayant and P. Noll, *Digital Coding of Waveforms,* Englewood Cliffs, N.J.: Prentice Hall, 1984, chap. 10.

G. V. Quinn, *The FAX Handbook,* Blue Ridge Summit, Pa.: Tab Books, 1989.

Y. Yasuda, Y. Yamazaki, T. Kamae, and K. Kobayashi, "Advances in FAX," *IEEE Proc.,* vol. 73 (April), pp. 706–730, 1985.

Further Information

C. Chamzas and D. L. Duttweiler, "Encoding facsimilie images for packet-switched networks," *IEEE Journal on Selected Areas in Communications,* vol. 7, pp. 857–864, June 1989.

G. Held, *Data Compression: Techniques and Applications: Hardware and Software Considerations,* 2nd ed., New York: John Wiley & Sons, 1987.

K. McConnell, D. Bodson, and R. Schaphorst, *FAX: Digital Facsimilie Technology and Applications,* Boston: Artech House, 1992.

75.3 Wireless Local-Area Networks

Mil Ovan

Wireless local-area networks (LANs) represent a new form of communications among personal computers inside buildings. To better understand its applicability, this paper defines the customer challenges in networking personal computers as well as specific product requirements for a wireless LAN. These insights were gained through market studies of over 1000 corporate and government entities surveyed through different marketing research techniques.

The Wireless In-Building Vision

To date, the evolution of wireless communications has been exemplified by the dramatic growth in **cellular communications**. Cellular has enabled customers to transcend the constraints of fixed telephony in communicating outside of buildings with portable and now personal communications devices.

There has been significant interest and publicity regarding wireless in-building communications lately, both for data and voice. Throughout the 1980s, we have seen the development of a significant range of in-building business communications problems that have been caused by changes in the technological, business, and regulatory environments. Because of these developments, buyers of telecommunications and data communications systems increasingly are having to face significant time, cost, and logistical problems associated with the installation, movement, and management of computing and communications equipment in dynamic office environments.

Over the next 20 years, society will witness a significant "wireless evolution" in both personal and professional communications and change the way we conduct our lives at home, on the road, and at work (see Fig. 75.3). New forms of wireless communications will free us from the "bonds" of wire that today restrict our movements or interaction.

Market Research

Beginning in the middle to late 1980s, a systematic evaluation of the technological and environmental attributes necessary to anticipate and define wireless in-building communications was undertaken. This included a

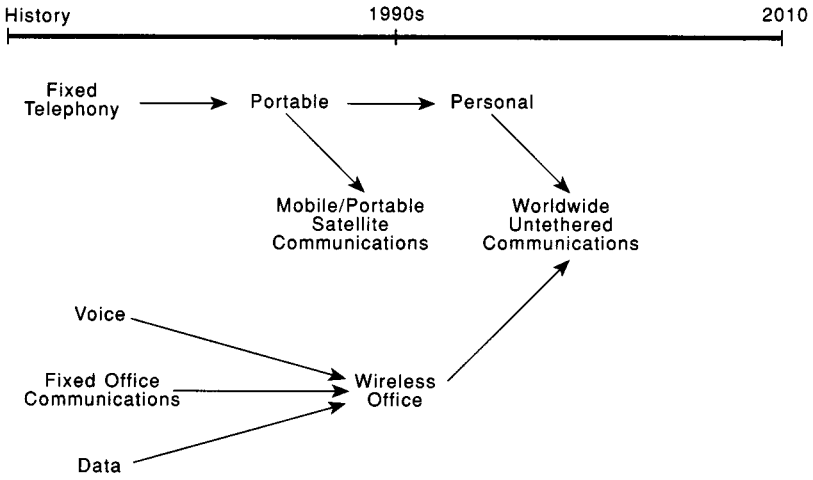


FIGURE 75.3 Evolution of wireless communications. (Courtesy of Motorola, Inc.)

comprehensive marketing needs assessment and research program. The overriding objective was to anticipate and identify customer needs and trends; that is, "What are the specific needs of various customer groups, and what type of product attributes will satisfy their needs?"

To determine answers to these and a whole host of other questions, a multiphased marketing research program was conducted. The overall aim of the program was to anticipate and ascertain the customer need, where this need existed currently, what were the market and customer environmental characteristics, and what product characteristics would be needed to provide an optimal wireless solution.

The remainder of this paper describes a higher-level overview of the results from these market research phases. This includes an overview of market needs, the problems/difficulties with current cabling methods, and a description of market requirements.

LAN Market Factors

Personal Computer Explosion

The move from mainframe and central information processing of the 1960s and 1970s provided an opportunity for minicomputers to enter the market. The minicomputer provided greater computer and applications access by employees. Throughout the 1980s the move to more intelligent desktop devices like personal computers was just that—personal. Organizations, in an effort to empower the worker, provided all types of applications, software, and hardware to the worker. The decremental costs of technology facilitated the distribution of personal computers. More importantly, projections state that business personal computer growth will continue its aggressive pace (see Fig. 75.4).

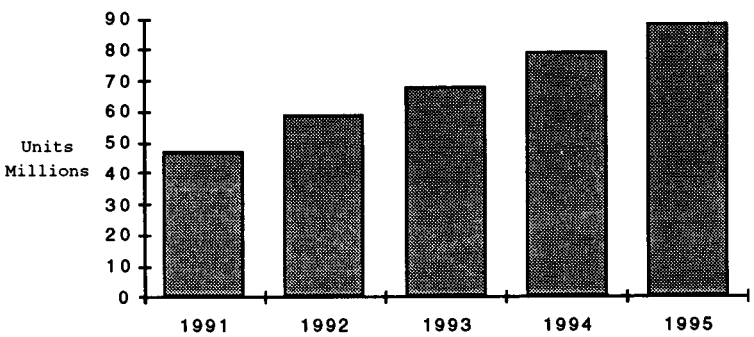


FIGURE 75.4 Worldwide business personal computer-installed base. (*Source:* International Data Corporation.)

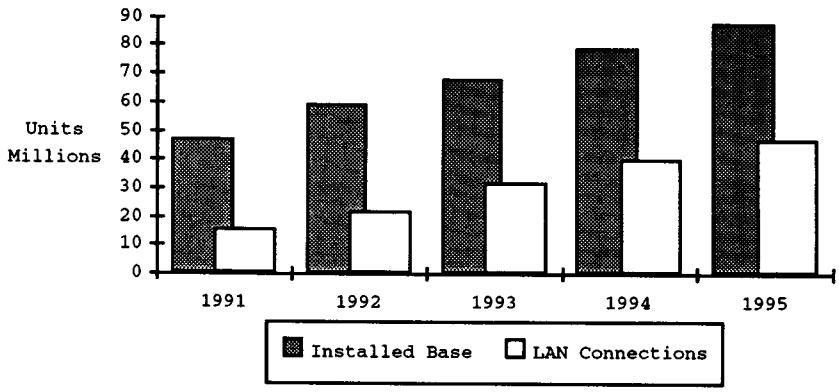


FIGURE 75.5 Worldwide business PCs and those PCs that are LAN connected. (*Source:* International Data Corporation.)

However, the growth of decentralized storage and computing created yet another problem—work groups needed to share information—and much of this information resided in individual hard disks. Furthermore, despite the declining costs of personal computers and associated technology, it was and still is considerably expensive to "fully load" the workforce with all of the applications it needs. The ability to share applications became desirable. It was these two trends which highlighted the need for LANs.

Information and Resource Sharing

The success of LAN computing was predictable. It started with the basic tenet of sharing resources and/or information. The need to amortize and justify the purchase of expensive resources, such as printers and storage, was an obvious factor which supported LAN growth. The need for knowledge workers to exchange data was and is imperative. Furthermore, the ability to share applications supported the growth of network computing.

LAN Growth

The success of LANs throughout the 1980s has been phenomenal. However, the projected growth throughout the 1990s is equally as impressive (see Fig. 75.5). This can be attributed not only to new installations of LANs, but also to the physical and logical segmentation of LANs as traffic and throughput degradations are observed.

Moves/Adds/Changes and Increasing Mobility

The world economies will continue to develop interdependencies and, likewise, global competition. The increasing competitive environment will demand greater worker mobility, changing assignments and reassignments, changing work groups, and mission mobility. The demand to have information how we want it, when we want it, and where we want it will be a strategic and competitive weapon. The need to improve efficiency and the growing need for information will accelerate the adoption of wireless communications.

Today's wired network, for all its great strides, is very restrictive. The cost to deploy and redeploy personnel and workgroups is time consuming and expensive. Cabling in today's environment inhibits the ability to attain efficiency and competitive advantages. The next section will highlight some of the author's market research findings.

Cabling Problems

As each phase of market investigation was conducted, several problems with today's wired networks were uncovered. Whether **copper twisted pair, coax, or optical fiber**, hard wiring for telecommunications and data communications systems within a building environment is expensive and troublesome to install, maintain, and, especially, change. Beneath today's increasingly dense office electronic environment lies a tangled, confusing, unmanageable maze of wiring.

What appeared to be very significant in the focus group research was how quickly the respondents stated the problems they have with wiring. Among the majority of respondents, the most favorable solution was to

free themselves of all wiring. Therefore, their first choice solution would be a wireless system, minimizing the time and effort of implementing a move, add, or change.

Moves/Adds/Changes: Cost and Frequency

A major portion of the cost of LANs is the cost of interconnecting them, which experts acknowledge can sometimes exceed the cost of computer hardware and software. Labor and material costs for wiring are almost always significant and can reach $1000 per node just for copper wire. Coax and optical fiber, not surprisingly, are considerably higher.

The news, however, gets even worse when it comes to maintenance. A study by the Frost and Sullivan[1] group quotes that LAN moves, adds, and changes (MACs) is the third largest cost component for LAN installation and hardware maintenance. They state that MACs account annually for almost $2 billion of a $12.2 billion LAN maintenance market, and that $2 billion does not even include the original cost to install cable.

Estimates of the cost to rewire range from $200 to $1000 per change. In fact, a survey by KPMG Peat Marwick[2] quoted that the average relocation cost for just rewiring a LAN station averages $300 per node. But those are just the direct costs; the time to effect the wired change is a significant problem as well. Moves, for example, often take weeks or longer to coordinate in addition to the time to actually make the physical wiring change.

Most of the research respondents were asked what proportion of their company's staff was involved in some kind of a move involving wiring or rewiring. The majority of the respondents, almost 80%, had some type of relocation or addition of personnel over the last year surveyed. Their responses ranged from as few as 20% per year up to as much as 200% annually. Furthermore, according to the KPMG Peat Marwick study, the average company moves its employees approximately 50% annually. Telecommunications consultant Richard Kuehn states that data terminals are moved as often as 1.5 to 3 times per year. The combined problems of the actual hard relocation costs, however, are just the beginning. Soft, or hidden, costs further exacerbate the cabling dilemma.

Hidden Costs

Significant problems arise when these moves or changes are implemented. There is always the disruption of the workers involved in the move or change, not to mention the loss in productivity. The problems, however, become much more involved when dealing with whole departments and more complex user equipment. In fact, surveyed firms responded that when a relocation takes place, over 60% of the time it involves the movement of an entire department.

The toll of wait time and down time on productivity varies greatly and is difficult to quantify, but certainly is significant and costly. In today's increasingly mobile working environment, it is likely to grow. The situation is exacerbated by relocations and additions which require reconstructions, thereby continuing to add to the effective cost of a move, add, or change.

Costs to rewire rise enormously with the age and complexity of the building. The majority of high-rise office space in large metropolitan areas presents major problems and expense for tenants trying to install, add, or move network wiring. Buildings more than 30 or 40 year old, with designs and construction that did not consider today's electronic office, poorly accommodate communications wiring. If asbestos insulation exists in the building, as it does even in many pre–health-safety regulated buildings, rewiring costs can take on huge proportions.

The coordination of personnel and the moving of one group out to prepare for the new group moving in is a very costly and labor intensive ordeal. In some cases, wiring had to be installed, or different cabling may have been needed to accommodate new or different types of users' equipment.

Cable Is Not Business Friendly

Although office planners, building managers, and network operators are well aware of the problems with wire, the limitations and huge costs of wire have not generated focused attention outside of this community. The general business world seems to accept wire as inevitable. Perhaps that is because there have been no real

[1] *PC Week Magazine*, "Maintenance Costs of LANs Keep Soaring," Frost & Sullivan, Inc.
[2] KPMG Peat Marwick Study, January 1991.

alternatives. Yet, as computing and telecommunications power continues to proliferate and becomes more widely distributed to the "knowledge worker," the problem will increase. Easy, quick, efficient movement of "people assets" within the working environment is also increasingly being recognized as essential to the productivity and competitiveness of a business. Wiring severely inhibits that movement.

The research indicated a need for a flexible, compatible, cost-effective, yet high-performance wireless alternative to extend and complement, if not replace, the capabilities of wire, cable, and fiber for in-building communications networks. More specifically, it is the convenience and flexibility that users need. In fact, the aggregate need for flexibility and convenience was found to be twice that of the perceived benefit for cost savings.

When research respondents were asked how they could improve upon their experiences when implementing a move, add, or change, many solutions were offered. These solutions ranged from having more compatibility among different vendors' equipment, to providing a better way to organize all the different cabling.

Structured Distribution Systems

A number of firms in the research study had deployed a **structured distribution system (SDS)**, a topology which advocates cabling saturation of a desired environment to accommodate all potential personnel movements and reconstructions within that office. SDS requires firms to invest large sums of capital initially on the assumption of not knowing how many telecommunications devices may be employed or where the devices are to be located. Consequently SDS usually plans for worst-case conditions, meaning that some or much of wiring systems capability may never be utilized.

However, many firms which have an SDS deployed also expressed those problems which stress their SDS investment. Some of the most frequently mentioned include:

- High equipment addition/relocations exceeding 40% annually
- Expansion and contraction of their workforce
- Changing technology and business support
- Continued investment and vigilance to maintaining the SDS and its intrinsic advantage
- Continued departmental LAN growth requirements

In short, the latter group of SDS respondents provided some notable requirements. A wireless system must:

- Extend the capabilities of their SDS system
- Facilitate the inherent advantages of the SDS
- Offer enhanced flexibility to nonserviced SDS portions of their building or occupancy

These points indicate that even in SDS environments, there is an opportunity to employ wireless devices. Wiring—the expense, time, and inflexibility of installing, moving, and changing—limits the way companies can productively use networks. To stay productive, these LANs have to move and change with the workforce they support. Therefore, a wireless offering must be a *complementary* solution for buildings with an SDS, in bringing wireless flexibility and extensibility to today's networks.

User Requirements Environment

Office Friendly

Several notable conclusions were derived from the marketing research. **Secondary market research** suggests that over 70% of LAN node installations were estimated to reside in an office environment (as opposed to factories and warehouses). Therefore, as an office-oriented offering, a wireless system would have to be, by definition, office friendly. A traditional office is composed of hard offices with opening and closing doors, furniture and personnel movement, cubicles, conference rooms, and walls of varying thickness and substance. Therefore, a wireless system must continually adapt to different and changing conditions and office layouts.

Optimized Service Area

The second wireless in-building need expressed by the office market is the manageability and reuse of any potential system. Unlike the signal propagation characteristics of many lower-frequency radio products, LAN

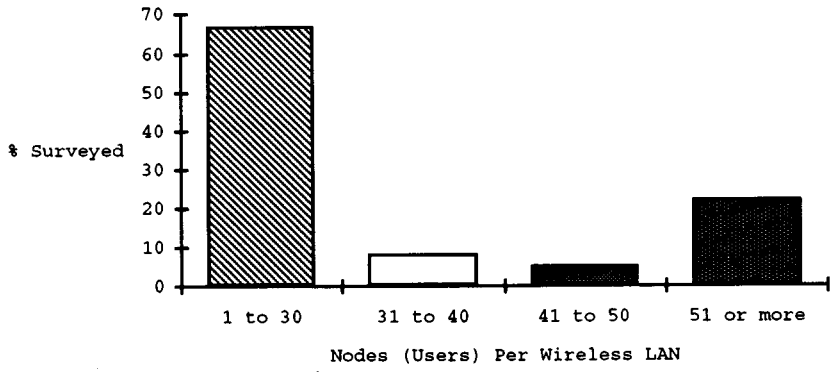


FIGURE 75.6 Survey results: forecast users per wireless LAN. (Courtesy of Motorola, Inc.)

administrators desired the ability to control or, more aptly, contain the coverage of a potential wireless system. The reasons were twofold:

- LAN managers wanted the ability to add different services to a new group of users. In fact these new users may very well be physically adjacent to another system, wireless or wired.
- These same managers wanted the flexibility to connect a new or existing user group to either a **backbone** or to create a stand-alone LAN.

LAN Workgroup Sizes

Respondents were asked as to where a wireless offering might be first installed. The market research indicated that approximately 70% of the installations would contain less than 30 users (see Fig. 75.6). Furthermore, the average LAN appeared to be in the 12- to-15-node range. This is further corroborated by the KPMG Peat Marwick study which found that the average LAN size is about 15 users per LAN.

Furthermore, the system must have the flexibility to manage the service area. That ability, to either incrementally add systems whether on the backbone or in a stand-alone configuration, must accommodate scalability within an organization.

It is interesting to note that these figures are consistent with good LAN administration practices for purposes of maintaining high throughput and fault isolation. As LANs become larger and traffic more intensive, there is a natural inclination to begin segmenting LANs into more logical and defined user areas/groups.

Coverage Area

To satisfy the majority of requirements, we determined that approximately 70% of LANs would be deployed in areas of less than 5000 ft^2 (see Fig. 75.7). This must take into consideration the fairly dense environment, made up of cubicles and apportioned hallway space. The market investigations indicated that a wireless offering must accommodate, at the least, 150 ft^2 per user. This is equivalent to 32 users/system in a 5,000 ft^2 area.

Product Requirements: End User Reaction

Transparency, Compatibility, and Performance

To justify the expense of a wireless system to end users, a wireless offering would have to provide reliable performance, as well as be practical and cost effective. Our market research indicated that the ideal system should be:

- Easy both to install and move, preferably by the user
- Able to coexist with both existing wire and cable, as well as with future optical fiber
- Easy to operate, virtually transparent to the user
- Almost universally applicable, suitable to replace any LAN cable or wire, in any office environment
- Secure, absolutely reliable, and cost effective

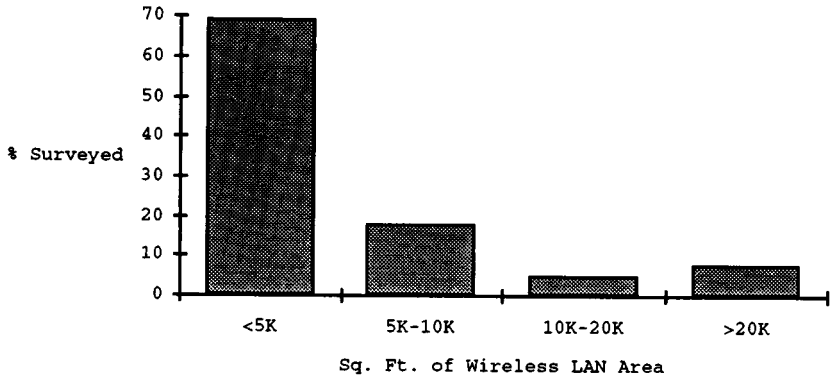


FIGURE 75.7 Survey results: forecast office area of wireless LAN. (Courtesy of Motorola, Inc.)

A wireless system must be totally transparent. If customers are to enjoy the attributes of wireless, the respondents indicated that the wireless implementation must not require the users to change the way they operate or interface with their personal computers. Also, a wireless offering must provide true compatibility. The wireless connection must be compatible with standards-based components such as operating systems and applications, LAN cards and other devices, as well as LAN wire that is already in place.

Security and Reliability

In addition, the market mandated that a wireless product offering be absolutely secure and reliable. Security was a requirement across several dimensions. To provide sufficient data security a wireless system should first prevent the effective capture of data by a receiver outside of the wireless system and, second, prevent capture of the data by unauthorized wireless hardware within the system. A wireless product must be secure from eavesdropping, either accidental or intentional.

Reliability was another important attribute. The users required absolute reliability. That is, users wanted a guaranteed packet delivery from the entry/exit wireline points—and they wanted it at least as error free as their current cabled environment.

Cost Effective

Finally, most businesses will place any capital or expense under rigorous financial analysis. As such, the acceptance of a new technology/application must pass the payback test for that business. Therefore, demonstrable payback and justification is needed to facilitate an organization's evaluation of any potential wireless offering.

Technology Alternatives in Meeting Customer Requirements

Infrared and Spread Spectrum Technologies Lack Performance

Developing the ideal system, obviously, is no trivial problem. Several wireless network products, to be sure, are available, but many suffer from limited performance and operational problems inherent to the technologies on which they are based. As such, they are perhaps interim point solutions, primarily for small networks but are not the long-term answer to the wire communications dilemma described earlier.

Two basic technologies characterize the wireless LANs currently available. Infrared (IR) systems use a part of the electromagnetic spectrum just below visible light as a transmission medium. IR, being light, travels in a straight, or line-of-sight, path. It is blocked by opaque objects and reflects well only off hard, mirror-like surfaces. This factor stands as a serious obstacle to IR systems for applications other than in open working environments.

Radio technologies form the platform for other wireless LAN products. Many of the radio LANs are based on spread spectrum technology. This technology, developed by the U.S. military, uses a combination of several small, narrow bands within a general region of this band as carrier frequencies. However, the commercial products operate with fewer frequencies than are available for military systems; hence interference rejection

and performance are lower. (In radio transmission, the wider the bandwidth or available frequencies on which to encode data, the higher the achievable total data rate.)

Another important issue in this ultrahigh frequency (UHF) environment is that radio frequency (RF) energy at these frequencies tends to propagate through and around obstacles, reaching beyond the confines of the network it is serving. That property makes this RF band suitable for receiving commercial broadcast signals from distant stations through the walls of buildings to receivers inside. It also makes it suitable for mobile cellular telephones, but it cannot be well contained within the confines of the **microcell** described earlier, which limits **spectrum reuse** and overall network capacity requirements absolutely essential in a viable wireless network communications system. Current UHF spread spectrum wireless in-building communications systems, then, suffer from critical bandwidth and spectrum reuse shortcomings that seem likely to prevent them from expanding beyond limited applications.

The 18- to 19-GHz Radio Band: An Ideal Choice for Wireless In-Building Microcellular Networks

A region of the electromagnetic spectrum above the kilohertz and megahertz radio bands, yet below the extremely high frequencies of the infrared band, offers two very compelling advantages. Specifically, the 18- to 19-GHz portion of this band fulfills the key requirements of spectrum reuse and bandwidth availability that eliminates most other frequencies from consideration.

Properties Right for Both Microcell Coverage and Confinement

The first major advantage of the 18-GHz band is its excellent propagation characteristics for a microcellular network. Indeed, the behavior and properties of these higher frequencies that are disadvantages for traditional long-range broadcast applications become critical advantages for wireless microcellular network applications. Propagation characteristics of 18-GHz radio waves make them well suited to diffuse thoroughly through a network microcell using only a minimum of transmitted power, yet still stay confined within it so that the same frequencies can be reused by another system within as little as 120 feet or so, or even on the other side of a dense, continuous barrier such as a cement floor. Typical microcells might encompass, for example, a level or floor, or portions of a floor, in a standard office building.

As one might expect, 18-GHz radio waves exhibit a blend of the characteristics of UHF frequencies below them and IR light above them. For example, 18-GHz waves act like light and unlike radio in that they are blocked and reflected by large structures such as concrete and steel. Reflecting back and forth would allow them to fill an area defined by concrete floors and walls with only very small amounts of transmitted power, yet not pass beyond. They also refract like light, penetrating tiny holes and cracks such as closed doors to diffuse and spread through the space beyond. What little radio signal that might escape the microcell would be rapidly dissipated.

Also, unlike lower frequency radio, 18-GHz radio is of a high enough frequency that not only office equipment but even high-energy factory equipment and processes do not interfere with it. Likewise, with its high-frequency and low-required transmitted power, the 18-GHz signals themselves from such a system would not interfere with other electronic systems or equipment. On the other hand, like radio and unlike light, 18-GHz signals still can pass through less dense materials such as drywall and interior office separators and, combined with their reflectivity, are thus not subject to "line-of-sight" limitations. They can also be modulated to carry information just as traditional radio signals are. Finally, since antenna size and design are largely a function of wavelength, which decreases as frequency increases, the antennas for the 16-mm wavelengths of an 18-GHz radio system would be relatively small and compact.

Plenty of Bandwidth in an Otherwise Crowded Spectrum

The second major advantage of 18-GHz radio is its available bandwidth. Few other areas of the electromagnetic spectrum are as interference-free, clear, and available as this band, certainly not the VHF and UHF bands, which must accommodate television, FM radio, cellular telephone, baby monitors, and more.

The reason for this clear band is largely that these higher frequencies have been difficult to work with. The particular technical properties of 18-GHz frequencies and the expense, size, and complexity of the equipment to use them have prevented them from being an attractive option for many commercial applications. As a result, the military has been the primary developer and user of the 15- to 300-GHz band, and the few commercial

uses that have emerged (weather, aircraft and police radar, point-to-point telecom transmission, etc.) use expensive technology pioneered by military-funded research.

Developing a comprehensive in-building radio system, however, had never been done until Motorola recently developed the Altair™ wireless ethernet network. Such an application required the creation of new, improved performance data handling and signal processing hardware and software, as well as a radio antenna system that could transmit and capture these data speeds on 18-GHz frequencies in an in-building environment.

Summary

The numerous problems with wiring will become even more acute in the office of the 1990s. This environment will be characterized by:

- The proliferation of decentralized computing resources
- Increased number of telephones and personal computers as an outgrowth of a country's economic shift toward service industries

As the penetration of personal computers nears a one-to-one relationship with phones in the office workplace, the limitations of separate voice and data networks will become even more evident. If these problems are not addressed, an organization's flexibility in redeploying "people assets" and ultimately competitiveness will be seriously hindered. The time it takes to move/add/change equipment and reconfigure communications wires will be the limiting factor in rapidly reorganizing workgroups and responding to new assignments. Wireless LANs will become an attractive solution in the office of the 1990s, interconnecting personal computers and offering data communications capabilities without the need for elaborate cabling methodologies. The obvious and inherent flexibility offered by wireless LANs is the obvious primary benefit. However, the ability to retrieve that investment, never retrievable until now, clearly presents a significant economical benefit.

Defining Terms

Backbone: Wiring which runs within and between floors of a building and connects local-area network segments together.

Cellular communications: Traditionally an outside-of-building radio telephone system that allows users to communicate from their car or from their portable telephone.

Copper twisted pair, coax, and optical fiber: Wired media which connects telephone and computer equipment.

Microcell A low-power radio network which transmits its signal over a confined distance.

Secondary marketing research: Market research conducted by other organizations.

Spectrum reuse: Reusing frequencies over and over again in a confined area, resulting in more efficient utilization and higher radio network capacity.

Structured distribution systems (SDS): A topology which advocates cabling saturation of a desired environment to accommodate all potential personnel movements and reconstructions within that office.

Wireless local-area networks: A method of connecting personal computers together without extensive cabling, allowing communications among these devices in an area such as a department or floor of a building.

Related Topic

72.3 Local-Area Networks

References

J. D. Gibson, *The Communications Handbook,* Boca Raton, Fla.: CRC Press, 1997.
N. J. Muller, *Wireless Data Networking,* Boston, Mass.: Artech House, 1995.

Further Information

Articles on LANS appear in *IEEE Communications Magazine* and *IEEE Network Magazine.*

75.4 Wireless PCS

Giridhar D. Mandyam

Personal Communications Services (**PCS**) promise to introduce a wide range of variety of digital wireless services; including high-speed data, improved voice services, and messaging (e-mail or paging). These services are also often identified with the part of the spectrum in which they are deployed, that is, the PCS band. In North America, this part of the spectrum lies between 1850 and 1990 MHz, and is divided into six blocks of either 5 or 15 MHz each. The Federal Communications Commission (FCC) of the United States has been auctioning these blocks since 1994. The DCS (Digital Cellular System) band, which also spans the range of 1.8 to 2 GHz, has been set aside for these advanced services in several parts of the world.

PCS can be contrasted with services already deployed in the cellular band — that is, the part of the spectrum ranging from 806 to 890 MHz. This band is divided into channels of 30 kHz apiece. In North America, the mature analog technology known as **AMPS** is widely deployed and provides the largest amount of coverage of all public wireless technologies available today. However, digital wireless does in fact exist in the cellular band. A depiction of spectral allocation can be found in Figure 75.8.

In addition, digital PCS technologies can be divided into two categories: **2nd generation** and **3rd generation** wireless systems. 3rd generation wireless systems, which have yet to be deployed, promise an improvement on 2nd generation (already deployed) systems in voice quality and data services. In particular, high rate packet data is a critical requirement of 3rd generation systems.

Cellular Band Systems

The first system to appear in the cellular band in the United States was AMPS. This system provided user traffic channels of 30 kHz, as part of a frequency division multiple access (**FDMA**) scheme. In addition, this system used the concept of frequency division duplexing (**FDD**) to provide different channels for an individual user to send and receive traffic. This system used analog frequency-modulation technology, and was primarily designed to provide voice service only, although some systems do exist that provide data services through AMPS. The first public service began operation in the Chicago area in 1983.

A problem with AMPS is the occupation of an entire 30-kHz channel by a single user. This affected the overall capacity of AMPS systems. Another problem is the lack of privacy in AMPS, which has led to a serious problem of phone cloning. In addition, the performance of analog FM in the mobile channel suffers from the threshold effect, where signal quality degrades rapidly once received signal levels fall below a threshold value.

Digital technologies take advantage of coding and modulation to increase signal quality when received signal levels are low. Moreover, digital technologies employ encryption, which addresses to a certain extent the problem of cloning phones. As a result, digital wireless systems were developed for use in the cellular band to address some of the problems with AMPS. A summary of the AMPS radio interface is given in Table 75.6.

A digital technology, which emerged in the United States in the late 1980s, was time division multiple access (**TDMA**). TDMA systems took advantage of time multiplexing different users into the same 30-kHz channel, which was used by AMPS. The **IS-54** public wireless standard introduces the concept of three-slot TDMA, in which three users were time multiplexed into a single 30-kHz channel. This has the effect of tripling the effective capacity of an AMPS network; therefore, sometimes TDMA is referred to as D-AMPS, for digital AMPS. The first commercial TDMA system to be launched in the United States was in 1991. The **IS-136** standard, which

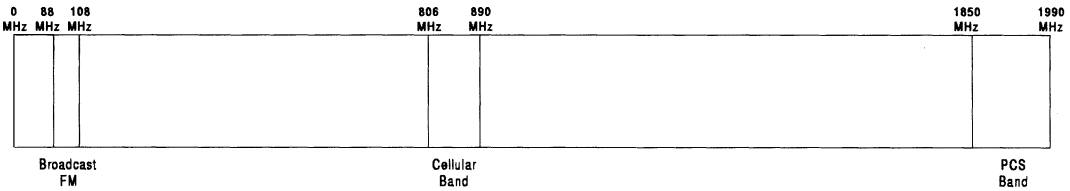


FIGURE 75.8 Frequency allocation.

TABLE 75.6 AMPS System Parameters

System Parameter	AMPS
Multiple access	FDMA
Channel bandwidth	30 kHz
Number of users per channel	1
Reverse (uplink) frequency range	824–849 MHz
Forward (downlink) frequency range	869–894 MHz
Voice modulation	FM
Voice peak frequency deviation	12 kHz
Control channel modulation	Binary FSK
Control channel peak frequency deviation	8 kHz
Control channel data rate	10 Kbps (Manchester encoded)
Control channel error correcting code	BCH

TABLE 75.7 IS-54/IS-136 System Parameters

System Parameter	IS-54/IS-136
Multiple access	TDMA
Channel bandwidth	30 kHz
Number of users per channel	3
Reverse (uplink) frequency range	824–849 MHz
Forward (downlink) frequency range	869–894 MHz
Modulation	π/4-DQPSK
Forward and reverse data rate	48.6 Kbps
Error correcting code	Rate ½ convolutional code, K = 7

was released in the mid-1990s, introduced enhancements to **IS-54,** including PCS functionality. The radio interface for IS-54 and **IS-136** is summarized in Table 75.7.

GSM is the European-originated digital **TDMA** standard. It also appears in the sub-1-GHz part of the spectrum, with the uplink band in the 890 to 915 MHz range and the downlink band in the 935 to 960 MHz range. However, GSM differs from IS-136 in several respects, including modulation (GSM uses GMSK), data rate (270.83 Kbps), and channel spacing (200 kHz). GSM systems exist in many parts of the world outside of Europe, including North America.

Another digital technology, introduced for cellular systems in the early 1990s, was code division multiple access (**CDMA**). This technology is based on the principles of spread spectrum, in which narrow-band user traffic is transformed into wideband signals resembling white noise over the resultant signal bandwidth. This is accomplished by modulating user traffic with a higher-rate spreading sequence, normally generated by maximum-length shift registers. This technology was used in military communications for its inherent security and resistance to jamming. The **IS-95** public wireless standard introduced CDMA, with each user occupying a 1.25-MHz bandwidth. The first commercial CDMA systems were launched in the United States in 1996. The radio interface for IS-95 is summarized in Table 75.8.

TABLE 75.8 IS-95 System Parameters

System Parameter	IS-95
Multiple access	CDMA
Channel bandwidth	1.25 MHz
Reverse (uplink) frequency range	824–849 MHz
Forward (downlink) frequency range	869–894 MHz
Modulation	BPSK — Quadrature Spread
Forward and reverse data rate	9.6 Kbps
Error correcting code	Rate ½ convolutional code, K = 9

It is of interest to note that in **IS-95,** the number of users per data channel is not specified in Table 75.8. This is due to the fact that such a quantity is not easy to derive for all conditions, due to a number of factors. The downlink for IS-95 provides two forms of user channelization, based on spreading sequence time offsets and Walsh codes. However, spreading sequence offsets alone provide uplink channelization. Moreover, both the uplink and downlink are interference-limited, due to multiple users occupying the same frequency band simultaneously.

PCS Services

Both the **IS-136** and IS-95 digital wireless standards have evolved to provide enhanced services for **PCS**. These enhancements include enhanced voice services, data services and short message services (SMS), and paging and e-mail.

Enhanced Voice Services

IS-136 employed an 8 Kbps speech coder, **VSELP**, as its initial codec for voice services. It takes as input 64 Kbps voice in pulse-coded modulation (PCM) format, and outputs 8 Kbps. Processing is performed on 20-ms intervals, or frames. The 13 Kbps **EFR** codec provides an evolution to PCS services, and has been introduced for **TDMA** systems, including IS-136 and **GSM.** This codec provides improvements in voice quality, and provides additional error correction under noisy conditions.

IS-95 also initially employed an 8 Kbps speech coder developed by Qualcomm Inc., **QCELP8.** In order to evolve to higher voice quality, the **QCELP13** 13 Kbps coder was introduced primarily for PCS applications. Both of these codecs employ variable-rate coding, choosing from four different compression rates for each 20-ms frame. This is done to enhance capacity, allowing individual users to occupy more of the channel when their voice activity was high. However, the evolution to 13 Kbps resulted in weakened error-corrective coding. As a result, another variable-rate 8 Kbps codec, **EVRC**, was introduced. This codec provides comparable voice quality to QCELP13 at a lower rate. EVRC uses only three of the original four coding rates used by QCELP8 for each frame.

Data Services

Both IS-95 and IS-136 systems presently support a "packet-over-circuit" approach to data services. This employs transmitting packet data originating from an application (usually running over TCP/IP) through a dedicated wireless traffic channel. Although such data transfer is straightforward, it is wasteful in that a user may be transmitting or receiving blank traffic frames while waiting for a packet burst from the application.

IS-136 also employs the **CDPD** method for packet data transfer over an existing **AMPS** channel. This method is efficient in that the user must "share" the channel with other users, so as to take advantage of the bursty nature of packet data.

IS-95 has also incorporated channel aggregation in recent revisions. As mentioned before, IS-95 supports two voice codec rates of 8 Kbps and 13 Kbps. As proposed in the recent IS-95-B wireless standard, a single user can send or receive multiple traffic channels, each of which either support 8 Kbps or 13 Kbps (but not both rates mixed). The maximum number of such channels is eight in either the uplink or downlink, and only one channel is variable rate.

SMS

Presently, both IS-95 and IS-136 support SMS. SMS is used to provide a host of teleservices, including over-the-air programming and alphanumeric messaging (typically less than 250 characters). SMS can also support low-rate data applications employing a transparent transport layer, such as UDP.

Paging and E-mail

Both IS-95 and IS-136 support user paging and wireless e-mail services. This is normally accomplished through generic data burst messaging on common control channels.

3rd Generation Enhancements

3rd Generation versions of IS-95 and IS-136 are currently being developed and standardized, with the projected deployment being in the PCS band. The requirements of these systems has in general been provided by the

International Telecommunications Union (ITU) as part of its IMT-2000 (International Mobile Telecommunications 2000) project. Enhancements provided to both systems encompass data rates of up to 2 Mbps, suitable performance under a variety of mobile channel conditions (indoor and outdoor), support of simultaneous user services, compatibility with existing **2nd generation** systems, and several other considerations. It is of interest to note that **GSM** and **IS-136** are both evolving to a common standard known as **EDGE**.

Both versions encompass a significant modification of their 2nd generation counterparts, including modulation, error correction coding, and user bandwidths. However, both versions take advantage of existing 2nd generation voice codecs.

In addition, another **CDMA** standard known as Wideband CDMA (**W-CDMA**) is under development in Europe and Japan. Although this standard is a CDMA standard, it is not backwards-compatible with **IS-95,** and is not designed in such a way that W-CDMA equipment can be used interchangeably with **3rd generation** IS-95 equipment.

The development of 3rd generation systems is ongoing, with equipment deployment forecast for the first part of the new millennium. Much work remains to be done before these standards can solidify sufficiently for equipment development and deployment to be completed.

Defining Terms

AMPS: Advanced Mobile Phone Services. A public, multiple-access wireless system that uses analog frequency-modulation technology. This service primarily appears in the cellular band (806–890 MHz).

CDMA: Code Division Multiple Access. A method of multiple access in which individual users are assigned unique code sequences while using a common frequency.

CDPD: Cellular Digital Packet Data. A packet data service over the 30 kHz analog channel. Can easily be overlaid over existing AMPS networks.

EDGE: Enhanced Data Rates for Global TDMA Evolution. 3rd Generation standard for both GSM and IS-136.

EFR: Enhanced Full Rate coding. 13-Kbps linear predictive coder used in IS-136 TDMA systems.

EVRC: Enhanced Variable Rate Coder. 8-Kbps variable-rate linear predictive coder used in IS-95 CDMA systems.

FDD: Frequency Division Duplexing. The practice of providing multiple frequency bands for an individual user. For example an individual user can receive traffic in one band and send traffic in a different band.

FDMA: Frequency Division Multiple Access. A method of multiple access in which individual users are assigned different frequencies.

1st Generation: Term referring to earliest deployed public wireless systems. AMPS is included in this category.

GSM: Global System for Mobile. A TDMA-based digital public wireless system first deployed in Europe.

IS-95: North American CDMA standard, developed by the Telecommunications Industry Association.

IS-54: North American TDMA standard, developed by the Telecommunications Industry Association. Replaced by IS-136.

IS-136: North American TDMA standard, developed by the Telecommunications Industry Association. Supplanted IS-54.

PCS: Personal Communications Services. Refers either to advanced digital wireless services, or to the frequency band where such services are normally deployed (1850–1990 MHz).

QCELP8: Qualcomm Code Excited Linear Predictive coding. 8 Kbps variable-rate linear predictive coder used in IS-95 CDMA systems.

QCELP13: Qualcomm Code Excited Linear Predictive coding. 13 Kbps variable-rate linear predictive coder used in IS-95 CDMA systems.

2nd Generation: Term referring to the first deployed digital public wireless systems. This category includes CDMA and TDMA technologies.

TDMA: Time Division Multiple Access. A method of multiple access where individual users are assigned time slots while using a common frequency.

3rd Generation: Term referring to digital public wireless systems that offer significant enhancements over 2nd generation systems. These enhancements include high-speed Internet access, realtime video communications, high-fidelity voice, multiple simultaneous services per user, broadcast capability, enhanced capacity, and many other desirable features.

VSELP: Vector Sum Excited Linear Predictive coding. 8 Kbps linear predictive coder used in IS-54/IS-136 TDMA systems.

W-CDMA: Wideband CDMA. 3rd Generation wireless standard currently being developed in Europe and Japan.

References

Garg, Vijay K., Kenneth Smolik, and Joseph E. Wilkes, *Applications of CDMA in Wireless/Personal Communications,* Upper Saddle River, NJ: Prentice-Hall, 1997.

Harte, Lawrence J., Adrian D. Smith, and Charles A. Jacobs, *IS-136 TDMA Technology, Economics, and Services,* Boston: Artech, 1998.

Rappaport, Theodore S., *Wireless Communications: Principles and Practice,* Upper Saddle River, NJ: Prentice-Hall, 1996.

Further Information

For further information on 1st generation wireless, see the AMPS standard:
EIA-553 Mobile Station — Land Station Compatibility Specification. Electronics Industry Association. September, 1989.

For further information on 2nd generation wireless, see:

TIA/EIA/IS-136-A TDMA Cellular/PCS-Radio Interface-Mobile Station-Base Station Compatibility, Telecommunications Industry Association, October, 1996, (TDMA standard).
TIA/EIA/IS-95-A Mobile Station — Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System, Telecommunications Industry Association, May, 1995 (CDMA standard).

For further information on 3rd generation wireless, see:

Dennett, Steve, *The cdma2000 ITU-R RTT Candidate Submission,* Telecommunications Industry Associations, July 28, 1998.
Meche, Paul, *Updated UWC-136 RTT,* Telecommunications Industry Association, September 28, 1998.
Ojanpera, Tero and Steven D. Gray, An Overview of cdma2000, WCDMA, and EDGE, *The Mobile Communications Handbook,* Ed. Jerry D. Gibson, Boca Raton, FL: CRC Press, 1999, Ch. 36.

Maddy, S.L., "Phase-Locked Loop"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

76

Phase-Locked Loop

76.1 Introduction
76.2 Loop Filter
76.3 Noise
76.4 PLL Design Procedures
76.5 Components
76.6 Applications

Steven L. Maddy
RLM Research

76.1 Introduction

A *phase-locked loop* (PLL) is a system that uses feedback to maintain an output signal in a specific phase relationship with a reference signal. PLLs are used in many areas of electronics to control the frequency and/or phase of a signal. These applications include frequency synthesizers, analog and digital modulators and demodulators, and clock recovery circuits. Figure 76.1 shows the block diagram of a basic PLL system. The *phase detector* consists of a device that produces an output voltage proportional to the phase difference of the two input signals. The *VCO* (voltage-controlled oscillator) is a circuit that produces an ac output signal whose frequency is proportional to the input control voltage. The *divide by N* is a device that produces an output signal whose frequency is an integer (denoted by *N*) division of the input signal frequency. The **loop filter** is a circuit that is used to control the PLL dynamics and therefore the performance of the system. The $F(s)$ term is used to denote the Laplace transfer function of this filter.

Servo theory can now be used to derive the equations for the output signal phase relative to the reference input signal phase. Because the VCO control voltage sets the frequency of the oscillation (rather than the phase), this will produce a pure integration when writing this expression. Several of the components of the PLL have a fixed gain associated with them. These are the **VCO** control voltage to output frequency conversion gain (K_v), the **phase detector** input signal phase difference to output voltage conversion gain (K_ϕ), and the feedback division ratio (N). These gains can be combined into a single factor called the loop gain (K). This loop gain is calculated using Eq. (76.1) and is then used in the following equations to calculate the loop transfer function.

$$K = \frac{K_\phi \times K_v}{N} \tag{76.1}$$

The closed-loop transfer function $[H(s)]$ can now be written and is shown in Eq. (76.2). This function is typically used to examine the frequency or time-domain response of a PLL and defines the relationship of the phase of the VCO output signal (θ_o) to the phase of the reference input (θ_i). It also describes the relationship of a change in the output frequency to a change in the input frequency. This function is low-pass in nature.

$$H(s) = \frac{\theta_o(s)}{\theta_i(s)} = \frac{KF(s)}{s + KF(s)} \tag{76.2}$$

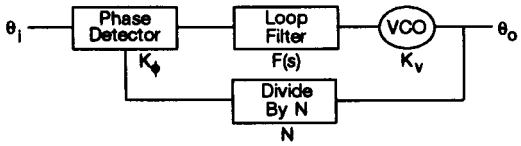


FIGURE 76.1 PLL block diagram.

The loop error function, shown in Eq. (76.3), describes the difference between the VCO phase and the reference phase and is typically used to examine the performance of PLLs that are modulated. This function is high-pass in nature.

$$\frac{\theta_i(s) - \theta_o(s)}{\theta_i(s)} = \frac{\theta_e(s)}{\theta_i(s)} = \frac{s}{s + KF(s)} \tag{76.3}$$

The open-loop transfer function $[G(s)]$ is shown in Eq. (76.4). This function describes the operation of the loop before the feedback path is completed. It is useful during the design of the system in determining the gain and phase margin of the PLL. These are indications of the stability of a PLL when the feedback loop is connected.

$$G(s) = \frac{KF(s)}{s} \tag{76.4}$$

These functions describe the performance of the basic PLL and can now be used to derive synthesis equations. The synthesis equations will be used to calculate circuit components that will give a desired performance characteristic. These characteristics usually involve the low-pass corner frequency and shape of the closed-loop response characteristic [Eq. (76.2)] and determine such things as the loop lock-up time, the ability to track the input signal, and the output signal noise characteristics.

76.2 Loop Filter

The loop filter is used to shape the overall response of the PLL to meet the design goals of the system. There are two implementations of the loop filter that are used in the vast majority of PLLs: the passive lag circuit shown in Fig. 76.2 and the active circuit shown in Fig. 76.3. These two circuits both produce a PLL with a second-order response characteristic.

The transfer functions of these loop filter circuits may now be derived and are shown in Eqs. (76.5) for the passive circuit (Fig. 76.2) and (76.6) for the active circuit (Fig. 76.3).

$$F_p(s) = \frac{sC_1R_2 + 1}{s(R_1 + R_2)C_1 + 1} \tag{76.5}$$

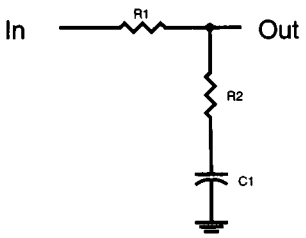


FIGURE 76.2 Passive loop filter.

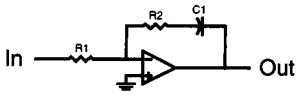


FIGURE 76.3 Active loop filter.

$$F_a(s) = \frac{sR_2C_1 + 1}{sR_1C_1} \tag{76.6}$$

These loop filter equations may now be substituted into Eq. (76.2) to form the closed-loop transfer functions of the PLL. These are shown as Eqs. (76.7) for the case of the passive filter and (76.8) for the active.

$$H_p(s) = \frac{s\dfrac{KR_2}{R_1 + R_2} + \dfrac{K}{(R_1 + R_2)C_1}}{s^2 + s\left[\dfrac{1}{(R_1 + R_2)C_1} + \dfrac{KR_2}{R_1 + R_2}\right] + \dfrac{K}{(R_1 + R_2)C_1}} \tag{76.7}$$

$$H_a(s) = \frac{s\dfrac{KR_2}{R_1} + \dfrac{K}{R_1C_1}}{s^2 + s\dfrac{KR_2}{R_1} + \dfrac{K}{R_1C_1}} \tag{76.8}$$

These closed-loop equations can also be written in the forms shown below to place the function in terms of the **damping factor** (ζ) and the loop natural frequency (ω_n). It will be shown later that these are very useful parameters in specifying PLL performance. Equation (76.9) is the form used for the PLL with a passive loop filter, and Eq. (76.10) is used for the active loop filter case.

$$H_p(s) = \frac{s[2\zeta\omega_n - (\omega_n^2/K)] + \omega_n^2}{s^2 + s2\zeta\omega_n + \omega_n^2} \tag{76.9}$$

$$H_a(s) = \frac{s2\zeta\omega_n + \omega_n^2}{s^2 + s2\zeta\omega_n + \omega_n^2} \tag{76.10}$$

Solving Eqs. (76.7) and (76.9) for R_1 and R_2 in terms of the loop parameters ζ and ω_n, we now obtain the synthesis equations for a PLL with a passive loop filter. These are shown as Eqs. (76.11) and (76.12).

$$R_2 = \frac{2\zeta}{\omega_n C} - \frac{1}{KC} \tag{76.11}$$

$$R_1 = \frac{K}{\omega_n^2 C} - R_2 \tag{76.12}$$

To maintain resistor values that are positive the passive loop filter PLL must meet the constraint shown in Eq. (76.13).

$$\zeta > \frac{\omega_n}{2K} \tag{76.13}$$

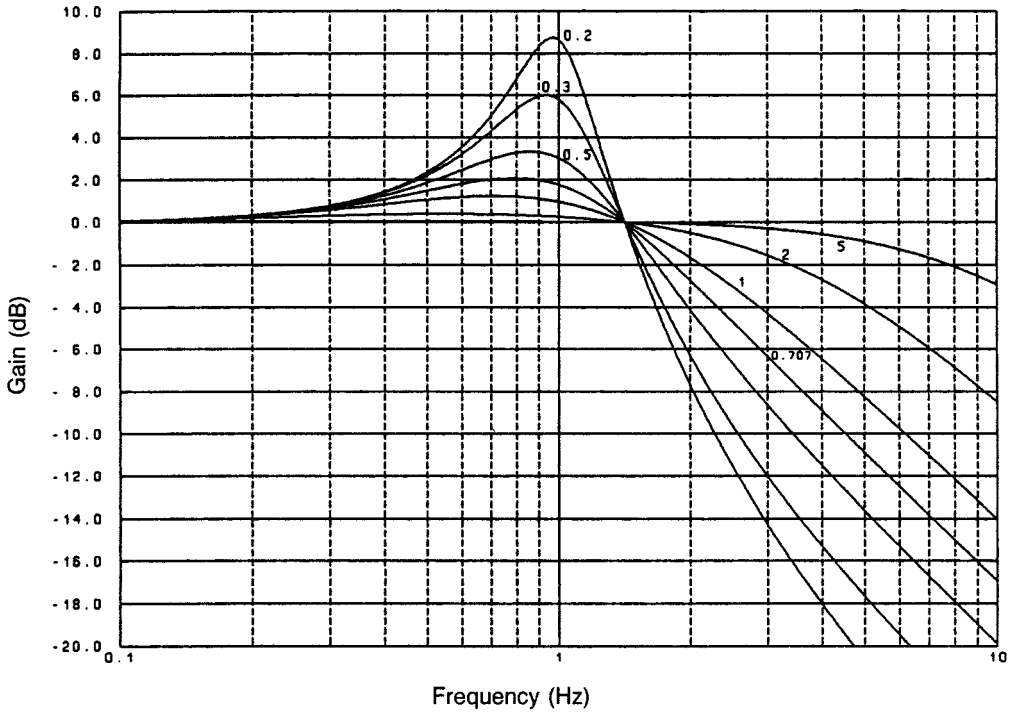


FIGURE 76.4 Closed-loop second-order type-2 PLL error response for various damping factors.

For the active loop filter case Eqs. (76.8) and (76.10) are solved and yield the synthesis equations shown in Eqs. (76.14) and (76.15). It can be seen that no constraints on the loop damping factor exist in this case.

$$R_1 = \frac{K}{\omega_n^2 C} \tag{76.14}$$

$$R_2 = \frac{2\zeta}{\omega_n C} \tag{76.15}$$

A typical design procedure for these loop filters would be, first, to select the loop damping factor and natural frequency based on the system requirements. Next, all the loop gain parameters are determined. A convenient capacitor value may then be selected. The remaining resistors can now be computed from the synthesis equations presented above.

Figure 76.4 shows the closed-loop frequency response of a PLL with an active loop filter [Eq. (76.10)] for various values of damping factor. The loop natural frequency has been normalized to 1 Hz for all cases.

Substituting Eq. (76.6) into (76.3) will give the loop error response in terms of damping factor. This function is shown plotted in Fig. 76.5. These plots may be used to select the PLL performance parameters that will give a desired frequency response shape.

The time response of a PLL with an active loop filter to a step in input phase was also computed and is shown plotted in Fig. 76.6.

76.3 Noise

An important design aspect of a PLL is the noise content of the output. The dominant resultant noise will appear as phase noise (jitter) on the output signal from the VCO. Due to the dynamics of the PLL some of these noise sources will be filtered by the loop transfer function [Eq. (76.2)] that is a low-pass characteristic.

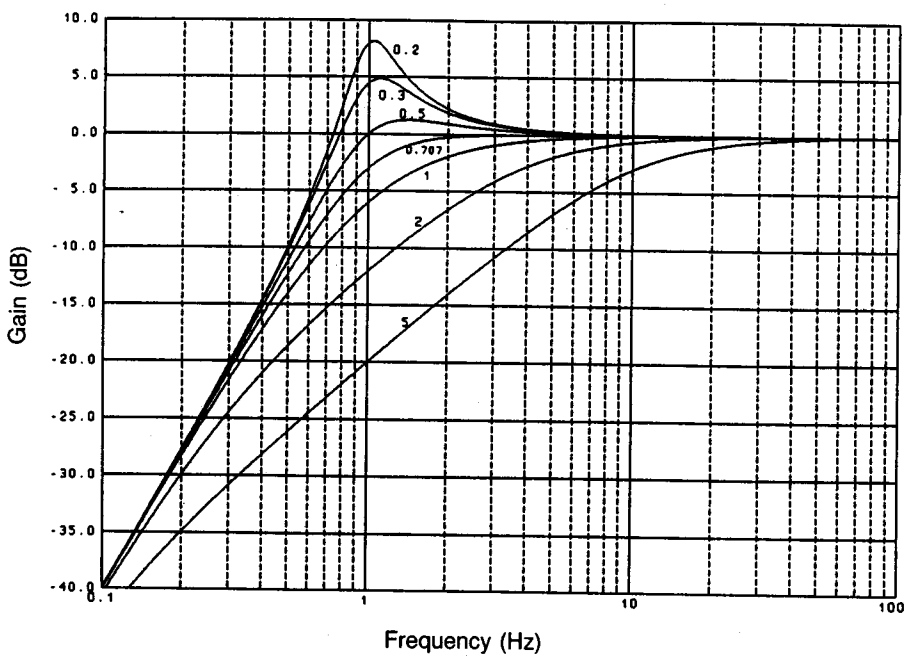


FIGURE 76.5 Closed-loop second-order type-2 PLL step response for various damping factors.

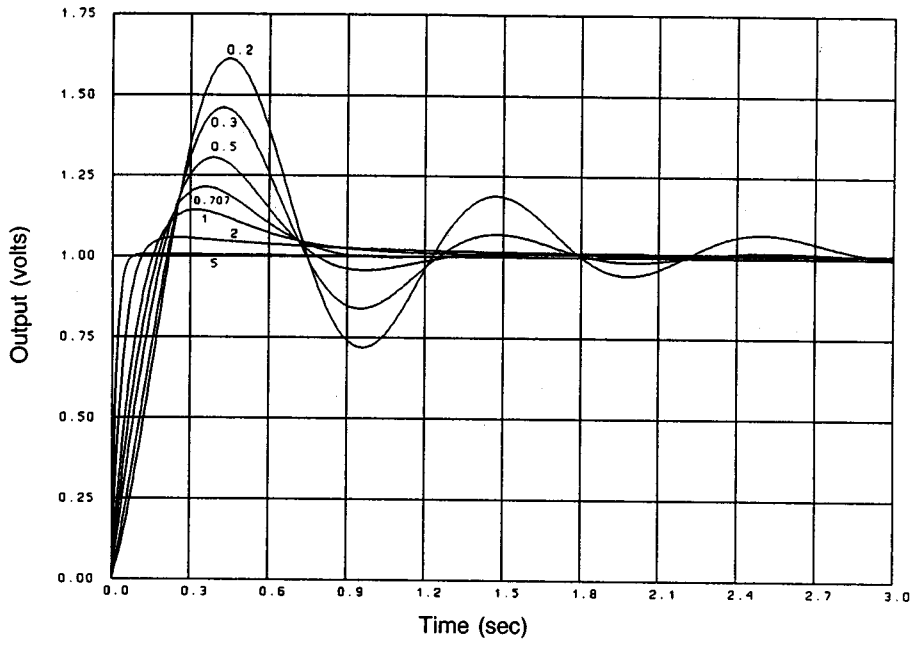


FIGURE 76.6 Closed-loop PLL response for various damping factors.

Others will be processed by the loop error function [Eq. (76.3)] that is a high-pass characteristic. Table 76.1 shows the major sources of noise in a PLL and the effect of the loop dynamics on this noise. All these factors must be combined to evaluate the complete noise performance of a PLL. Often it will be found that one particular noise source will be dominant and the PLL performance can then be adjusted to minimize the output noise.

TABLE 76.1 PLL Noise Sources

Noise Source	Filter Function
Reference oscillator phase noise	Low pass
Phase detector noise	Low pass
Active loop filter input noise	Low pass
Digital divider noise	Low pass
Active loop filter output noise	High pass
VCO free-running phase noise	High pass

A PLL is frequently used to enhance the noise performance of an oscillator by taking advantage of these noise-filtering properties. For example, a crystal oscillator typically has very good low-frequency noise characteristics, and a free-running LC oscillator can be designed with very good high-frequency noise performance but will exhibit poor low-frequency noise characteristics. By phase-locking an LC oscillator to a crystal oscillator and setting the loop response corner frequency to the noise crossover point between the two oscillators, the desirable characteristics of both oscillators are realized.

When designing frequency synthesizers using PLLs, care must be taken to prevent noise from the PLL components from introducing excessive noise. The divider ratio (N) used in the feedback of the loop has the effect of multiplying any noise that appears at the input or output of the phase detector by this factor. Frequently, a large value of N is required to achieve the desired output frequencies. This can cause excessive output noise. All these effects must be taken into account to achieve a PLL design with optimum noise performance.

76.4 PLL Design Procedures

The specific steps used to design a PLL depend on the intended application. Typically the architecture of the loop will be determined by the output frequency agility required (frequency synthesizer) and the reference sources available. Other requirements such as size and cost play important factors, as well as available standard components. Once the topology has been determined, then the desired loop transfer function must be synthesized. This may be dictated by noise requirements as discussed above or other factors such as loop lock-up time or input signal tracking ability. The design Eqs. (76.11) through (76.15) may then be used to determine the component values required in the loop filter.

Frequently several of these factors must be balanced or traded off to obtain an acceptable design. A design that requires high performance in several of these areas usually can be realized at the expense of design complexity or increased component cost.

76.5 Components

The development of large-scale integrated circuits over the past several years has made the design and implementation of PLLs and frequency synthesizers much cheaper and easier. Several major manufacturers (Motorola, Signetics, National, Plessey, etc.) currently supply a wide range of components for PLL implementation. The most complex of these are the synthesizer circuits that provide a programmable reference divider, programmable divide by N, and a phase detector. Several configurations of these circuits are available to suit most applications. Integrated circuits are also available to implement most of the individual blocks shown in Fig. 76.1.

A wide variety of phase detector circuits are available, and the optimum type will depend on the circuit requirements. An analog multiplier (or mixer) may be used and is most common in applications where the comparison frequency must be very high. This type of phase detector produces an output that is the multiplication of the two input signals. If the inputs are sine waves, the output will consist of a double-frequency component as well as a dc component that is proportional to the cosine of the input phase difference. The double-frequency component can be removed with a **low-pass filter**, leaving only the dc component. The analog multiplier has a somewhat limited phase range of ±90 degrees. The remainder of the phase detector types discussed here are digital in nature and operate using digital edges or transitions of the signals to be compared.

The sample-and-hold phase detector is widely used where optimum noise performance is required. This circuit operates by using one of the phase detector inputs to sample the voltage on the other input. This latter input is usually converted to a triangle wave to give a linear phase detector characteristic. Once the input is sampled, its voltage is held using a capacitor. The good noise performance is achieved since most of the time the phase detector output is simply a stored charge on this capacitor. The phase range of the sample-and-hold phase detector depends on the type of waveform shaping used and can range from ±90 to ±180 degrees.

One of the simplest types of phase detectors to implement uses an exclusive OR gate to digitally multiply the two signals together. The output must then be low-pass filtered to extract only the dc component. The main drawback to this circuit is the large component that exists in the output at twice the input frequency. This requires a large amount of low-pass filtering and may restrict the PLL design. The phase range of this type of circuit is ±90 degrees.

One of the main drawbacks of all the above types of phase detectors is that they only provide an output that is proportional to phase and not to a frequency difference in the input signals. For many applications the PLL input signals are initially not on the same frequency. Several techniques have been used in the past to resolve this such as sweeping the VCO or using separate circuitry to first acquire the input frequency. The sequential (sometimes called phase/frequency) phase detector has become the most commonly used solution due to its wide availability in integrated form. This type of phase detector produces pulses with the width of the pulses indicating the phase difference of the inputs. It also has the characteristic of providing the correct output to steer the VCO to the correct frequency. The noise characteristic of this type of phase detector is also quite good since either no or very narrow pulses are produced when the inputs are in phase with each other. The phase range of this type of circuit is ±360 degrees.

Digital dividers are widely available and may either have programmable or fixed division ratios depending on the application. For optimum noise performance a synchronous type of divider should be used. When a programmable divider is required to operate at a high frequency (>50 MHz), a dual modulus circuit is normally used. This circuit uses a technique called *pulse swallowing* to extend the range of normal programmable divider integrated circuits by using a dual modulus prescaler (usually ECL). The dual modulus prescaler is a high-frequency divider that can be programmed to divide by only two sequential values. A second programmable divider section is then used to control the prescaler. Further details of this type of divider are available from component manufacturers' data sheets as well as in the references.

The voltage-controlled oscillator is typically the most critical circuit in determining the overall noise performance of a PLL. For this reason it is often implemented using discrete components, especially at the higher frequencies. Some digital integrated circuits exist for lower-frequency VCOs, and microwave integrated circuit VCOs are now available for use to several gigahertz. The major design parameters for a VCO include the operating frequency, tuning range, tuning linearity, and phase noise performance. Further information on the design of VCOs is contained in the references.

Loop filters used in PLLs may be either active or passive depending on the specific application. Active filters are normally used in more critical applications when superior control of loop parameters and reference frequency suppression is required. The loop filter is typically followed by a low-pass filter to remove any residual reference frequency component from the phase detector. This low-pass filter will affect the calculated loop response and will typically appear to reduce the loop damping factor as its corner frequency is brought closer to the loop natural frequency. To avoid this degradation the corner frequency of this filter should be approximately one order of magnitude greater than the loop natural frequency. In some cases a notch filter may be used to reduce the reference frequency when it is close to the reference frequency.

76.6 Applications

Phase-locked loops are used in many applications including frequency synthesis, modulation, demodulation, and clock recovery. A frequency synthesizer is a PLL that uses a programmable divider in the feedback. By selecting various values of division ratio, several output frequencies may be obtained that are integer multiples of the reference frequency (Fref). Frequency synthesizers are widely used in radio communications equipment to obtain a stable frequency source that may be tuned to a desired radio channel. Since the output frequency is an integer multiple of the reference frequency, this will determine the channel spacing obtained. The main

design parameters for a synthesizer are typically determined by the required channel change time and output noise.

Transmitting equipment for radio communications frequently uses PLLs to obtain frequency modulation (FM) or phase modulation (PM). A PLL is first designed to generate a radio frequency signal. The modulation signal (i.e., voice) is then applied to the loop. For FM the modulating signal is added to the output of the loop filter. The PLL will maintain the center frequency of the VCO, while the modulation will vary the VCO frequency about this center. The frequency response of the FM input will exhibit a high-pass response and is described by the error function shown in Eq. (76.3). Phase modulation is obtained by adding the modulation signal to the input of the loop filter. The modulation will then vary the phase of the VCO output signal. The frequency response of the PM input will be a low-pass characteristic described by the closed-loop transfer function shown in Eq. (76.2).

A communications receiver must extract the modulation from a radio frequency carrier. A PLL may be used by phase locking a VCO to the received input signal. The loop filter output will then contain the extracted FM signal, and the loop filter input will contain the PM signal. In this case the frequency response of the FM output will be a low-pass function described by the closed-loop transfer function and the PM output response will be a high-pass function described by the error function.

In digital communications (modems) it is frequently necessary to extract a coherent clock signal from an input data stream. A PLL is often used for this task by locking a VCO to the input data. Depending on the type of data encoding that is used, the data may first need to be processed before connecting the PLL. The VCO output is then used as the clock to extract the data bits from the input signal.

Defining Terms

Capture range: The range of input frequencies over which the PLL can acquire phase lock.

Damping factor: A measure of the ability of the PLL to track an input signal step. Usually used to indicate the amount of overshoot present in the output to a step perturbation in the input.

Free-run frequency: The frequency at which the VCO will oscillate when no input signal is presented to the PLL. Sometimes referred to as the rest frequency.

Lock range: The range of input frequencies over which the PLL will remain in phase lock once acquisition has occurred.

Loop filter: The filter function that follows the phase detector and determines the system dynamic performance.

Loop gain: The combination of all dc gains in the PLL.

Low-pass filter: A filter that usually follows the loop filter and is used to remove the reference frequency components generated by the phase detector.

Natural frequency: The characteristic frequency of the PLL dynamic performance. The frequency of the closed-loop transfer function dominant pole.

Phase detector gain: The ratio of the dc output voltage of the phase detector to the input phase difference. This is usually expressed in units of volts/radian.

VCO gain: The ratio of the VCO output frequency to the dc control input level. This is usually expressed in units of radians/second/volt.

Related Topics

10.3 The Ideal Linear-Phase Low-Pass Filter • 25.3 Application-Specific Integrated Circuits • 73.2 Noise

References

AFDPLUS Reference Manual, Boulder, Colo.: RLM Research, 1991 (software used to generate the graphs in this section).

R. G. Best, *Phase-Locked Loops—Theory, Design & Applications*, New York: McGraw-Hill, 1984.

A. Blanchard, *Phase-Locked Loops, Application to Coherent Receiver Design*, New York: Wiley Interscience, 1976.

W. F. Egan, *Frequency Synthesis by Phase Lock*, New York: Wiley Interscience, 1981.
F. M. Gardner, *Phaselock Techniques*, New York: Wiley, 1979.
J. Gorski-Popiel, *Frequency Synthesis; Techniques & Applications*, Piscataway, N.J.: IEEE Press, 1975.
W.C. Lindsey and M.K. Simon, *Phase-Locked Loops & Their Applications*, Piscataway, N.J.: IEEE Press, 1978.
V. Manassewtsch, *Frequency Synthesizers: Theory and Design*, New York: Wiley Interscience, 1980.
U. L. Rhode, *Digital PLL Frequency Synthesizers Theory and Design*, Englewood Cliffs, N.J.: Prentice-Hall, 1983.

Further Information

Recommended periodicals that cover the subject of PLLs include *IEEE Transactions on Communications*, *IEEE Transactions on Circuits and Systems*, and *IEEE Transactions on Signal Processing*. Occasionally articles dealing with PLLs may also be found in *EDN*, *Electronic Design*, *RF Design*, and *Microwaves and RF Magazine*. A four-part PLL tutorial article titled *PLL Primer*, by Andrzej B. Przedpelski, appeared in *RF Design Magazine* in the March/April 1983, May/June 1983, July/August 1983, and November 1987 issues.

Another good source of general PLL design information can be obtained from application notes available from various PLL component manufacturers. *Phase-Locked Loop Design Fundamentals*, by Garth Nash, is available from Motorola, Inc. as AN-535 and gives an excellent step-by-step synthesizer design procedure.

Hoeppner, C.H. “Telemetry”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

77

Telemetry

77.1 Introduction to Telemetry
77.2 Measuring and Transmitting
77.3 Applications of Telemetry
Power Sources • Power Plants
77.4 Limitations of Telemetry
77.5 Transmitters and Batteries
77.6 Receivers and Discriminators
77.7 Antennas and Total System Operation
77.8 Calibration
77.9 Telemetry Frequency Allocations
77.10 Telemetry Antennas
77.11 Measuring and Transmitting
77.12 Modulating and Multiplexing
77.13 Passive Telemeters
77.14 The Receiving Station

Conrad H. Hoeppner
The Johns Hopkins University

77.1 Introduction to Telemetry

Telemetry, or measurement at a distance, takes many and varied forms. It may use the principles of radio, electricity, optics, mechanics, or hydraulics to convey measurements made at one place to indicators, actuators, recorders, or computers at another. By far the most popular telemetry systems are electrical and use radio or wire links to convey information. In this respect, all of the considerations in the foregoing chapters on communications apply, as well as considerations of antennas, power supplies and convertors, heat removal, and radio frequency interference. Additional considerations that are unique to telemetry are treated here.

The deeper an instrumented vehicle probes into the remote reaches of outer space, the more technologically spectacular seem the achievements of telemetry. There is still something exciting and uncanny about performing measurements of a physical quantity at a distant location and precisely reproducing them at a more convenient place for reading or recording them.

Yet the vast distances spanned by telemetry signals are less challenging technically than the stubborn problems of almost sheer inaccessibility in some industrial applications to the quantities being measured. Signals from a missile-launched space probe soaring toward the sun are often easier to obtain than measurements from inside a stolid, earthbound motor only a foot or two away. To find the temperature of the spinning rotor, housed in a steel casing and surrounded by a strong alternating magnetic field, may require more ingenuity to transcend the operating environment than taking measurements from the most distant instrument payload speeding through the unaccommodating environment of space.

The technology that has produced missile and space telemetry is also spawning new forms of industrial radio telemetry, capitalizing on the development of new transducers, powerful miniature radio transmitters, improved self-contained power sources, and better techniques of environmental protection.

Simply enough, to telemeter is to measure at a distance. First, at the remote point, is needed a transducer, a device that converts the physical quantity being measured into a signal, usually an electrical one, so that it

can be more conveniently transmitted. Then a connecting link between the location where the measurement is being made and the point where one can read or record the signal being sent is required. This link can be either an electrical circuit—there have been wired telemetry systems since long before the turn of the century—or pneumatic or hydraulic lines, a beam of light, or now, more practically, a radio carrier.

A radio telemetry system comprises (1) transducers that convert measurements into electrical systems, (2) a subcarrier oscillator modulated by the transducers, (3) a radio transmitter modulated by the subcarrier, (4) a transmitting antenna, (5) a receiving antenna, (6) a radio receiver, and (7) a subcarrier discriminator. The radio link can transmit an analog of the continuous variable being measured, or, with pulse-code methods, it sends the measurement data digitally as a finite number of symbols representing a finite number of possible values of the measurement signal at the time it is sampled. The range of a radio link is limited by the power radiated toward the receiver from the transmitter and by the sensitivity of the receiver. The wider the bandwidth, the more the effect from noise, and therefore the more transmitted power required for a detectable signal.

Optical, mechanical, and hydraulic telemeters represent a smaller segment of the telemetry field than do electrical and radio telemeters; they will be given only brief treatment here.

Optical telemeters use light transmitted through space or through optical fibers, the light being modulated by the measurement signal. The modulation may be produced either electrically or mechanically. Electrically, light-emitting diodes, lasers, or electroluminescent material are used to convert the electricity to light. The light may then be modulated with the measurement information by modulating the electricity that produces the light or it may be polarized and rotated by Kerr or Pockels cells, absorbed by electrically activated chromofors, or converted to another wavelength by electrically controlled nonlinear elements, all of which are activated by the modulating signal.

Hydraulic telemeters are generally used in conjunction with hydraulic sensors and hydraulic displays, such as pressure gauges. They are immune to all electrical and optical interference and hence find application in unfavorable electrical and optical environments. A typical hydraulic telemeter is used to measure load lifted and boom angle on a crane. Here a hydraulic piston is activated by the tension of a lifting rope pushed sidewise by the piston roller. An increase in tension produced by the load tends to straighten the deflected rope and press the piston into its cylinder. The change in pressure is communicated through fluid in a tube to a remote indicator. The hydraulic telemeter measures the boom angle by simply placing a fluid reservoir on the boom, which when it is raised provides increased pressure through its tube to a second remote indicator. In this way, with remote indicators, the operator monitors the crane to prevent overloading and/or overturning.

Electrical telemeters proliferate through (1) space, (2) battlefields, and (3) industrial sites, varying in size, configuration, and information-carrying capacity with their various applications. Space research and missile development used the first significant multichannel telemeters. Telemeters developed at the Naval Research Laboratories and built by the Raytheon Company were first used to explore outside the earth's atmosphere in German V-2 rockets launched at the White Sands Proving Ground. These telemeters used 1000-MHz pulse position modulated signals at ranges greater than 100 miles. Conrad H. Hoeppner designed the equipment and managed the installations and operation. From 1945 onward, telemetry developed rapidly and found its way to the various missile ranges also being developed.

To permit tests to be made interchangeably at all ranges, it was necessary to standardize types of telemeters at the ranges. To this end the Department of Defense Research and Development Board formed the Guided Missiles Committee, which in turn formed the Working Group on Telemetry. This later became the Inter-Range Instrumentation Group (IRIG), which has published telemetry standards that are widely accepted.

Meanwhile, industrial telemetry has developed along different lines, producing miniaturized complete capsules for applications to process control, detection of defects, and machine design. Medical science is currently using telemetry in experimental, clinical, and diagnostic applications. Some of the particular body characteristics telemetered include heartbeat, brain waves, blood pressure, temperature, voice patterns, heart sounds, respiration sounds, and muscle tensions. Similar studies are being pursued in the biological and psychological fields, where more experimental latitude permits embedding of transmitters within living animals.

The basic telemetry system consists of three building blocks. These are (1) input transducer, (2) the transmitter, and (3) the receiving station. Transducers convert the measured physical quantity into a usable form for transmission. The conversion of the desired information into a form capable of being transmitted to the receiver is a function of the type of transducer employed. Transducers convert the physical quantities to be

measured into electrical, light, pneumatic, or hydraulic energy. The type of energy conversion is determined by the type of transmission desired. In a radio telemetry system, the transmitter and receiver have much in common with communications equipment. The transducers, however, are unique to telemetry and will be described in some detail here.

One of the most common types of transducers generates electrical signals as a function of the changing physical quantity, and one of the most common varieties of this type is the resistance wire strain gauge. In this transducer, the ability of the wire to change its dimension as it is stressed causes a corresponding change in its electrical resistance. A decrease in wire diameter generally results in greater resistance to the flow of electricity. Similarly, temperature-sensitive materials that have electrical characteristics changing with temperature make temperature detection possible.

In most transducers, the electrical output is varied as a function of changes in the physical parameter. These electrical changes can be transmitted by wire direct to a control center, data display area, or to a data analysis section for evaluation. The difficulties with the use of wire in many applications have given rise to wireless telemetry. In order to transmit the transducer information through the air, it is necessary to apply this information to a high-frequency electrical carrier, as is commonly done in radio. Application of the transducer information to a high-frequency carrier is commonly called modulation. High-frequency or rapidly changing electricity has the capability of being propagated through space, whereas low-frequency or battery, nonchanging voltage does not possess this ability.

The technique used for applying or modulating the high-frequency carrier by the transducer output involves any one of three different methods. It is possible to modulate a carrier by a change in amplitude, a change in frequency, or a change in the carrier phase. The last technique is similar to the modulation used in transmitting color by television. In color TV the brightness signal is transmitted as amplitude modulation (AM), the sound as frequency modulation (FM), and the color as phase modulation (PM), or pulse coding. Pulse coding is used to modulate the radio frequency carrier in either AM, FM, or PM.

A common and extremely useful technique for increasing the information-carrying capability of a single transmitting telemetry line is called multiplexing. When it is desirable to monitor different physical parameters, such as temperature and pressure, it may be wasteful to have duplicating telemetry transmission lines. Multiplexing techniques can usually be considered to be of two types: frequency division multiplexing and time division multiplexing. In the frequency division multiplexing system, different subcarrier frequencies are modulated by their respective changing physical parameter; these subcarrier frequencies are then used to modulate the carrier frequency, enabling the transmission of all desired channels of information, simultaneously by one carrier. At the receiver, these subcarrier frequencies must be individually removed. This is accomplished by filters that allow any one of the respective subcarrier frequencies to pass. Each subcarrier frequency is then converted back to a voltage by the discriminator. The discriminator voltages can be used to actuate recorders and/or similar devices. Time division telemetry systems may use pulse modulation or pulse code modulation. In these systems the information signal is applied, in time sequence, to modulate the radio carrier. The characteristics of a pulse signal can be affected by modulating its amplitude, frequency, or phase.

Telemetry began as a wire communication technique between two remotely located stations. As science extends its domains into the realm of space, telemetry will be the essential communicating link among satellites, spaceships, robots, and other scientific devices yet to be designed.

The range of a radio link is limited by the strength of the signal radiated by the transmitter toward the receiver and by the sensitivity of that receiver. A 10-microwatt (μW) output will transmit data easily 100 feet with a bandwidth of 100 kHz. The wider the bandwidth, the more the effect from noise, and therefore the more transmitting power required for an acceptable signal.

At the receiving station, there are usually no space restrictions in accommodating large antennas, sensitive radio tuners and recorders, and an ample power supply, but the transmitting station often must be small, possibly doughnut-size, but sometimes no bigger than a pea, and must be self-sufficient, carrying its own power or perhaps receiving it by radio.

On the surface, industrial radio telemetry seems to be simply a matter of hardware. It almost is, except that the functional requirements are a lot different from those in missile and space telemetry. Distances are much shorter, a matter of a few feet to a few hundred yards; signal power can be radiated directly from the transmitter circuitry or from an antenna as simple as an inch or two of wire. Most tests are repeatable—no missile blowing

up on the pad here, taking with it valuable instruments and invaluable records of the events leading up to that failure.

Quantities can be measured one or two at a time, rather than requiring an enormous amount of information to be transmitted at once. This results in relatively inefficient use of the radio link but enables simpler circuitry at both the transmitting and receiving ends.

Surprisingly, environment plays the most critical role in industrial telemetry. It makes by far the largest difference between telemetry operations from missiles and spacecraft and those used in industrial remote measurement. While missile telemetry equipment is expected to withstand accelerations of 10 to 20 *g*, the rotating applications of telemetry in industry, such as the embedding of a transducer in a spinning shaft, require immunity to 10,000 or 20,000 *g* centrifugal accelerations.

The environmental extremes under which industrial telemeters must work are considered normal operating conditions by their users. Unlike missile telemetry equipment, which is shielded and insulated against extremes of temperature, shock, and vibrations and which is carefully calibrated for weeks before it is used only once in an actual shot, industrial telemeters must operate repeatedly without adjustment and calibration. Used outdoors, they are often subjected to a temperature range of –40 to +140°F. They must operate when immersed in hot or cold fluids, and thus it is almost mandatory that they be completely encapsulated to be impervious not only to humidity and water but to many other chemical fluids and fumes. Many lubricating oils operate at temperatures of 300 or 350°F.

We know that missile telemetry components must be small and light, yet an order of magnitude reduction in size and weight has been necessary to make telemetry suitable for high-speed rotating shafts or for biological implants. They must be so reliable that no maintenance is required, for there are no service centers set up to handle this kind of equipment, and it must work without failure to continue to gain industrial acceptance.

Information theory has been used extensively to develop space telemetry for the most efficient data transmission over a maximum distance with a minimum of transmitted power. Inefficiencies, being of no real consequence in industrial telemetry, make for less elaborate, less costly equipment. Radio channels are used in a relatively inefficient manner, and the distances between transmitter and receiver are usually so short that there are few problems of weak signals. In many cases, measurement and testing via telemetry links takes place in completely shielded buildings or in metal housings.

Although telemetry is usually defined as measurement at a distance, it has also gradually begun to embody the concept of control from a distance. In telemetry—the transmission of the value of a quantity from a remote point—it may serve merely to communicate the reading on an instrument at a distance. The output of the instrument can also be fed into a control mechanism, however, such as a relay or an alarm, so that the telemetered signal can activate, stop, or otherwise regulate a process. Measurement may be taken at one location, indication provided at a second location, and the remote control function initiated at one of those two locations or at a third point.

For example, a motor might be pumping oil from one location while oil pressure is being measured at another. When the pressure reading is telemetered to a control station, a decision can be made there to reduce pump motor speed when the pressure is too high, or a valve can be opened at still another location to direct the oil to flow in another path. The decision-making controller may be an experienced pipeline dispatcher or an automatic device.

77.2 Measuring and Transmitting

Telemetry, then, really begins with measurement. A physical quantity is converted to a signal for transmission to another point. The transducer that converts this physical quantity into an electrical signal may be a piezoelectric crystal, a variable resistance, or perhaps an accelerometer. Telemetered information need be no less accurate than that obtained directly under laboratory conditions. For instance, in telemetering strain measurements, it is possible to achieve accuracies of a few microinches per inch or greater. The only limitation is usually the degree of stability in the bond of the strain gauge to the specimen, and not the strain gauge itself.

If great accuracy in temperature measurement is desired, it can be attained by choosing a transducer that provides a large variation of output signal over a small range of process property variation. The resolution which this provides may be translated into true accuracy by careful transducer calibration. Accuracy is reduced,

of course, if a wider range of temperature needs to be detected. Typical single-channel analog telemetry links maintain a measurement accuracy of 1–5%. This is not a limitation of the total system, however, since 1% of a 100° temperature change would only be 1°, so several telemetry channels can easily share the total temperature range to be measured, say a 100°F range divided into four 25°F ranges, to produce an accuracy of one-fourth of a degree.

Special temperature probes have been produced for the range of 70 to 400°F and higher to maximize the stability and accuracy of temperature telemetry. These probes, when used with the proper choice of transmitters and receivers, can provide temperature measurements to closer than 0.05°F.

One of the limitations to accuracy and to repeatability in telemetry is the output level of the transducer. The low electrical levels produced by thermocouples and strain gauges (millivolts) are much more difficult to telemeter than a higher-voltage level of, say, 5 V. At low signal levels, extraneous electrical noises produce great degradation. This noise may be thermally generated, caused by atmospheric effects, or generated by nearby electrical equipment. When low-level transducers are used, stable amplifiers are required to raise the signal voltage to useful modulation levels.

There may be great variations in the strength of the radio signal received because of variations in distance between transmitter and receiver or because of the interposition of metallic objects. In industrial radio telemetry transmission, these effects can be prevented from disturbing the data by resorting to FM of both the subcarrier and the carrier so that the telemetered signal is unchanged by undesirable amplitude variations. This method is called FM/FM telemetry.

If FM is used in the subcarrier of the transmitter, the transducer signal modulates the frequency of the subcarrier oscillator. This can be done by a simple resonant circuit that produces a given frequency in the audio range, say 1,000 Hz, which is varied above or below by the signal from the transducer as it responds to the variable it is measuring. If the signal were fed to a loudspeaker, a rising or falling tone would be heard. The subcarrier oscillator then modulates a radio frequency carrier, varying its frequency (FM) or its amplitude (AM) in accordance with the subcarrier signal. The radio frequency in FM industrial radio telemetry links is usually in the 88- to 108-MHz band. At the receiving end of the link, the radio receiver demodulates the signal, removing the carrier and feeding the subcarrier to a special discriminator circuit that removes the modulation and precisely reproduces the original measurement signal for calibrated indication or recording.

Multiple measurements can also be transmitted over the carrier by sampling the output of several transducers in rapid sequence, a technique called time-division multiplexing. This technique has been employed to handle as many as a million samples per second. It provides for simple data displays and easier separation of channels for recording or analysis, and it is free of cross talk. If possible, it is advantageous to use no multiplexing at all for concurrent data taking, but to use separate radio carriers for each measurement being transmitted.

Many and varied kinds of modulation have been used in telemetry systems. All have general usefulness, with cost and application being the drivers. Synchronized modulation is generally used with other systems, being synchronized to them to give additional information such as range. Typically, a command control uplink to an aircraft is used to synchronize a telemetry downlink with the delay being proportional to distance or the length of the link. Signal-to-noise advantages are also achieved. Another example is one in which telemetry is tacked on to a radar transponder to add additional pulses to indicate altitude, heading identification, or other conditions of the vehicle carrying the transponder. This is usually accomplished with pulse position modulation.

In many instances, a reconnaissance vehicle will carry a television camera. Its signals may be recorded on board but are often telemetered for real-time observations. Data may be placed on the same carrier using a few lines of the TV picture or an additional subcarrier. The much greater bandwidth of the TV signal seriously compromises the combination of range, transmitter power, and antenna directivity, and typically signal-to-noise ratio is reduced as much as 30 dB.

77.3 Applications of Telemetry

High-voltage transmission lines are an excellent example of how inaccessible an object of measurement may be. These lines vibrate in the wind, and the stresses and strains require measurement under the dynamic conditions that contribute to fatigue failure. Strain tests to determine fatigue will show quickly whether the endurance limit of the line has been exceeded, and only if it is exceeded need we be concerned about fatigue

failure. Therefore, it is necessary to measure the number and magnitude of the strain reversals in order to predict the time of failure. Telemetry techniques permit dynamic testing under actual service conditions rather than by simulated laboratory conditions or static tests.

While the transducer that produces an electrical signal proportional to strain may have an output of 0.01 V, the live transmission line to which it is attached may be at a potential of several hundred thousand volts. The problem is to detect this hundredth of a volt in the presence of a very large signal. In the language of the telemetry engineer, this is rejection of a common mode voltage on the order of 10^8 to 1. Then why not deenergize the line? It's a simple matter of economics—an idle line transmits no power, and the wind forces that cause the line to vibrate are neither predictable nor constant. So, weeks or months may be spent in gathering measurements for a particular set of spans. However, a radio telemetry link makes it possible to transmit the strain signal even while power is being carried.

A self-contained FM radio transmitter is attached to the transmission line at a point adjacent to a strain gauge. All remain at the same electrical potential as the line, much like a bird sitting safely on the wire, transmitting the strain gauge output to a radio receiver and recorder located at some convenient point on the ground, where vibration analysis can be made. As a result, armor rods may be placed around the line at the vulnerable points, or vibration absorbers of the correct resonant frequency can be installed at the proper points on the line.

More down to earth, but equally inaccessible to measurement, is strain on the chain belt of an earth mover. Too light a chain will quickly fail from fatigue caused by the alternating stresses imposed by the full and empty buckets it transports. Measurements made under actual operating conditions of the earth hoist mean attaching strain gauges to a chain traveling at 500 ft per minute, subjecting them to violent shock and vibration. On this kind of moving equipment, slip rings and wire-link remote measurements will not work. Here again, radio telemetry is now providing the dynamic measurements needed to test the earth-moving equipment at work. A transducer and a small, rugged transmitter are attached to points along the chain—strain varies from link to link, depending on the proximity to the bucket—until the most vulnerable part of the chain is found. It is preferable to use several transducers and multiple-channel telemetry equipment for such measurements to simplify correlation between load and the resulting strain at various links.

Telemetry can also determine water levels and flow rates of rivers to provide vital data for flood control or for efficient hydroelectric power generation. Data on the potential watershed into rivers can be obtained by analyzing the water content of snow that would eventually melt and feed them. One requirement is to measure the depth and water content of snow in the mountains, then transmit these data from remote points to a central receiving station. The snow-measuring transducer may consist of a radioactive source atop a tall pole and a radiation intensity meter on the ground beneath the snow. The gamma-ray intensity reaching the meter is a function of the height and water content of the intervening snow. Both the meter and the transmitting equipment can be powered by a storage battery and controlled by a clock timer that sets the time of transmission to a few seconds per day.

In the design of machinery, one of the most difficult factors to cope with is alternating fatigue-producing stresses that occur at some parts of the machine. It has long been the custom to measure stress in equipment with bonded strain gauges to predict the failure limits before actual failure occurs. This had only been possible on those portions of the equipment that could be connected by wires. With radio telemetry, it is not possible on all members. Costly fatigue failures are now avoidable through installation of miniature telemetry components that are reliable, rugged, and accurate in heretofore inaccessible locations and environments. Industrial uses are virtually limitless; systems can be built to specifications and encapsulated to withstand the most adverse conditions. Low-cost measurement and telemetry systems have been applied to read internal vibrations and strain in rotating equipment, chains, vehicles, and projectiles—eliminating slip rings and wires. Measurement can be made under operating conditions of vibration, acceleration, strain, temperature, pressure, magnetic fields, electrical current, and voltage, under such adverse conditions as in a field of high electrical potential, in fluids, in steam, or in high-velocity gases.

Power Sources

Power sources for the transmitter in industrial telemetry applications are seldom a problem. Batteries can be used for temporary applications and at temperatures below 200°F. Small and light, rechargeable and expendable

batteries are available solidly encapsulated in epoxy resin to withstand almost as rugged environments as the telemeter itself.

In a moving or rotating application, stationary magnets can be placed so that they generate electricity in a moving coil and are used to provide automatic power generation. If this method is not feasible, a stationary coil can be placed in the vicinity of the transmitter and fed electrical energy at a high frequency, so that its field can easily couple into a moving coil in almost any environment. The stationary coil ring may be large, even encompassing a whole room; usually only one turn of wire is necessary. The stationary coil may also be made extremely small, 1/4 to 1/2 in. in diameter, and coupled to the end of a rotating shaft. These power supplies and coil configurations are standard available units.

Power Plants

In power plants, coal is fed in turn to a number of hoppers by conveyor belt. A tripper on the conveyor belt diverts the coal into a particular hopper until it is full. Either an operator or a mechanical sensing device determines when the hopper is full, and a signal is transmitted to the conveyor to move onto the next hopper. Before telemetering equipment was in use, costly accidents could occur if the operator should be away momentarily or if the sensor failed to function. As much as six tons of coal a minute could overflow onto the power station floor.

To prevent this, pressure switches are installed in the tripper chute to activate a radio transmitter if coal backs up into the tripper. The transmitter sends its signal to a receiver located at the conveyor belt and sounds an alarm. This type of control is difficult if not impossible to achieve by wired power connections because the tripper is moving and because the corrosive coal dust atmosphere attacks the wires. For this reason, a radio transmitter equipped with long-life batteries is mounted on the tripper. The receiver at the control end is powered by ac. Subcarrier tone (frequency) coding is used to eliminate the effects of interference and noise, giving positive protection at all times.

77.4 Limitations of Telemetry

The preceding paragraphs describe a number of the requirements placed upon telemetry systems by the transducers and quantities being measured. Unfortunately, the development of telemetry has not been such as to satisfy all requirements, and in many cases the telemetry system seriously limits the measurement. A compromise is therefore required between telemetry capabilities and the requirements of measurement. The shortcomings and limitations of the telemetry system place restrictions upon measurements above and beyond those encountered in the laboratory when the telemeter is not used. In the first place, an electrical output from the measuring device is required in order that the measurement may be placed on a radio link. Consequently, transducers that produce an electrical output on one form or another are necessary. Also, the telemetry system may not be perfectly stable down to zero frequency (dc), and transducers and methods of measurement must be chosen to minimize the effects of drift. Overmodulating the subcarrier, or the time-division multiplexer, may also affect adjacent channels, as well as produce erroneous data in its own channel. If various measuring devices are switched, the switching transients must be minimized, or the accuracy of the telemetry system may be impaired. When mechanical commutators or time multiplexers are used, the measurement of the time occurrence of the event, such as the impact of cosmic particles or the receipt of a guidance pulse, is made more difficult and the time ambiguity of the multiplexed system is a serious limitation.

The measurement of a large number of parameters requires extensive and bulky equipment, unless the parameters can be combined in groups of similar inputs to minimize the signal conditioning required. This fact generally dictates a relatively standard transducer rather than an optimum one for each particular measurement.

The bandwidth of the measurement, or the frequency with which the measured quantity changes, is also seriously limited by the telemeter. In the FM/FM telemeter, the permissible bandwidth varies from a relatively low value on the lower-frequency subcarriers to a reasonably high value on the high-frequency subcarriers. The bandwidth of the measurement must not exceed the subcarrier bandwidth limitations, or sidebands will be generated in adjacent channels, thereby reducing the accuracy of other measurements (if multiplexed), or interference with adjacent RF signals will be caused.

In a time-multiplexed system, the problem of "folded data" is present whenever the rate of data change is faster than one-half the sampling rate. When this occurs, it is not known whether the measured quantity has reversed itself several times between samples or if there has been no reversal at all. It is considered desirable to limit the bandwidth of the data so that this ambiguity is not present; however, with refined techniques of analysis, this is not a rigid requirement. The form in which the data is displayed or recorded is also a limitation on measurement. In general, time-history plots of the measured quantity are desired. In this case, the speed at which the recording medium moves is often a severe limitation. If sampling is not regular, demultiplexing difficulties are magnified.

77.5 Transmitters and Batteries

The transmitter is made up of two components: the subcarrier oscillator and the radio frequency oscillator. The subcarrier can be bridge controlled (BCO) or voltage controlled (VCO). The subcarrier center frequency is 4,000 Hz, which can be modulated ±400 Hz by the strain or voltage being measured. Using BCOs, a strain as large as 2,500 microinches per inch (μin./in.) and as small as 2 μin./in. can be measured and transmitted. The temperature measurement range of the VCO is from –200 to 4,000°F. With a copper-constantan thermocouple, a temperature change as small as 2°F can be sensed and transmitted.

The single-resistance strain gauge transmitter does not have a subcarrier oscillator and can be used from –40 to 212°F. It has only a radio frequency oscillator, which is modulated by the sensor signal. For this reason, it is not suitable for static strain measurements and must be used for dynamic strain measurements only. It has a frequency response to 25,000 Hz or greater. A static strain signal transmitted by this device will drift. It is provided with self-contained rechargeable nickel-cadmium batteries. Pins protruding through the epoxy case are used for all electrical connections. Only one screw adjustment is provided, and this is used to set the radio frequency.

Rechargeable nickel-cadmium batteries are used with the BCO and the VCO. The BCO batteries have useful lives of 4 and 9 hr. A VCO battery has 40 hr useful life. The single-resistance strain gauge transmitter has a built-in nickel-cadmium battery with a life of 4 hr.

77.6 Receivers and Discriminators

A typical industrial receiver has a tuning range of 88 to 108 MHz. When the transmitter is used in its greatest sensitivity mode, the output of the discriminator is approximately 1 V for a 25-μin./in. strain with a single active gauge in the bridge. At the most insensitive mode 1 V is obtained for an approximately 500-μin./in. strain. The discriminator can withstand a 500% overload, which means that a 5-V signal will be obtained from a strain of 125-μin./in. at the maximum sensitivity and from 2,500-μin./in. at the minimum sensitivity.

77.7 Antennas and Total System Operation

A nickel-cadmium battery supplies the power to the transmitter. For the BCO, the resistance change of the strain gauge changes the frequency of the subcarrier. In the case of the VCO, the millivolt output of the thermocouple changes the frequency of the subcarrier. This change modulates the radio frequency transmitted by the antenna. The receiving antenna picks up the signal and conducts it by wire link to the radio receiver, which is tuned to the transmitting frequency. The radio receiver demodulates the FM carrier to reproduce the subcarrier signal. The subcarrier signal is then fed to the discriminator, which demodulates this signal to obtain a dc voltage, which is then amplified by the dc amplifier and recorded on the oscillograph. The oscillograph record, properly calibrated, is then a display of the strain in microinches per inch for the BCOs, or the temperature in degrees for the VCO. At the same time the dc signal can be read on a VTVM and can be used as a check on the oscillograph.

The transmitter subcarrier oscillators are factory set to operate at a center frequency of about 4,000 Hz. They have a frequency range of ±400 Hz about the 4,000-Hz center frequency. The center frequency is set with a

counter at the time of testing. The change of ±400 Hz is the information frequency change brought about by the change in strain or temperature measured by the sensor. It is this information frequency change that the discriminators isolate as a dc voltage change, which is proportional to the measured strain and is recorded on the oscillograph.

77.8 Calibration

Batteries are calibrated under simulated service conditions for voltage drop versus time. Bridge-controlled transmitters are calibrated for strain subcarrier frequency change using a cantilever beam instrumented with resistance strain gauges. The beam is calibrated for load versus strain using a strain indicator. It is then used to calibrate the bridge-controlled transmitters statically, by measuring the subcarrier frequency change as a function of strain. A dynamic calibration can also be made by using a second cantilever beam driven by a vibration generator. Two resistance strain gauges are mounted back-to-back on the second beam and calibrated. One of the gauges is monitored through the telemetry system and the other by wire link to the oscillograph, and the two signals are then compared. The single-resistance strain gauge transmitter is similarly calibrated, but in this case, the beam is fixed in a fatigue machine operating at 30 Hz. Calibrations are performed at various strain levels. Again, two calibrated gauges are monitored and compared, one using the telemetry systems and the other using wire link.

The effect of temperature on a transmitter and battery is measured at temperatures from 65 to 135°F by placing both in an air-circulating oven, with the receiving equipment and the calibration beams to room temperature outside the oven.

The voltage-controlled transmitter is calibrated for temperature subcarrier frequency change from 78 to 640°F. Two calibrated thermocouples, welded next to one another on a piece of stainless steel, are heated simultaneously. After determining by wire link instrumentation that both thermocouples are indicating the same temperature, the millivolt output of one is fed into the transmitter, and the output of the other is fed by wire into a precision potentiometer. The subcarrier frequency change is determined as a function of temperature, and the radio signal is recorded on the calibrated oscillograph, with a galvanometer determining its deflection as a function of temperature. The data obtained by wire link and radio are then compared to establish the calibration. The effect of thermocouple lengths can also be investigated in the same test setup. The receivers and discriminators are calibrated before these tests.

Cold junction compensation may be investigated from –40 to 258°F by cooling the transmitter and battery, with leads shorted and with a 20-mV input, in a cold chamber below room temperature and by heating in an air-circulating oven to above room temperature. The 20-mV input is imposed with a dc power supply kept outside the temperature chamber.

The discriminators are calibrated with the transmitters. The subcarrier frequency, which is the input to the discriminator, is monitored with a digital counter as the calibration beam is loaded. The voltage output corresponding to the frequency change can be monitored with a vacuum tube voltmeter. The strain, subcarrier frequency change, and the voltage output of the discriminator are then correlated. A digital frequency counter is used to set the transmitter center frequency.

77.9 Telemetry Frequency Allocations

Frequency bands for telemetry have been allocated as follows:

88–108 MHz	Low power, noninterference
216–260 MHz	General telemetry
400–475 MHz	Command destruct
1435–1540 MHz	General telemetry
1710–1850 MHz	Video telemetry
2.2–2.3 GHz	General telemetry

The low-power, 88 to 108 MHz, band is shared with FM broadcast stations. Telemetry is allowed to be used, but it must not interfere with broadcast reception. Transmitter power and antennas are limited to provide a signal strength no greater than 50 μV/m at 50 ft from the antenna and/or transmitter. In use, the telemetry transmitters are generally tuned to operate at frequencies between local FM broadcast stations.

The remaining frequency bands are used mainly with aircraft, unmanned vehicles, space vehicles, and for military applications. Equipment for these applications is rigidly constrained for stability, low spurious emissions, low cross talk, good linearity, etc.

77.10 Telemetry Antennas

When transmitter and receiver are stationery, antenna considerations for telemetry are no different than for communications. The usual case, however, is that the transmitter is moving, both rotating and translating and often obscuring the transmission with reflective material. This poses problems in both receiving the signal and tracking the moving transmitter with a directive receiving antenna. In many cases it is necessary to have two or more receiving antennas to receive from a single transmitter.

Transmitting antennas may be either conformal or protruding. The protruding antennas are usually cheaper and simpler. The radiation pattern must include downward directivity if an aircraft or space vehicle will fly directly overhead. A vertical whip antenna does not provide this coverage. A simple choice to provide smooth coverage from the nadir to the horizon is circular polarization at the nadir and elliptic polarization in between. A circular polarized receiving antenna is used to receive the signals over the complete pattern. It must be polarized in the same circular direction as the transmitting antenna. A spiral or helix antenna is usually used on the ground. Two complete radio frequency systems are generally used to receive signals of any polarization. One system uses a circular polarized right-hand antenna and the other uses a circular polarized left-hand antenna. In this manner, as the vehicle tilts or spins, signals are received continuously unless the transmitting antenna is occluded. To receive occluded signals, diversity reception is required. Each receiving station is located such that one fills in the occluded pattern of the other.

In short distance telemetry the same problems are encountered but from a different cause. If the transmitter is occluded at one position by the shaft of a rotating machine, it would be expected that reflections from nearby objects or walls would fill the gap. In practice it is the usual occurrence that a gap is generated once per revolution. This is caused by multiple signal cancellation. There are two methods of overcoming this signal drop-out: (1) with diverse polarization and (2) by locating the receiving antenna close to the transmitting antenna and effectively surrounding the shaft with it.

While information theory has been used extensively to develop space telemetry for the most efficient data transmission over a maximum distance with a minimum of transmitted power, the very inefficiencies permitted in industrial telemetry make for less elaborate, less costly equipment.

Radio channels are used in a relatively inefficient manner, and the distances between transmitter and receiver are usually so short that there are few problems of weak signals. In many cases, measurement and testing via telemetry links take place in completely shielded buildings or in metal housings.

Although telemetry is usually defined as measurement at a distance, it has also gradually begun to embody the concept of control from a distance. In a telemeter—the transmission of the value of a quantity from a remote point—it may only be necessary to observe the reading of an instrument to determine the temperature, pressure, or vibration of a distant or inaccessible object. One can also feed the output of the instrument into a control mechanism, however, such as a relay or an alarm device, so that the telemetered signal may activate or stop a controllable process. Measurement may be performed at one location, indication provided at a second location, and the remote control function initiated at one of the first two locations or even at a third point.

Take, for example, an oil pipeline in which a motor is pumping oil from one location and oil pressure is being measured at a second location. The pressure reading is telemetered to a station where a decision can be made to reduce the speed of the pump motor when the pressure is too high, or a valve may be opened at still another location to cause the oil flow in another path. The decision-making element may be human, an experienced pipeline dispatcher, or an automatic controller. Human or automatic device—either one telemeters a command to the control points.

77.11 Measuring and Transmitting

Telemetry, then, really begins with measurement. A physical quantity is converted to a signal for transmission to another point. The transducer that converts the physical quantity into an electric signal typically may be a piezoelectric crystal, a variable resistance, or perhaps an accelerometer.

Telemetering the measurement signal of the best transducers in no way degrades the measurement below accuracies attainable under laboratory conditions. For instance, in strain measurement it is possible to achieve accuracies of a few microinches per inch or greater, but the limitation is usually the degree of stability in the bond of the strain gauge to the specimen.

If one wants accuracy in temperature measurement, it can be attained by choosing a transducer that provides a large variation in output signal over a small range of temperature. The resolution that this provides may be translated to true accuracy by careful transducer calibration. Typical analog telemetry links maintain a measurement accuracy on a single channel to 1%. This is not a limitation of the total system, since 1% of a 100-degree temperature change would only be 1 degree, so several telemetry channels can easily share the total temperature range to be measured, say, a 100 range divided into four 25 ranges to produce an accuracy of 1/4 degree.

One of the limitations to accuracy and repeatability in telemetry is the output level of the transducer. The low electrical levels produced by thermocouples and strain gauges (0.010 V) are more difficult to telemeter than high-voltage levels of 5 V. At low signal levels, extraneous electrical noises produce greater degradation. These may be thermally generated or caused by atmospheric effects or generated by nearby electrical equipment. When low-level transducers are used, stable amplifiers are required to raise their signal voltage to useful modulation levels.

77.12 Modulating and Multiplexing

The transducer signal modulates the frequency of the subcarrier oscillator. This is simply a resonant circuit that produces a given frequency in the audio range, say 100 Hz, and is varied plus or minus this center frequency by the signal from the transducer as it responds to the variable that it is measuring. When the signal is fed to a loudspeaker, one actually hears a rising or falling tone. The subcarrier oscillator modulates a radio frequency carrier, varying its frequency in accordance with the subcarrier voltage signal. The radio frequency in FM industrial radio telemetry links is usually in the 88- to 108-MHz band, permitting the use of high-grade radio tuners already mass produced for the high-fidelity market. The radio receiver demodulates the signal, removing the carrier and feeding it to a special discriminator circuit that removes the double modulation and reproduces an analog of the original measurement signal for calibrated indication or recording.

There can be great variations in the strength of the radio signal received because of variations in distance between transmitter and receiver or because of the interposition of metallic objects. In industrial radio telemetry transmission, these effects are prevented from disturbing the data by resorting to FM of both the subcarrier and the carrier so that the telemetered signal is unchanged by undesirable amplitude variations. This method is called FM/FM telemetry. There are other methods of carrier modulation, such as pulse amplitude modulation, phase modulation, and pulse duration modulation. Each has its proper place in missile and space telemetry, where great distances must be spanned with a maximum of data over crowded and often noisy communication channels. Pulse code modulation of an FM link, however, may be expected to become more widespread in industrial telemetry.

Particularly in missile telemetry, it is important that multiple measurements be transmitted over a single carrier to save power and minimize electronic equipment and antennas. Such simultaneous transmission of signals over a common path, called multiplexing, is sometimes used in industrial telemetry. When concurrent data about several simultaneous events are transmitted by several subcarriers, each subcarrier oscillator has a distinctive reference frequency and swings from this center frequency toward arbitrary maximum and minimum frequencies in response to signals from a corresponding transducer. Thus a number of separate audio frequency bands are sent over the radio frequency carrier. This is called frequency division multiplexing. The frequency division multiplex requires careful adjustment of subcarrier frequencies and the corresponding filters at the receiver and strong suppression of harmonics to avoid cross talk or interaction between channels.

Multiple measurements may also be transmitted over the carrier by sampling the output of each transducer in rapid sequence, a technique called time-division multiplexing. The technique has been used to handle as many as a million samples per second. It provides for very simple data displays, easier separation of channels for recording or analysis, and is free of cross talk. If possible, though, it is advantageous to use no multiplexing at all for concurrent data talking but to use separate radio carriers for each measurement being transmitted. The multiplex telemeter requires careful adjustment of subcarrier frequencies and precisely tuned filters to separate them at the receiver. This adds to the cost of the equipment and requires considerable experience of an operator.

The telemetry data received may be recorded in a number of ways, but such records must preserve the accuracy of the entire system. For example, if one is monitoring a 1% system and can distinguish 1/64th in. on a paper graph, the minimum graph size for full scale should be approximately 2 in. Similarly, numeric data should be printed to enough decimal places to preserve the accuracy of the system.

A single channel of industrial FM/FM telemetry equipment may cost between $1000 and $2000, depending on the flexibility required and the measurements being made. It buys everything needed for a given remote measurement—transducer, radio link, power supply, and simple indicator.

77.13 Passive Telemeters

Passively powered telemeters offer some interesting advantages. When a number of telemeters are used, they can be powered in sequence or only when measurements are required, thus preventing radio frequency congestion. In medical applications of telemetry, passive devices eliminate the danger involved in swallowing or implanting batteries.

Most passive telemeters are essentially an inductance and a capacitance coupled as a resonant circuit. Either of these components may be pressure sensitive or temperature sensitive. A nearby magnetic coil coupled to this circuit can, by means of a varying frequency, determine the resonant point of the telemeter, which can then be a function of the temperature or pressure being measured.

77.14 The Receiving Station

The industrial telemetry receiving station differs vastly in purpose and principle from the transmitting station. Its usual environment is no more difficult to cope with, in terms of ambient temperature, shock, and vibrations, than an automobile radio. It receives signals over relatively short distances in which the subcarrier frequencies are so widely spaced that harmonics and drift are no problem.

In the FM broadcast band, available professional-grade high-fidelity tuners have 1 or 2 μV sensitivity for 30 dB of quieting of extraneous noises and automatic frequency control circuits, which compensate for both transmitter and receiver drift. They feed telemetry phase-lock discriminators, which lock the receiver into the frequency and phase of the incoming signal.

The transmitters usually radiate from the resonant elements themselves, avoiding elaborate antennas that might be required for longer distance transmission; the receivers use simple dipole or commercial TV antennas. Thus, industrial radio telemetry has become a carefully engineered blend of the borrowed and the new.

Related Topics

38.1 Wire • 69.1 Modulation and Demodulation • 69.2 Radio

Tranter, W.H., Kosbar, K.L. "Computer-Aided Design and Analysis of Communication Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

78

Computer-Aided Design and Analysis of Communication Systems

78.1 Introduction
78.2 The Role of Simulation
78.3 Motivation for the Use of Simulation
78.4 Limitations of Simulation
78.5 Simulation Structure
78.6 The Interdisciplinary Nature of Simulation
78.7 Model Design
78.8 Low-Pass Models
78.9 Pseudorandom Signal and Noise Generators
78.10 Transmitter, Channel, and Receiver Modeling
78.11 Symbol Error Rate Estimation
78.12 Validation of Simulation Results
78.13 A Simple Example Illustrating Simulation Products
78.14 Conclusions

William H. Tranter
University of Missouri–Rolla

Kurt L. Kosbar
University of Missouri–Rolla

78.1 Introduction

It should be clear from the preceding chapters that communication systems exist to perform a wide variety of tasks. The demands placed on today's communication systems necessitate higher data rates, greater flexibility, and increased reliability. Communication systems are therefore becoming increasingly complex, and the resulting systems cannot usually be analyzed using traditional (pencil and paper) analysis techniques. In addition, communication systems often operate in complicated environments that are not analytically tractable. Examples include channels that exhibit severe bandlimiting, multipath, fading, interference, non-Gaussian noise, and perhaps even burst noise. The combination of a complex system and a complex environment makes the design and analysis of these communication systems a formidable task. Some level of computer assistance must usually be invoked in both the design and analysis process. The appropriate level of computer assistance can range from simply using numerical techniques to solve a differential equation defining an element or subsystem to developing a **computer simulation** of the end-to-end communication system.

There is another important reason for the current popularity of computer-aided analysis and simulation techniques. It is now practical to make extensive use of these techniques. The computing power of many personal computers and workstations available today exceeds the capabilities of many large mainframe computers of only a decade ago. The low cost of these computing resources make them widely available. As a result, significant computing resources are available to the communications engineer within the office or even the home environment.

FIGURE 78.1 Basic communication link.

Personal computers and workstations tend to be resources dedicated to a specific individual or project. Since the communications engineer working at his or her desk has control over the computing resource, lengthy simulations can be performed without interfering with the work of others. Over the past few years a number of software packages have been developed that allow complex communication systems to be simulated with relative ease [Shanmugan, 1988]. The best of these packages contains a wide variety of subsystem models as well as integrated graphics packages that allow waveforms, spectra, histograms, and performance characteristics to be displayed without leaving the simulation environment. For those motivated to generate their own simulation code, the widespread availability of high-quality C, Pascal, and FORTRAN compilers makes it possible for large application-specific simulation programs to be developed for personal computers and workstations. When computing tools are both available and convenient to use, they will be employed in the day-to-day efforts of system analysts and designers.

The purpose of this chapter is to provide a brief introduction to the subject of **computer-aided design and analysis** of communication systems. Since computer-aided design and analysis almost always involves some level of simulation, we focus our discussion on the important subject of the simulation of communication systems.

Computer simulations can, of course, never replace a skilled engineer, although they can be a tremendous help in both the design and analysis process. The most powerful simulation program cannot solve all the problems that arise, and the process of making trade-off decisions will always be based on experience. In addition, evaluating and interpreting the results of a complex simulation require considerable skill and insight. While these remarks seem obvious, as computer-aided techniques become more powerful, one is tempted to replace experience and insight with computing power.

78.2 The Role of Simulation

The main purposes of simulation are to help us understand the operation of a complex communication system, to determine acceptable or optimum parameters for implementation of a system, and to determine the performance of a communication system. There are basically two types of systems in which communication engineers have interest: **communication links** and communication networks.

A communication link is usually a single source, a single user, and the components and channel between source and user. A typical link architecture is shown in Fig. 78.1. The important performance parameter in a digital communication link is typically the reliability of the communication link as measured by the symbol or bit error rate (BER). In an analog communication link the performance parameter of interest is typically the signal-to-noise ratio (SNR) at the receiver input or the mean-square error of the receiver output. The simulation is usually performed to determine the effect of system parameters, such as filter bandwidths or code rate, or to determine the effect of environmental parameters, such as noise levels, noise statistics, or power spectral densities.

A communication network is a collection of communication links with many signal sources and many users. Computer simulation programs for networks often deal with problems of routing, flow and congestion control, and the network delay. While this chapter deals with the communication link, the reader is reminded that network simulation is also an important area of study. The simulation methodologies used for communication networks are different from those used on links because, in a communication link simulation, each waveform present in the system is sampled using a constant sampling frequency. In contrast, network simulations are event-driven, with the important events being such quantities as the time of arrival of a message.

Simulations can be developed to investigate either transient phenomena or steady-state properties of a system. The study of the acquisition time of a phase-lock loop receiver is an example of a transient phenomenon. Simulations that are performed to study transient behavior often focus on a single subsystem such as a receiver synchronization system. Simulations that are developed to study steady-state behavior often model the entire system. An example is a simulation to determine the BER of a system.

78.3 Motivation for the Use of Simulation

As mentioned previously, simulation is a reasonable approach to many design and analysis problems because complex problems demand that computer-based techniques be used to support traditional analytical approaches. There are many other motivations for making use of simulation.

A carefully developed simulation is much like having a breadboard implementation of the communication system available for study. Experiments can be performed using the simulation much like experiments can be performed using hardware. System parameters can be easily changed, and the impact of these changes can be evaluated. By continuing this process, parameteric studies can easily be conducted and acceptable, or perhaps even optimum, parameter values can be determined. By changing parameters, or even the system topology, one can play "what if" games much more quickly and economically using a simulation than with a system realized in hardware.

It is often overlooked that simulation can be used to support analysis. Many people incorrectly view simulation as a tool to be used only when a system becomes too complex to be analyzed using traditional analysis techniques. Used properly, simulation goes hand in hand with traditional techniques in that simulation can often be used to guide analysis. A properly developed simulation provides insight into system operation. As an example, if a system has many parameters, these can be varied in a way that allows the most important parameters, in terms of system performance, to be identified. The least important parameters can then often be discarded, with the result being a simpler system that is more tractable analytically. Analysis also aids simulation. The development of an accurate and efficient simulation is often dependent upon a careful analysis of various portions of the system.

78.4 Limitations of Simulation

Simulation, useful as it is, does have limitations. It must be remembered that a system simulation is an approximation to the actual system under study. The nature of the approximations must be understood if one is to have confidence in the simulation results. The accuracy of the simulation is limited by the accuracy to which the various components and subsystems within the system are modeled. It is often necessary to collect extensive experimental data on system components to ensure that simulation models accurately reflect the behavior of the components. Even if this step is done with care, one can only trust the simulation model over the range of values consistent with the previously collected experimental data. A main source of error in a simulation results because models are used at operating points beyond which the models are valid.

In addition to modeling difficulties, it should be realized that the digital simulation of a system can seldom be made perfectly consistent with the actual system under study. The simulation is affected by phenomena not present in the actual system. Examples are the aliasing errors resulting from the sampling operation and the finite word length (quantization) effects present in the simulation. Practical communication systems use a number of filters, and modeling the analog filters present in the actual system by the digital filters required by the simulation involves a number of approximations. The assumptions and approximations used in modeling an analog filter using impulse-invariant digital filter synthesis techniques are quite different from the assumptions and approximations used in bilinear z-transform techniques. Determining the appropriate modeling technique requires careful thought.

Another limitation of simulation lies in the excessive computer run time that is often necessary for estimating performance parameters. An example is the estimation of the system BER for systems having very low nominal bit error rates. We will expand on this topic later in this chapter.

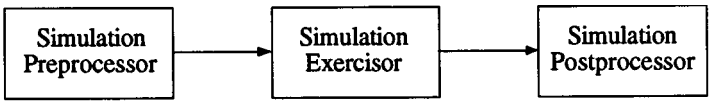


FIGURE 78.2 Typical structure of a simulation program.

78.5 Simulation Structure

As illustrated in Fig. 78.1, a communication system is a collection of subsystems such that the overall system provides a reliable path for information flow from source to user. In a computer simulation of the system, the individual subsystems must first be accurately modeled by signal processing operations. The overall simulation program is a collection of these signal processing operations and must accurately model the overall communication system. The important subject of subsystem modeling will be treated in a following section.

The first step in the development of a simulation program is to define the topology of the system, which specifies the manner in which the individual subsystems are connected. The subsystem models must then be defined by specifying the signal processing operation to be performed by each of the various subsystems. A simulation structure may be either fixed topology or free topology. In a fixed topology simulation, the basic structure shown in Fig. 78.1 is modeled. Various subsystems can be bypassed if desired by setting switches, but the basic topology cannot be modified. In a free topology structure, subsystems can be interconnected in any way desired and new additional subsystems can be added at will.

A simulation program for a communication system is a collection of at least three operations, shown in Fig. 78.2, although in a well-integrated simulation these operations tend to merge together. The first operation, sometimes referred to as the *preprocessor,* defines the parameters of each subsystem and the intrinsic parameters that control the operation of the simulation. The second operation is the *simulation exercisor,* which is the simulation program actually executed on the computer. The third operation performed in a simulation program is that of *postprocessing.* This is a collection of routines that format the simulation output in a way which provides insight into system operations and allows the performance of the communication system under study to be evaluated. A postprocessor usually consists of a number of graphics-based routines, allowing the user to view waveforms and other displays generated by the simulation. The postprocessor also consists of a number of routines that allow estimation of the bit error rate, signal-to-noise ratios, histograms, and power spectral densities.

When faced with the problem of developing a simulation of a communication system, the first fundamental choice is whether to develop a custom simulation using a general-purpose high-level language or to use one of the many special-purpose communication system simulation languages available. If the decision is made to develop a dedicated simulation using a general-purpose language, a number of resources are needed beyond a quality compiler and a mathematics library. Also needed are libraries for filtering routines, software models for each of the subsystems contained in the overall system, channel models, and the waveform display and data analysis routines needed for the analysis of the simulation results (postprocessing). While at least some of the required software will have to be developed at the time the simulation is being written, many of the required routines can probably be obtained from digital signal processing (DSP) programs and other available sources. As more simulation projects are completed, the database of available routines becomes larger.

The other alternative is to use a **dedicated simulation language**, which makes it possible for one who does not have the necessary skills to create a custom simulation using a high-level language to develop a communication system simulation. Many simulation languages are available for both personal computers and workstations [Shanmugan, 1988]. While the use of these resources can speed simulation development, the user must ensure that the assumptions used in developing the models are well understood and applicable to the problem of interest. In choosing a dedicated language from among those that are available, one should select a language that has an extensive model library, an integrated postprocessor with a wide variety of data analysis routines, on-line help and documentation capabilities, and extensive error-checking routines.

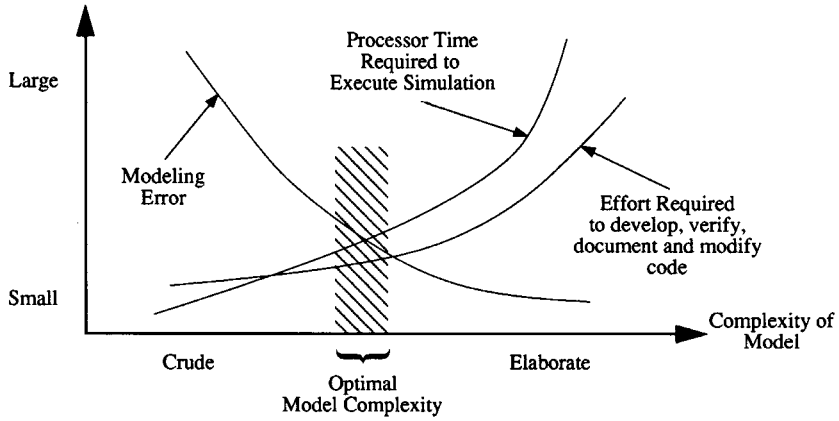


FIGURE 78.3 Design constraints and trade-offs.

78.6 The Interdisciplinary Nature of Simulation

The subject of computer-aided design and analysis of communication systems is very much interdisciplinary in nature. The major disciplines that bear on the subject are communication theory, DSP, numerical analysis, and stochastic process theory. The roles played by these subjects is clear. The simulation user must have knowledge of the behavior of communication theory if the simulation results are to be understood. The analysis techniques of communication theory allow simulation results to be verified. Since each subsystem in the overall communication system is a signal processing operation, the tools of DSP provide the algorithms to realize filters and other subsystems. Numerical analysis techniques are used extensively in the development of signal processing algorithms. Since communication systems involve random data signals, as well as noise and other disturbances, the concepts of stochastic process theory are important in developing models of these quantities and also for determining performance estimates.

78.7 Model Design

Practicing engineers frequently use models to investigate the behavior of complex systems. Traditionally, models have been physical devices or a set of mathematical expressions. The widespread use of powerful digital computers now allows one to generate computer programs that model physical systems. Although the detailed development and use of computer models differs significantly from their physical and mathematical counterparts, the computer models share many of the same design constraints and trade-offs. For any model to be useful one must guarantee that the response of the model to stimuli will closely match the response of the target system, the model must be designed and fabricated in much less time and at significantly less expense than the target system, and the model must be reasonably easy to validate and modify. In addition to these constraints, designers of computer models must assure that the amount of processor time required to execute the model is not excessive. The optimal model is the one that appropriately balances these conflicting requirements. Figure 78.3 describes the typical design trade-off faced when developing computer models. A somewhat surprising observation is that the optimal model is often not the one that most closely approximates the target system. A highly detailed model will typically require a tremendous amount of time to develop, will be difficult to validate and modify, and may require prohibitive processor time to execute. Selecting a model that achieves a good balance between these constraints is as much an art as a science. Being aware of the trade-offs which exist, and must be addressed, is the first step toward mastering the art of modeling.

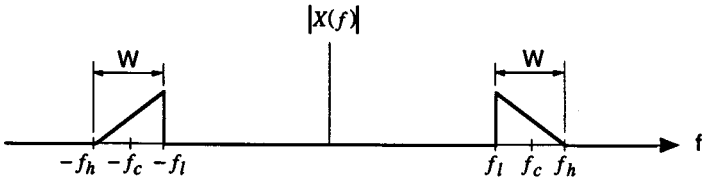


FIGURE 78.4 Amplitude spectrum of a bandpass signal.

78.8 Low-Pass Models

In most cases of practical interest the physical layer of the communication system will use continuous time (CT) signals, while the simulation will operate in discrete time (DT). For the simulation to be useful, one must develop DT signals and systems that closely match their CT counterparts. This topic is discussed at length in introductory DSP texts. A prominent result in this field is the Nyquist sampling theorem, which states that if a CT signal has no energy above frequency f_h Hz, one can create a DT signal that contains *exactly* the same information by sampling the CT signal at any rate in excess of 2 f_h samples per second. Since the execution time of the simulation is proportional to the number of samples it must process, one naturally uses the lowest sampling rate possible. While the Nyquist theorem should not be violated for arbitrary signals, when the CT signal is bandpass one can use low-pass equivalent (LPE) waveforms that contain all the information of the CT signal but can be sampled slower than 2 f_h.

Assume the energy in a bandpass signal is centered about a carrier frequency of f_c Hz and ranges from f_l to f_h Hz, resulting in a bandwidth of $f_h - f_l = W$ Hz, as in Fig. 78.4. It is not unusual for W to be many orders of magnitude less than f_c. The bandpass waveform $x(t)$ can be expressed as a function of two low-pass signals. Two essentially equivalent LPE expansions are known as the envelope/phase representation [Davenport and Root, 1958],

$$x(t) = A(t)\cos[2\pi f_c t + \theta(t)] \tag{78.1}$$

and the quadrature representation,

$$x(t) = x_c(t)\cos(2\pi f_c t) - x_s(t)\sin(2\pi f_c t) \tag{78.2}$$

All four real signals $A(t)$, $\theta(t)$, $x_c(t)$, and $x_s(t)$ are low pass and have zero energy above $W/2$ Hz. A computer simulation that replaces $x(t)$ with a pair of LPE signals will require far less processor time since the LPE waveforms can be sampled at W as opposed to 2 f_h samples per second. It is cumbersome to work with two signals rather than one signal. A more mathematically elegant LPE expansion is

$$x(t) = \mathrm{Re}\{v(t)e^{j2\pi fct}\} \tag{78.3}$$

where $v(t)$ is a low-pass, *complex-time domain signal* that has no energy above $W/2$ Hz. Signal $v(t)$ is known as the complex envelope of $x(t)$ [Haykin, 1983]. It contains all the information of $x(t)$ and can be sampled at W samples per second without aliasing. This notation is disturbing to engineers accustomed to viewing all time domain signals as real. However, a complete theory exists for complex time domain signals, and with surprisingly little effort one can define convolution, Fourier transforms, analog-to-digital and digital-to-analog conversions, and many other signal processing algorithms for complex signals. If f_c and W are known, the LPE mapping is one-to-one so that $x(t)$ can be completely recovered from $v(t)$. While it is conceptually simpler to sample the CT signals at a rate in excess of $2f_h$ and avoid the mathematical difficulties of the LPE representation, the tremendous difference between f_c and W makes the LPE far more efficient for computer simulation. This type

of trade-off frequently occurs in computer simulation. A careful mathematical analysis of the modeling problem conducted *before* any computer code is generated can yield substantial performance improvements over a conceptually simpler, but numerically inefficient approach.

The fundamental reason the LPE representation outlined above is popular in simulation is that one can easily generate **LPE models** of linear time-invariant bandpass filters. The LPE of the output of a bandpass filter is merely the convolution of the LPE of the input signal and the LPE of the impulse response of the filter. It is far more difficult to determine a LPE model for nonlinear and time-varying systems. There are numerous approaches that trade off flexibility and simplicity. If the system is nonlinear and time invariant, a Volterra series can be used. While this series will exactly represent the nonlinear device, it is often analytically intractable and numerically inefficient. For nonlinear devices with a limited amount of memory the AM/AM, AM/PM [Shimbo, 1971] LPE model is useful. This model accurately describes the response of many microwave amplifiers including traveling-wave tubes, solid-state limiting amplifiers, and, under certain conditions, devices which exhibit hysteresis. The Chebyshev transform [Blachman, 1964] is useful for memoryless nonlinearities such as hard and soft limiters. If the nonlinear device is so complex that none of the conventional LPE models can be used, one may need to convert the LPE signal back to its bandpass representation, route the bandpass signal through a model of the nonlinear device, and then reconvert the output to a LPE signal for further processing. If this must be done, one has the choice of increasing the sampling rate for the entire simulation or using different sampling rates for various sections of the simulation. The second of these approaches is known as a *multirate simulation* [Cochiere and Rabiner, 1983]. The interpolation and decimation operations required to convert between sampling rates can consume significant amounts of processor time. One must carefully examine this trade-off to determine if a multirate simulation will substantially reduce the execution time over a single, high sampling rate simulation. Efficient and flexible modeling of nonlinear devices is in general a difficult task and continues to be an area of active research.

78.9 Pseudorandom Signal and Noise Generators

The preceding discussion was motivated by the desire to efficiently model filters and nonlinear amplifiers. Since these devices often consume the majority of the processor time, they are given high priority. However, there are a number of other subsystems that do not resemble filters. One example is the data source that generates the message or waveform which must be transmitted. While signal sources may be analog or digital in nature, we will focus exclusively on binary digital sources. The two basic categories of signals produced by these devices are known as *deterministic* and *random*. When performing worst-case analysis, one will typically produce known, repetitive signal patterns designed to stress a particular subsystem within the overall communication system. For example, a signal with few transitions may stress the symbol synchronization loops, while a signal with many regularly spaced transitions may generate unusually wide bandwidth signals. The generation of this type of signal is straightforward and highly application dependent. To test the nominal system performance one typically uses a random data sequence. While generation of a truly random signal is arguably impossible [Knuth, 1981], one can easily generate pseudorandom (PN) sequences. PN sequence generators have been extensively studied since they are used in Monte Carlo integration and simulation [Rubinstein, 1981] programs and in a variety of wideband and secure communication systems. The two basic structures for generating PN sequences are binary shift registers (BSRs) and linear congruential algorithms (LCAs).

Digital data sources typically use BSRs, while noise generators often use LCAs. A logic diagram for a simple BSR is shown in Fig. 78.5. This BSR consists of a clock, six *D-type* flip-flops (F/F), and an exclusive OR gate denoted by a modulo-two adder. If all the F/F are initialized to 1, the output of the device is the waveform shown in Fig. 78.6. Notice that the waveform is periodic with period $63 = 2^6 - 1$, but within one cycle the output has many of the properties of a random sequence. This demonstrates all the properties of the BSR, LCA, and more advanced PN sequence generators. All PN generators have memory and must therefore be initialized by the user before the first sample is generated. The initialization data is typically called the seed. One must choose this seed carefully to ensure the output will have the desired properties (in this example, one must avoid setting all F/F to zero). All PN sequence generators will produce periodic sequences. This may or may not be

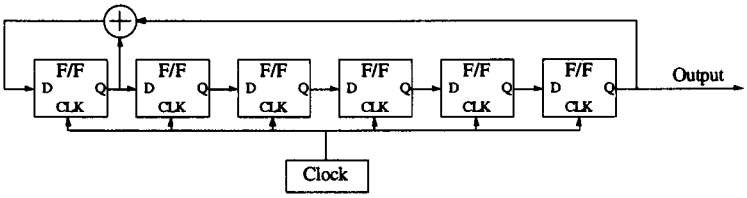


FIGURE 78.5 Six-stage binary shift register PN generator.

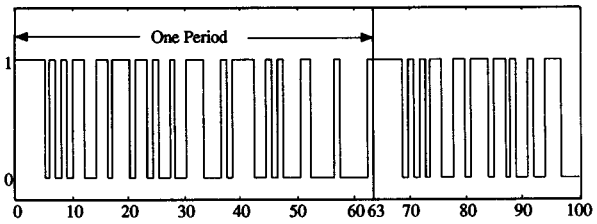


FIGURE 78.6 Output of a six-stage maximal length BSR.

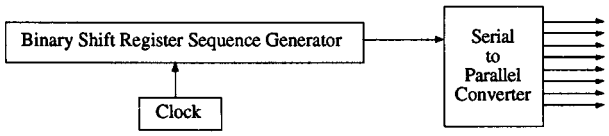


FIGURE 78.7 *M*-ary PN sequence generator.

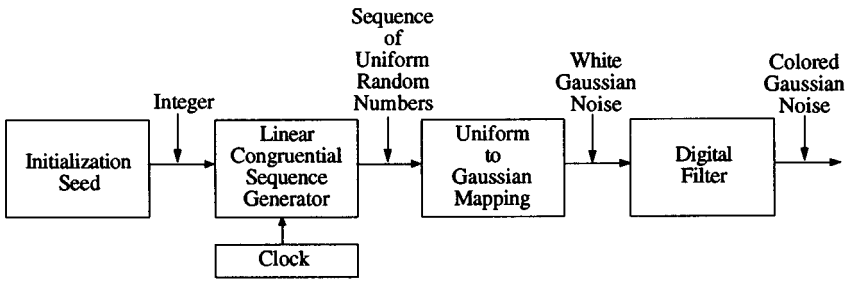


FIGURE 78.8 Generation of Gaussian noise.

a problem. If it is a concern, one should ensure that one period of the PN sequence generator is longer than the total execution time of the simulation. This is usually not a significant problem, since one can easily construct BSRs that have periods greater than 10^{27} clock cycles. The final concern is how closely the behavior of the PN sequence generator matches a truly random sequence. Standard statistical analysis algorithms have been applied to many of these generators to validate their performance.

Many digital communication systems use m bit *(M-ary) sources where* $m > 1$. Figure 78.7 depicts a simple algorithm for generating a M-ary random sequence from a binary sequence. The clock must now cycle through m cycles for every generated symbol, and the period of the generator has been reduced by a factor of m. This may force the use of a longer-period BSR. Another common application of PN sequence generators is to produce

samples of a continuous stochastic process, such as Gaussian noise. A structure for producing these samples is shown in Fig. 78.8. In this case the BSR has been replaced by an LCA [Knuth, 1981]. The LCA is very similar to BSR in that it requires a seed value, is clocked once for each symbol generated, and will generate a periodic sequence. One can generate a white noise process with an arbitrary first-order probability density function (pdf) by passing the output of the LCA through an appropriately designed nonlinear, memoryless mapping. Simple and well-documented algorithms exist for the uniform to Gaussian mapping. If one wishes to generate a nonwhite process, the output can be passed through the appropriate filter. Generation of a wide-sense stationary Gaussian stochastic process with a specified power spectral density is a well-understood and -documented problem. It is also straightforward to generate a white sequence with an arbitrary first-order pdf or to generate a specified power spectral density if one does not attempt to control the pdf. However, the problem of generating a noise source with an arbitrary pdf *and* an arbitrary power spectral density is a significant challenge [Sondhi, 1983].

78.10 Transmitter, Channel, and Receiver Modeling

Most elements of transmitters, channels, and receivers are implemented using standard DSP techniques. Effects that are difficult to characterize using mathematical analysis can often be included in the simulation with little additional effort. Common examples include gain and phase imbalance in quadrature circuits, nonlinear amplifiers, oscillator instabilities, and antenna platform motion. One can typically use LPE waveforms and devices to avoid translating the modulator output to the carrier frequency. Signal levels in physical systems often vary by many orders of magnitude, with the output of the transmitters being extremely high energy signals and the input to receivers at very low energies. To reduce execution time and avoid working with extremely large and small signal level simulations, one often omits the effects of linear amplifiers and attenuators and uses normalized signals. Since the performance of most systems is a function of the signal-to-noise ratio, and not of absolute signal level, normalization will have no effect on the measured performance. One must be careful to document the normalizing constants so that the original signal levels can be reconstructed if needed. Even some rather complex functions, such as error detecting and correcting codes, can be handled in this manner. If one knows the uncoded error rate for a system, the coded error rate can often be closely approximated by applying a mathematical mapping. As will be pointed out below, the amount of processor time required to produce a meaningful error rate estimate is often inversely proportional to the error rate. While an uncoded error rate may be easy to measure, the coded error rate is usually so small that it would be impractical to execute a simulation to measure this quantity directly. The performance of a coded communication system is most often determined by first executing a simulation to establish the channel **symbol error rate**. An analytical mapping can then be used to determine the decoded BER from the channel symbol error rate.

Once the signal has passed though the channel, the original message is recovered by a receiver. This can typically be realized by a sequence of digital filters, feedback loops, and appropriately selected nonlinear devices. A receiver encounters a number of clearly identifiable problems that one may wish to address independently. For example, receivers must initially synchronize themselves to the incoming signal. This may involve detecting that an input signal is present, acquiring an estimate of the carrier amplitude, frequency, phase, symbol synchronization, frame synchronization, and, in the case of spread spectrum systems, code synchronization. Once acquisition is complete, the receiver enters a steady-state mode of operation, where concerns such as symbol error rate, mean time to loss of lock, and reaction to fading and interference are of primary importance. To characterize the system, the user may wish to decouple the analysis of these parameters to investigate relationships that may exist.

For example, one may run a number of acquisition scenarios and gather statistics concerning the probability of acquisition within a specified time interval or the mean time to acquisition. To isolate the problems faced in synchronization from the inherent limitation of the channel, one may wish to use perfect synchronization information to determine the minimum possible BER. Then the symbol or carrier synchronization can be held at fixed errors to determine sensitivity to these parameters and to investigate worst-case performance. Noise processes can be used to vary these parameters to investigate more typical performance. The designer may also

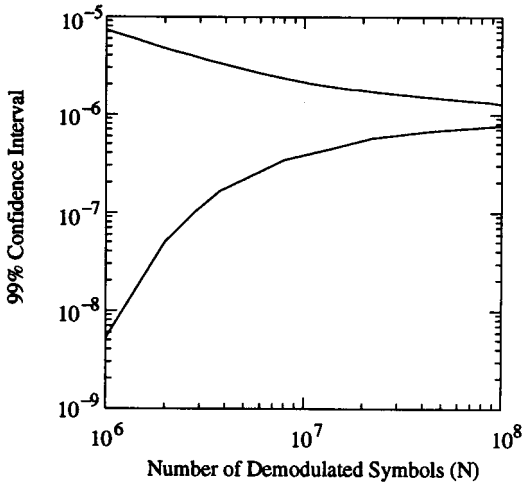


FIGURE 78.9 Typical confidence interval (BER) point estimate = 10^{-6}.

wish to investigate the performance of the synchronization system to various data patterns or the robustness of the synchronization system in the face of interference. The ability to measure the system response to one parameter while a wide range of other parameters are held fixed and the ability to quickly generate a wide variety of environments are some of the more significant advantages that simulation enjoys over more conventional hardware and analytical models.

78.11 Symbol Error Rate Estimation

One of the most fundamental parameters to measure in a digital communication system is the steady-state BER. The simplest method for estimating the BER is to perform a **Monte Carlo (MC) simulation**. The simulation conducts the same test one would perform on the physical system. All data sources and noise sources produce typical waveforms. The output of the demodulator is compared to the output of the message source, and the BER is estimated by dividing the number of observed errors by the number of bits transmitted. This is a simple technique that will work with any system that has ergodic [Papoulis, 1965] noise processes. The downside of this approach is that one must often pass a very large number of samples through the system to produce a reliable estimate of the BER. The question of how many samples must be collected can be answered using confidence intervals. The confidence interval gives a measure of how close the true BER will be to the estimate produced by the MC simulation. A typical confidence interval curve is shown in Fig. 78.9. The ratio of the size of the confidence interval to the size of the estimate is a function of the number of errors observed. Convenient rules of thumb for this work are that after one error is observed the point estimate is accurate to within 3 orders of magnitude, after 10 errors the estimate is accurate to within a factor of 2, and after 100 errors the point estimate will be accurate to a factor of 1.3. This requirement for tens or hundreds of errors to occur frequently limits the usefulness of MC simulations for systems that have low error rates and has motivated research into more efficient methods of estimating BER.

Perhaps the fastest method of BER estimation is the semi-analytic (SA) or quasi-analytic technique [Jeruchim, 1984]. This technique is useful for systems that resemble Fig. 78.10. In this case the mean of the decision metric is a function of the transmitted data pattern and is independent of the noise. All other parameters of the pdf of the decision metric are a function of the noise and are independent of the data. This means that one can analytically determine the conditional pdf of the decision metric given the transmitted data pattern. By using total probability one can then determine the unconditional error rate. The problem with conventional mathematical analysis is that when the channel has a significant amount of memory or the nonlinearity is rather

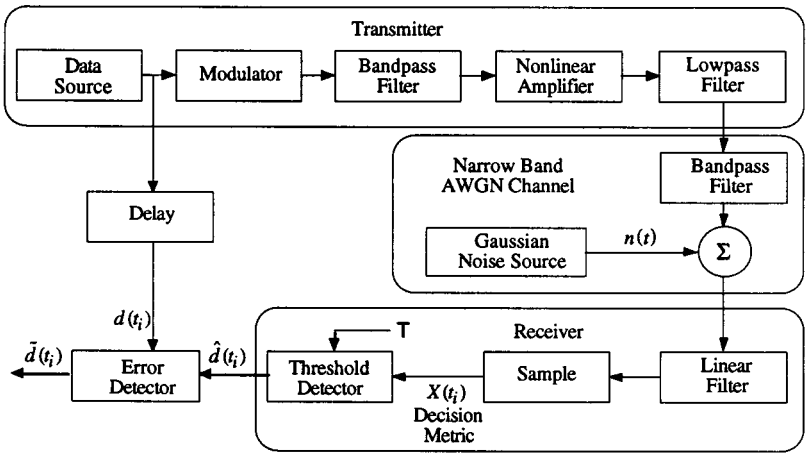


FIGURE 78.10 Typical digital communication system.

complex, one must compute a large number of conditional density functions. Simulation can easily solve this problem for most practical systems. A noise-free simulation is executed, and the value of the decision metric is recorded in a data file. Once the simulation is complete, this information can be used to reconstruct the conditional and ultimately the unconditional error rate. This method generates highly accurate estimates of the BER and makes very efficient use of computer resources, but can only be used in the special cases where one can analytically determine the conditional pdf.

The MC and SA techniques fall at the two extremes of BER estimation. MC simulations require no *a priori* information concerning the system performance or architecture but may require tremendous amounts of computer time to execute. SA techniques require an almost trivial amount of computer time for many cases but require the analyst to have a considerable amount of information concerning the system. There is a continuing search for algorithms that fall in between these extremes. These variance reduction algorithms all share the property of making a limited number of assumptions concerning system performance and architecture, then using this information to reduce the variance of the MC estimate. Popular techniques are summarized in [Jeruchim, 1984] and include importance sampling, large deviation theory, extremal statistics, and tail extrapolation. To successfully use one of these techniques one must first understand the basic concept behind the technique. Then one should carefully determine what assumptions were made concerning the system architecture to determine if the system under study satisfies the requirements. This can be a difficult task since it is not always clear what assumptions are required for a specified technique to be applicable. Finally, one should always determine the accuracy of the measurement through some technique similar to confidence interval estimation.

78.12 Validation of Simulation Results

One often constructs a simulation to determine the value of a single parameter, such as the system BER. However the estimate of this parameter has little or no value unless one can ensure that the simulation model closely resembles the physical system. A number of methods can be used to validate a simulation. Individually, none of them will guarantee that the simulation results are accurate, but taken as a group, they form a convincing argument that the results are realistic. Seven methods of validation are mathematical analysis, comparison with hardware, bounding techniques, degenerate case studies, reasonable relationship tests, subsystem tests, and redundant simulation efforts.

If one has access to mathematical analysis or hardware that predicts or approximates the performance of the system, one should obviously compare the simulation and mathematical results. Unfortunately, in most cases these results will not be available. Even though exact mathematical analysis of the system is not possible, it may be possible to develop bounds on the system performance. If these bounds are tight, they may accurately

characterize the system performance, but even loose bounds will be useful since they help verify the simulation results. Most systems have parameters that can be varied. While it may be mathematically difficult to determine the performance of the system for arbitrary values, it is often possible to mathematically determine the results when parameters assume extreme or degenerate values.

Other methods of validation are decidedly less mathematical. One may wish to vary parameters and ascertain whether the performance parameter changes in a reasonable manner. For example, small changes in SNR rarely cause dramatic changes in system performance. When constructing a simulation, each subsystem, such as filters, nonlinear amplifiers, and noise and data sources, should be thoroughly tested before being included in a larger simulation. Be aware, however, that correct operation of all the various subsystems that make up a communication system does not imply that the overall system performs correctly. If one is writing his or her own code, one must verify that there are no software bugs or fundamental design errors. Even if one purchases a commercial software package, there is no guarantee that the designer of the software models made the same assumptions the user will make when using the model. In most cases it will be far easier to test a module before it is inserted into a simulation than it will be to isolate a problem in a complex piece of code. The final check one may wish to perform is a redundant simulation. There are many methods of simulating a system. One may wish to have two teams investigate a problem or have a single team implement a simulation using two different techniques to verify that the results are reasonable.

78.13 A Simple Example Illustrating Simulation Products

To illustrate the output that is typically generated by a communication system simulation, a simple example is considered. The system is that considered in Fig. 78.10. An OQPSK (offset quadrature phase-shift keyed) modulation format is assumed so that one of four waveforms is transmitted during each symbol period. The data source may be viewed as a single binary source, in which the source symbols are taken two at a time when mapped onto a transmitted waveform, or as two parallel data sources, with one source providing the direct channel modulation and the second source providing the quadrature channel modulation. The signal constellation at the modulator output appears as shown in Fig. 78.11(a), with the corresponding eye diagram appearing as shown in Fig. 78.11(b). The eye diagram is formed by overlaying successive time intervals of a time domain waveform onto a single graph, much as would be done with a common oscilloscope. Since the simulation sampling frequency used in generating Fig. 78.11(b) was 10 samples per data symbol, it is easily seen that the eye diagram was generated by retracing every 2 data symbols or 20 simulation samples. Since Fig. 78.11(a) and (b) correspond to the modulator output, which has not yet been filtered, the transitions between binary states occur in one simulation step. After filtering, the eye diagram appears as shown in Fig. 78.11(c). A seventh-order Butterworth bilinear z-transform digital filter was assumed with a 3-dB bandwidth equal to the bit rate. It should be noted that the bit transitions shown in Fig. 78.11(c) do not occur at the same times as the bit transitions shown in Fig. 78.11(b). The difference is due to the group delay of the filter. Note in Fig. 78.10 that the transmitter also involves a nonlinear amplifier. We will see the effects of this component later in this section.

Another interesting point in the system is within the receiver. Since the communication system is being modeled as a baseband system due to the use of the complex-envelope representation of the bandpass waveforms generated in the simulation, the data detector is represented as an integrate-and-dump detector. The detector is then modeled as a sliding-average integrator, in which the width of the integration window is one bit time. The integration is therefore over a single bit period when the sliding window is synchronized with a bit period. The direct-channel and quadrature-channel waveforms at the output of the sliding-average integrator are shown in Fig. 78.12(a). The corresponding eye diagrams are shown in Fig. 78.12(b). In order to minimize the error probability of the system, the bit decision must be based on the integrator output at the time for which the eye opening is greatest. Thus the eye diagram provides important information concerning the sensitivity of the system to timing errors.

The signal constellation at the sliding integrator output is shown in Fig. 78.12(c) and should be carefully compared to the signal constellation shown in Fig. 78.11(a) for the modulator output. Three effects are apparent. First, the signal points exhibit some scatter, which, in this case, is due to intersymbol interference resulting

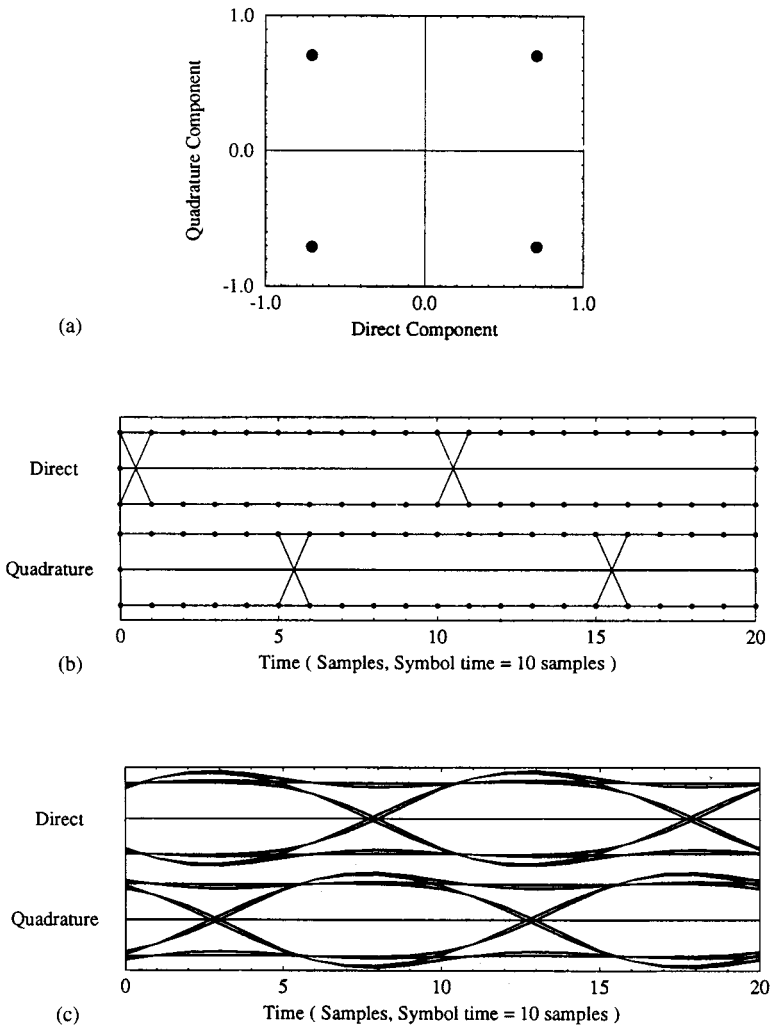
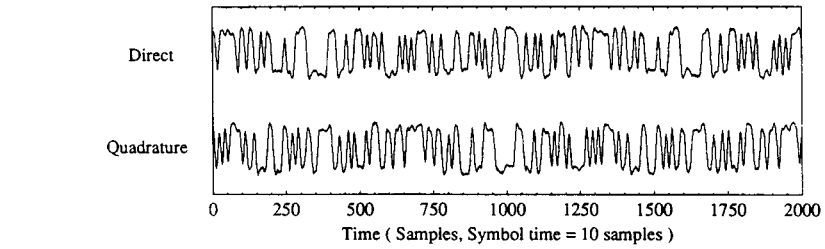


FIGURE 78.11 Transmitter signal constellation and eye diagrams: (a) OQPSK signal constellation; (b) eye diagram of modulator output; (c) eye diagram of filtered modulator output.

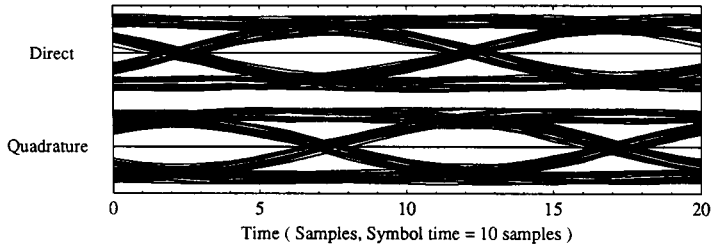
from the transmitter filter and additive noise. It is also clear that the signal is both compressed and rotated. These effects are due to the nonlinear amplifier that was mentioned previously. For this example simulation the nonlinear amplifier is operating near the saturation point, and the compression of the signal constellation is due to the AM/AM characteristic of the nonlinearity and the rotation is due to the AM/PM characteristic of the nonlinearity.

The performance of the overall communication system is illustrated in Fig. 78.12(d). The error probability curve is perhaps the most important simulation product. Note that both uncoded and coded results are shown. The coded results were calculated analytically from the uncoded results assuming a (63, 55) Reed–Solomon code. It should be mentioned that **semi-analytic simulation** was used in this example since, as can be seen in Fig. 78.10, the noise is injected into the system on the receiver side of the nonlinearity so that linear analysis may be used to determine the effects of the noise on the system performance.

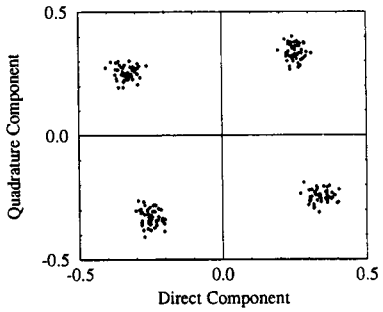
This simple example serves to illustrate only a few of the possible simulation products. There are many other possibilities including histograms, correlation functions, estimates of statistical moments, estimates of the power spectral density, and estimates of the signal-to-noise ratio at various points in the system.



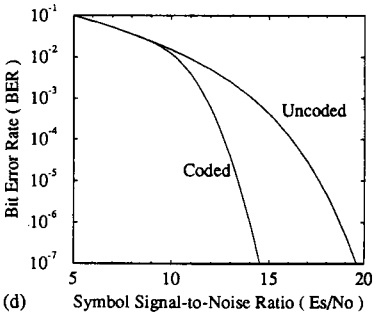
(a)



(b)



(c)



(d)

FIGURE 78.12 Integrator output signals and system error probability: (a) sliding integrator output signals; (b) sliding integrator output eye diagram; (c) sliding integrator output signal constellation; (d) error probability.

A word is in order regarding spectral estimation techniques. Two basic techniques can be used for spectral estimation: Fourier techniques and model-based techniques. In most simulation problems one is blessed with a tremendous amount of data concerning sampled waveforms but does not have a simple model describing how these waveforms are produced. For this reason model-based spectral estimation is typically not used. The most common form of spectral estimation used in simulation is the Welch periodogram. While this approach is straightforward, the effects of windowing the data sequence must be carefully considered, and tens or even hundreds of data windows must be averaged to achieve an accurate estimate of the power spectral density.

78.14 Conclusions

We have seen that the analysis and design of today's complex communication systems often requires the use of computer-aided techniques. These techniques allow the solution of problems that are otherwise not tractable and provide considerable insight into the operating characteristics of the communication system.

Defining Terms

Communication link: A point-to-point communication system that typically involves a single information source and a single user. This is in contrast to a communications network, which usually involves many sources and many users.

Computer-aided design and analysis: The process of using computer assistance in the design and analysis of complex systems. In our context, the design and analysis of communication systems, computer-aided design and analysis often involves the extensive use of simulation techniques. Computer-aided techniques often allow one to address design and analysis problems that are not tractable analytically.

Computer simulation: A set of computer programs which allows one to imitate the important aspects of the behavior of the specific system under study. Simulation can aid the design process by, for example, allowing one to determine appropriate system design parameters or aid the analysis process by, for example, allowing one to estimate the end-to-end performance of the system under study.

Dedicated simulation language: A computer language, either text based or graphics based, specifically developed to facilitate the simulation of the various systems under study, such as communication systems.

Low-pass equivalent (LPE) model: A method of representing bandpass signals and systems by low-pass signals and systems. This technique is extremely useful when developing discrete time models of bandpass continuous-time systems. It can substantially reduce the sampling rate required to prevent aliasing and does not result in any loss of information. This in turn reduces the execution time required for the simulation. This modeling technique is closely related to the quadrature representation of bandpass signals.

Monte Carlo simulation: A technique for simulating systems that contain signal sources producing stochastic or random signals. The signal sources are modeled by pseudorandom generators. Performance measures, such as the symbol error rate, are then estimated by time averages. This is a general-purpose technique that can be applied to an extremely wide range of systems. It can, however, require large amounts of computer time to generate accurate estimates.

Pseudorandom generator: An algorithm or device that generates deterministic waveforms which in many ways resemble stochastic or random waveforms. The power spectral density, autocorrelation, and other time averages of pseudorandom signals can closely match the time and ensemble averages of stochastic processes. These generators are useful in computer simulation where one may be unable to generate a truly random process, and they have the added benefit of providing reproducible signals.

Semi-analytic simulation: A numerical analysis technique that can be used to efficiently determine the symbol error rate of digital communication systems. It can be applied whenever one can analytically determine the probability of demodulation error given a particular transmitted data pattern. Although this technique can only be applied to a restricted class of systems, in these cases it is far more efficient, in terms of computer execution time, than Monte Carlo simulations.

Simulation validation: The process of certifying that simulation results are reasonable and can be used with confidence in the design or analysis process.

Symbol error rate: A fundamental performance measure for digital communication systems. The symbol error rate is estimated as the number of errors divided by the total number of demodulated symbols. When the communication system is ergodic, this is equivalent to the probability of making a demodulation error on any symbol.

Related Topics

69.3 Television Systems • 73.1 Signal Detection • 102.1 Avionics Systems • 102.2 Communications Satellite Systems: Applications

References

K. Shanmugan, "An update on software packages for simulation of communication systems (links)," *IEEE J. Selected Areas Commun.,* no. 1, 1988.

W. Davenport and W. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.

S. Haykin, *Communication Systems*, New York: Wiley, 1983.

O. Shimbo, "Effects of intermodulation, AM-PM conversion, and additive noise in multicarrier TWT systems," *Proc. IEEE*, no. 2, 1971.

N. Blachman, "Bandpass nonlinearities," *IEEE Trans. Inf. Theory*, no. 2, 1964.

R. Cochiere and L. Rabiner, *Multirate Digital Signal Processing,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.

D. Knuth, *The Art of Computer Programming,* vol. 2, *Seminumerical Algorithms,* 2nd ed., Reading, Mass.: Addison-Wesley, 1981.

R. Rubinstein, *Simulation and the Monte Carlo Method,* New York: Wiley, 1981.

M. Sondhi, "Random processes with specified spectral density and first-order probability density," *Bell Syst. Tech. J.,* vol. 62, 1983.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1965.

M. Jeruchim, P. Balaban, and K. Shanmugan, *Simulation of Communication Systems*, New York: Plenum, 1992.

M. Jeruchim, "Techniques for estimating the bit error rate in the simulation of digital communication systems," *IEEE J. Selected Areas Commun.,* no. 1, January 1984.

P. Bratley, B. L. Fox, and L. E. Schrage, *A Guide to Simulation,* New York: Springer-Verlag, 1987.

P. Balaban, K. S. Shanmugan, and B. W. Stuck (eds.), "Special issue on computer-aided modeling, analysis and design of communication systems," *IEEE J. Selected Areas Commun.,* no. 1, 1984.

P. Balaban, E. Biglieri, M. C. Jeruchim, H. T. Mouftah, C. H. Sauer, and K. S. Shanmugan (eds.), "Computer-aided modeling, analysis and design of communication systems II," *IEEE J. Selected Areas Commun.,* no. 1, 1988.

H. T. Mouftah, J. F. Kurose, and M. A. Marsan (eds.), "Computer-aided modeling, analysis and design of communication networks I," *IEEE J. Selected Areas Commun.,* no. 9, 1990.

H. T. Mouftah, J. F. Kurose, and M. A. Marsan (eds.), "Computer-aided modeling, analysis and design of communication networks II," *IEEE J. Selected Areas Commun.,* no. 1, 1991.

J. Gibson, *The Mobile Communications Handbook,* Boca Raton, Fla.: CRC Press, 1996.

J. Gagliardi, *Optical Communication,* New York: Wiley, 1995.

S. Haykin, *Communication Systems,* New York: Wiley, 1994.

R. L. Freeman, *Telecommunications Systems Engineering,* New York: Wiley, 1996.

N. D. Sherali, "Optimal Location of Transmitters," *IEEE J. on Selected Areas in Communications,* pp. 662–673, May 1996.

Further Information

Until recently the subject of computer-aided analysis and simulation of communication systems was a very difficult research area. There were no textbooks devoted to the subject, and the fundamental papers were scattered over a large number of technical journals. While a number of excellent books treated the subject of simulation of systems in which random signals and noise are present [Rubinstein, 1981; Bratley et al., 1987], none of these books specifically focused on communication systems.

Starting in 1984, the *IEEE Journal on Selected Areas in Communications (JSAC)* initiated the publication of a sequence of issues devoted specifically to the subject of computer-aided design and analysis of communication systems. A brief study of the contents of these issues tells much about the rapid development of the discipline. The first issue, published in January 1984 [Balaban et al., 1984], emphasizes communication links, although there are a number of papers devoted to networks. The portion devoted to links contained a large collection of papers devoted to simulation packages.

The second issue of the series was published in 1988 and is roughly evenly split between links and networks [Balaban et al., 1988]. In this issue the emphasis is much more on techniques than on simulation packages. The third part of the series is a two-volume issue devoted exclusively to networks [Mouftah et al., 1990, 1991].

As of this writing, the book by Jeruchim et al. is the only comprehensive treatment of the simulation of communication links [Jeruchim, 1984]. It treats the component and channel modeling problem and the problems associated with using simulation techniques for estimating the performance of communication systems in considerable detail. This textbook, together with the previously cited JSAC issues, gives a good overview of the area.

Sandige, R.S. "Section VIII – Digital Devices"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

LCD projection televisions, such as this "Television of the Future" could replace the CRT-based television if certain performance criteria are obtained. One criteria being addressed is the need for a very bright, optically efficient point source for projecting an image through a small LCD. If new lighting technologies are developed, such as an illumination source that can provide a brighter image with better colors than current CRT technology, then the consumer television market will certainly change. (Photo courtesy of Thomson Multimedia.)

VIII

Digital Devices

79 Logic Elements *G.L. Moss, P. Graham, R.S. Sandige, H.S. Hinton*
IC Logic Family Operation and Characteristics • Logic Gates (IC) • Bistable Devices • Optical Devices

80 Memory Devices *W.D. Pricer, R.H. Katz, P.A. Lee, M. Mansuripur*
Integrated Circuits (RAM, ROM) • Basic Disc System Architectures • Magnetic Tape • Magneto-Optical Disk Data Storage

81 Logical Devices *F.P. Preparata, R.S. Sandige, B.R. Bannister, D.G. Whitehead, M. Bolton, B.D. Carroll*
Combinational Networks and Switching Algebra • Logic Circuits • Registers and their Applications • Programmable Arrays • Arithmetic Logic Units

82 Microprocessors *J. Staudhammer, S.-L. Chen, P.J. Windley, J.F. Frenzel*
Practical Microprocessors • Applications

83 Displays *J.E. Morris, A. Martin, L.F. Weber*
Light-Emitting Diodes • Liquid-Crystal Displays • The Cathode Ray Tube • Color Plasma Displays

84 Data Acquisition *D. Kurumbalapitiya, S.R.H. Hoole*
The Analog and Digital Signal Interface • Analog Signal Conditioning • Sample-and-Hold and A/D Techniques in Data Acquisition • The Communication Interface of a Data Acquisition System • Data Recording • Software Aspects

85 Testing *M. Serra, B.I. Dervisoglu*
Digital IC Testing • Design for Test

Richard S. Sandige
University of Wyoming

ELECTRONIC DESIGNERS have placed increasing significance on digital devices since the late 1960s. This is due primarily to the greater reliability and improved accuracy gained when using electronic devices in a two-level mode (binary mode) as compared to using electronic devices in a continuous mode (analog mode). As silicon integrated circuits (ICs) became denser and more consistently reproducible over the past few decades, so did digital electronic devices. Today digital circuits and digital systems produced from digital devices can be found in every walk of life ranging from children's toys, kitchen appliances, laboratory instruments, personal and workstation computers to space shuttle and satellite applications.

The intent of this section is to present topics related to the utilization and application of digital devices. Chapter 79 establishes the foundation for digital logic elements beginning with IC, logic gates, logical families, bistable devices, and optical devices. Discussed in the next chapter are memory devices, which include integrated circuits (RAM, ROM), disk systems, magnetic tape, and optical disks. Chapter 81 on logical devices discusses switching algebra, logic circuits, registers, programmable arrays (PAL, FPGA), and arithmetic logic units.

The next chapter explains the microprocessor, perhaps the best-known digital device. The topics covered include practical microprocessors and microprocessor applications. Chapter 83 on displays consists of the light-emitting diode, the liquid-crystal display, the cathode ray tube, and the plasma display. The gathering of digital

information, referred to as data acquisition, is discussed next. No digital system is released to production without extensive testing. Chapter 85 presents methods of testing and design for testing.

The variety of topics presented in this section should provide readers with a contemporary overview of digital devices and their applications. To obtain additional information, the reader may refer to the References and Further Information in each chapter.

Nomenclature

Symbol	Quantity	Unit	Symbol	Quantity	Unit
A	area	m^2	λ	radiation wavelength	nm
a	average absorption coefficient		m	magnification factor	
			μ_n	electron mobility	
B	luminance off the projection screen		n	aperture	
			ν	photon frequency	Hz
C	brightness contrast		p	photon momentum	$kg \cdot m/s$
CMRR	common-mode rejection ratio		R	recombination rate	ns
			R_i	reflectivity	
C_R	contrast ratio		S	emitting screen surface	m^2
d	diameter	m	SNR	signal-to-noise ratio	
E_g	band gap energy	eV	t_{add}	add time	ns
ε	screen efficiency	lumen/W	t_h	hold time	ns
f	focal length	m	t_{pd}	propagation delay time	ns
F	maximum flux	lumen	T	transmission ratio	
h	Planck's constant	6.626×10^{-34} J · s	T	transmission of faceplate	
			τ	lifetime	ns
η	quantum efficiency		q_c	critical angle	degree
I	beam current	amp	V_B	accelerating voltage	V
L	luminance	cd/m^2	V_s	screen voltage	V
L	raster luminance				

Moss, G.L., Graham, P., Sandige, R.S., Hinton, H.S. “Logic Elements”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

79

Logic Elements

Gregory L. Moss
Purdue University

Peter Graham
Florida Atlantic University (Retired)

Richard S. Sandige
University of Wyoming

H. S. Hinton
University of Colorado

79.1 IC Logic Family Operation and Characteristics
IC Logic Families and Subfamilies • TTL Logic Family • CMOS Logic Family • ECL Logic Family • Logic Family Circuit Parameters • Interfacing Between Logic Families

79.2 Logic Gates (IC)
Gate Specification Parameters • Bipolar Transistor Gates • Complementary Metal-Oxide Semiconductor (CMOS) Logic • Choosing a Logic Family

79.3 Bistable Devices
Basic Latches • Gated Latches • Flip-Flops • Edge-Triggered Flip-Flops • Special Notes on Using Latches and Flip-Flops

79.4 Optical Devices
All-Optical Devices • Optoelectronic Devices • Limitations

79.1 IC Logic Family Operation and Characteristics

Gregory L. Moss

Digital logic circuits can be classified as belonging to one of two categories, either combinational (also called combinatorial) or sequential logic circuits. The output logic level of a combinatorial circuit depends only on the current logic levels present at the circuit's inputs. Sequential logic circuits, on the other hand, have a memory characteristic so the sequential circuit's output is dependent not only on the current input conditions but also on the current output state of the circuit. The primary building block in combinational circuits is the logic gate. The three simplest logic gate functions are the inverter (or NOT), AND, and OR. Other common basic logic functions are derived from these three. Table 79.1 gives **truth table** definitions of the various types of logic gates. The memory elements used to construct sequential logic circuits are called latches and flip-flops.

The integrated circuit switching logic used in modern digital systems will generally be from one of three families: transistor-transistor logic (TTL), complementary metal-oxide semiconductor logic (CMOS), or emitter-coupled logic (ECL). Each of the logic families has its advantages and disadvantages. The three major families are also divided into various subfamilies derived from performance improvements in integrated circuit (IC) design technology. Bipolar transistors provide the switching action in both TTL and ECL families, while enhancement-mode MOS transistors are the basis for the CMOS family. Recent improvements in switching circuit performance are also attained using BiCMOS technology, the merging of bipolar and CMOS technologies on a single chip. A particular logic family is usually selected by digital designers based on such criteria as

1. Switching speed
2. Power dissipation
3. PC board area requirements (levels of integration)
4. Output drive capability (**fan-out**)
5. Noise immunity characteristics
6. Product breadth
7. Sourcing of components

TABLE 79.1 Defining Truth Tables for Logic Gates

1-Input Function		2-Input Functions							
Input	Output	Inputs		Output Functions					
A	NOT	A	B	AND	OR	NAND	NOR	XOR	XNOR
0	1	0	0	0	0	1	1	0	1
1	0	0	1	0	1	1	0	1	0
		1	0	0	1	1	0	1	0
		1	1	1	1	0	0	0	1

TABLE 79.2 Logic Families and Subfamilies

Family and Subfamily	Description
TTL	Transistor-transistor logic
74xx	Standard TTL
74Lxx	Low-power TTL
74Hxx	High-speed TTL
74Sxx	Schottky TTL
74LSxx	Low-power Schottky TTL
74ASxx	Advanced Schottky TTL
74ALSxx	Advanced low-power Schottky TTL
74Fxx	Fast TTL
CMOS	Complementary metal-oxide semiconductor
4xxx	Standard CMOS
74Cxx	Standard CMOS using TTL numbering system
74HCxx	High-speed CMOS
74HCTxx	High-speed CMOS—TTL compatible
74FCTxx	Fast CMOS—TTL compatible
74ACxx	Advanced CMOS
74ACTxx	Advanced CMOS—TTL compatible
74AHCxx	Advanced high-speed CMOS
74AHCTxx	Advanced high-speed CMOS-TTL compatible
ECL (or CML)	Emitter-coupled (current-mode) logic
10xxx	Standard ECL
10Hxxx	High-speed ECL

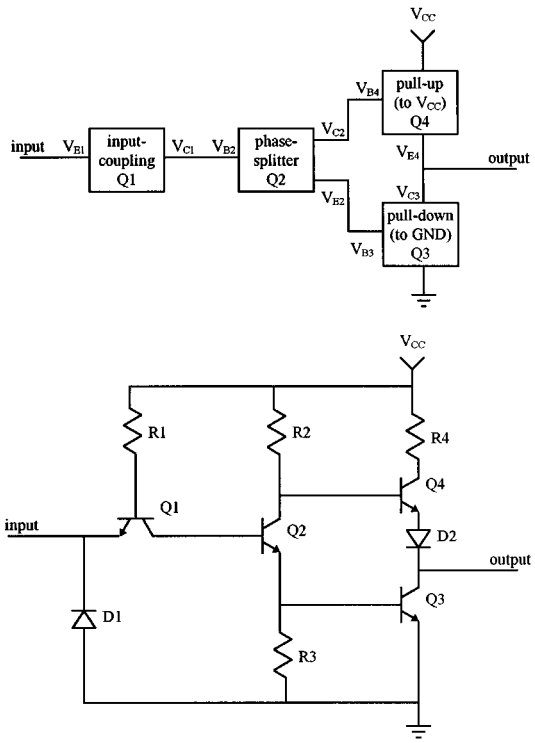
IC Logic Families and Subfamilies

The integrated circuit logic families actually consist of several subfamilies of ICs that differ in various performance characteristics. The TTL logic family has been the most widely used family type for applications that employ small-scale integration (SSI) or medium-scale integration (MSI) integrated circuits. Lower power consumption and higher levels of integration are the principal advantages of the CMOS family. The ECL family is generally used in applications that require high-speed switching logic. Today, the most common device numbering system used in the TTL and CMOS families has a prefix of 54 (generally used in military applications and having an operating temperature range of –55 to 125°C) and 74 (generally used in industrial/commercial applications and having an operating temperature range of 0 to 70°C). Table 79.2 identifies various logic families and subfamilies.

TTL Logic Family

The TTL family has been the most widely used logic family for many years in applications that use SSI and MSI. It is relatively fast and offers a great variety of standard chips.

The active switching element used in all TTL family circuits is the *npn* bipolar junction transistor (BJT). The transistor is turned on when the base is approximately 0.7 V more positive than the emitter and there is a sufficient amount of base current flowing. The turned on transistor (in non-Schottky subfamilies) is said to



input	V_{C1}	Q2	V_{C2}	V_{E2}	Q3	V_{C3}	Q4	V_{E4}	output
hi	hi	on	low	hi	on	low	off	open	low
low	low	off	hi	low	off	open	on	hi	hi

FIGURE 79.1 TTL inverter circuit block diagram and operation.

be in saturation and, ideally, acts like a closed switch between the collector and emitter terminals. The transistor is turned off when the base is not biased with a high enough voltage (with respect to the emitter). Under this condition, the transistor acts like an open switch between the collector and emitter terminals.

Figure 79.1 illustrates the transistor circuit blocks used in a standard TTL inverter. Four transistors are used to achieve the inverter function. The input to the gate connects to the emitter of transistor Q1, the input coupling transistor. A clamping diode on the input prevents negative input voltage spikes from damaging Q1. The collector voltage (and current) of Q1 controls Q2, the phase splitter transistor. Q2, in turn, controls the Q3 and Q4 transistors forming the output circuit, which is called a totem-pole arrangement. Q4 serves as a pull-up transistor to pull the output high when it is turned on. Q3 does just the opposite to the output and serves as a pull-down transistor. Q3 pulls the output low when it is turned on. Only one of the two transistors in the totem pole may be turned on at a time, which is the function of the phase splitter transistor Q2.

When a high **logic level** is applied to the inverter's input, Q1's base-emitter junction will be reverse biased and the base-collector junction will be forward biased. This circuit condition will allow Q1 collector current to flow into the base of Q2, saturating Q2 and thereby providing base current into Q3, turning it on also. The collector voltage of Q2 is too low to turn on Q4 so that it appears as an open in the top part of the totem pole. A diode between the two totem-pole transistors provides an extra voltage drop in series with the base-emitter junction of Q4 to ensure that Q4 will be turned off when Q2 is turned on. The saturated Q3 transistor brings the output near ground potential, producing a low output result for a high input into the inverter.

When a low logic level is applied to the inverter's input, Q1's base-emitter junction will be forward biased and the base-collector junction will be reverse biased. This circuit condition will turn on Q1 so that the collector terminal is shorted to the emitter and, therefore, to ground (low level). This low voltage is also on the base of Q2 and turns Q2 off. With Q2 off, there will be insufficient base current into Q3, turning it off also. Q2 leakage current is shunted to ground with a resistor to prevent the partial turning on of Q3. The collector voltage of

Q2 is pulled to a high potential with another resistor and, as a result, turns on Q4 so that it appears as a short in the top part of the totem pole. The saturated Q4 transistor provides a low resistance path from V_{CC} to the output, producing a high output result for a low input into the inverter.

A TTL NAND gate is very similar to the inverter circuit, with the exception that the input coupling transistor Q1 is constructed with multiple emitter-base junctions and each input to the NAND is connected to a separate emitter terminal. Any of the transistor's multiple emitters can be used to turn on Q1. The TTL NAND gate thus functions in the same manner as the inverter in that if any of the NAND gate inputs are low, the same circuit action will take place as with a low input to the inverter. Therefore, any time a low input is applied to the NAND gate it will produce a high ouput. Only if all of the NAND gate inputs are simultaneously high will it then produce the same circuit action as the inverter with its single input high, and the resultant output will be low. Input coupling transistors with up to eight emitter-base junctions, and therefore, eight input NAND gates, are constructed.

Storage time (the time it takes for the transistor to come out of saturation) is a major factor of propagation delay for saturated BJT transistors. A long storage time limits the switching speed of a standard TTL circuit. The propagation delay can be decreased and, therefore, the switching speed can be increased, by placing a Schottky diode between the base and collector of each transistor that might saturate. The resulting Schottky-clamped transistors do not go into saturation (effectively eliminating storage time) since the diode shunts current from the base into the collector before the transistor can achieve saturation. Today, digital circuit designs implemented with TTL logic almost exclusively use one of the Schottky subfamilies to take advantage of the significant improvement in switching speed.

CMOS Logic Family

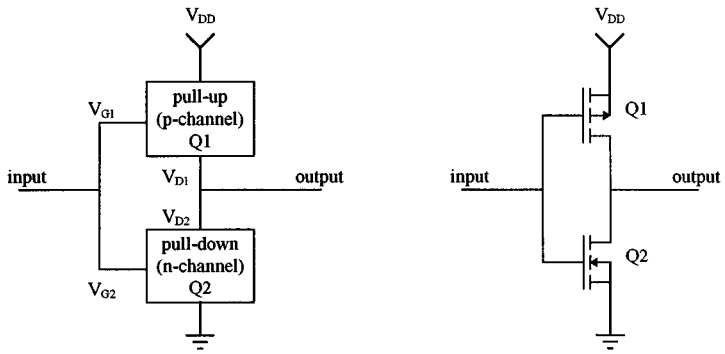
The active switching element used in all CMOS family circuits is the metal-oxide semiconductor field-effect transistor (MOSFET). CMOS stands for complementary MOS transistors and refers to the use of both types of MOSFET transistors, *n*-channel and *p*-channel, in the design of this type of switching circuit. While the physical construction and the internal physics of a MOSFET are quite different from that of the BJT, the circuit switching action of the two transistor types is quite similar. The MOSFET switch is essentially turned off and has a very high channel resistance by applying the same potential to the gate terminal as the source. An *n*-channel MOSFET is turned on and has a very low channel resistance when a high voltage with respect to the source is applied to the gate. A *p*-channel MOSFET operates in the same fashion but with opposite polarities; the gate must be more negative than the source to turn on the transistor.

A block diagram and schematic for a CMOS inverter circuit is shown in Fig. 79.2. Note that it is a simpler and much more compact circuit design than that for the TTL inverter. That fact is a major reason why MOSFET integrated circuits have a much higher circuit density than BJT integrated circuits and is one advantage that MOSFET ICs have over BJT ICs. As a result, CMOS is used in all levels of integration, from SSI through VLSI (very large scale integration).

When a high logic level is applied to the inverter's input, the *p*-channel MOSFET Q1 will be turned off and the *n*-channel MOSFET Q2 will be turned on. This will cause the output to be shorted to ground through the low resistance path of Q2's channel. The turned off Q1 has a very high channel resistance and acts nearly like an open.

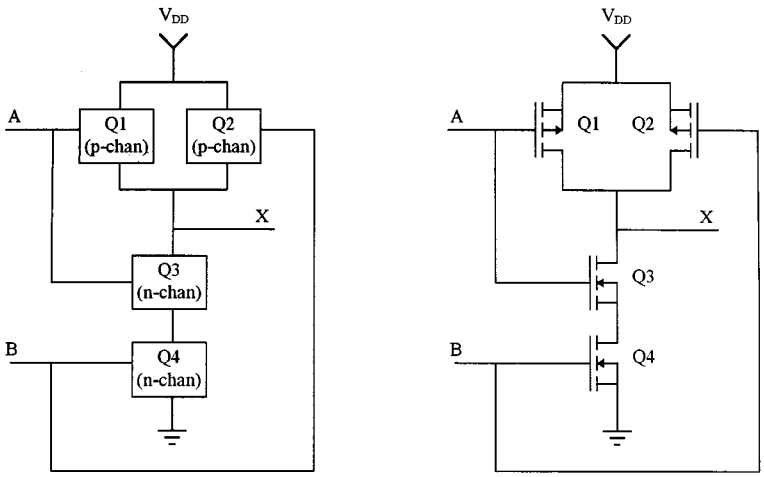
When a low logic level is applied to the inverter's input, the *p*-channel MOSFET Q1 will be turned on and the *n*-channel MOSFET Q2 will be turned off. This will cause the output to be shorted to V_{DD} through the low resistance path of Q1's channel. The turned off Q2 has a very high channel resistance and acts nearly like an open.

CMOS NAND gates are constructed by paralleling *p*-channel MOSFETs, one for each input, and putting in series an *n*-channel MOSFET for each input, as shown in the block diagram and schematic of Fig. 79.3. The NAND gate will produce a low output only when both Q3 and Q4 are turned on, creating a low resistance path from the output to ground through the two series channels. This can be accomplished by having a high on both input A and input B. This input condition will also turn off Q1 and Q2 . If either input A or input B or both is low, the respective parallel MOSFET will be turned on, providing a low resistance path for the output to V_{DD}. This will also turn off at least one of the series MOSFETs, resulting in a high resistance path for the output to ground.



input	Q1	V_{D1}	Q2	V_{D2}	output
hi	off	open	on	low	low
low	on	hi	off	open	hi

FIGURE 79.2 CMOS inverter circuit block diagram and operation.



inputs		transistors				output
A	B	Q1	Q2	Q3	Q4	X
low	low	on	on	off	off	hi
low	hi	on	off	off	on	hi
hi	low	off	on	on	off	hi
hi	hi	off	off	on	on	low

FIGURE 79.3 CMOS two-input NAND circuit block diagram and operation.

ECL Logic Family

ECL is a higher-speed logic family. While it does not offer as large a variety of IC chips as are available in the TTL family, it is quite popular for logic applications requiring high-speed switching.

The active switching element used in the ECL family circuits is also the *npn* BJT. Unlike the TTL family, however, which switches the transistors into saturation when turning them on, ECL switching is designed to prevent driving the transistors into saturation. Whenever bipolar transistors are driven into saturation, their switching speed will be limited by the charge carrier storage delay, a transistor operational characteristic. Thus, the switching speed of ECL circuits will be significantly higher than for TTL circuits. ECL operation is based

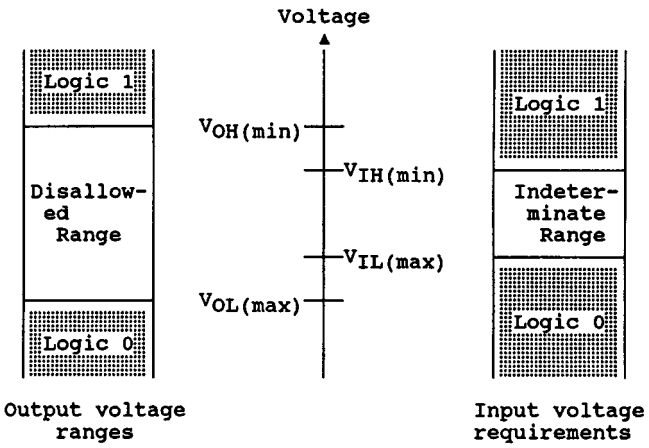


FIGURE 79.4 Switching device logic levels.

TABLE 79.3 Logic Signal Voltage Parameters for Selected Logic Subfamilies (in Volts)

Subfamily	$V_{OH(min)}$	$V_{OL(max)}$	$V_{IH(min)}$	$V_{IL(max)}$
74xx	2.4	0.4	2.0	0.8
74LSxx	2.7	0.5	2.0	0.8
74ASxx	2.5	0.5	2.0	0.8
74ALSxx	2.5	0.4	2.0	0.8
74Fxx	2.5	0.5	2.0	0.8
74HCxx	4.9	0.1	3.15	0.9
74HCTxx	4.9	0.1	2.0	0.8
74ACxx	3.8	0.4	3.15	1.35
74ACTxx	3.8	0.4	2.0	0.8
74AHCxx	4.5	0.1	3.85	1.65
74AHCTxx	3.65	0.1	2.0	0.8
10xxx	–0.96	–1.65	–1.105	–1.475
10Hxxx	–0.98	–1.63	–1.13	–1.48

on switching a fixed amount of bias current that is less than the saturation amount between two different transistors. The basic circuit found in the ECL family is the differential amplifier. One side of the differential amplifier is controlled by a bias circuit and the other is controlled by the logic inputs to the gate. This logic family is also referred to as current-mode logic (CML) because of its current switching operation.

Logic Family Circuit Parameters

Digital circuits and systems operate with only two states, logic 1 and 0, usually represented by two different voltage levels, a *high* and a *low*. The two logic levels actually consist of a range of values with the numerical quantities dependent upon the specific family that is used. Minimum high logic levels and maximum low logic levels are established by specifications for each family. Minimum device output levels for a logic high are called $V_{OH(min)}$ and minimum input levels are called $V_{IH(min)}$. The abbreviations for maximum output and input low logic levels are $V_{OL(max)}$ and $V_{IL(max)}$, respectively. Figure 79.4 shows the relationships between these parameters. Logic voltage level parameters are illustrated for selected prominent logic subfamilies in Table 79.3. As seen in this illustration, there are many operational incompatibilities between major logic family types.

Noise margin is a quantitative measure of a device's **noise immunity.** High-level noise margin (V_{NH}) and low-level noise margin (V_{NL}) are defined in Eqs. (79.1) and (79.2).

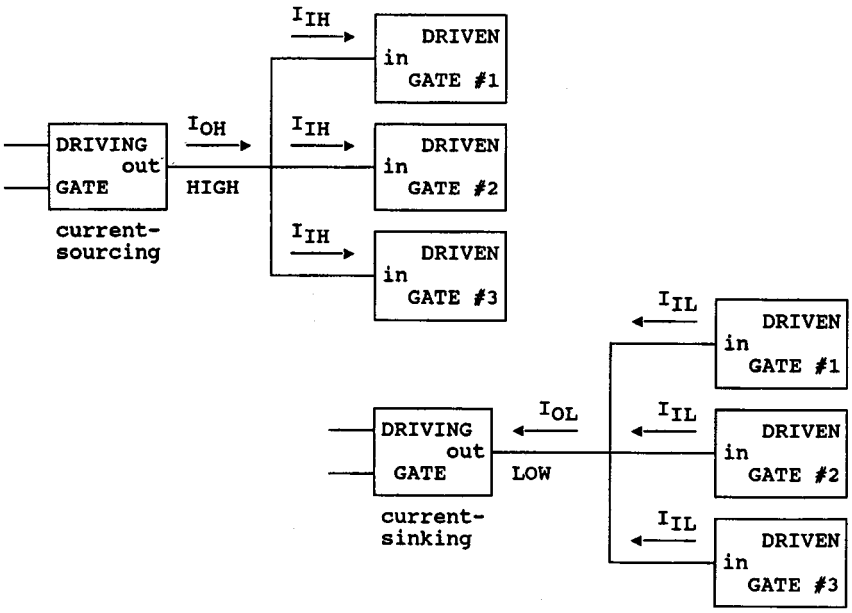


FIGURE 79.5 Current loading of driving gates.

TABLE 79.4 Worst Case Current Parameters for Selected Logic Subfamilies

Subfamily	$I_{OH(max)}$	$I_{OL(max)}$	$I_{IH(max)}$	$I_{IL(max)}$
74xx	–400 μA	16 mA	40 μA	–1.6 μA
74LSxx	–400 μA	8 mA	20 μA	–400 μA
74ASxx	–2 mA	20 mA	200 μA	–2 mA
74ALSxx	–400 μA	8 mA	20 μA	–100 μA
74Fxx	–1 mA	20 mA	20 μA	–0.6 mA
74HCxx	–4 mA	4 mA	1 μA	–1 μA
74HCTxx	–4 mA	4 mA	1 μA	–1 μA
74ACxx	–24 mA	24 mA	1 μA	–1 μA
74ACTxx	–24 mA	24 mA	1 μA	–1 μA
74AHCxx	–8 mA	8 mA	1 μA	–1 μA
74AHCTxx	–8 mA	8 mA	1 μA	–1 μA
10xxx	50 mA	–50 mA	–265 μA	500 nA
10Hxxx	50 mA	–50 mA	–265 μA	500 nA

$$V_{NH} = V_{OH(min)} - V_{IH(min)} \tag{79.1}$$

$$V_{NL} = V_{IL(max)} - V_{OL(max)} \tag{79.2}$$

Using the logic voltage values given in Table 79.3 for the selected subfamilies reveals that highest noise immunity is obtained with logic devices in the CMOS family, while lowest noise immunity is endemic to the ECL family.

Switching circuit outputs are loaded by the inputs of the devices that they are driving, as illustrated in Fig. 79.5. Worst case input loading current levels and output driving current capabilities are listed in Table 79.4 for various logic subfamilies. The fan-out of a driving device is the ratio between its output current capabilities at each logic level and the corresponding gate input current loading value. Switching circuits based on bipolar transistors have fan-out limited primarily by the current-sinking and current-sourcing capabilities of the driving device.

TABLE 79.5 Speed-Power Comparison for Selected Logic Subfamilies

Subfamily	Propagation Delay Time, ns (ave.)	Static Power Dissipation, mW (per gate)	Speed-Power Product, pJ
74xx	10	10	100
74LSxx	9.5	2	19
74ASxx	1.5	2	13
74ALSxx	4	1.2	5
74Fxx	3	6	18
74HCxx	8	0.003	24×10^{-3}
74HCTxx	14	0.003	42×10^{-3}
74ACxx	5	0.010	50×10^{-3}
74ACTxx	5	0.010	50×10^{-3}
74AHCxx	5.5	0.003	16×10^{-3}
74AHCTxx	5	0.003	14×10^{-3}
10xxx	2	25	50
10Hxxx	1	25	25

CMOS switching circuits are limited by the charging and discharging times associated with the output resistance of the driving gate and the input capacitance of the load gates. Thus, CMOS fan-out depends on the frequency of switching. With fewer (capacitive) loading inputs to drive, the maximum switching frequency of CMOS devices will increase.

The switching speed of logic devices is dependent on the device's **propagation delay time.** The propagation delay of a logic device limits the frequency at which it can be operated. There are two propagation delay times specified for logic gates: t_{PHL}, delay time for the output to change from high to low, and t_{PLH}, delay time for the output to change from low to high. Average typical propagation delay times for a single gate are listed for several logic subfamilies in Table 79.5. The ECL family has the fastest switching speed.

The amount of power required by an IC is normally specified in terms of the amount of current I_{CC} (TTL family), I_{DD} (CMOS family), or I_{EE} (ECL family) drawn from the power supply. For complex IC devices, the required supply current is given under specified test conditions. For TTL chips containing simple gates, the average power dissipation $P_{D(ave)}$ is normally calculated from two measurements, I_{CCH} (when all gate outputs are high) and I_{CCL} (when all gate outputs are low). Table 79.5 compares the static power dissipation of several logic subfamilies. The ECL family has the highest power dissipation, while the lowest is attained with the CMOS family. It should be noted that power dissipation for the CMOS family is directly proportional to the gate input signal frequency. For example, one would typically find that the power dissipation for a CMOS logic circuit would increase by a factor of 100 if the input signal frequency is increased from 1 kHz to 100 kHz.

The **speed-power product** is a relative figure of merit that is calculated by the formula given in Eq. (79.3). This performance measurement is normally expressed in picojoules (pJ).

$$\text{Speed-power product} = (t_{PHL} + t_{PLH})/2 \times P_{D(ave)} \tag{79.3}$$

A low value of speed-power product is desirable to implement high-speed (and, therefore, low propagation delay time) switching devices that consume low amounts of power. Because of the nature of transistor switching circuits, it is difficult to attain high-speed switching with low power dissipation. The continued development of new IC logic families and subfamilies is largely due to the trade-offs between these two device switching parameters. The speed-power product for various subfamilies is also compared in Table 79.5.

Interfacing Between Logic Families

The interconnection of logic chips requires that input and output specifications be satisfied. Figure 79.6 illustrates voltage and current requirements. The driving chip's $V_{OH(min)}$ must be greater than the driven circuit's $V_{IH(min)}$, and the driver's $V_{OL(max)}$ must be less than $V_{IL(max)}$ for the loading circuit. Voltage level shifters must be

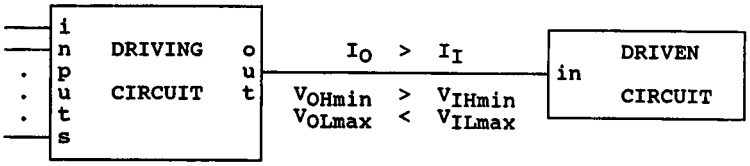


FIGURE 79.6 Circuit interfacing requirements.

used to interface the circuits together if these voltage requirements are not met. Of course, a driving circuit's output must not exceed the maximum and minimum allowable input voltages for the driven circuit. Also, the current sinking and sourcing ability of the driver circuit's output must be greater than the total current requirements for the loading circuit. Buffer gates or stages must be used if current requirements are not satisfied. All chips within a single logic family are designed to be compatible with other chips in the same family. Mixing chips from multiple subfamilies together within a single digital circuit can have adverse effects on the overall circuit's switching speed and noise immunity.

Defining Terms

Fan-out: The specification used to identify the limit to the number of loading inputs that can be reliably driven by a driving device's output.

Logic level: The high or low value of a voltage variable that is assigned to be a 1 or a 0 state.

Noise immunity: A logic device's ability to tolerate input voltage fluctuation caused by noise without changing its output state.

Propagation delay time: The time delay from when the input logic level to a device is changed until the resultant output change is produced by that device.

Speed-power product: An overall performance measurement that is used to compare the various logic families and subfamilies.

Truth table: A listing of the relationship of a circuit's output that is produced for various combinations of logic levels at the inputs.

Related Topic

25.3 Application-Specific Integrated Circuits

References

A. P. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Design,* Boston: Kluwer Academic, 1995.

D. J. Comer, *Digital Logic and State Machine Design,* 2nd ed., Philadelphia: Saunders College Publishing, 1990.

S. H. K. Embabi, A. Bellaouar, and M. I. Elmasry, *Digital BiCMOS Integrated Circuit Design,* Boston: Kluwer Academic, 1993.

T. L. Floyd, *Digital Fundamentals,* 5th ed., Columbus, Ohio: Merrill Publishing Company, 1994.

K. Gopalan, *Introduction to Digital Microelectronic Circuits,* Chicago: Irwin, 1996.

J. D. Greenfield, *Practical Digital Design Using ICs,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1994.

R. J. Prestopnik, *Digital Electronics: Concepts and Applications for Digital Design,* Philadelphia: Saunders College Publishing, 1990.

R. S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990.

M. Shoji, *Theory of CMOS Digital Circuits and Circuit Failures,* Princeton, N.J.: Princeton University Press, 1992.

R. J. Tocci, *Digital Systems: Principles and Applications,* 6th ed., Englewood Cliffs, N.J.: Prentice-Hall, 1995.

S. H. Unger, *The Essence of Logic Circuits,* 2nd ed., New York: IEEE Press, 1996.

J. F. Wakerly, *Digital Design: Principles and Practices,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1994.

Further Information

Data Books and Device Index:

D. M. Howell, Ed. *IC Master,* Garden City, NY: Hearst Business Communications, annual.

Engineering Staff, *Advanced BiCMOS Technology Data Book,* Dallas: Texas Instruments, 1994.

Engineering Staff, *Advanced High-Speed CMOS Logic Data Book,* Dallas: Texas Instruments, 1996.

Engineering Staff, *ALS/AS Logic Data Book,* Dallas: Texas Instruments, 1995.

Engineering Staff, *ECLinPS Data,* Phoenix: Motorola, 1995.

Engineering Staff, *FACT Advanced CMOS Logic Databook,* Santa Clara, Calif: National Semiconductor Corporation, 1993.

Engineering Staff, *FACT Data,* Phoenix: Motorola, 1996.

Engineering Staff, *FACT & LS TTL Data,* Phoenix: Motorola, 1992.

Engineering Staff, *Low-Voltage Logic Data Book,* Dallas: Texas Instruments, 1996.

Engineering Staff, *MECL Data,* Phoenix: Motorola, 1993.

Journals and Trade Magazines:

EDN, Highlights Ranch, Colo.: Cahners Publishing.

Electronic Design, Cleveland, Ohio: Penton Publishing.

Electronic Engineering Times, Manhasset, N.Y.: CMP Publications.

IEEE Journal of Solid-State Circuits, New York: Institute of Electrical and Electronic Engineers.

IEEE Transactions on Circuits and Systems, Part I: Fundamental Theory and Applications, New York: Institute of Electrical and Electronic Engineers.

Internet Addresses for Digital Device Data Sheets:

Motorola, Inc.	http://Design-net.com
National Semiconductor Corp.	http://www.national.com/design/index.html
Texas Instruments, Inc.	http://www.ti.com/sc/docs/schome.htm

79.2 Logic Gates (IC)[1]

Peter Graham

This section introduces and analyzes the electronic circuit realizations of the basic gates of the three technologies: transistor-transistor logic (TTL), emitter-coupled logic (ECL), and complementary metal-oxide semiconductor (CMOS) logic. These circuits are commercially available on small-scale integration chips and are also the building blocks for more elaborate logic systems. The three technologies are compared with regard to speed, power consumption, and noise immunity, and parameters are defined which facilitate these comparisons. Also included are recommendations which are useful in choosing and using these technologies.

Gate Specification Parameters

Theoretically almost any logic device or system could be constructed by wiring together the appropriate configuration of the basic gates of the selected technology. In practice, however, the gates are interconnected during the fabrication process to produce a desired system on a single chip. The circuit complexity of a given chip is described by one of the following four rather broad classifications:

- **Small-Scale Integration (SSI).** The inputs and outputs of every gate are available for external connection at the chip pins (with the exception that exclusive OR and AND-OR gates are considered SSI).
- **Medium-Scale Integration (MSI).** Several gates are interconnected to perform somewhat more elaborate logic functions such as flip-flops, counters, multiplexers, etc.

[1]Based on P. Graham, "Gates," in *Handbook of Modern Electronics and Electrical Engineering,* C. Belove, Ed., New York: Wiley-Interscience, 1986, pp. 864–876. With permission.

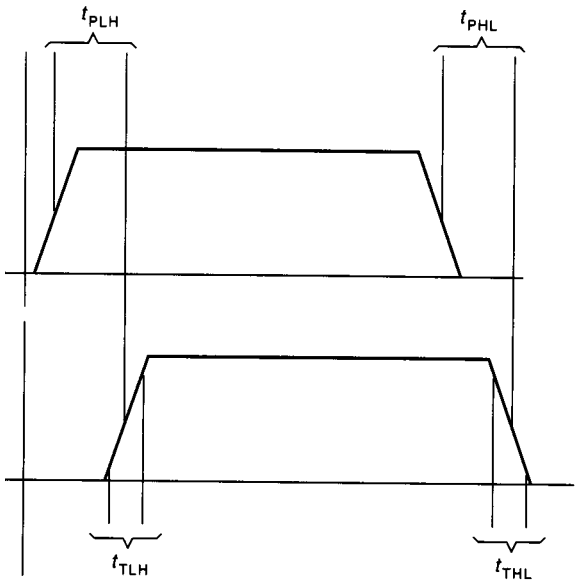


FIGURE 79.7 Definitions of switching times.

- **Large-Scale Integration (LSI).** Several of the more elaborate circuits associated with MSI are interconnected within the integrated circuit to form a logic system on a single chip. Chips such as calculators, digital clocks, and small microprocessors are examples of LSI.
- **Very-Large-Scale Integration (VLSI).** This designation is usually reserved for chips having a very high density, 1000 or more gates per chip. These include the large single-chip memories, gate arrays, and microcomputers.

Specifications of logic speed require definitions of switching times. These definitions can be found in the introductory pages of most data manuals. Four of them pertain directly to gate circuits. These are (see also Fig. 79.7):

- **LOW-to-HIGH Propagation Delay Time (t_{PLH}).** The time between specified reference points on the input and output voltage waveforms when the output is changing from low to high.
- **HIGH-to-LOW Propagation Delay Tune (t_{PHL}).** The time between specified reference points on the input and output voltage waveforms when the output is changing from high to low.
- **Propagation Delay Time (t_{PD}).** *The average of the two propagation delay times:* $t_{PD} = (t_{PD} + t_{PHL})/2$.
- **LOW-to-HIGH Transition Time (t_{TLH}).** *The rise time between specified reference points on the LOW-to-HIGH shift of the output waveform.*
- **HIGH-to-LOW Transition Time (t_{THL}).** *The fall time between specified reference points on the HIGH-to-LOW shift of the output waveform. The reference points usually are 10 and 90% of the voltage level difference in each case.*

Power consumption, driving capability, and effective loading of gates are defined in terms of currents.

- **Supply Current, Outputs High (I_{xxH}).** The current delivered to the chip by the power supply when all outputs are open and at the logical 1 level. The xx subscript depends on the technology.
- **Supply Current, Outputs Low (I_{xxL}).** The current delivered to the chip by the supply when all outputs are open and at the logical 0 level.
- **Supply Current, Worst Case (I_{xx}).** When the output level is unspecified, the input conditions are assumed to correspond to maximum supply current.

- **Input HIGH Current (I_{IH}).** The current flowing into an input when the specified HIGH voltage is applied.
- **Input LOW Current (I_{IL}).** The current flowing into an input when the specified LOW voltage is applied.
- **Output HIGH Current (I_{OH}).** The current flowing into the output when it is in the HIGH state. I_{OHmax} is the largest I_{OH} for which $V_{OH} \geq V_{OHmin}$ is guaranteed.
- **Output LOW Current (I_{OL}).** The current flowing into the output when it is in the LOW state. I_{OLmax} is the largest I_{OL} for which $V_{OL} \geq V_{OLmax}$ is guaranteed.

The most important voltage definitions are concerned with establishing ranges on the logical 1 (HIGH) and logical 0 (LOW) voltage levels.

- **Minimum High-Level Input Voltage (V_{IHmin}).** The least positive value of input voltage guaranteed to result in the output voltage level specified for a logical 1 input.
- **Maximum Low-Level Input Voltage (V_{ILmax}).** The most positive value of input voltage guaranteed to result in the output voltage level specified for a logical 0 input.
- **Minimum High-Level Output Voltage (V_{OHmin}).** The guaranteed least positive output voltage when the input is properly driven to produce a logical 1 at the output.
- **Maximum Low-Level Output Voltage (V_{OLmax}).** The guaranteed most positive output voltage when the input is properly driven to produce a logical 0 at the output.
- **Noise Margins.** $NM_H = V_{OHmin} - V_{IHmin}$ is how much larger the guaranteed least positive output logical 1 level is than the least positive input level that will be interpreted as a logical 1. It represents how large a negative-going glitch on an input 1 can be before it affects the output of the driven device. Similarly, $NM_L = V_{ILmax} - V_{OLmax}$ is the amplitude of the largest positive- going glitch on an input 0 that will not affect the output of the driven device.

Finally, three important definitions are associated with specifying the load that can be driven by a gate. Since in most cases the load on a gate output will be the sum of inputs of other gates, the first definition characterizes the relative current requirements of gate inputs.

- **Load Factor (LF).** Each logic family has a reference gate, each of whose inputs is defined to be a unit load in both the HIGH and the LOW conditions. The respective ratios of the input currents I_{IH} and I_{IL} of a given input to the corresponding I_{IH} and I_{IL} of the reference gate define the HIGH and LOW load factors of that input.
- **Drive Factor (DF).** A device output has drive factors for both the HIGH and the LOW output conditions. These factors are defined as the respective ratios of I_{OHmax} and I_{OLmax} of the gate to I_{OHmax} and I_{OLmax} of the reference gate.
- **Fan-Out.** For a given gate the fan-out is defined as the maximum number of inputs of the same type of gate that can be properly driven by that gate output. When gates of different load and drive factors are interconnected, fan-out must be adjusted accordingly.

Bipolar Transistor Gates

A logic circuit using bipolar junction transistors (BJTs) can be classified either as saturated or as nonsaturated logic. A saturated logic circuit contains at least one BJT that is saturated in one of the stable modes of the circuit. In nonsaturated logic circuits none of the transistors is allowed to saturate. Since bringing a BJT out of saturation requires a few additional nanoseconds (called the storage time), nonsaturated logic is faster. The fastest circuits available at this time are emitter-coupled logic (ECL), with transistor-transistor logic (TTL) having Schottky diodes connected to prevent the transistors from saturating (Schottky TTL) being a fairly close second. Both of these families are nonsaturated logic. All TTL families other than Schottky are saturated logic.

Transistor-Transistor Logic

TTL evolved from resistor-transistor logic (RTL) through the intermediate step of diode-transistor logic (DTL). All three families are catalogued in data books published in 1968, but of the three only TTL is still available.

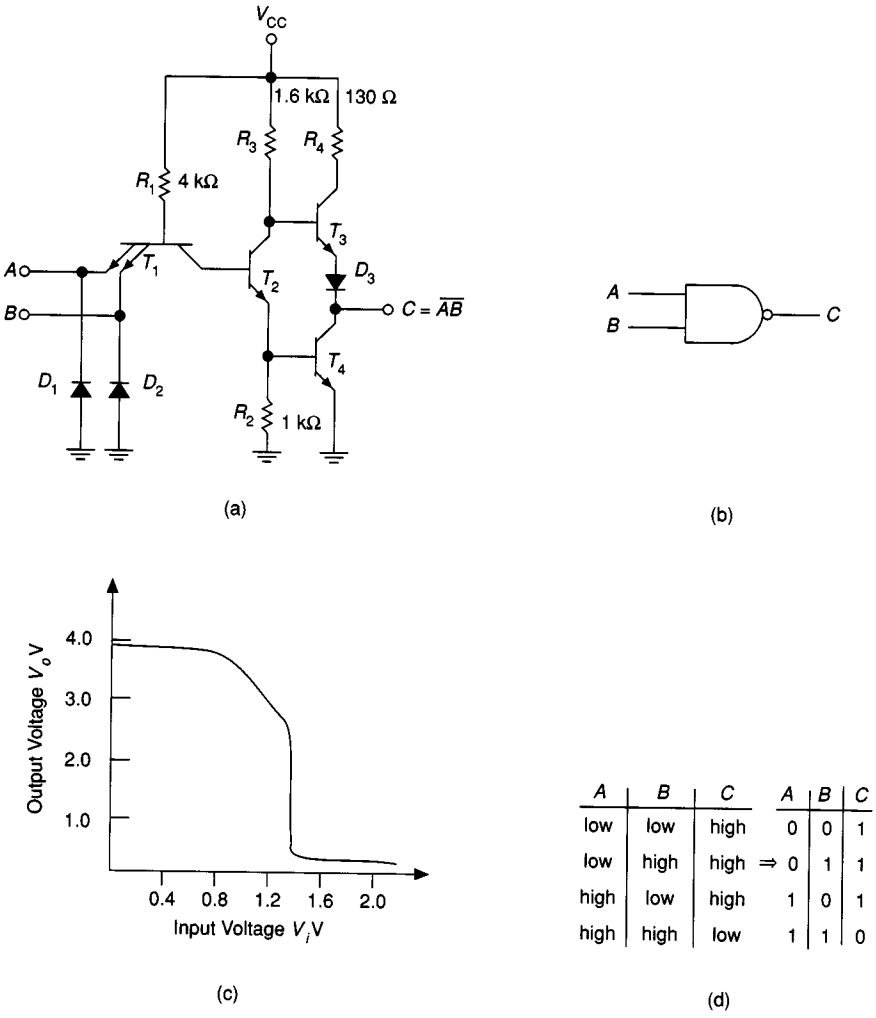


FIGURE 79.8 Two-input transistor-transistor logic (TTL) NAND gate type 7400: (a) circuit, (b) symbol, (c) voltage transfer characteristic (V_i to both inputs), (d) truth table.

The basic circuit of the standard TTL family is typified by the two-input NAND gate shown in Fig. 79.8(a). To estimate the operating levels of voltage and current in this circuit, assume that any transistor in saturation has $V_{CE} = 0.2$ and $V_{BE} = 0.75$ V. Let drops across conducting diodes also be 0.75 V and transistor current gains (when nonsaturated) be about 50. As a starting point, let the voltage levels at both inputs A and B be high enough that T_1 operates in the reversed mode. In this case the emitter currents of T_1 are negligible, and the current into the base of T_1 goes out the collector to become the base current of T_2. This current is readily calculated by observing that the base of T_1 is at $3 \times 0.75 = 2.25$ V so there is a 2.75-V drop across the 4-kΩ resistor. Thus $I_{B1} = I_{B2} = 0.7$ mA, and it follows that T_2 is saturated. With T_2 saturated, the base of T_3 is at $V_C + V_{BE4} = 0.95$ V. If T_4 is also saturated, the emitter of T_3 will be at $V_{D3} + V_{CE4} = 0.95$ V, and T_3 will be cut off. The voltage across the 1.6-kΩ resistor is $5 - 0.95 = 4.05$ V, so the collector current of T_2 is about 2.5 mA. This means the emitter current of T_2 is 3.2 mA. Of this, 0.75 mA goes through the 1-kΩ resistor, leaving 2.45 mA as the base current of T_4. Since the current gain of T_4 is about 50, it will be well into saturation for any collector current less than 100 mA, and the output at C is a logic 0. The corresponding minimum voltage levels required at the inputs are estimated from $V_{BE4} + V_{ECI}$, or about 1.7 V.

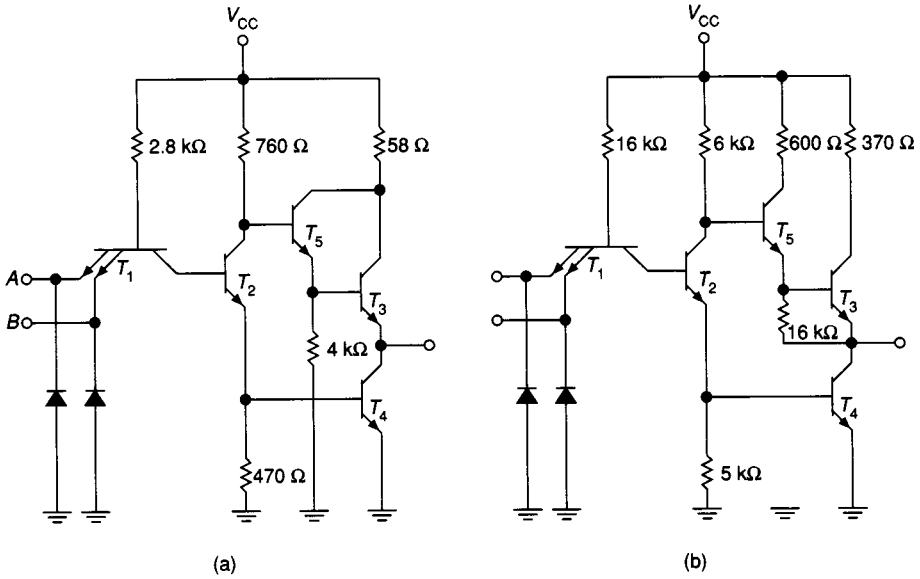


FIGURE 79.9 Modified transistor-transistor logic (TTL) two-input NAND states: (a) type 74Hxx, (b) type 74L00.

Now let either or both of the inputs be dropped to 0.2 V. T_1 is then biased to saturation in the normal mode, so the collector current of T_1 extracts the charge from the base region of T_2. With T_2 cut off, the base of T_4 is at 0 V and T_4 is cut off. T_3 will be biased by the current through the 1.6-kΩ resistor (R_3) to a degree regulated by the current demand at the output C. The drop across R_3 is quite small for light loads, so the output level at C will be $V_{CC} - V_{BE3} - V_{D3}$, which will be about 3.5 V corresponding to the logical 1.

The operation is summarized in the truth table in Fig. 79.8(d), identifying the circuit as a two-input NAND gate. The derivation of the input-output voltage transfer characteristic [Fig. 79.8(c)], where V_i is applied to inputs A and B simultaneously, can be found in most digital circuit textbooks. The sloping portion of the characteristic between $V_i = 0.55$ and 1.2 V corresponds to T_2 passing through the active region in going from cutoff to saturation.

Diodes D_1 and D_2 are present to damp out "ringing" that can occur, for example, when fast voltage level shifts are propagated down an appreciable length (20 cm or more) of microstripline formed by printed circuit board interconnections. Negative overshoots are clamped to the 0.7 V across the diode.

The series combination of the 130-Ω resistor, T_3, D_3, and T_4 in the circuit of Fig. 79.8(a), forming what is called the totem-pole output circuit, provides a low impedance drive in both the source (output $C = 1$) and sink (output $C = 0$) modes and contributes significantly to the relatively high speed of TTL. The available source and sink currents, which are well above the normal requirements for steady state, come into play during the charging and discharging of capacitive loads. Ideally T_3 should have a very large current gain and the 130-Ω resistor should be reduced to 0. The latter, however, would cause a short-circuit load current which would overheat T_3, since T_3 would be unable to saturate. All TTL families other than the standard shown in Fig. 79.8(a) use some form of Darlington connection for T_3, providing increased current gain and eliminating the need for diode D_3. The drop across D_3 is replaced by the base emitter voltage of the added transistor T_5. This connection appears in Fig. 79.9(a), an example of the 74Hxx series of TTL gates that increases speed at the expense of increased power consumption, and in Fig. 79.9(b), a gate from the 74Lxx series that sacrifices speed to lower power dissipation.

A number of TTL logic function implementations are available with open collector outputs. For example, the 7403 two-input NAND gate shown in Fig. 79.10 is the open collector version of Fig. 79.8(a). The open collector output has some useful applications. The current in an external load connected between the open collector and V_{CC} can be switched on and off in response to the input combinations. This load, for example,

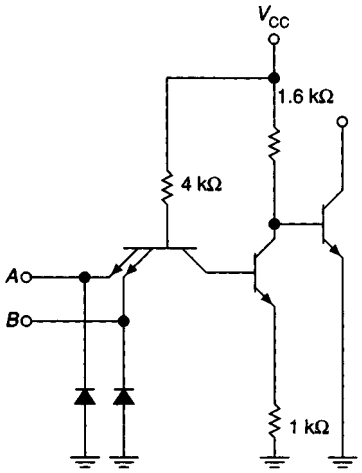


FIGURE 79.10 Open collector two-input NAND gate.

might be a relay, an indicator light, or an LED display. Also, two or more open collector gates can share a common load, resulting in the anding together of the individual gate functions. This is called a "wired-AND connection." In any application, there must be some form of load or the device will not function. There is a lower limit to the resistance of this load which is determined by the current rating of the open collector transistor. For wired-AND applications the resistance range depends on how many outputs are being wired and on the load being driven by the wired outputs. Formulas are given in the data books. Since the open collector configuration does not have the speed enhancement associated with an active pull-up, the low to high propagation delay (t_{PLH}) is about double that of the totem-pole output. It should be observed that totem-pole outputs should not be wired, since excessive currents in the active pull-up circuit could result.

Nonsaturated TTL. Two TTL families, the Schottky (74Sxx) and the low-power Schottky (74LSxx), can be classified as nonsaturating logic. The transistors in these circuits are kept out of saturation by the connection of Schottky diodes, with the anode to the base and the cathode to the collector.

Schottky diodes are formed from junctions of metal and an *n*-type semiconductor, the metal fulfilling the role of the *p*-region. Since there are thus no minority carriers in the region of the forward-biased junction, the storage time required to bring a *pn* junction out of saturation is eliminated. The forward-biased drop across a Schottky diode is around 0.3 V. This clamps the collector at 0.3 V less than the base, thus maintaining V_{CE} above the 0.3-V saturation threshold. Circuits for the two-input NAND gates 74LS00 and 74S00 are given in Fig. 79.11(a) and (b). The special transistor symbol is a short-form notation indicating the presence of the Schottky diode, as illustrated in Fig. 79.11(c).

Note that both of these circuits have an active pull-down transistor T_6 replacing the pull-down resistance connected to the emitter of T_2 in Fig. 79.9. The addition of T_6 decreases the turn-on and turn-off times of T_4. In addition, the transfer characteristic for these devices is improved by the squaring off of the sloping region between $V_i = 0.55$ and 1.2 V [see Fig. 79.8(c)]. This happens because T_2 cannot become active until T_6 turns on, which requires at least 1.2 V at the input.

The diode AND circuit of the 74LS00 in place of the multi-emitter transistor will permit maximum input levels substantially higher than the 5.5-V limit set for all other TTL families. Input leakage currents for 74LSxx are specified at $V_i = 10$ V, and input voltage levels up to 15 V are allowed. The 74LSxx has the additional feature of the Schottky diode D_1 in series with the 100-Ω output resistor. This allows the output to be pulled up to 10 V without causing a reverse breakdown of T_5. The relative characteristics of the several versions of the TTL two-input NAND gate are compared in Table 79.6. The 74F00 represents one of the new technologies that have introduced improved Schottky TTL in recent years.

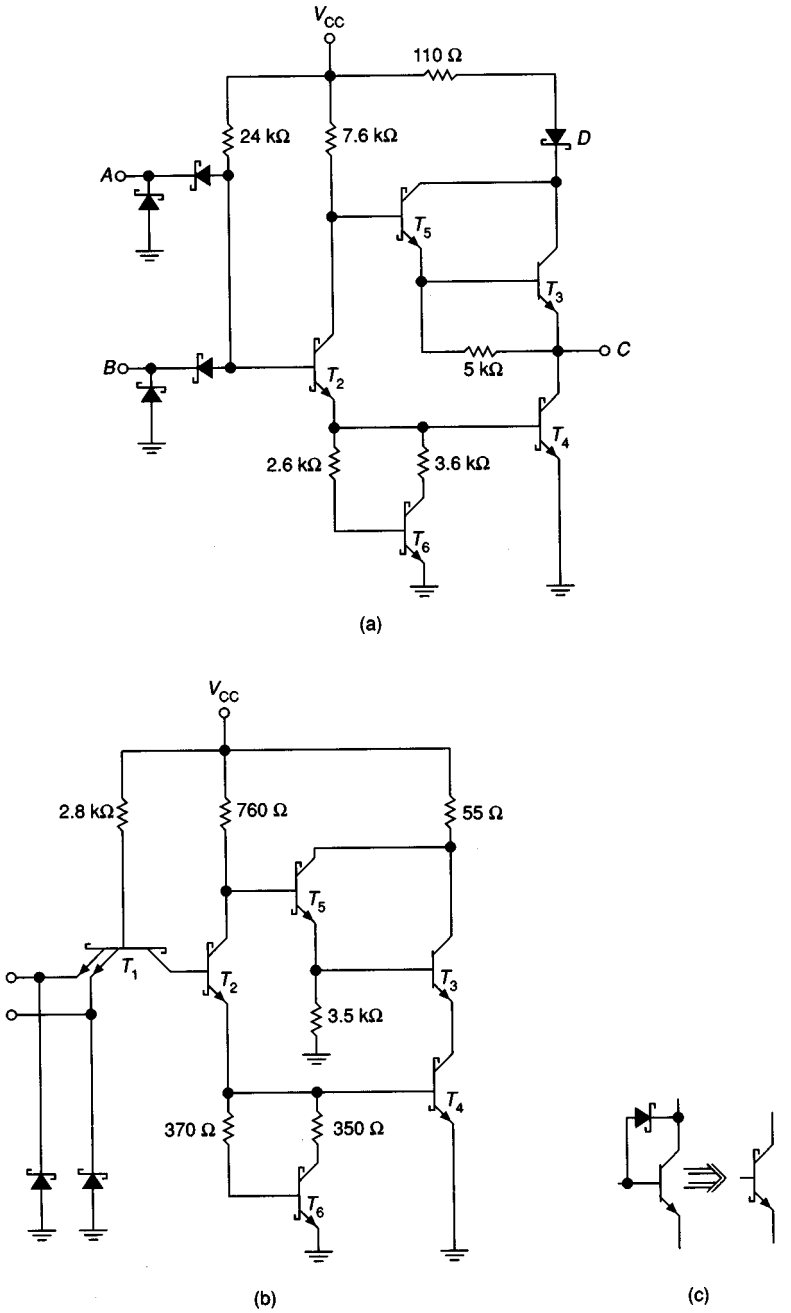


FIGURE 79.11 Transistor-transistor logic (TTL) nonsaturated logic. (a) Type 74LS00 two-input NAND gate, (b) type 74S00 two-input NAND gate, (c) significance of the Schottky transistor symbol.

TTL Design Considerations. Before undertaking construction of a logic system, the wise designer consults the information and recommendations provided in the data books of most manufacturers. Some of the more significant tips are provided here for easy reference.

1. **Power supply, decoupling, and grounding.** The power supply voltage should be 5 V with less than 5% ripple factor and better than 5% regulation. When packages on the same printed circuit board are

TABLE 79.6 Comparison of TTL Two-Input NANDGates

TTL Type	Supply Current I_{CCH}[a] (mA)	Supply Current I_{CCL} (mA)	Propagation Delay Time t_{PLH} (ns)	Propagation Delay Time t_{PHL} (ns)	Noise Margins NM_H (V)	Noise Margins NM_L (V)	Load Factor, H/L	Drive Factor, H/L	Fan-out
74F00	2.8	10.2	2.9	2.6	0.7	0.3	0.5/0.375	25/12.5	33
74S00	10	20	3	3	0.7	0.3	1.25/1.25	25/12.5	10
74H00	10	26	5.9	6.2	0.4	0.4	1.25/1.25	12.5/12.5	10
74LS00	0.8	2.4	9	10	0.7	0.3	0.5/0.25	10/5	20
7400	4	12	11	7	0.4	0.4	1/1	20/10	10
74L00	0.44	1.16	31	31	0.4	0.5	0.24/0.1125	5/2.25	20

[a]See text for explanation of abbreviations.

supplied by a bus there should be a 0.05-μF decoupling capacitor between the bus and the ground for every five to ten packages. If a ground bus is used, it should be as wide as possible, and should surround all the packages on the board. Whenever possible, use a ground plane. If a long ground bus is used, both ends must be tied to the common system ground point.

2. **Unused gates and inputs.** If a gate on a package is not used, its inputs should be tied either high or low, whichever results in the least supply current. For example, the 7400 draws three times the current with the output low as with the output high, so the inputs of an unused 7400 gate should be grounded. An unused input of a gate, however, must be connected so as not to affect the function of the active inputs. For a 7400 NAND gate, such an input must either be tied high or paralleled with a used input. It must be recognized that paralleled inputs count as two when determining the fan-out. Inputs that are tied high can be connected either to V_{CC} through a 1-kΩ or more resistance (for protection from supply voltage surges) or to the output of an unused gate whose input will establish a permanent output high. Several inputs can share a common protective resistance. Unused inputs of low-power Schottky TTL can be tied directly to V_{CC}, since 74LSxx inputs tolerate up to 15 V without breakdown. If inputs of low-power Schottky are connected in parallel and driven as a single input, the switching speed is decreased, in contrast to the situation with other TTL families.
3. **Interconnection.** Use of line lengths of up to 10 in. (5 in. for 74S) requires no particular precautions, except that in some critical situations lines cannot run side by side for an appreciable distance without causing cross talk due to capacitive coupling between them. For transmission line connections, a gate should drive only one line, and a line should be terminated in only one gate input. If overshoots are a problem, a 25- to 50-Ω resistor should be used in series with the driving gate input and the receiving gate input should be pulled up to 5 V through a 1-kΩ resistor. Driving and receiving gates should have their own decoupling capacitors between the V_{CC} and ground pins. Parallel lines should have a grounded line separating them to avoid cross talk.
4. **Mixing TTL subfamilies.** Even synchronous sequential systems often have asynchronous features such as reset, preset, load, and so on. Mixing high-speed 74S TTL with lower speed TTL (74LS for example) in some applications can cause timing problems resulting in anomalous behavior. Such mixing is to be avoided, with rare exceptions which must be carefully analyzed.

Emitter-Coupled Logic

ECL is a nonsaturated logic family where saturation is avoided by operating the transistors in the common collector configuration. This feature, in combination with a smaller difference between the HIGH and LOW voltage levels (less than 1 V) than other logic families, makes ECL the fastest logic available at this time. The circuit diagram of a widely used version of the basic two-input ECL gate is given in Fig. 79.12. The power supply terminals V_{CC1}, V_{CC2}, V_{EE}, and V_{TT} are available for flexibility in biasing. In normal operation, V_{CC1} and V_{CC2} are connected to a common ground, V_{EE} is biased to –5.2 V, and V_{TT} is biased to –2 V. With these values the nominal voltage for the logical 0 and 1 are, respectively, –1.75 and –0.9 V. Operation with the V_{CC} terminals grounded maximizes the immunity from noise interference.

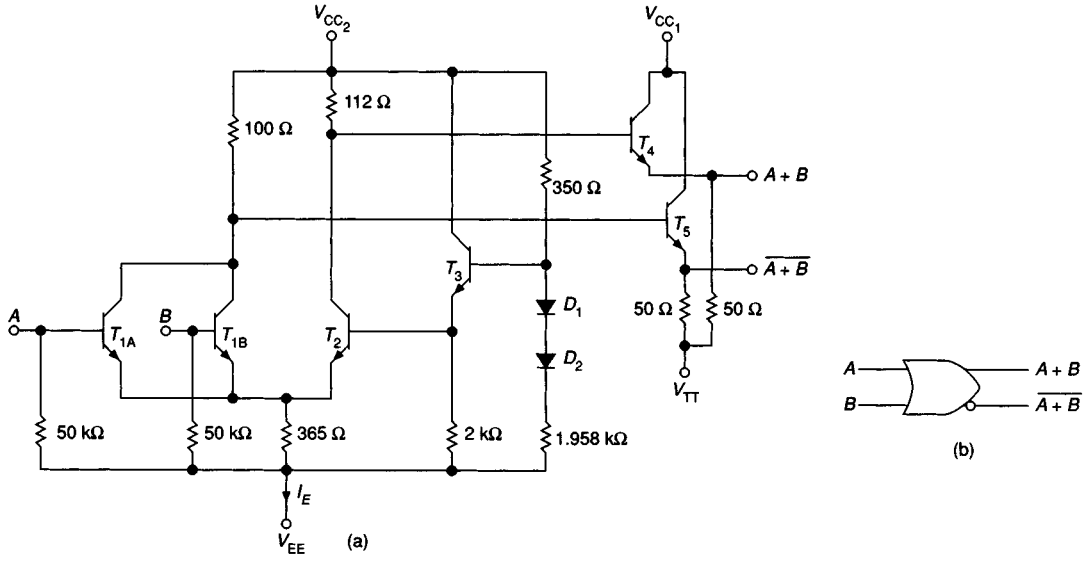


FIGURE 79.12 Emitter-coupled logic basic gate (ECL 10102): (a) circuit, (b) symbol.

A brief description of the operation of the circuit will verify that none of the transistors saturates. For the following discussion, V_{CC1} and V_{CC2} are grounded, V_{EE} is –5.2 V, and V_{TT} is –2 V. Diode drops and base-emitter voltages of active transistors are 0.8 V.

First, observe that the resistor-diode (D_1 and D_2) voltage divider establishes a reference voltage of –0.55 V at the base of T_3, which translates to –1.35 V at the base of T_2. When either or both of the inputs A and B are at the logical 1 level of –0.9 V, the emitters of T_{1A}, T_{1B}, and T_2 will be 0.8 V lower, at –1.7 V. This establishes the base-emitter voltage of T_2 at –1.35 – (–1.7) = 0.35 V, so T_2 is cut off. With T_2 off, T_4 is biased into the active region, and its emitter will be at about –0.9 V, corresponding to a logical 1 at the $(A + B)$ output. Most of the current through the 365-Ω emitter resistor, which is [–1.7 – (–5.2)]/0.365 = 9.6 mA, flows through the 100-Ω collector resistor, dropping the base voltage of T_5 to –0.96 V. Thus the voltage level at the output terminal designated $(A + B)$ is –1.76 V, corresponding to a logical 0.

When both A and B inputs are at the LOW level of –1.75 V, T_2 will be active, with its emitter voltage at –1.35 – 0.8 = –2.15 V. The current through the 365-Ω resistor becomes [–2.15 – (–5.2)]/0.365 = 8.2 mA. This current flows through the 112-Ω resistor pulling the base of T_4 down to –0.94 V, so that the $(A + B)$ output will be at the LOW level of –1.75 V. With T_{1A} and T_{1B} cut off, the base of T_5 is close to 0.0 V, and the $(A + B)$ output will therefore be at the nominal HIGH level of –0.9 V.

Observe that the output transistors T_4 and T_5 are always active and function as emitter followers, providing the low-output impedances required for driving capacitive loads. As T_{1A} and/or T_{1B} turn on, and T_2 turns off as a consequence, the transition is accomplished with very little current change in the 365-Ω emitter resistor. It follows that the supply current from V_{EE} does not undergo the sudden increases and decreases prevalent in TTL, thus eliminating the need for decoupling capacitors. This is a major reason why ECL can be operated successfully with the low noise margins which are inherent in logic having a relatively small voltage difference between the HIGH and LOW voltage levels (see Table 79.7). The small level shifts between LOW and HIGH also permit low propagation times without excessively fast rise and fall times. This reduces the effects of residual capacitive coupling between gates, thereby lessening the required noise margin. For this reason the faster ECL (100xxx) should not be used where the speed of the 10xxx series is sufficient. A comparison of three ECL series is given in Table 79.7. The propagation times t_{PLH} and t_{PHL} and transition times t_{TLH} and t_{THL} are defined in Fig. 79.7. Transitions are between the 20 and 80% levels.

TABLE 79.7 Comparison of ECL Quad Two-Input NOR Gates ($V_{TT} = V_{EE} = 5.2$ V, $V_{CC1} = 0$ V)

ECL Type	Power Supply Terminal V_{EE} (V)	Power Supply Current I_E (mA)	Propagation Delay Time t_{PLH}[a] (ns)	Propagation Delay Time t_{PHL} (ns)	Transition Time t_{TLH}[b] (ns)	Noise Margins t_{THL}[b] (ns)	Noise Margins NM_H (V)	Noise Margins NM_L (V)	Test Load
ECL II									
1012	–5.2	18[c]	5	4.5	4	6	0.175	0.175	Fan-out of 3
95102	–5.2	11	2	2	2	2	0.14	0.145	50 Ω
10102	–5.2	20	2	2	2.2	2.2	0.135	0.175	50 Ω
ECLIII									
1662	–5.2	56[c]	1	1.1	1.4	1.2	0.125	0.125	50 Ω
100102[d]	–4.5	55	0.75	0.75	0.7	0.7	0.14	0.145	50 Ω
11001[e]	–5.2	24	0.7	0.7	0.7	0.7	0.145	0.175	50 Ω

[a] See text for explanation of abbreviations.
[b] 20 to 80% levels.
[c] Maximum value (all other typical).
[d] Quint 2-input NOR/OR gate.
[e] Dual 5/4-input NOR/OR gate.

The 50-Ω pull-down resistors shown in Fig. 79.12 are connected externally. The outputs of several gates can therefore share a common pull-down resistor to form a wired-OR connection. The open emitter outputs also provide flexibility for driving transmission lines, the use of which in most cases is mandatory for interconnecting this high-speed logic. A twisted pair interconnection can be driven using the complementary outputs $(A + B)$ and $(A + B)$ as a differential output. Such a line should be terminated in an ECL line receiver (10114).

Since ECL is used in high-speed applications, special techniques must be applied in the layout and interconnection of chips on circuit boards. Users should consult design handbooks published by the suppliers before undertaking the construction of an ECL logic system.

While ECL is not compatible with any other logic family, interfacing buffers, called translators, are available. In particular, the 10124 converts TTL output levels to ECL complementary levels, and the 10125 converts either single-ended or differential ECL outputs to TTL levels. Among other applications of these translators, they allow the use of ECL for the highest speed requirements of a system while the rest of the system uses the more rugged TTL. Another translator is the 10177, which converts the ECL output levels to *n*-channel metal-oxide semiconductor (NMOS) levels. This is designed for interfacing ECL with *n*-channel memory systems.

Complementary Metal-Oxide Semiconductor (CMOS) Logic

Metal-oxide semiconductor (MOS) technology is prevalent in LSI systems due to the high circuit densities possible with these devices. *p*-Channel MOS was used in the first LSI systems, and it still is the cheapest to produce because of the higher yields achieved due to the longer experience with PMOS technology. PMOS, however, is largely being replaced by NMOS (*n*-channel MOS), which has the advantages of being faster (since electrons have greater mobility than holes) and having TTL compatibility. In addition, NMOS has a higher function/chip area density than PMOS, the highest density in fact of any of the current technologies. Use of NMOS and PMOS, however, is limited to LSI and VLSI fabrications. The only MOS logic available as SSI and MSI is CMOS (complementary MOS).

CMOS is faster than NMOS and PMOS, and it uses less power per function than any other logic. While it is suitable for LSI, it is more expensive and requires somewhat more chip area than NMOS or PMOS. In many respects it is unsurpassed for SSI and MSI applications. Standard CMOS (the 4000 series) is as fast as low-power TTL (74Lxx) and has the largest noise margin of any logic type.

A unique advantage of CMOS is that for all input combinations the steady-state current from V_{DD} to V_{SS} is almost zero because at least one of the series FETs is open. Since CMOS circuits of any complexity are interconnections of the basic gates, the quiescent currents for these circuits are extremely small, an obvious advantage which becomes a necessity for the practicality of digital watches, for example, and one which alleviates

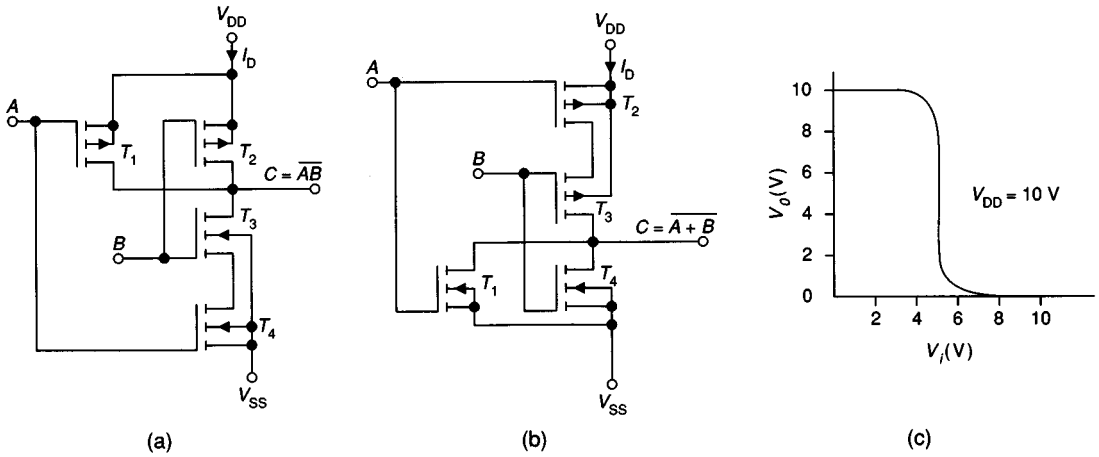


FIGURE 79.13 (a) Complementary metal-oxide semiconductor (CMOS) NAND gate, (b) NOR gate, and (C) inverter transfer characteristic.

heat dissipation problems in high-density chips. Also a noteworthy feature of CMOS digital circuits is the absence of components other than FETs. This attribute, which is shared by PMOS and NMOS, accounts for the much higher function/chip area density than is possible with TTL or ECL. During the time the output of a CMOS gate is switching there will be current flow from V_{DD} to V_{SS}, partly due to the charging of junction capacitances and partly because the path between V_{DD} and V_{SS} closes momentarily as the FETs turn on and off. This causes the dc supply current to increase in proportion to the switching frequency in a CMOS circuit. Manufacturers specify that the supply voltage for standard CMOS can range over $3\ V \le V_{DD} - V_{SS} \le 18\ V$, but switching speeds are slower at the lower voltages, mainly due to the increased resistances of the "on" transistors. The output switches between low and high when the input is midway between V_{DD} and V_{SS}, and the output logical 1 level will be V_{DD} and the logical 0 level V_{SS} [Fig. 79.13(c)]. If CMOS is operated with $V_{DD} = 5$ V and $V_{SS} = 0$ V, the V_{DD} and V_{SS} levels will be almost compatible with TTL except that the TTL totem-pole output high of 3.4 V is marginal as a logical 1 for CMOS. To alleviate this, when CMOS is driven with TTL a 3.3-kΩ pull-up resistor between the TTL output and the common V_{CC}, V_{DD} supply terminal should be used. This raises V_{OH} of the TTL output to 5 V.

All CMOS inputs are diode protected to prevent static charge from accumulating on the FET gates and causing punch-through of the oxide insulating layer. A typical configuration is illustrated in Fig. 79.14. Diodes D_1 and D_2 clamp the transistor gates between V_{DD} and V_{SS}. Care must be taken to avoid input voltages that would cause excessive diode currents. For this reason manufacturers specify an input voltage constraint from $V_{SS} - 0.5$ V to $V_{DD} + 0.5$ V. The resistance R_s helps protect the diodes from excessive currents but is introduced at the expense of switching speed, which is deteriorated by the time constant of this resistance and the junction capacitances.

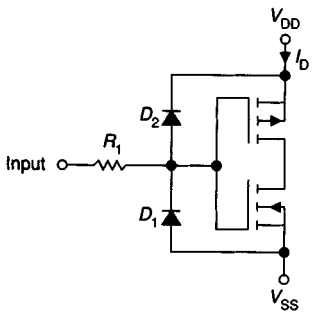


FIGURE 79.14 Diode protection of input transistor gates. $200\ \Omega < R_s < 1.5\ k\Omega$.

Advanced versions of CMOS have been developed which are faster than standard CMOS. The first of these to appear were designated 74HCxx and 74HCTxx. The supply voltage range for this series is limited to $2\ V \le V_{DD} - V_{SS} \le 6\ V$. The pin numbering of a given chip is the same as its correspondingly numbered TTL device. Furthermore, gates with the HCT code have skewed transfer characteristics which match those of its TTL cousin, so that these chips can be directly interchanged with low-power Schottky TTL.

TABLE 79.8 Comparison of Standard, High-Speed, and Advanced High-Speed CMOS

			Standard CMOS NORGates		High-Speed CMOS Inverter		Advanced CMOS Inverter	
Parameter	Symbol	Unit	4001B	4011UB	74HC04	74HCT04	74AC04	74ACT04
Supply voltage	V_{DD}-V_{SS}	V	15	15	6	5.5	5.5	5.5
Input voltage	V_{IHmin}	V	11	12.5	4.2	2	3.85	2
thresholds	V_{ILmax}	V	4	2.5	1.8	0.8	1.65	0.8
Guaranteed output	V_{OHmin}	V	13.5	13.5	5.9	4.5	4.86	4.76
levels at	V_{OLmax}	V	1.5	1.5	0.1	0.26	0.32	0.37
maximum IO								
Maximum	I_{OH}	mA	–8.8	–3.5	–4	–4	–24	–24
output currents	I_{OL}	mA	8.8	8.8	4	4	24	24
Noise	NM_L	V	2.5	2.5	1.7	0.54	1.33	.43
margins	NM_H	V	2.5	2.5	1.7	2.5	1.01	1.24
Propagation	t_{PLH}	ns	40	40	16	15	4	4.3
times	t_{PHL}	ns	40	40	16	17	3.5	3.9
Max input	I_{INmax}	μA	0.1	0.1	0.1	0.1	0.1	0.1
current leakage								
D-flip-flop			4013B		74HC374	74HCT374A	74AC374	74ACT374
max frequency	f_{max}	MHz	7.0	N.A.	35	30	100	100
(guaranteed								
minimum)								

More recently, a much faster CMOS has appeared and carries the designations 74ACxx and 74ACTxx. These operate in the same supply voltage range and bear the same relationship with TTL as the HCMOS. The driving capabilities (characterized by I_{OH} and I_{OL}) of this series are much greater, such that they can be fanned out to 10 low-power Schottky inputs.

The three types of CMOS are compared in Table 79.8. The relative speeds of these technologies are best illustrated by including in the table the maximum clock frequencies for *D* flip-flops. In each case, the frequency given is the maximum for which the device is guaranteed to work. It is worth noting that a typical maximum clocking of 160 MHz is claimed for the 74ACT374 *D* flip-flop.

CMOS Design Considerations

Design and handling recommendations for CMOS, which are included in several of the data books, should be consulted by the designer using this technology. A few selected recommendations are included here to illustrate the importance of such information.

1. All unused CMOS inputs should be tied either to V_{DD} or V_{SS}, whichever is appropriate for proper operation of the gate. This rule applies even to inputs of unused gates, not only to protect the inputs from possible static charge buildup, but to avoid unnecessary supply current drain. Floating gate inputs will cause all the FETs to be conducting, wasting power and heating the chip unnecessarily.
2. CMOS inputs should never be driven when the supply voltage V_{DD} is off, since damage to the input-protecting diodes could result. Inputs wired to edge connectors should be shunted by resistors to V_{DD} or V_{SS} to guard against this possibility.
3. Slowly changing inputs should be conditioned using Schmitt trigger buffers to avoid oscillations that can arise when a gate input voltage is in the transition region.
4. Wired-AND configurations cannot be used with CMOS gates, since wiring an output HIGH to an output LOW would place two series FETs in the "on" condition directly across the chip supply.
5. Capacitive loads greater than 5000 pF across CMOS gate outputs act as short circuits and can overheat the output FETs at higher frequencies.
6. Designs should be used that avoid the possibility of having low impedances (such as generator outputs) connected to CMOS inputs prior to power-up of the CMOS chip. The resulting current surge when V_{DD} is turned on can damage the input diodes.

While this list of recommendations is incomplete, it should alert the CMOS designer to the value of the information supplied by the manufacturers.

Choosing a Logic Family

A logic designer planning a system using SSI and MSI chips will find that an extensive variety of circuits is available in all three technologies: TTL, ECL, and CMOS. The choice of which technology will dominate the system is governed by what are often conflicting needs, namely, speed, power consumption, noise immunity, cost, availability, and the ease of interfacing. Sometimes the decision is easy. If the need for a low static power drain is paramount, CMOS is the only choice. It used to be the case that speed would dictate the selection; ECL was high speed, TTL was moderate, and CMOS low. With the advent of advanced TTL and, especially, advanced CMOS the choice is no longer clear-cut. All three will work at 100 MHz or more. ECL might be used since it generates the least noise because the transitions are small, yet for that same reason it is more susceptible to externally generated noise. Perhaps TTL might be the best compromise between noise generation and susceptibility. Advanced CMOS is the noisiest because of its rapid rise and fall times, but the designer might opt to cope with the noise problems to take advantage of the low standby power requirements.

A good rule is to use devices which are no faster than the application requires and which consume the least power consistent with the needed driving capability. The information published in the manufacturers' data books and designer handbooks is very helpful when choice is in doubt.

Defining Term

Logic gate: Basic building block for logic systems that controls the flow of pulses.

Related Topics

25.3 Application-Specific Integrated Circuits • 81.2 Logic Circuits

References

Advanced CMOS Logic Designers Handbook, Dallas: Texas Instruments, Inc., 1987.
C. Belove and D. Schilling, *Electronic Circuits, Discrete and Integrated,* 2nd ed., New York: McGraw-Hill, 1979.
FACT Data, Phoenix: Motorola Semiconductor Products, Inc., 1989.
Fairchild Advanced Schottky TTL, California: Fairchild Camera and Instrument Corporation, 1980.
W. I. Fletcher, *An Engineering Approach to Digital Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1980.
High Speed CMOS Logic Data, Phoenix: Motorola Semiconductor Products, Inc., 1989.
P. Horowitz and W. Hill, *The Art of Electronics,* 2nd ed., New York: Cambridge University Press, 1990.
MECL System Design Handbook, Phoenix: Motorola Semiconductor Products, Inc., 1988.
H. Taub and D. Schilling, *Digital Integrated Electronics,* New York: McGraw-Hill, 1977.
The TTL Data Book for Design Engineers, Dallas: Texas Instruments, Inc., 1990.

Further Information

An excellent presentation of the practical design of logic systems using SSI and MSI devices is developed in the referenced book *An Engineering Approach to Digital Design* by William I. Fletcher. The author pays particular attention to the importance of device speed and timing.

The Art of Electronics by Horowitz and Hill is particularly helpful for its practical approach to interfacing digital with analog.

Everything one needs to know about digital devices and their interconnection can be found somewhere in the data manuals, design handbooks, and application notes published by the device manufacturers. Unfortunately, no single publication has it all, so the serious user should acquire as large a collection of these sources as possible.

79.3 Bistable Devices

Richard S. Sandige

This section deals with bistable devices which are also commonly referred to as **bistables, latches,** or **flip-flops.** Bistable devices are **memory elements.** Each bistable provides storage for only 1-bit, i.e., it can store a 1 or a 0. Figure 79.15 shows a graphic classification of bistable devices.

Manufacturers supply integrated circuit (IC) packages containing several bistable devices. One data book for the transistor-transistor logic (TTL) circuit technology lists 4-, 8-, 9-, and 10-bit latches in one IC package. The same data book lists 2-, 4-, 6-, 8-, 9-, and 10-bit flip-flops in one IC package. While a 1-bit bistable can only store 1 bit of information, 8-bit bistables are capable of storing 8 bits of information. Bistable devices implemented with logic gates are **volatile devices.** When power is first applied the first stored value of the bistable is random (it can store a 1 or a 0), and when power is removed the bistable loses its storage capability. Certain memories (also called stores) are nonvolatile and therefore retain their data when power is removed. These devices will not be discussed in this section.

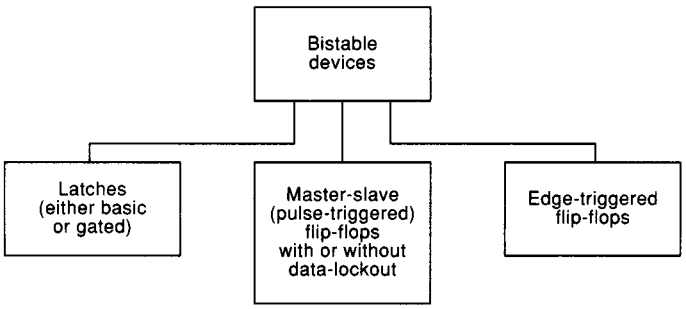


FIGURE 79.15 Graphic classification of bistable devices. (*Source:* Modified from R. S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 467. With permission.)

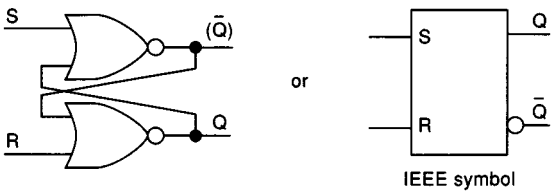


FIGURE 79.16 Basic *S-R* NOR latch implementation. (*Source:* Modified from R. S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 448. With permission.)

Basic Latches

A latch can be either basic or gated. Figure 79.16 is an example of a basic *S-R* NOR latch implementation using two cross-coupled NOR gates. The logic symbol recommended for the *S-R* NOR latch by the Institute of Electrical and Electronics Engineers (IEEE) is shown to the right of the logic circuit implementation.

The input signal named *S* stands for set while the input signal named *R* stands for reset. Manufacturers often select *Q* as the output signal name for bistable devices in their data books. The *Q*s on the outputs are added for clarity and are not part of the IEEE symbol. The *S-R* NOR latch shown in Fig. 79.16 is a basic latch circuit since the *S* and *R* inputs are not gated with a control signal. The **reduced characteristic table** in Table 79.9 shows the operation of the *S-R* NOR latch circuit.

For $S\,R = 00$, $Q = Q_0$, illustrating that the output for the next state *Q* is the same as the present state output Q_0. For $S\,R = 01$, $Q = 0$, specifying

TABLE 79.9 Reduced Characteristic Table for the *S-R* NOR Latch

S	*R*	*Q*	Operation
0	0	Q_0	no change
0	1	0	reset
1	0	1	set
1	1	0	not normally allowed

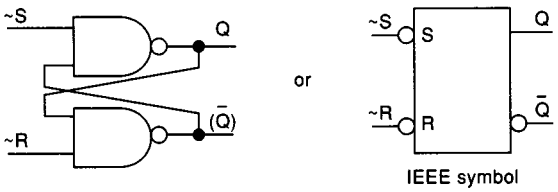


FIGURE 79.17 Basic *S-R* NAND latch implementation. (*Source:* Modified from R. S. Sandige. *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 449. With permission.)

TABLE 79.10 Reduced Characteristic Table for the *S-R* NAND Latch

~*S*	~*R*	*Q*	Operation
0	0	1	not normally allowed
0	1	1	set
1	0	0	reset
1	1	Q_0	no change

that the output for the next state is reset. For $S\,R = 10$, $Q = 1$, indicating that the output for the next state is set. In most cases the input conditions $S\,R = 11$ are not allowed for two reasons. If $S\,R = 11$, then the alternate output for the bistable, shown in parentheses as $\overline{Q}$ in Fig. 79.16, is not logically correct as it is for all other input combinations. The second reason is more subtle since the next state of the bistable can be set or reset due to a **critical race** condition when the inputs are changed from 11 to 00. Such unpredictability is not desirable and therefore the $S\,R = 11$ condition is generally not allowed. Latches and flip-flops that contain a Q and a $\overline{Q}$ output (complementary outputs) provide double-rail outputs.

The *S-R* NAND latch implementation shown in Fig. 79.17 uses two cross-coupled NAND gates. The tildes shown in the logic circuit diagram preceding *S* and *R* represent inline symbols for the logical complements of *S* and *R*, respectively, as recommended by the IEEE. Data books usually refer to the *S-R* NAND latch as the $\overline{S}$-$\overline{R}$ latch. The logic symbol recommended for the *S-R* NAND latch by IEEE is shown to the right of the logic circuit diagram.

The ~*S* and ~*R* on the inputs and *Q*s on the outputs of the IEEE symbol are added for clarity and are not part of the IEEE symbol. The reduced characteristic table illustrated in Table 79.10 shows the operation of the *S-R* NAND latch circuit in Fig. 79.17.

In most cases the input conditions $\sim S\sim R = 00$ ($S\,R = 11$) are not allowed for the same reasons provided above for the *S-R* NOR latch. For $\sim S\sim R = 01$ ($S\,R = 10$), $Q = 1$, indicating that the output for the next state is set. For $\sim S\sim R = 10$ ($S\,R = 01$), $Q = 0$, specifying that the output for the next state is reset. For $\sim S\sim R = 11$ ($S\,R = 00$), $Q = Q_0$, illustrating that the output for the next state Q is the same as the present state output Q_0.

Gated Latches

All other gate level latches and flip-flops are functionally equivalent to either the configuration of the cross-coupled NOR latch circuit or the cross-coupled NAND latch circuit. A gated *S-R* NOR latch circuit and a gated *S-R* NAND latch are illustrated in Fig. 79.18 along with the recommended IEEE symbol. The reduced characteristic table for both of these circuits is provided in Table 79.11.

In each circuit both the *S* and *R* inputs are gated with a control signal *C*. Notice in the reduced characteristic table that the *S* and *R* inputs are only enabled, and thus have an effect on the output, when $C = 1$ (**transparent mode**).

Whatever value the output has when *C* goes to 0 is latched, captured, or stored (memory mode). Like the basic latches, the input conditions for $S\,R = 11$ are not generally allowed for the gated *S-R* latches when *C* goes to 1. The gated *D* latch circuit is perhaps the most used latch circuit since the added Inverter shown in the circuit diagram in Fig. 79.19 ensures that the input conditions for $S\,R = 11$ cannot occur when *C* goes to 1. The reduced characteristic table for the gated *D* latch circuit is shown in Table 79.12.

Flip-Flops

We will use the term *flip-flop* to distinguish between the bistable device called a latch and the bistable device that allows feed-back without oscillation. Early types of flip-flops were of the master-slave (pulse-triggered) variety that had no data-lockout circuitry and caused a storage error if improperly used due to 1s and 0s catching. To prevent 1s and 0s catching, data-lockout (also called variable-skew) circuitry was added to a few

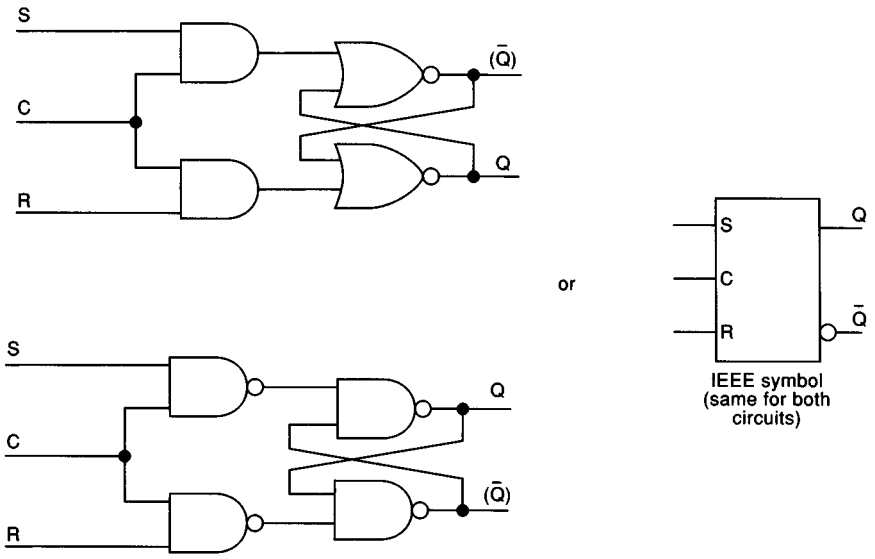


FIGURE 79.18 Gated *S-R* NOR and gated *S-R* NAND latch circuit. (*Source:* Modified from R. S. Sandige. *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 468. With permission.)

TABLE 79.11 Reduced Characteristic Table for the Gated *S-R* Latches

C	*S*	*R*	*Q*	Operation
0	0	0	Q_0	no change
0	0	1	Q_0	no change
0	1	0	Q_0	no change
0	1	1	Q_0	no change
1	0	0	Q_0	no change
1	0	1	0	reset
1	1	0	1	set
1	1	1	0,1	reset (*S-R* NOR), set (*S-R* NAND)

TABLE 79.12 Reduced Characteristic Table for the Gated *D* Latch

C	*D*	*Q*	Operation
0	0	Q_0	no change
0	1	Q_0	no change
1	0	0	reset
1	1	1	set

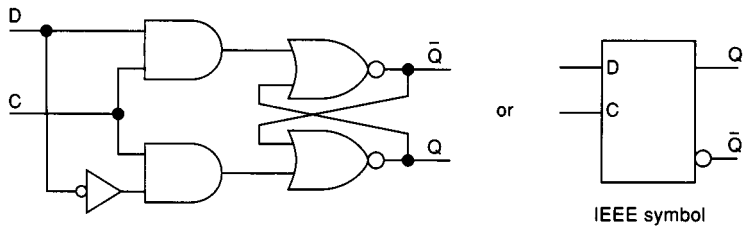


FIGURE 79.19 Gated *D* latch circuit. (*Source:* Modified from R. S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 470. With permission.)

master-slave flip-flop types. Due to the better design features and popularity of **edge-triggered** flip-flops, master-slave flip-flops are not recommended for newer designs and in some cases have been made obsolete by manufacturers, making them difficult to obtain even for repair parts. For this reason only edge-triggered flip-flops will be discussed.

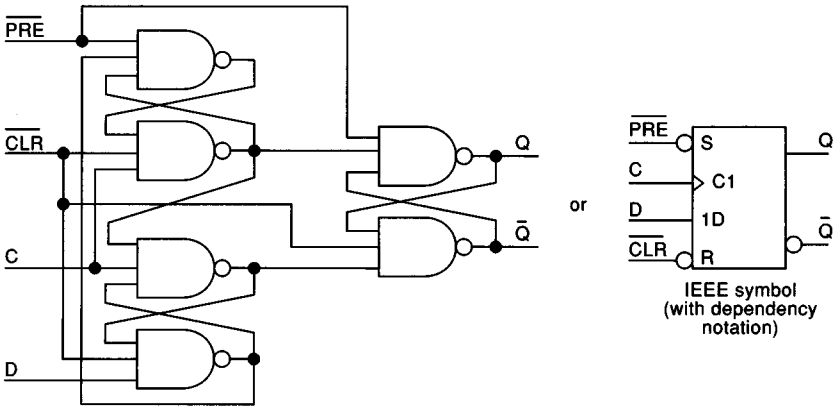


FIGURE 79.20 Positive edge-triggered *D* flip-flop circuit. (*Source:* Modified from R. S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 490. With permission.)

TABLE 79.13 Reduced Characteristic Table for Positive Edge-Triggered *D* Flip-Flop

$\overline{PRE}$	$\overline{CLR}$	*C*	*D*	*Q*	Operation
0	0	X	X	1	not normally allowed
0	1	X	X	1	preset
1	0	X	X	0	clear
1	1	↑	1	1	set
1	1	↑	0	0	reset
1	1	0	X	Q_0	no change

TABLE 79.14 Reduced Characteristic Table for Negative Edge-Triggered *J-K* Flip-Flop

$\overline{PRE}$	$\overline{CLR}$	*C*	*J*	*K*	*Q*	Operation
0	0	X	X	X	1	not normally allowed
0	1	X	X	X	1	preset
1	0	X	X	X	0	clear
1	1	↓	0	0	Q_0	no change
1	1	↓	1	0	1	set
1	1	↓	0	1	0	reset
1	1	↓	1	1	$\overline{Q}_0$	toggle
1	1	1	X	X	Q_0	no change

Edge-Triggered Flip-Flops

Two types of edge-triggered flip-flops are predominantly used in modern designs. These are the *D* type and *J-K* type. The *D* type is perhaps the most used because its circuitry generally takes up less real estate on an IC chip and because most engineers consider it an easier device with which to design. An example of a positive edge-triggered *D* flip-flop circuit is shown in Fig. 79.20. The reduced characteristic table illustrating the operation of this flip-flop is shown in Table 79.13.

The main difference between a latch and an edge-triggered flip-flop is the question of transparency. The gated *D* latch is transparent (the *Q* output follows the *D* input when the control input $C = 1$) and it latches, captures, or stores the value at the *D* input at the time the control input *C* goes to 0. The *positive edge-triggered D* flip-flop is never transparent from its data input *D* to its output *Q*. When the control input *C* is 0 the output *Q* does not follow the *D* input and remains unchanged; however, the value at the *D* input is latched, captured, or stored at the time the *control input C makes a 0 to 1 transition.* The characteristic that makes edge-triggered flip-flops desirable for feedback applications is that, due to their nontransparent property, their outputs can be fed back as inputs to the device without causing oscillation. This is true for all types of edge-triggered flip-flops. A negative edge-triggered *J-K* flip-flop circuit is shown in the circuit diagram in Fig. 79.21 with its corresponding IEEE symbol. Notice that the *J-K* flip-flop requires eight logic gates compared to only six logic gates for the *D* flip-flop in Fig. 79.20. The reduced characteristic table for this negative edge-triggered flip-flop is shown in Table 79.14.

Notice in the reduced characteristic table (Table 79.14 for the *J-K* flip-flop) when the *J* and *K* inputs are both 1 and the control input *C* makes a 1 to 0 transition, the flip-flop **toggles,** i.e., the next state output *Q*

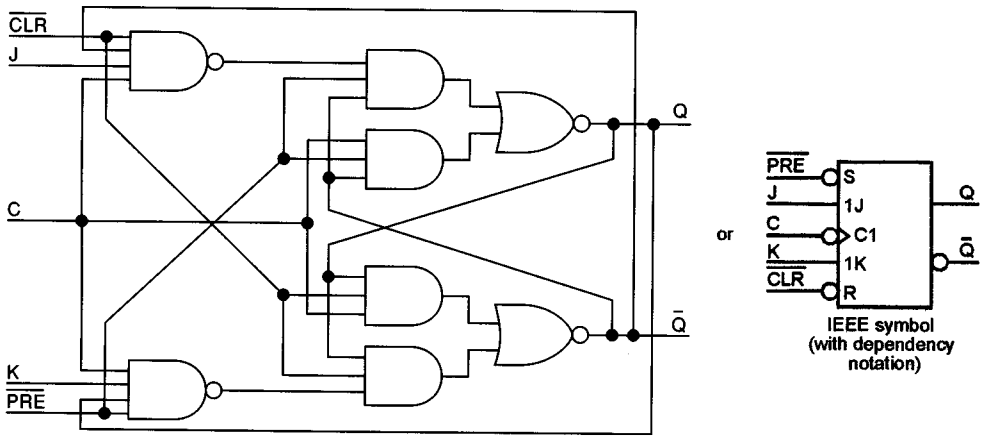


FIGURE 79.21 Negative edge-triggered *J-K* flip-flop circuit. (*Source:* Modified from R. S. Sandige. *Modern Digital Design,* New York: McGraw-Hill, 1990, p. 493. With permission.)

changes to the complement of the present state output Q_0. By simply connecting *J* and *K* together and renaming it *T* for toggle, one can obtain a negative edge-triggered *T* flip-flop.

Special Notes on Using Latches and Flip-Flops

Since bistable devices are asynchronous **fundamental mode** sequential logic circuits, only one input is allowed to change at a time. This means that for proper operation for a basic latch, only one of the data inputs *S* or *R* for a *S-R* NOR latch (~*S* or ~*R* for a *S-R* NAND latch) may be changed at one time. For a gated latch this means for proper operation the data inputs *S* and *R* or data input *D* must meet a **minimum setup** (t_{su}) **and hold time** (t_h) **requirement,** i.e., the data input(s) must be stable for a minimum time period, prior to the control input *C* changing the latch from the transparent mode to the memory mode. For proper operation of an edge-triggered flip-flop this means that the data input *D* or data inputs *J* and *K* must meet a minimum setup time and hold time requirement relative to the control input *C* changing from 0 to 1 (positive edge-triggered) or from 1 to 0 (negative edge-triggered). In manufacturers' data books, the control input *C* is often named the enable input for latches and the clock (*CLK*) input for flip-flops.

Defining Terms

Bistable, latch, and flip-flop: Names used in place of the term **bistable device.**
Critical race: A change in two input variables that results in an unpredictable output value for a bistable device.
Edge-triggered: Term used to describe the edge of a positive or negative pulse applied to the control input of a nontransparent bistable device to latch, capture, or store the value indicated by the data input(s).
Fundamental mode: Operating mode of a circuit that allows only one input to change at a time.
Memory element: A bistable device or element that provides data storage for a logic 1 or a logic 0.
Reduced characteristic table: A tabular representation used to illustrate the operation of various bistable devices.
Setup and hold time requirement: Setup time (hold time) is the time required for the data input(s) to be held stable prior to (or after) the control input *C* changes to latch, capture, or store the value indicated by the data input(s).
Toggle: Change of state from logic 0 to logic 1 or from logic 1 to logic 0 in a bistable device.
Transparent mode: Mode of a bistable device where an output responds to data input signal changes.
Volatile device: A memory or storage device that loses its storage capability when power is removed.

Related Topics

25.3 Application-Specific Integrated Circuits • 81.3 Resistors and Their Applications

References

ANSI/IEEE Std 91-1984, *IEEE Standard Graphic Symbols for Logic Functions,* New York: Institute of Electrical and Electronics Engineers.

ANSI/IEEE Std 991-1986, *IEEE Standard for Logic Circuit Diagrams,* New York: Institute of Electrical and Electronics Engineers.

D. L. Dietmeyer, *Logic Design of Digital Systems,* 2nd ed., Boston: Allyn and Bacon, 1988.

F. J. Hill and G. R. Peterson, *Introduction to Switching Theory & Logical Design,* 3rd ed., New York: John Wiley, 1981.

E. L. Johnson and M. A. Karim, *Digital Design Pragmatic Approach,* Boston: Prindle, Weber & Schmidt Publishers, 1987.

I. Kampel, *A Practical Introduction to the New Logic Symbols,* 2nd ed., London: Butterworths, 1986.

C. H. Roth, Jr., *Fundamentals of Logic Design,* 4th ed., St. Paul: West Publishing, 1992.

R. S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990.

Texas Instruments, *The TTL Data Book,* vol. 3, Advanced Low-Power Schottky, Advanced Schottky, Dallas: Texas Instruments, 1984.

Further Information

The monthly magazine *IEEE Transactions on Computers* presents papers discussing bistable devices, for example, "A Simulation-Based Method for Generating Tests for Sequential Circuits" in its December 1990 issue, pp. 1456–1463.

Another monthly magazine, *IEEE Transactions on Computer-Aided Design,* sometimes presents papers discussing bistable devices, for example, "Schematic Generation with an Expert System" in its December 1990 issue, pp. 1289-1306.

79.4 Optical Devices

H. S. Hinton

Since the first demonstration of optical logic devices in the late 1970s, there have been many different experimental devices reported. Figure 79.22 categorizes optical logic devices into four main classes. The first division is between all-optical and optoelectronic devices. All-optical devices are devices that do not use electrical currents to create the nonlinearity required by digital devices. These devices can be either single-pass devices (light passes through the nonlinear material once) or they can use a resonant cavity to further enhance the optical nonlinearity (multiple passes through the same nonlinear material). Optoelectronic devices, on the other hand, use electrical currents and electronic devices to process a signal that has gone through an optical-to-electrical conversion process. The output of these devices is either provided by electrically driving an optical source such as a laser or LED (detect/emit) or by modulating some external light source (detect/modulate). Below each of these categories are listed some of the devices that have been experimentally demonstrated.

All-Optical Devices

To create an all-optical logic device requires a medium that will allow one beam of light to affect another. This phenomenon can arise from the cubic response to the applied field. These third-order processes can lead to purely dielectric phenomena, such as irradiance-dependent refractive indices. By exploiting purely dielectric third-order nonlinearities, such as the optical Kerr effect, changes can be induced in the optical constants of the medium which can be read out directly at the same wavelength as that inducing them. This then opens up

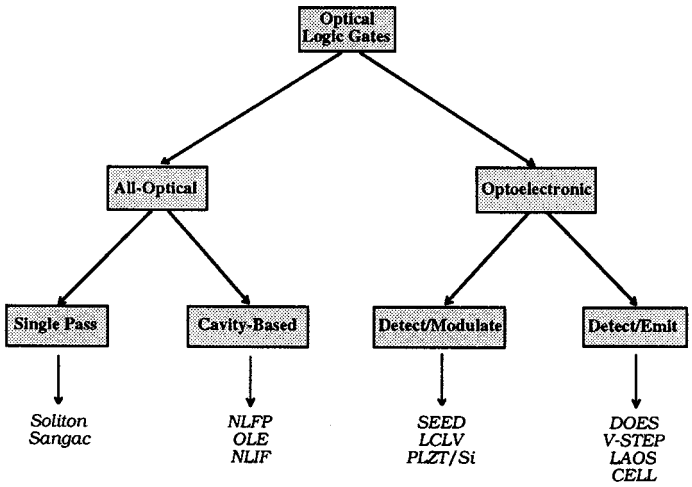


FIGURE 79.22 Classification of optical logic devices.

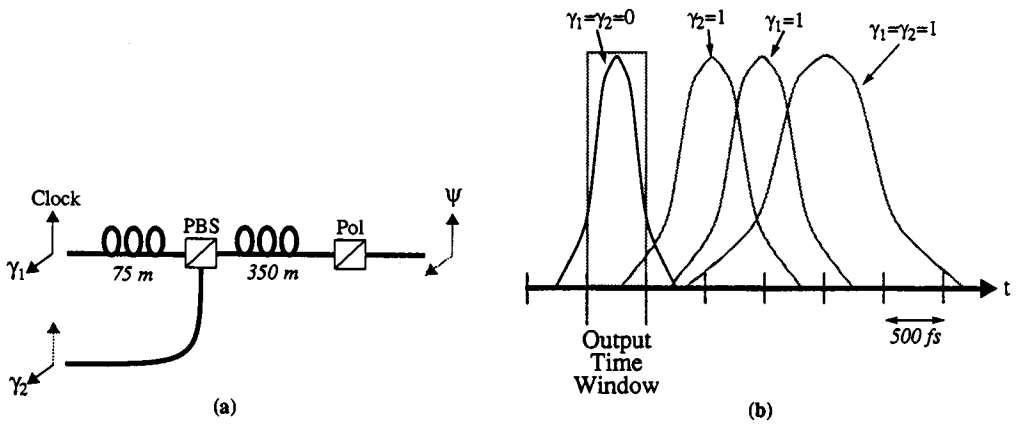


FIGURE 79.23 Soliton NOR gate: (a) physical implementation, (b) timing diagram.

the possibilities for digital optical circuitry based on cascadable all-optical logic gates. Although there have been many different all-optical gates demonstrated, this section will only briefly review the **soliton** gate (single-pass) and one example of the **nonlinear Fabry-Perot** structures (cavity-based).

Single-Pass Devices

An example of an all-optical single-pass optical logic gate is the soliton NOR gate. It is an all-fiber logic gate based on time shifts resulting from soliton dragging. A NOR gate consists of two birefringent fibers connected through a polarizing beamsplitter with the output filtered by a polarizer as shown in Fig. 79.23. The clock pulse, which provides both gain and logic level restoration, propagates along one principal axis in both fibers. For the NOR gate the fiber length is trimmed so that in the absence of any signal the entering clock pulse will arrive within the output time window corresponding to a "1." When either or both of the input signals are incident, they interact with the clock pulse through soliton dragging and shift the clock pulse out of the allowed output time window creating a "0" output. In soliton dragging two temporally coincident, orthogonally polarized pulses interact in the fiber through cross-phase modulation and shift each other's velocities. This velocity shift converts into a time shift after propagating some distance in the fiber. To implement the device, the two input signal pulses γ_1 and γ_2 are polarized orthogonal to the clock. The signals are timed so that γ_1 and the clock

pulse coincide at the input to the first fiber and γ_2 and the clock pulse coincide (in the absence of γ_1) at the input to the second fiber. At the output the two input signals are blocked by the polarizer, allowing only the temporally modified clock pulse to pass. In a prototyped demonstration this all-optical NOR gate required 5.8 pJ of signal energy and provided an effective gain of 6.

Cavity-Based Devices

Cavity-based optical logic devices are composed of two highly reflective mirrors that are separated by a distance d [Fig. 79.24(a)]. The volume between the mirrors, referred to as the cavity of the etalon, is filled with a nonlinear material possessing an index of refraction that varies with intensity according to $n_c = n_0 + n_2 \gamma_c$ where n_0 is the linear index of refraction, n_2 is the nonlinear index of refraction, and γ_c is the intensity of light within the cavity. In the ideal case, the characteristic response of the reflectivity of a Fabry-Perot cavity, R_{fp}, is shown in Fig. 79.24(b). At low intensities, the cavity resonance peak is not coincident with the wavelength of the incident light; thus the reflectivity is high, which allows little of the incident light to be transmitted [solid curves in Fig. 79.24(b)]. As the intensity of the incident light γ increases, so does the intercavity light intensity which shifts the resonance peak [dotted curve in Fig. 79.24(b)]. This shift in the resonant peak increases the transmission which in turn reduces the reflectivity. This reduction in ψ will continue with increasing γ until a minimum value is reached. It should be noted that in practice all systems of interest have both intensity-dependent absorption and n_2.

To implement a two-input NOR gate using the characteristic curve shown in Fig. 79.24(c) requires a third input which is referred to as the *bias beam*, γ_b. This energy source biases the etalon at a point on its operating curve such that any other input will exceed the nonlinear portion of the curve moving the etalon from the high reflection state. This is illustrated in Fig. 79.24(c) where the γ_b combines with the inputs γ_1 and γ_2 to exceed the threshold of the nonlinear characteristic curve.

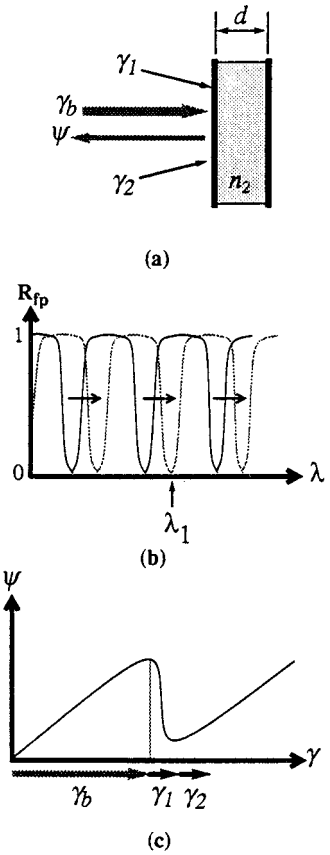


FIGURE 79.24 (a) Nonlinear Fabry-Perot etalon, (b) reflection peaks of NLFP, and (c) NLFP in reflection (NOR).

The first etalon-based optical logic device was in the form of a nonlinear interference filter (NLIF). A simple interference filter has a general form similar to a Fabry-Perot etalon, being constructed by depositing a series of thin layers of transparent material of various refractive indices on a transparent substrate. The first several layers deposited form a stack of alternating high and low refractive indices, all of optical thickness equal to one quarter of the operating wavelength. The next layer is a low integer (1–20) number of half wavelengths thick and finally a further stack is deposited to form the filter. The two outer stacks have the property of high reflectivity at one wavelength, thus playing the role of mirrors forming a cavity. A high finesse cavity is usually formed when both mirrors are identical, i.e., of equal reflectivity. However, unlike a Fabry-Perot etalon with a nonabsorptive material in the cavity, matched (equal) stack reflectivities do not give the optimum cavity design to minimize switch power because of the absorption in the spacer (which may be necessary to induce nonlinearity). A balanced design which takes into account the effective decrease in back mirror reflectivity due to the double pass through the absorbing cavity is preferable and also results in greater contrast between bistable states. The balanced design is easily achieved by varying one or all of the available parameters: number of periods, thickness and refractive index of each layer within either stack.

Optoelectronic Devices

Optoelectronic devices take advantage of both the digital processing capabilities of electronics and communications capabilities of the optical domain. This section will review both the SEED-based optical logic gates and the *pnpn* structures that have demonstrated optical logic.

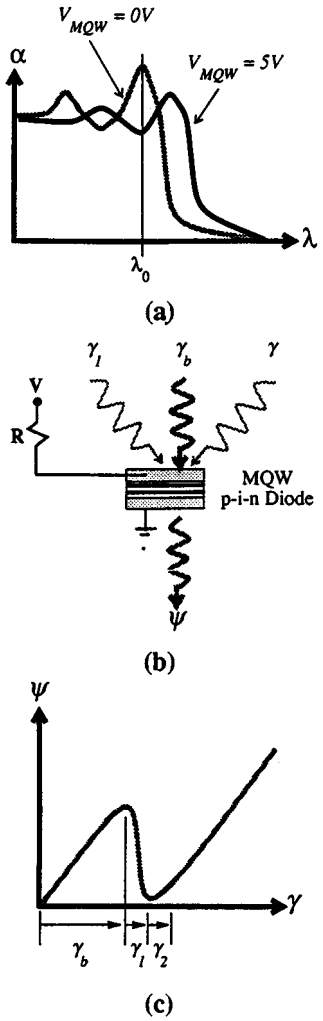


FIGURE 79.25 (a) Absorption spectra of MQW material for both 0 and 5 V, (b) schematic of MQW *pin* diode, (c) input/output characteristics of MQW *pin* diode.

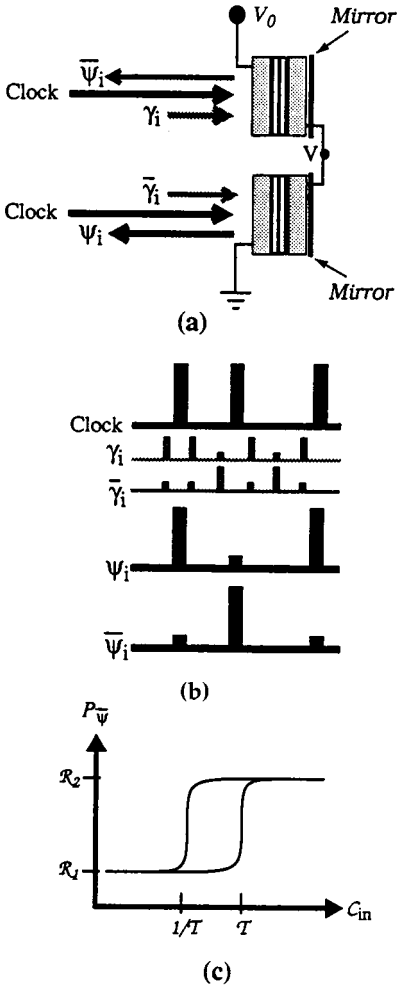


FIGURE 79.26 Symmetric self-electro-optic effect device (S-SEED). (a) S-SEED with inputs and outputs, (b) power transfer characteristics, and (c) optically enabled S-SEED.

Detect/Modulate Devices

In the most general terms the self-electro-optic effect device **(SEED) technology** corresponds to any device based on **multiple quantum well** (MQW) modulators. The basic physical mechanism used by this technology is the quantum confined Stark effect. This mechanism creates a shift in the bandedge of a semiconductor with an applied voltage. This is illustrated in Fig. 79.25(a). This shift in the bandedge is then used to vary the absorption of incident light on the MQW material. When this MQW material is placed in the intrinsic region of a *pin* diode and electrically connected to a resistor as shown in Fig. 79.25(b) the characteristic curve shown in Fig. 79.25(c) results. When the incident intensity, γ_i, is low there is no current flowing through the *pin* diode or resistor; thus the majority of the voltage is across the *pin* diode. If the device is operating at the wavelength λ_0, the device will be in a low absorptive state. As the incident intensity increases so does the current flowing in the *pin* diode; this in turn reduces the voltage across the diode which increases the absorption and current flow. This state of increasing absorption creates the nonlinearity in the output signal, ψ, shown in Fig. 79.25(c). Optical logic gates can be formed by optically biasing the R-SEED close to the nonlinearity, γ_b, and then applying lower level data signals γ_1 and γ_2 to the device.

The S-SEED, which behaves like an optical inverting *S-R* latch, is composed of two electrically connected MQW *pin* diodes as illustrated in Fig. 79.26(a). In this figure, the device inputs include the signal, γ_i (Set), and

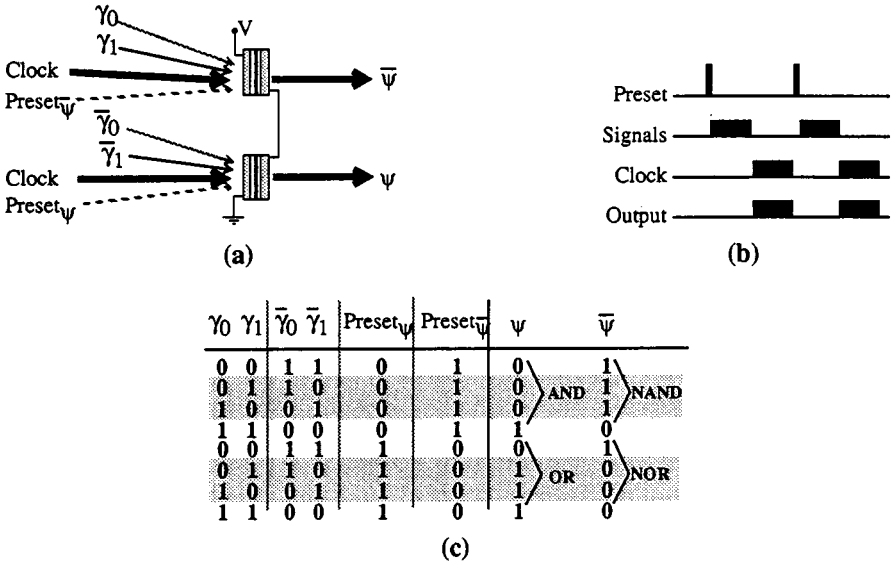


FIGURE 79.27 Logic using S-SEED devices.

its complement, $\bar{\gamma}_i$ (Reset), and a clock signal. To operate the S-SEED the γ_i and $\bar{\gamma}_i$ inputs are also separated in time from the clock inputs as shown in Fig. 79.25(b). The γ_i and $\bar{\gamma}_i$ inputs, which represent the incoming data and its complement, are used to set the state of the device. When $\bar{\gamma}_i > \gamma_i$, the S-SEED will enter a state where the upper MQW *pin* diode will be reflective, forcing the lower diode to be absorptive. When $\gamma_i > \bar{\gamma}_i$ the opposite condition will occur. Low switching intensities are able to change the device's state when the clock signals are not present. After the device has been put into its proper state, the clock beams are applied to both inputs. The ratio of the power between the two clock beams should be approximately one, which will prevent the device from changing states. These higher energy clock pulses, on reflection, will transmit the state of the device to the next stage of the system. Since the inputs γ_i and $\bar{\gamma}_i$ are low-intensity pulses and the clock signals are high-intensity pulses, a large differential gain may be achieved. This type of gain is referred to as time-sequential gain.

The operation of an S-SEED is determined by the power transfer characteristic shown in Fig. 79.26(c). The optical power reflected by the ψ_i window, when the clock signal is applied, is plotted against the ratio of the total optical signal power impinging on the γ_i and $\bar{\gamma}_i$ windows (when the clock signal is not applied). Assuming the clock power incident on both signal windows, γ_i and $\bar{\gamma}_i$, the output power is proportional to the reflectivity, R_i. The ratio of the input signal powers is defined as the input contrast ratio $C_{in} = P_\gamma/P_{\bar{\gamma}i}$. As C_{in} is increased from zero, the reflectivity of the ψ_i window switches from a low value, R_1, to a high value, R_2, at a C_{in} value approximately equal to the ratio of the absorbances of the two optical windows: $T = (1 - R_1)/(1 - R_2)$. Simultaneously, the reflectivity of the other window (ψ_i) switches from R_2 to R_1. The return transition point (ideally) occurs when $C_{in} = (1 - R_2)/(1 - R_1) = 1/T$. The ratio of the two reflectivities, R_2/R_1, is the output contrast, C_{out}. Typical measured values of the preceding parameters include $C_{out} = 3.2$, $T = 1.4$, $R_2 = 50\%$ and $R_1 = 15\%$. The switching energy for these devices has been measured at ~7 fJ/μm^2.

The S-SEED is also capable of performing optical logic functions such as NOR, OR, NAND, and AND. The inputs will also be differential, thus still avoiding any critical biasing of the device. A method of achieving logic gate operation is shown in Fig. 79.27. The logic level of the inputs will be defined as the ratio of the optical power on the two optical windows. When the power of the signal incident on the γ_i input is greater than the power of the signal on the $\bar{\gamma}_i$ input, a logic "1" will be present on the input. On the other hand, when the power of the signal incident on the γ_i input is less than the power of the signal on the $\bar{\gamma}_i$ input, a logic "0" will be incident on the input.

For the noninverting gates, OR and AND, we can represent the output logic level by the power of the signal coming from the ψ output relative to the power of the signal coming from the $\bar{\psi}$ output. As before, when the power of the signal leaving the ψ output is greater than the power of the signal leaving the $\bar{\psi}$ output, a logic

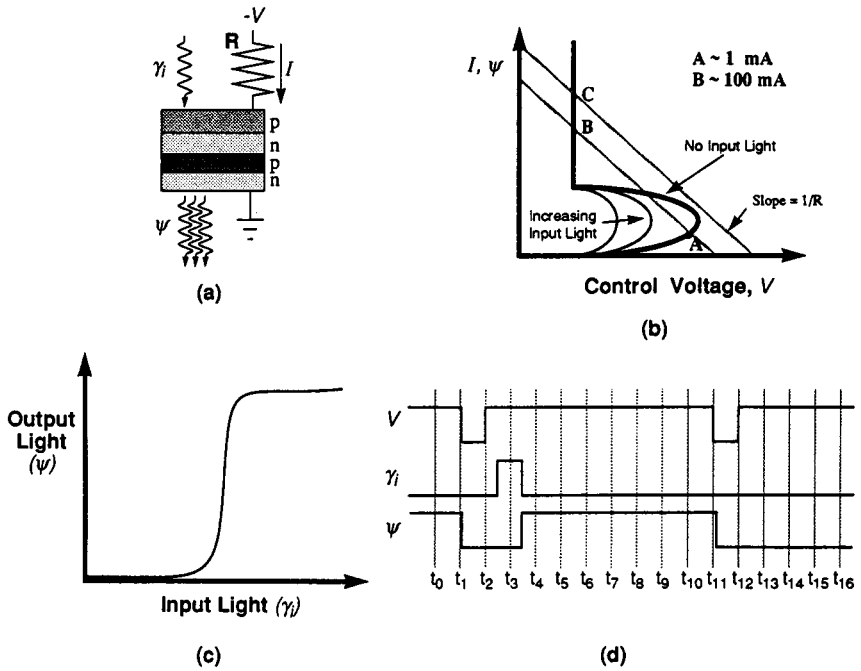


FIGURE 79.28 *pnpn* devices: (a) basic structure, (b) voltage/output characteristics, (c) input/output characteristics, and (d) timing diagram of device operation.

"1" will be represented on the output. To achieve AND operation, the device is initially set to its "off" or logic "0" state (i.e., ψ low and $\overline{\psi}$ high) with preset pulse, $Preset_{\psi}$ incident on only one *pin* diode as shown in Fig. 79.27. If both input signals have logic levels of "1" (i.e., set = 1, reset = 0), then the S-SEED AND gate is set to its "on" state. For any other input combination, there is no change of state, resulting in AND operation. After the signal beams determine the state of the device, the clock beams are then set high to read out the state of the AND gate. For NAND operation, the logic level is represented by the power of the $\overline{\psi}$ output signal relative to the power of the ψ output signal. That is, when the power of the signal leaving the $\overline{\psi}$ output is greater than the power of the signal leaving the ψ output, a logic "1" is present on the output. The operation of the OR and NOR gates is identical to the AND and NAND gates, except that preset pulse $Preset_{\psi}$ is used instead of the preset pulse $Preset_{\overline{\psi}}$. Thus, a single array of devices can perform any or all of the four logic functions and memory functions with the proper optical interconnections and preset pulse routing.

Detect/Emit Devices

Detect/emit devices are optoelectronic structures that detect the incoming signal, process the information, and then transfer the information off the device through the use of active light emitters such as LEDs or lasers. An example of a detect/emit device is the "thyristor-like" *pnpn* device as illustrated in Fig. 79.28(a). It is a digital active optical logic device with "high" and "low" light-emitting optical output states corresponding to electrical states of high impedance (low optical output) or low impedance (high optical output). The device can be driven from one state to the other either electrically or optically. The optical output can be either a lasing output or light-emitting diode output. There are several devices that are based on this general structure. The double heterostructure optoelectronic switch (DOES) is actually an *npnp* structure that is designed as an integrated bipolar inversion channel heterojunction field-effect transistor (BICFET) phototransistor controlling and driving either an LED or microlaser output. The second device is a *pnpn* structure referred to as a vertical-to-surface transmission electrophotonic device (VSTEP).

The operation of these *pnpn* structures can be illustrated through the use of load lines. For the simplest device, the load consists of a resistor and a power supply. In Fig. 79.28(b), we see that for small amounts of light, the device will be at point *A*. Point *A* is in a region of high electrical impedance with little or no optical output. As the input light intensity increases, there is no longer an intersection point near *A* and the device

will switch to point *B* [Fig. 79.28(c)]. At this point the electrical impedance is low and light is emitted. When the input light is removed, the operating point returns via the origin to point *A* by momentarily setting the electrical supply to zero [Fig. 79.28(d)]. These devices can be used as either optical OR or AND gates using a bias beam and several other optical inputs.

The device can also be electrically switched. Assuming no input intensity, the initial operating point is at point *A*. By increasing the power supply voltage, the device will switch to point *C*. Point *C* like point *B* is in the region of light emission and low impedance. To turn off the device, the power supply must then be reduced to zero, after which it may be increased up to some voltage where switching occurs.

A differential *pnpn* device made by simply connecting two *pnpn* devices in parallel and connecting that combination in series with a resistive load is illustrated in Fig. 79.29. The operation of the device can be described as follows. When the device is biased below threshold, that is, with the device unilluminated, both optical switches are "off." When the device is illuminated, the one with the highest power is switched "on." The increase in current leads to a voltage drop across the resistor which in turn leads to a lowering of the voltage across both optical switches. Therefore, the one with the lower input cannot be switched "on." Unless both inputs were illuminated with precisely the same power and both devices had identical characteristics (both of these are impossible), only one of the two optical switches will emit light.

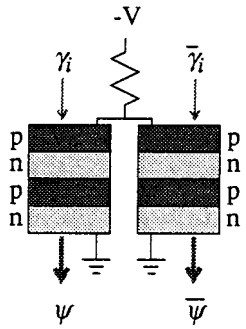


FIGURE 79.29 Differential *pnpn* device.

The required input optical switching energy density can be quite low if the device without light is biased critically just below threshold. Since incoherent light from an LED cannot be effectively collected from small devices or focused onto small devices, a lasing *pnpn* is needed. Microlaser-based structures are also required to reduce the total power dissipation to acceptable levels. Surface-emitting microlasers provide an ideal laser because of their small size, single-mode operation, and low thresholds. The surface-emitting microlasers consist of two AlAs/AlGaAs dielectric mirrors with a thin active layer in between. This active layer typically consists of one or a few MQWs. The material can be etched vertically into small posts, typically 1–5 μm in diameter. Thresholds are typically on the order of milliwatts.

The switching speed of these devices is limited by the time it takes the photogenerated carriers to diffuse into the light-emitting region. Optical turn-off times are also limited by the RC time constant. For devices made so far, the RC time constants are in the range of 1–10 ns, and optical switch-on times were ~10 ns. Performance of the devices is expected to improve as the areas are reduced; switching times comparable to the best electronic devices (~10 ps) are possible, although the optical turn-on times of at least the surface-emitting LED devices will continue to be slower since this time is determined by diffusion effects and not device capacitance and resistance. Lasing devices should offer improved optical turn-on times.

Another approach to active devices is to combine lasers/modulators with electronics and photodiodes as has been proposed for optical interconnections of electronic circuits. Since the logic function is implemented with electronic circuitry, any relatively complex functionality can be achieved. Several examples of logic gates have been made using GaAs circuitry and light-emitting diodes. Again surface-emitting microlasers provide an ideal emitter for this purpose, because of their small size and low threshold current. However, the integration of these lasers with the required electrical components has yet to be demonstrated.

Limitations

In the normal operating regions of most devices, a fixed amount of energy, the switching energy, is required to make them change states. This switching energy can be used to establish a relationship between both the switching speed and the power required to change the state of the device. Since the power required to switch the device is equal to the switching energy divided by the switching time, a shorter switching time will require more power. As an example, for a photonic device with an area of 100 μm^2 and a switching energy of 1 $fJ/\mu m^2$ to change states in 1 ps requires 100 mW of power instead of the 100 μW that would be required if the device were to switch at 1 ns. Thus, for high power signals the device will change states rapidly, while low power signals yield a slow switching response.

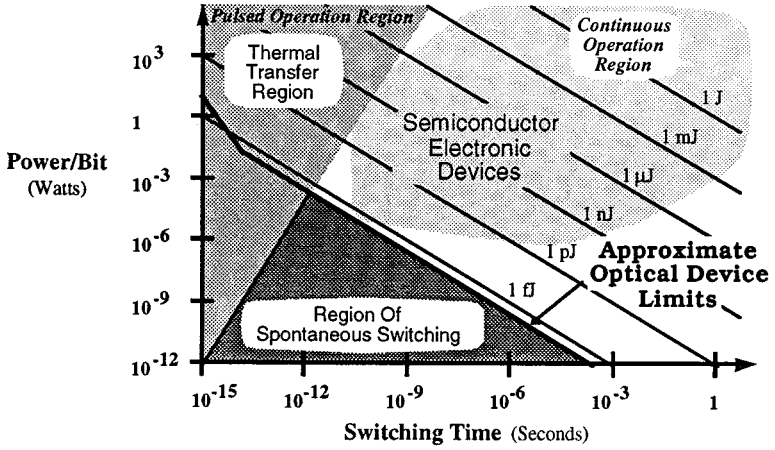


FIGURE 79.30 Fundamental limitations of optical logic devices.

Some approximate limits on the possible switching times of a given device, whether optical or electrical, are illustrated in Fig. 79.30. In this figure the time required to switch the state of a device is on the abscissa while the power/bit required to switch the state of a device is on the ordinate. The region of spontaneous switching is the result of the background thermal energy that is present in a device. If the switching energy for the device is too low, the background thermal energy will cause the device to change states spontaneously. To prevent these random transitions in the state of a device, the switching energy required by the device must be much larger than the background thermal energy. To be able to differentiate statistically between two states, this figure assumes that each bit should be composed of at least 1000 photons. Thus, the total energy of 1000 photons sets the approximate boundary for this region of spontaneous switching. For a wavelength of 850 nm, this implies a minimum switching energy on the order of 0.2 fJ. For the thermal transfer region, it is assumed that for continuous operation, the thermal energy present in the device cannot be removed any faster than 100 W/cm^2 (1 μW/μm^2). There has been some work done to indicate that this value could be as large as 1000 W/cm^2. This region also assumes that there will be no more than an increase of 20°C in the temperature of the device. Devices can be operated in this region using a pulsed rather than continuous mode of operation. Thus, high energy pulses can be used if sufficient time is allowed between pulses to allow the absorbed energy to be removed from the devices. The cloud in Fig. 79.30 represents the performance capabilities of current electronic devices. This figure illustrates that optical devices will not be able to switch states orders of magnitude faster than electronic devices when the system is in a continuous rather than a pulsed mode of operation. There are, however, other considerations in the use of photonic switching devices than how fast a single device can change states. Assume that several physically small devices need to be interconnected so that the state information of one device can be used to control the state of another device. To communicate this information, there needs to be some type of interconnection with a large bandwidth that will allow short pulses to travel between the separated devices. Fortunately, the optical domain can support the bandwidth necessary to allow bit rates in excess of 100 Gb/s, which will allow high-speed communication between these individual switching devices. In the electrical domain, the communications bandwidth between two or more devices is limited by the resistance, capacitance, and inductance of the path between the different devices. Therefore, even though photonic devices cannot switch orders of magnitude faster than their electronic counterparts, the communications capability or transmission bandwidth present in the optical domain should allow higher data rate systems than are possible in the electrical domain.

Defining Terms

Light-amplifying optical switch (LAOS): Vertically integrated heterojunction phototransistor and light-emitting diode which has latching thyristor-type current-voltage characteristics.

Liquid-crystal light valve (LCLV): Optical controlled spatial light modulator based on liquid crystals.

Multiple quantum well (MQW): Collection of alternating thin layers of semiconductors (e.g., GaAs and AlGaAs) that results in strong peaks in the absorption spectrum which can be shifted with an applied voltage.

Nonlinear Fabry-Perot (NLFP): Fabry-Perot etalon or interferometer that has an optically nonlinear medium in its cavity.

Optical logic etalon (OLE): Pulsed nonlinear Fabry-Perot etalon that requires two wavelengths (λ_1 = signal, λ_2 = clock).

PLZT/Si: Technology based on conventional silicon electronics using silicon detectors for the device inputs and PLZT modulators for the outputs.

Sagnac logic gate: An all-optical gate based on a Sagnac interferometer. A Sagnac interferometer is composed of two coils of optical fiber arranged so that light from a single source travels clockwise in one and counterclockwise in the other.

SEED technology: Any device based on multiple quantum well (MQW) modulators.

Soliton: Any isolated wave that propagates without dispersion of energy.

Surface-emitting laser logic (CELL): Device that integrates a phototransistor with a low threshold vertical-cavity surface-emitting laser.

Related Topic

43.1 Introduction

References

H. S. Hinton, "Architectural consideration for photonic switching networks," *IEEE Journal on Selected Areas in Communications,* 6, 1988.

M. N. Islam et al., "Ultrafast all-optical fiber-soliton gates," in *Proceedings on Photonic Switching,* vol. 8, H. S. Hinton and J. W. Goodman, eds., Washington, D.C.: Optical Society of America, 1991, pp. 98–104.

J. L. Jewell et al., "Use of a single nonlinear Fabry-Perot etalon as optical logic gates," *Applied Physics Letters,* 44, 1984.

K. Kasahara et al., "Double heterostructure optoelectronic switch as a dynamic memory with lowpower consumption," *Applied Physics Letters,* 52, 1988.

A. L. Lentine et al., "Symmetric self-electrooptic effect device: Optical set-reset latch, differential logic gate, and differential modulator/detector," *IEEE Journal of Quantum Electronics,* 25, 1989.

J. E. Midwinter, "Digital optics, optical logic or smart interconnect or optical logic," *Physics in Technology,* 19, 1988.

D. A. B. Miller, "Quantum well self-electro-optic effect devices," *Optical and Quantum Electronics,* 22, 1990.

P. W. Smith, "On the physical limits of digital optical switching and logic elements," *Bell System Technical Journal,* 61, 1982.

S. D. Smith, "Optical bistability, photonic logic, and optical computation," *Applied Optics,* 25, 1986.

G. W. Taylor et al., "A new double heterostructure optoelectronic device using molecular beam epitaxy," *Journal of Applied Physics,* 59, 1986.

Further Information

Books which cover this material in more detail include:

H. H. Arsenault, T. Szoplik, and B. Macukow, *Optical Processing and Computing,* New York: Academic Press, 1989.

H. M. Gibbs, *Optical Bistability: Controlling Light with Light,* New York: Academic Press, 1985.

M. N. Islam, *Ultrafast Fiber Switching Devices and Systems,* London: Cambridge University Press, 1992.

A. D. McAulay, *Optical Computer Architectures,* New York: John Wiley, 1991.

B. S. Wherrett and F. A. P. Tooley, *Optical Computing,* Scottish Universities Summer School in Physics, 1989.

Pricer, W.D., Katz, R.H., Lee, P.A., Mansuripur, M. “Memory Devices”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

80

Memory Devices

W. David Pricer
IBM

Randy H. Katz
University of California, Berkeley

Peter A. Lee
Department of Trade and Industry, London

M. Mansuripur
University of Arizona, Tucson

80.1 Integrated Circuits (RAM, ROM)
Dynamic RAMs (DRAMs) • Static RAMs (SRAMs) • Nonvolatile Programmable Memories • Read-Only Memories (ROMs)

80.2 Basic Disk System Architectures
Basic Magnetic Disk System Architecture • Characterization of I/O Workloads • Extensions to Conventional Disk Architectures

80.3 Magnetic Tape
A Brief Historical Review • Introduction • Magnetic Tape • Tape Format • Recording Modes

80.4 Magneto-Optical Disk Data Storage
Preliminaries and Basic Definitions • The Optical Path • Automatic Focusing • Automatic Tracking • Thermomagnetic Recording Process • Magneto-Optical Readout • Materials of Magneto-Optical Data Storage

80.1 Integrated Circuits (RAM, ROM)

W. David Pricer

The major forms of semiconductor memory in descending order of present economic importance are

1. Dynamic Random-Access Memories (DRAMs)
2. Static Random-Access Memories (SRAMs)
3. Nonvolatile Programmable Memories (PROMs, EEPROMs, EAROMs, EPROMs)
4. Read-Only Memories (ROMs)

DRAMs and SRAMs differ little in their applications. DRAMs are distinguished from SRAMs in that no bistable electronic circuit internal to the storage cell maintains the information. Instead DRAM information is stored "dynamically" as charge on a capacitor. All modern designs feature one field-effect transistor (FET) to access the information for both reading and writing and a thin film capacitor for information storage. SRAMs maintain their bistability, so long as power is applied, by a cross-coupled pair of inverters within each storage cell. Almost always two additional transistors serve to access the internal nodes for reading and writing. Most modern cell designs are CMOS, with two P-channel and four N-channel FETs.

Programmable memories operate much like read-only memories with the important attribute that they can be programmed at least once, and some can be reprogrammed a million times or more. Storage is almost always by means of a floating-gate FET. Information in such storage cells is not indefinitely nonvolatile. The discharge time constant is on the order of ten years. ROMs are generally programmed by a custom information mask within the fabrication sequence. As the name implies, information thence can only be read. The information thus stored is truly nonvolatile, even when power is removed. This is the most dense form of semiconductor storage (and the least flexible). Other forms of semiconductor memories, such as associative memories and charge-coupled devices, are used rarely.

Dynamic RAMs (DRAMs)

The universally used storage cell circuit of one transistor and one capacitor has remained unchanged for over 20 years. The physical implementation, however, has undergone much diversity and many refinements. The innovation in physical implementation is driven primarily by the need to maintain a nearly constant value of capacitance while the surface area of the cell has decreased. A nearly fixed value of capacitance is needed to meet two important design goals. The cell has no internal amplification. Once the information is accessed, the stored voltage is vastly attenuated by the much larger bit line capacitance (see Fig. 80.1). The resulting signal must be kept larger than the resolution limits of the sensing amplifier. DRAMs in particular are also sensitive to a problem called soft errors. These are typically initiated by atomic events such as the incidence of a single alpha particle. An alpha particle can cause a spurious signal of 50,000 electrons or more. All modern DRAM designs resolve this problem by constructing the capacitor in space out of the plane of the transistors (see Fig. 80.2 for examples). Placing the capacitor in space unusable for transistor fabrication has allowed great strides in DRAM density, generally at the expense of fabrication complexity. DRAM chip capacity has increased by about a factor of four every three years.

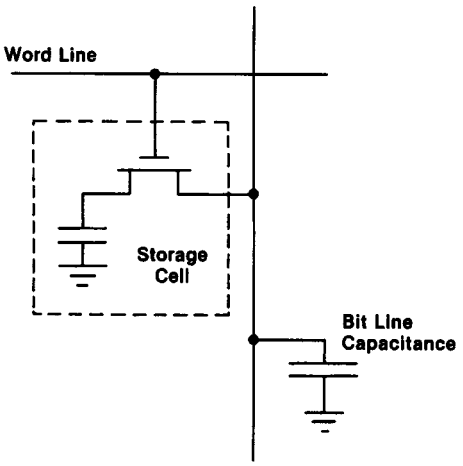
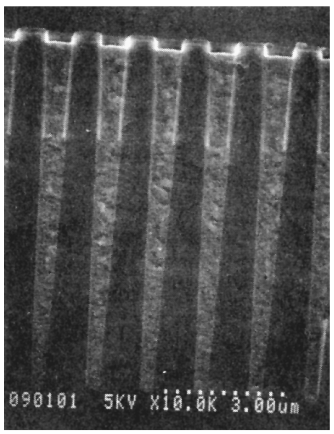


FIGURE 80.1 Cell and bit line capacitance.

DRAMs are somewhat slower than SRAMs. This relationship derives directly from the smaller signal available from DRAMs and from certain constraints put on the support circuitry by the DRAM array. DRAMs also



(a)

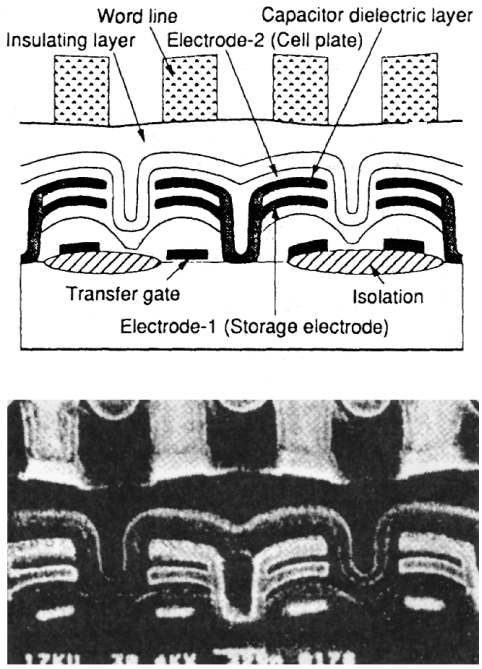
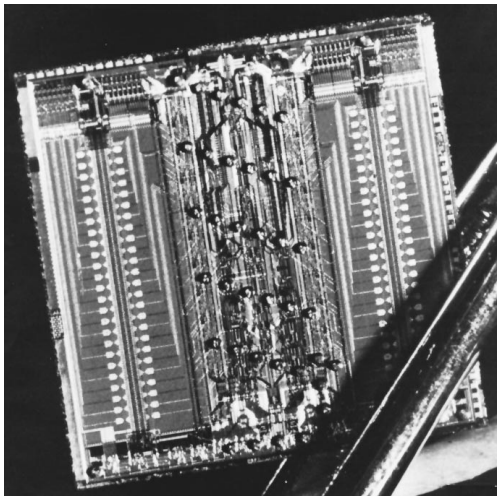


FIGURE 80.2 (a) Cross section of "trench capacitors" etched vertically into the semiconductor surface of a DRAM integrated circuit. (Courtesy of IBM.) (b) Cross section of "stacked" capacitors fabricated above the semiconductor surface of a DRAM integrated circuit. (*Source:* M. Taguchi et al., "A 40-ns 64-b parallel data bus architecture," *IEEE J. Solid State Circuits*, vol. 26, no. 11, p. 1495. © 1991 IEEE. With permission.)

The Revolution of Electronics Technology

The last three decades have witnessed a revolution in electrical and, especially, electronics technology. This revolution was paced by changes in solid-state electronics that greatly expanded capabilities while at the same time radically reduced costs. The entire field of electrical engineering has grown far beyond the boundaries that characterized it just a generation ago. Electrical engineers have become the creators and masters of the most pervasive technology of our time, with profound effects on society and on their profession.

The effects of the electronics revolution are complex. For the profession, the most obvious impact has been explosive growth. The increase in the number of students studying in the field continues to be dramatic and shows no signs of slowing. The electrical engineering community represents the largest single technical group in the world, and the members of the IEEE make up the world's largest engineering society. (Courtesy of the IEEE Center for the History of Electrical Engineering.)



This 64-kB random access memory chip, developed by IBM in 1978, was one of the densest of its time. It could store as many as 64,000 bits of information—roughly equivalent to 1,000 eight-letter words. (Photo courtesy of the IEEE Center for the History of Electrical Engineering.)

require periodic intervals to "refresh" lost charge from the capacitor. This charge is lost primarily across the semiconductor junctions and must be replenished every few milliseconds. The manufacturer usually supplies these "housekeeping" functions with on-chip circuitry.

Signal detection and amplification remain a critical focus of good DRAM design. Figure 80.3 illustrates an arrangement called a "folded bit line." This design cancels many of the noise sources originating in the array and decreases circuit sensitivity to manufacturing process variations. It also achieves a high ratio of storage cells per sense amplifier. Note the presence of the dummy cells, which create a reference signal midway between a "one" and a "zero" for the convenience of the sense amplifier. The stored reference voltage in this case is created by shorting two driven bit lines after one of the storage cells has been written.

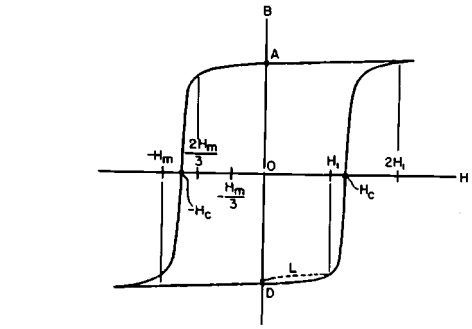
Large DRAM integrated circuit chips frequently provide other features that users may find useful. Faster access is provided between certain adjacent addresses, usually along a common word line. Some designs feature on-chip buffer memories, low standby power modes, or error correction circuitry. A few DRAM chips are designed to mesh with the constraints of particular applications such as image support for CRT displays. Some on-chip features are effectively hidden from the user. These may include redundant memory addresses which the maker activates by laser to improve manufacturing yield.

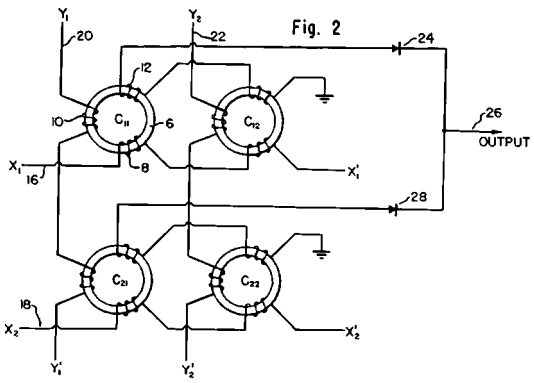
The largest single market for DRAMs is with microprocessors in personal computers. Rapid microprocessor performance improvements have led DRAM manufacturers to offer improvements especially designed for the "PC" environment. Extended Data Out mode (EDO) keeps the data accessed from a DRAM valid over a longer period of the DRAM cycle. EDO mode is intended to ease the synchronization problem between a DRAM and the increasingly higher speed microprocessor. Synchronous DRAM (SDRAM) allows the rapid sequential

Multicoordinate Digital Information Storage Device

Jay W. Forrester
Patented February 28, 1956
#2,736,880

Up to this time, digital data storage was generally done by encoding binary data on rotating magnetic drums or other means where data had to be stored and retrieved sequentially. This patent describes a system whereby data could be stored and retrieved randomly by a simple addressing scheme. It used tiny doughnut-shaped ferromagnetic cores with windings to magnetically polarize the material in one direction or the other. This was about one hundred times faster than rotating drums and took up perhaps 2% of the volume. A 4-Kbyte core memory module would take up about 60 cubic inches and could access data in less than one millisecond. Random access memory (RAM) was born. Core memory (as it has become known) was non-volatile; that is, the information would not be lost when power was cut. Modern non-volatile "flash" memory is yet again thousands of times faster and achieves data density of over 100,000 times greater than the breakthrough magnetic core memory described by Forrester. (Copyright © 1995, DewRay Products, Inc. Used with permission.)





transfer of large blocks of data between the microprocessor and the DRAM without extensive signal "handshaking". While SDRAMs do nothing to improve the access time to first data, they greatly improve the "bandwidth" between microprocessor and DRAM.

Static RAMs (SRAMs)

The primary advantages of SRAMs as compared to DRAMs are high speed and ease of use. In addition, SRAMs fabricated in CMOS technology exhibit extremely low standby power. This later feature is effectively used in much portable equipment like pocket calculators. Bipolar SRAMs are generally faster but less dense than FET versions. Figure 80.4 illustrates two cells. SRAM performance is dominated by the speed of the support circuits, leading some manufacturers to design bipolar support circuits to FET arrays.

Bipolar designs frequently incorporate circuit consolidation unavailable in FET technology, such as the multiemitter cell shown in Fig. 80.4(a). Here one of the two lower emitters is normally forward biased, turning one inverter on and the other off for bistability. The upper emitters can be used either to extract a differential signal

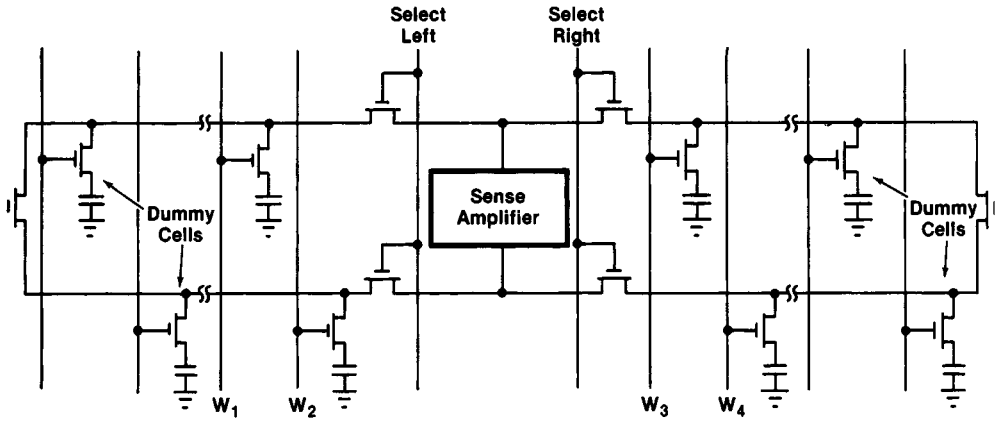


FIGURE 80.3 Folded bit line array.

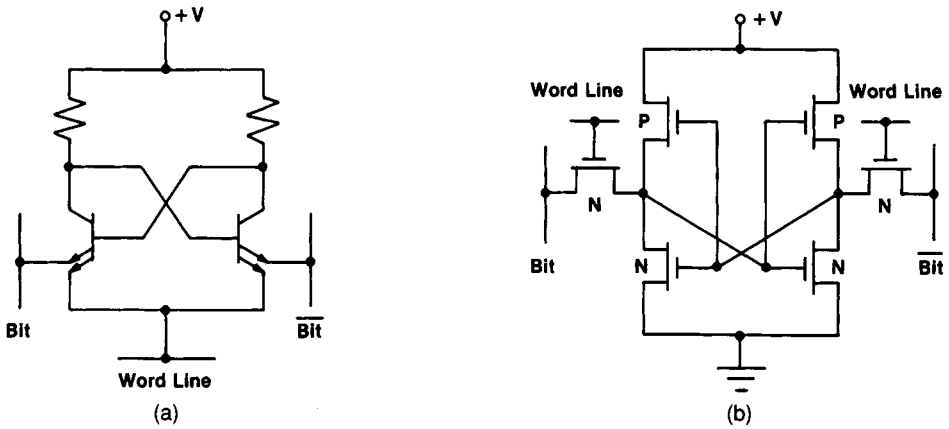


FIGURE 80.4 (a) Bipolar SRAM cell. (b) CMOS SRAM cell.

or to discharge one collector towards ground in order to write the cell. The word line is pulsed positive to both read and write the cell.

A few RAMs use **polysilicon** load resistors of very high resistance value in place of the two P-channel transistors shown in Fig. 80.4(b). Most are full CMOS designs like the one shown. Sometimes the P-channel transistors are constructed by thin film techniques and are physically placed over the N-channel transistors to improve density. When both P- and N-channel transistors are fabricated in the same plane of the single-crystal semiconductor, the standby current can be extremely low. Typically this can be microamps for megabit chips. The low standby current is possible because each cell sources and sinks only that current needed to overcome the actual node leakage within the cell.

Selecting the proper transconductance for each transistor is an important focus of the designer. The accessing transistors should be large enough to extract a large read signal but insufficiently large to disturb the stored information. During the write operation, these same transistors must be capable of overriding the current drive of at least one of the internal CMOS inverters.

The superior performance of SRAMs derives from their larger signal and the absence of a need to refresh the stored information as in a DRAM. As a result, SRAMs need fewer sense amplifiers. Likewise these amplifiers are not constrained to match the cell pitch of the array. SRAM design engineers have exploited this freedom to realize higher-performance sense amplifiers.

Practical SRAM designs routinely achieve access times of a few nanoseconds to a few tens of nanoseconds. Cycle time typically equals access time, and in at least one pipelined design, cycle time is actually less than access time.

SRAM integrated circuit chips have fewer special on-chip features than DRAM chips, primarily because no special performance enhancements are needed. By contrast, many other integrated circuit chips feature on-chip SRAMs. For example, many **ASICs (application-specific integrated circuits)** feature on-chip RAMs because of their low power and ease of use.

All modern microprocessors include one or more on-chip "cache" SRAM memories which provide a high speed link between processor and memory.

Nonvolatile Programmable Memories

A few nonvolatile memories are programmable just once. These have arrays of diodes or transistors with fuses or **antifuses** in series with each semiconductor cross point. Aluminum, titanium, tungsten, platinum silicide, and polysilicon have all been successfully used as fuse technology (see Fig. 80.5).

Most nonvolatile cells rely on trapped charge stored on a floating gate in an FET. These can be rewritten many times. The trapped charge is subject to very long term leakage, on the order of ten years. The number of times the cell may be rewritten is limited by programming stress-induced degradation of the dielectric. Charge reaches the floating gate either by **tunneling** or by **avalanche injection** from a region near the drain. Both phenomena are induced by over-voltage conditions and hence the degradation after repeated erase/write cycles. Commercially available chips typically promise 100 to 100,000 write cycles. Erasure of charge from the floating gate may be by tunneling or by exposure to ultraviolet light. Asperities on the polysilicon gate and silicon-rich oxide have both been shown to enhance charging and discharging of the gate. The nomenclature used is not entirely consistent throughout the industry. However, EPROM is generally used to describe cells which are electronically written but UV erased. EEPROM is used to describe cells which are electronically both written and erased.

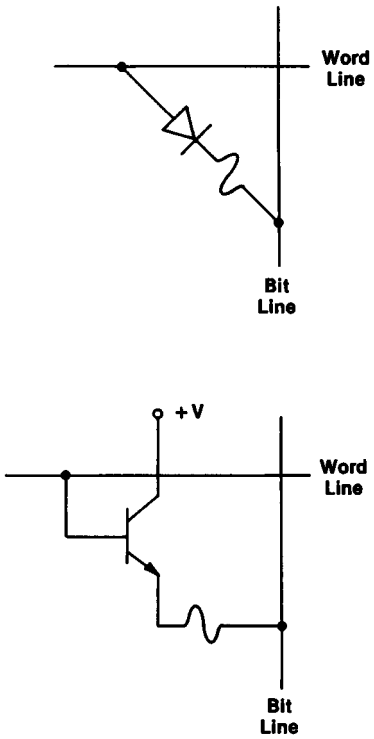


FIGURE 80.5 PROM cells.

Cells are of either a two- or a one-transistor design. Where two transistors are used, the second transistor is a conventional **enhancement mode** transistor (see Fig. 80.6). The second transistor works to minimize the disturb of unselected cells. It also removes some constraints on the writing limits of the programmable transistor, which in one state may be **depletion mode**. The two transistors in series then assume the threshold of the second (enhancement) transistor, or a very high threshold as determined by the programmable transistor. Some designs are so cleverly integrated that the features of the two transistors are merged.

Flash EEPROMs describe a family of single-transistor cell EPPROMs. Cell sizes are about half that of two-transistor EEPROMs, an important economic consideration. Care must be taken that these cells are not programmed into the depletion mode. An array of depletion mode cells would confound the read operation by providing multiple signal paths. Programming to enhancement only thresholds can be accomplished by a sequence of partial program and then monitor subcycles, until the threshold is brought to compliance with specification limits. Flash EEPROMs require bulk erasure of large portions of the array.

NVRAM is a term used to describe a SRAM or DRAM with nonvolatile circuit elements. The cell is built to operate as a RAM with normal power applied. On command or with power failure imminent, the EEPROM elements can be activated to capture the last state of the RAM cell. The nonvolatile information is restored to a SRAM cell by normal internal cell regeneration when power is restored.

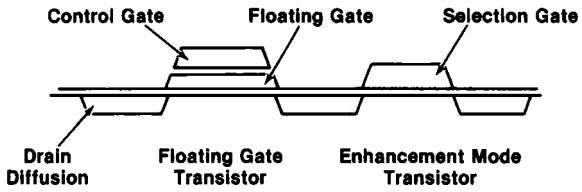


FIGURE 80.6 Cross section of two-transistor EEPROM cells.

Read-Only Memories (ROMs)

ROMs are the only form of semiconductor storage which is permanently nonvolatile. Information is retained without power applied, and there is not even very gradual information loss as in EEPROMs. It is also the most dense form of semiconductor storage. ROMs are, however, less used than RAMs or EEPROMs. ROMs must be personalized by a mask in the fabrication process. This method is cumbersome and expensive unless many identical parts are to be made. Furthermore it seems much "permanent" information is not really permanent and must be occasionally updated.

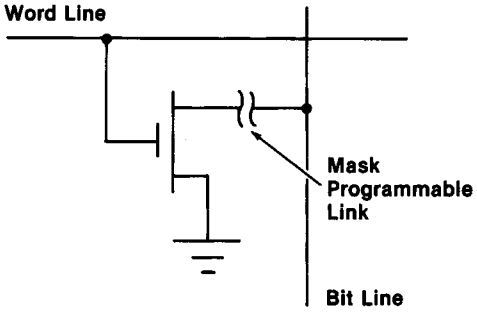


FIGURE 80.7 ROM cell.

ROM cells can be formed as diodes or transistors at every intersection of the word and bit lines of a ROM array (see Fig. 80.7). One of the masks in the chip fabrication process programs which of these devices will be active. Clever layout and circuit techniques may be used to obtain further density. Two such techniques are illustrated in Figs. 80.8 and 80.9. The X array shares bit and virtual ground lines. The AND array places many ROM cells in series. Each of these series AND ROM cells is either

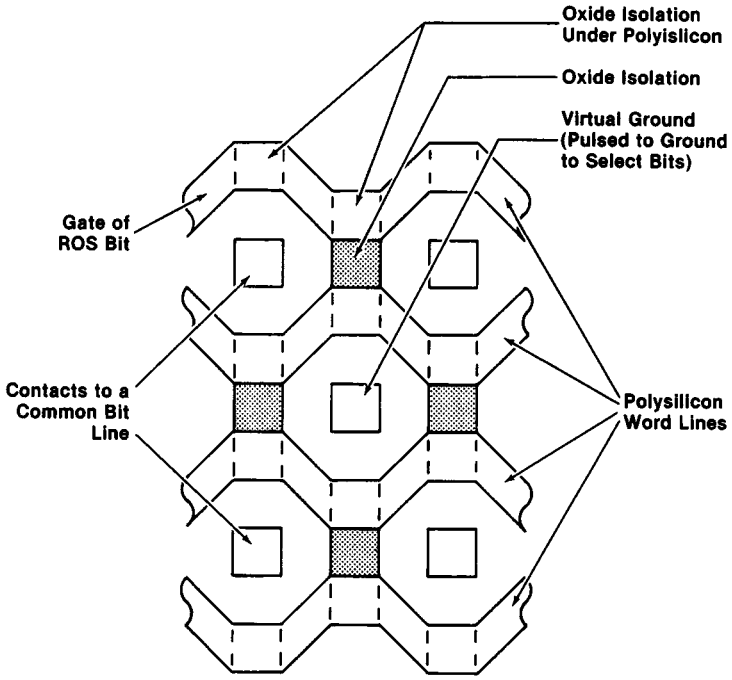


FIGURE 80.8 Layout of ROS X array.

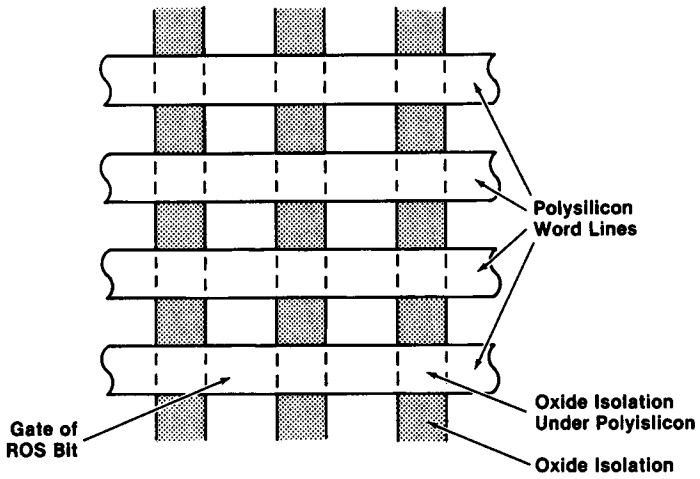


FIGURE 80.9 Layout of ROS AND array.

an enhancement or a depletion channel of an FET. Sensing is accomplished by pulsing the gates of all series cells positive except the gate which is to be interrogated. Current will flow through all series channels only if the interrogated channel is depletion mode.

ROM applications include look-up tables, machine-level instruction code for computers, and small arrays used to perform logic (see PLA in Section 81.4 of this handbook).

Defining Terms

Antifuse: A fuse-like device which when activated becomes low impedance.

Application-specific integrated circuits (ASICs): Integrated circuits specifically designed for one particular application.

Avalanche injection: The physics whereby electrons highly energized in avalanche current at a semiconductor junction can penetrate into a dielectric.

Depletion mode: An FET which is on when zero volts bias is applied from gate to source.

Enhancement mode: An FET which is off when zero volts bias is applied from gate to source.

Polysilicon: Silicon in polycrystalline form.

Tunneling: A physical phenomenon whereby an electron can move instantly through a thin dielectric.

Related Topic

25.3 Application-Specific Integrated Circuits

References

H. Kalter et al., "A 50 nsec 16 Mb DRAM with 10 nsec data rate and on-chip ECC," *IEEE Journal of Solid-State Circuits*, vol. SC 25, no. 5, 1990.

H. Kato, "A 9 nsec 4 Mb BiCMOS SRAM with 3.3 V operation," *Digest of Technical Papers ISSCC*, vol 35, 1992.

H. Kawague, and N. Tsuji, "Minimum size ROM structure compatible with silicon-gate E/D MOS LSI," *IEEE Journal of Solid State Circuits,* vol. SC 11, no. 2, 1976.

Further Information

W. Donoghue et al., "A 256K H CMOS ROM using a four state cell approach," *IEEE Journal of Solid-State Circuits*, vol. SC20, no. 2, 1985.

D. Frohmann-Bentchkowsky, "A fully decoded 2048 bit electronically programmable MOS-ROM," *Digest of Technical Papers ISSCC*, vol. 14, 1971.
L. A. Glasser and D. W. Dobberpuhl, *The Design and Analysis of VLSI Circuits*, Reading, Mass.: Addison-Wesley, 1985.
F. Masuoka, "Are you ready of the next generation dynamic RAM chips," *IEEE Spectrum Magazine*, vol. 27, no. 11, 1990.
R. D. Pashley and S. K. Lai, "Flash memories: The best of two worlds," *IEEE Spectrum Magazine*, vol. 26, no. 12, 1989.

80.2 Basic Disk System Architectures

Randy H. Katz

Architects of high-performance computers have long been forced to acknowledge the existence of a large gap between the speed of the CPU and the speed of its attached I/O devices. A number of techniques have been developed in an attempt to narrow this gap, and we shall review them in this chapter.

A key measure of magnetic disk technology is the growth in the maximum number of bits that can be stored per square inch, i.e., the bits per inch in a disk **track** times the number of tracks per inch of media. Called MAD, for **maximal areal density**, the "First Law in Disk Density" predicts [Frank, 1987]:

$$MAD = 10^{(\text{Year-}1971)/10} \tag{80.1}$$

This is plotted against several real disk products in Fig. 80.10. Magnetic disk technology has doubled capacity and halved price every three years, in line with the growth rate of semiconductor memory. Between 1967 and 1979 the growth in disk capacity of the average IBM data processing system more than kept up with its growth in main memory, maintaining a ratio of 1000:1 between disk capacity and physical memory size [Stevens, 1981].

In contrast to primary memory technologies, the performance of conventional magnetic disks has improved only modestly. These *mechanical* devices, the elements of which are described in more detail in the next section, are dominated by seek and rotation delays: from 1971 to 1981, the raw seek time for a high-end IBM disk improved by only a factor of two while the rotation time did not change [Harker et al., 1981]. Greater recording density translates into a higher transfer rate once the information is located, and extra positioning actuators for the read/write **heads** can reduce the average seek time, but the raw seek time only improved at a rate of 7% per year. This is to be compared to a doubling in processor power every year, a doubling in memory density every two years, and a doubling in disk density every three years. The gap between processor performance and disk speeds continues to widen, and there is no reason to expect a radical improvement in raw disk performance in the near future.

To maintain balance, computer systems have been using even larger main memories or solid-state disks to buffer some of the I/O activity. This may be an acceptable solution for applications whose I/O activity has locality of reference and for which volatility is not an issue, but applications dominated by a high rate of random requests for small pieces of data (e.g., transaction processing) or by a small number of sequential requests for massive amounts of data (e.g., supercomputer applications) face a serious performance limitation.

The rest of the chapter is organized as follows. In the next section, we will briefly review the fundamentals of disk system architecture. The third section describes the characteristics of the applications that demand high I/O system performance. Conventional ways to improve disk performance are discussed in the last section.

Basic Magnetic Disk System Architecture

We will review here the basic terminology of magnetic disk devices and controllers and then examine the disk subsystems of three manufacturers (IBM, Cray, and DEC). Throughout this section we are concerned with technologies that support random access, rather than sequential access (e.g., magnetic tape). A more detailed discussion, focusing on the structure of small dimension disk drives, can be found in Vasudeva [1988]. The basic concepts are illustrated in Fig. 80.11. A spindle consists of a collection of platters. Platters are metal disks

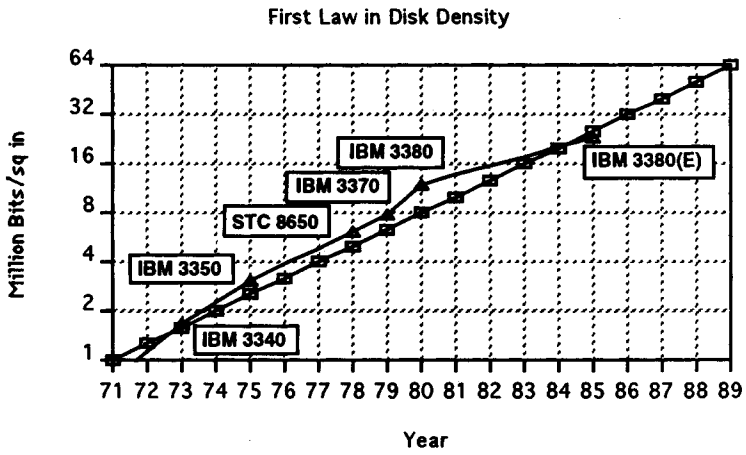


FIGURE 80.10 Maximal areal density law. Squares represent predicted density; triangles are the MAD reported for the indicated products.

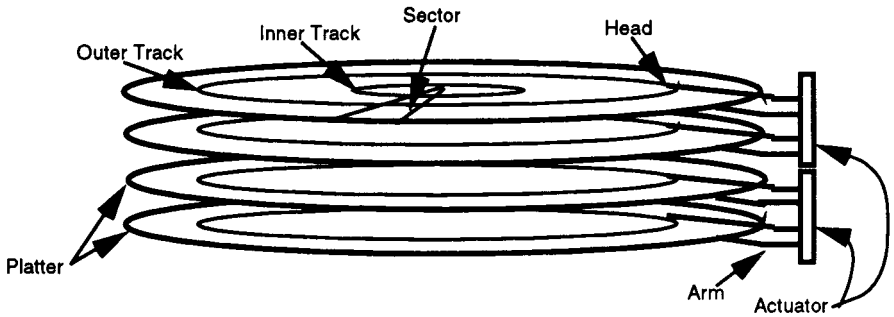


FIGURE 80.11 Disk terminology. Heads reside on arms which are positioned by actuators. Tracks are concentric rings on platters. A sector is the basic unit of read/write. A cylinder is a stack of tracks at one actuator position. An HDA is everything in the figure plus the air-tight casing. In some devices it is possible to transfer from multiple surfaces simultaneously. The collection of heads that participate in a single logical transfer that is spread over multiple surfaces is called a head group.

covered with a magnetic material for recording information. Each platter contains a number of circular recording *tracks*. A sector is a unit of a track that is physically read or written at the same time. In traditional magnetic disks, the constant angular rotation of the platters dictates that sectors on inner tracks are recorded more densely than sectors on the outer tracks. Thus, the platter can spin at a constant rate and the same amount of data can be recorded on the inner and outer tracks.[1] Some modern disks use zone recording techniques to more densely record data on the outer tracks, but this requires more sophisticated read/write electronics.

The read/write *head* is an electromagnet that produces switchable magnetic fields to read and record bit streams on a platter's track. It is associated with a disk **arm**, attached to an actuator. The head "flies" close to, but never touches, the rotating platter (except perhaps when powered down). This is the classical definition of a **Winchester disk.** The actuator is a mechanical assembly that positions the head electronics over the appropriate track. It is possible to have multiple read/write mechanisms per surface, e.g., multiple heads per arm—at one extreme, one could have a head-per-track position, that is, the disk equivalent of a magnetic drum—or

[1]Some optical disks use a technique called constant linear velocity (CLV), where the platter rotates at different speeds depending on the relative position of the track. This allows more data to be stored on the outer tracks than the inner tracks, but because it takes more delay to vary the speed of rotation, the technique is better suited to sequential rather than random access.

multiple arms per surface through multiple actuators. Due to costs and technical limitations, it is usually uneconomical to build a device with a large number of actuators and heads.

A **cylinder** is a stack of tracks at one actuator position. A **head disk assembly** (HDA) is the collection of platters, heads, arms, and actuators, plus the air-tight casing. A *disk drive* is an HDA plus all associated electronics. A *disk* might be a platter, an actuator, or a drive depending the context.

We can illustrate these concepts by describing two first-generation supercomputer disks, the Cray DD-19 and the CDC 819 [Bucher and Hayes, 1980]. These were state-of-the-art disks around 1980. Each disk has 40 recording surfaces (20 platters), 411 cylinders, and 18 (DD-19) or 20 (CDC 819) 512-byte sectors per track. Both disks possess a limited "parallel read-out" capability. A given data word is actually byte interleaved over four surfaces. Rather than a single set of read/write electronics for the actuator, these disks have four sets, so it is possible to read or write with four heads at a time. Four heads on adjacent arms are called a *head group*. A disk track is thus composed of the stacked recording tracks of four adjacent surfaces, and there are 10 tracks per cylinder, spread over 40 surfaces. The advances over the last decade can be illustrated by the Cray DD-49, which is a typical high-end supercomputer disk of today. It consists of 16 recording surfaces (9 platters), 886 cylinders, 42 4096-byte sectors per track, with 32 read/write heads organized into eight head groups, four groups on each of two independent actuators. Each actuator can sweep the entire range of tracks, and by "scheduling" the arms to position the actuator closest to the target track of the pending request, the average seek time can be reduced. The DD-49 has a capacity of 1.2 Gbytes of storage and can transfer at a sustained rate of 9.6 Mbytes/s.

A variety of standard and proprietary interfaces are defined for transferring the data recorded on the disk to or from the host. We concentrate on industry standards here. On the disk surface, information is represented as alternating polarities of magnetic fields. These signals need to be sensed, amplified, and decoded into synchronized pulses by the read electronics. For example, the pulse-level protocol ST506/412 standard describes the way pulses can be extracted from the alternating flux fields. The bit-level ESDI, SMD, and IPI-2 standards describe the bit encoding of signals. At the packet level, these bits must be aligned into bytes, error correcting codes need to be applied, and the extracted data must be delivered to the host. These "intelligent" standards include SCSI (small computer standard interface) and IPI-3.

The ST506 is a low-cost but primitive interface, most appropriate for interfacing floppy disks to personal computers and low-end workstations. For example, the controller must perform data separation on its own; this is not done for it by the disk device. As a result, its transfer rate is limited to 0.625 Mbytes/s. The SMD interface is higher performance and is used extensively in connecting disks to mainframe disk controllers. ESDI is similar, but geared more towards smaller disk systems. One of its innovations over the ST506 is its ability to specify a seek to a particular track number rather than requiring track positioning via step-by-step pulses. Its performance is in the range of 1.25–1.875 Mbytes/s. SCSI has so far been used primarily with workstations and minicomputers, but offers the highest degree of integration and intelligence. Implementations with performance at the level of 1.5–4 Mbytes/s are common. The newer IPI-3 standard has the advantages of SCSI, but provides even higher performance at a higher cost. It is beginning to make inroads into mainframe systems. However, because of the very widespread use of SCSI, many believe that SCSI-2, an extension of SCSI to wider signal paths, will become the de facto standard for high-performance small disks.

The connection pathway between the host and the disk device varies widely depending on the desired level of performance. A low-end workstation or personal computer would use a SCSI interface to directly connect the device to the host. A higher end file server or minicomputer would typically use a separate disk controller to manage several devices at the same time. These devices attach to the controller through SMD interfaces. It is the controller's responsibility to implement error checking and corrections and direct memory transfer to the host.

Mainframes tend to have more devices and more complex interconnection schemes to access them. In IBM terminology [Buzen and Shum, 1986], the *channel path*, i.e., the set of cables and associated electronics that transfer data and control information between an I/O device and main memory, consists of a *channel*, a *storage director*, and a *head of string* (see Fig. 80.12). The collection of disks that share the same pathway to the head of string is called a *string*.

In earlier IBM systems, a channel path and channel are essentially the same thing. The channel processor is the hardware that executes channel programs, which are fetched from the host's memory. A *subchannel* is the

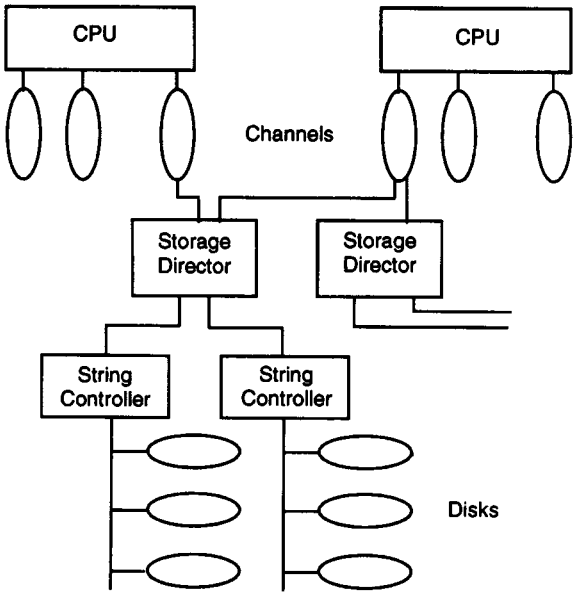


FIGURE 80.12 Host-to-device pathways. For large IBM mainframes, the connection between host and device must pass through a channel, storage director, and string controller. Note that multiple storage directors can be attached to a channel, multiple string controllers per storage director, and multiple devices per string controller. This multipathing approach makes it possible to share devices among hosts and to provide alternative pathways to better utilize the drives and controllers. While logically correct, the figure does not reflect the true physical components of high-end IBM systems (308X, 3090). The concept of channel has disappeared from these systems and has been replaced by a channel path.

execution environment of a channel program, similar to a process on a conventional CPU. Formerly, a subchannel was statically assigned for execution to a particular channel, but a major innovation in high-end IBM systems (308X and 3090) allows subchannels to be dynamically switched among channel paths. This is like allocating a process to a new processor within a multiprocessor system every time it is rescheduled for execution.

I/O program control statements, e.g., *transfer in channel*, are interpreted by the channel, while the storage director (also known as the *device controller* or *control unit*) handles seek and data-transfer requests. Besides these control functions, it may also perform certain datapath functions, such as error detection/correction and mapping between serial and parallel data. In response to requests from the storage director, the device will position the access mechanism, select the appropriate head, and perform the read or write. If the storage director is simply a control unit, then the datapath functions will be handled by the head of string (also known as a *string controller*).

To minimize the **latency** caused by copying into and out of buffers, the IBM I/O system uses little buffering between the device and memory.[1] In a high-performance environment, devices spend a good deal of time waiting for the pathway's resources to become free. These resources are used for time periods related to disk transfer speeds, measured in milliseconds. One possible method for improving utilization is to support disconnect/reconnect. A subchannel can connect to a device, issue a seek, disconnect to free the channel path for other requests, and reconnect later to perform the transfer when the seek is completed. Unfortunately, not all reconnects can be serviced immediately, because the control units are busy servicing other devices. These *RPS misses* (to be described in more detail in the next section) are a major source of delay in heavily utilized IBM storage subsystems [Buzen and Shum, 1987]. Performance can be further improved by providing multiple paths between memory and devices. To this purpose, IBM's high-end systems support *dynamic path reconnect*, a

[1]Only the most recent generation of storage directors (e.g., IBM 3880, 3990) incorporate disk caches, but care must be taken to avoid cache management-related delays [Buzen, 1982].

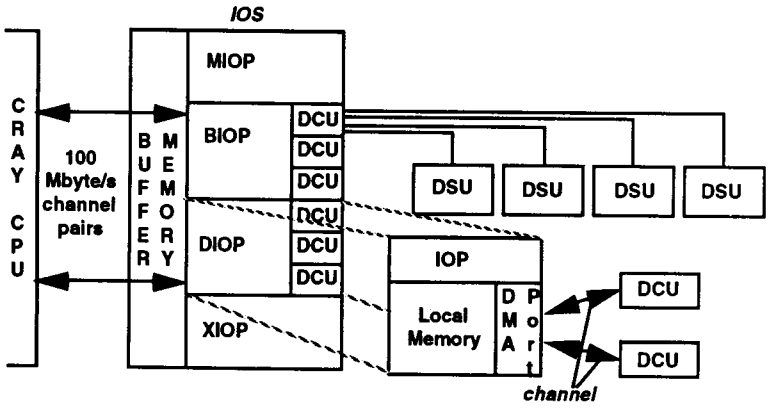


FIGURE 80.13 Elements of the Cray I/O system for the Y-MP. An IOS contains up to four IOPs. The MIOP connects to the operator workstation and performs mainly maintenance functions. The XIOP supports block multiplexing and is most appropriate for controlling relatively slow speed devices, such as tapes. The BIOP and DIOP are designed for controlling high-speed devices like disks. Up to four disk storage units (DSUs) can be attached through the disk control unit (DCU) to the IOP. Three DCUs can be connected to each of the BIOP and DIOP, leading to a total of 24 disks per IOS. The Y-MP can be configured with two IOSs, for a system total of 48 devices.

mechanism that allows a subchannel to change its channel path each time it cycles through a disconnect/reconnect with a given device. Rather than wait for its currently allocated path to become free, it can be assigned to another available path.

Turning to supercomputer I/O systems, we will now examine the I/O architecture of the Cray machines. Because the Cray I/O system (IOS) varies from model to model, the following discussion concentrates on the IOS found on the Cray X-MP and Y-MP [Cray, 1988]. In general, the IOS consists of two to four I/O processors (IOPs), each with its own local memory and sharing a common buffer memory with the other IOPs. The IOP is designed to be a simple, fast machine for controlling data transfers between devices and the central memory of the Cray main processors. Since it executes the control statements of an I/O program, it is not unlike the IBM channel processor in terms of its functionality, except that IO programs reside in its local memory rather than in the host's. An IOP's local memory is connected through a high-speed communications interface, called a *channel* in Cray terminology, to a disk control unit (DCU). A given port into the local memory can be time multiplexed among multiple channels. Data is transferred back and forth between devices and the main processors through the IOP's local memory, which is interfaced to central memory through a 100-Mbyte/s channel pair (one pathway for each direction of transfer).

The DCU provides the interface between the IOP and the disk drives and is similar in functionality to IBM's storage director. It oversees the data transfers between devices and the IOP's local memory, provides speed matching buffer storage, and transmits control signals and status information between the IOP and the devices. Disk storage units (DSUs) are attached to the DCU through point-to-point connections. The DSU contains the disk device and is responsible for dealing with its own defect management, by using a technique called sector slipping. Figure 80.13 summarizes the elements of the Cray I/O system.

Digital Equipment Corporation's high-end I/O strategy is described in terms of the digital storage architecture (DSA) and is embodied in system configurations such as the VAXCluster shared disk system (see Fig. 80.14). The architecture provides a rigorous definition of how storage subsystems and host computers interact. It achieves this by defining a client/server message-based model for I/O interaction based on device-independent interfaces [Massiglia, 1986; Kronenberg et al., 1986]. A *mass storage subsystem* is viewed at the architectural level as consisting of logical block machines capable of storing and retrieving fixed blocks of data, i.e., the I/O system supports the transfer of logical blocks between CPUs and devices given a logical block number. From the viewpoint of physical components, a subsystem consists of *controllers* which connect computers to *drives.*

The software architecture is divided into four levels: the *Operating System Client* (also called the Class Driver), the *Class Server* (Controller), the *Device Client* (Data Controller), and the *Device Server* (Device). The Disk Class Driver, resident on a host CPU, accepts requests for disk I/O service from applications, packages these

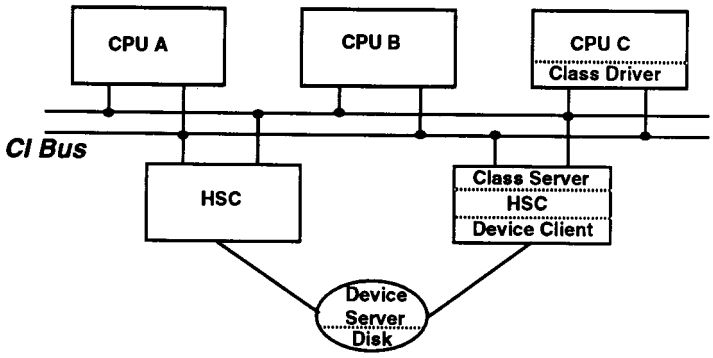


FIGURE 80.14 VAXCluster architecture. CPUs are connected to HSCs (hierarchical storage controllers) through a dual CI (computer interconnect) bus. Thirty-one hosts and 31 HSCs can be connected to a CI. Up to 32 disks can be connected to an HSC-70.

requests into messages, and transmits them via a communications interface (such as the *Computer Interconnect* port driver) to the Disk Class Server resident within a controller in the I/O subsystem. The command set supported by the Class Server includes such relatively device-independent operations as read logical block, write logical block, bring on-line, and request status. The Disk Class Server[1] interprets the transmitted commands, handles the scheduling of command execution, tracks their progress, and reports status back to the Class Driver. Note the absence of seek or select head commands. This interface can be used equally well for solid-state disks as for conventional magnetic disks. Device-specific commands are issued at a lower level of the architecture, i.e., between the Device Client (disk controller) and Device Server (disk device). The former provides the path for moving commands and data between hosts and drives, and it is usually realized physically by a piece of hardware that corresponds to the device controller. The latter coincides with the physical drives used for storing and retrieving data.

It is interesting to contrast these proprietary approaches with an industry standard approach like SCSI, admittedly targeted for the low to mid range of performance. SCSI defines the logical and physical interface between a host bus adapter (HBA) and a disk controller, usually embedded within the assembly of the disk device. The HBA accepts I/O requests from the host, initiates I/O actions by communicating with the controllers, and performs direct memory access transfers between its own buffers and the memory of the host. Requesters of service are called initiators, while providers of service are called targets. Up to eight nodes can reside on a single SCSI string, sharing a common pathway to the HBA. The embedded controller performs device handling and error recovery. Physically, the interface is implemented with a single daisy-chained cable, and the 8-bit datapath is used to communicate control and status information, as well as data. SCSI defines a layered communications protocol, including a message layer for protocol and status, and a command/status layer for target operation execution. The HBA roughly corresponds to the function of the IBM channel processor or Cray IOP, while the embedded controller is similar to the IBM storage director/string controller or the Cray DCU. Despite the differences in terminology, the systems we have surveyed exhibut significant commonality of function and similar approaches for partitioning these functions among hardware components.

Characterization of I/O Workloads

Before characterizing the I/O behavior of different workloads, it is necessary to first understand the elements of disk performance. Disk performance is a function of the service time, which consists of three main components: *seek time*, *rotational latency*, and *data transfer time*.[2] **Seek time** is the time needed to position the heads

[1]Other kinds of class servers are also supported, such as for tape drives.

[2]In a heavily utilized system, delays waiting for a device can match actual disk service times, which in reality is composed of device queuing, controller overhead, seek, rotational latency, reconnect misses, error retries, and data transfer.

to the appropriate track position containing the desired data. It is a function of a substantial initial start-up cost to accelerate the disk head (on the order of 6 ms) as well as the number of tracks that must be traversed. Typical average seek times, i.e., the time to traverse between two randomly selected tracks (approximately 28% of the data band), are in the range of 10 to 20 ms. The track-to-track seek time is usually below 10 ms and as low as 2 ms.

The second component of service time is **rotational latency**. It takes some time for the desired sector to rotate under the head position before it can be read or written. Today's devices spin at a rate of approximately 3600 rpm, or 60 revolutions per second (we expect to see rotation speeds increase to 5400 rpm in the near future). For today's disks, a full revolution is 16 ms, and the average latency is 8 ms. Note that the worst-case latencies are comparable to average seeks.

The last component is the **transfer time,** i.e., the time to physically transfer the bytes from disk to the host. While the transfer time is a strong function of the number of bytes to be transferred, seek and rotational latencies times are independent of the transfer blocksize. If data is to be read or written in large chunks, it makes sense to choose a large blocksize, since the "fixed cost" of seek and latency is better amortized across a large data transfer.

A low-performance I/O system might dedicate the pathway between the host and the disk for the entire duration of the seek, rotate, and transfer times. Assuming small blocksizes, transfer time is a small component of the overall service time, and these pathways can be better utilized if they are shared among multiple devices. Thus, higher performance systems support independent seeks, in which a device can be directed to detach itself from the pathway while seeking to the desired track (recall the discussion of dynamic path reconnect in the previous section). The advantage is that multiple seeks can be overlapped, reducing overall I/O latency and better utilizing the available I/O **bandwidth**.

However, to make it possible for devices to reattach to the pathway, the I/O system must support a mechanism called **rotational position sensing,** i.e., the device interrupts the I/O controller when the desired sector is under the heads. If the pathway is currently in use, the device must pay a full rotational delay before it can again attempt to transfer. These rotational positional reconnect miss delays (RPS delays) represent a major source of degradation in many existing I/O systems [Buzen and Shum, 1987]. This arises from the lack of device buffering and the real-time service requirements of magnetic disks. At the time that these architectures were established, buffer memories were expensive and the demands for high I/O performance were less pressing with slower speed CPUs. An alternative, made more attractive by today's relative costs of electronic and mechanical components, is to associate a **track buffer** with the device that can be filled immediately. This can then be used as the source of the transfer when the pathway becomes available [Houtekamer, 1985].

I/O intensive applications vary widely in the demand they place on the I/O system. They run the gamut from processing small numbers of bulk I/Os that must be handled with minimum delay (supercomputer I/O) to large numbers of simple tasks that touch small amounts of data (transaction processing). An important design challenge is to develop an I/O system that can handle the performance needs of these diverse workloads.

A given workload's demand for I/O service can be specified in terms of three metrics: *throughput*, *latency*, and *bandwidth*. **Throughput** refers to the number of requests for service made per unit time. Latency measures how long it takes to service an individual request. Bandwidth gauges the amount of data flowing between service requesters (i.e., applications) and service providers (i.e., devices).

As observed by Bucher and Hayes [1980], supercomputer I/O can be characterized almost entirely by sequential I/O. Typically, computation parameters are moved in bulk from disk to in-memory data structures, and results are periodically written back to disk. These workloads demand large bandwidth and minimum latency, but are characterized by low throughput. Contrast this with transaction processing, which is characterized by enormous numbers of random accesses, relatively small units of work, and a demand for moderate latency with very high throughput.

Figure 80.15 shows another way of thinking about the varying demands of I/O intensive applications. It shows the percentage of time different applications spend in the three components of I/O service time. Transaction processing systems spend the majority of their service time in seek and rotational latency; thus technological advances which reduce the transfer time will not affect their performance very much. On the other hand, scientific applications spend a more equal amount of time in seek and data transfer, and their performance is sensitive to any improvement in disk technology.

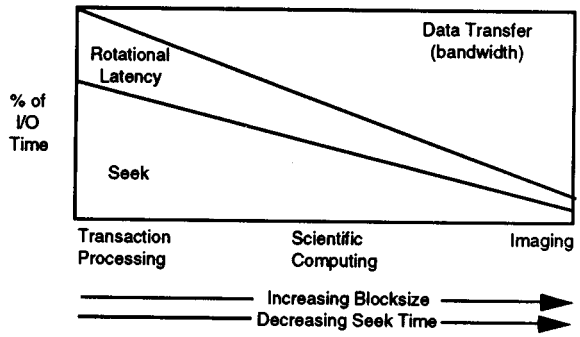


FIGURE 80.15 I/O system parameters as a function of application. Transaction processing applications are seek and rotational latency limited, since only small blocks are usually transferred from disk. Image-processing applications, on the other hand, transfer huge blocks and thus spend most of their I/O time in data transfer. Scientific computing applications tend to fall in between. (*Source:* I. Y. Bucher and A. H. Hayes, "I/O performance measurement on Cray-1 and CDC 7600 computers," *Proc. Cray Users Group Conference*, October 1980. With permission.)

Extensions to Conventional Disk Architectures

In this subsection, we will focus on techniques for improving the performance of conventional disk systems, i.e., methods which allow us to reduce the seek time, rotational latency, or transfer time of conventional disks. By reducing disk service times, we also decrease device queuing delays. These techniques include fixed-head disks, parallel transfer disks, increased disk density, solid-state disks, disk caches, and disk scheduling.

Fixed-Head Disk

The concept of a fixed-head disk is to place a read/write head at every track position. The need for positioning the heads is eliminated, thus eliminating the seek time altogether. The approach does not assist in reducing rotational latencies, nor does it lessen the transfer time. Fixed-head disks were often used in the early days of computing systems as a back-end store for virtual memory. However, since modern disks have hundreds of tracks per surface, placing a head at every position is no longer viewed as an economical solution.

Parallel Transfer Disks

Some high-performance disk drives make it possible to read or write from multiple disk surfaces at the same time. For example, the Cray DD-19 and DD-49 disks described in the second section have a parallel transfer capability. The advantage is that much higher transfer rates can be achieved, but no assistance is provided for seek or rotational latency. Thus transfer units are correspondingly larger in these systems.

A number of economic and technological issues limit the usefulness of parallel transfer disks. From the economic perspective, providing more than one set of read/write electronics *per actuator* is expensive. Further, current disks use sophisticated control systems to lock onto an individual track, and it is difficult to do this simultaneously across tracks within the same cylinder. Hence, the Cray strategy is limiting head groups to only four surfaces. There appears to be a fundamental trade-off between track density and the number of platters: as the track density increases, it becomes ever more difficult to lock onto tracks across many platters, and the number of surfaces that can participate in a parallel transfer is reduced. For example, current Cray track densities are around 980 tracks/inch, and require a rather sophisticated closed-loop track-following servo system to position the heads accurately with finely controlled voice coil actuators. A lower cost ($/megabyte) high-performance disk system can be constructed from several standard drives than from a single parallel transfer device, in part because of the relatively small sales volume of parallel transfer devices compared to standard drives.

Increasing Disk Density

As described in the first section, the improvements in disk recording density are likely to continue. Higher bit densities are achieved through a combination of the use of thinner films on the disk platters (e.g., densities improve from 16,000 bpi to 21,000 bpi when thick iron oxide is replaced with thin film materials), smaller gaps between the poles of the read/write head's electromagnet, and heads which fly closer to the disk surface.

While vertical recording techniques have long been touted as the technology of the future, advances in head technology make it possible to continue using conventional horizontal methods, but still keep disks on the MAD curve. These *magneto-resistive heads* employ noninductive methods for reading, which work well with dense horizontal recording fields. However, a more conventional head is needed for writing, but this dual-head organization permits separate optimizations for read and write.

Also, the choice of coding technique can have a significant effect on density. Standard modified frequency modulation techniques require approximately one flux change per bit, while more advanced run-length limited codes can increase density by an additional factor of 50%. Densities as high as 31,429 bpi can be attained with these techniques. As the recording densities increase, the transfer times decrease, as more bits transit beneath the heads per unit time. Of course, this approach provides no improvement in seek and latency times. Most of the increase in density comes from increases in the number of tracks per inch, which does not improve (and may actually reduce) performance.

Although increased densities are inevitable, the problem is primarily economic. Increasing the tracks per inch may make seeks slower as it becomes more time consuming for the heads to correctly "lock" onto the appropriate track. The sensing electronics get more complex and thus more expensive. Once again, it can be argued that higher capacity can be achieved at lower cost by using several smaller disks rather than one expensive high-density disk.

Solid-State Disks

Solid-state disks (SSD), constructed from relatively slow memory chips, can be viewed either as a kind of large and slow main memory or as a small and high-speed disk. When viewed as large main memory, the SSD is often called expanded storage (ES). The expanded storage found in the IBM 3090 class machines [Buzen and Shum, 1986] supports operations for paging data blocks from and to main memory. Usually, the expanded storage looks to the system more like memory than an I/O device: it is directly attached to main memory through a high-speed bus rather than an I/O controller. The maximum transfer bandwidth on the IBM 3090 between expanded store and memory is two orders of magnitude faster than conventional devices: approximately 216 Mbytes/s—one word each 18.5 ns!

Further, unlike conventional devices, a transfer between memory and expanded storage is performed synchronously with the CPU. This is viewed as acceptable, because the transfer requires so little time and does not involve the usual operating system overheads of I/O set-up and interrupts. Note that to transfer data from ES to disk requires the data to be first staged into main memory.

The Cray X-MP and Y-MP also support SSDs, which can come in configurations of up to 4096 Mbytes, approximately four times the capacity of the DD-49. The SSD has the potential for enormous bandwidth. It can be attached to the Cray IO system or directly to the CPU through up to two 1000-Mbyte/s channels. Access can be arranged in one of three ways [Reinhardt, 1988]. The first alternative is to treat the SSD as a logical disk, with users responsible for staging heavily accessed files to it. Unfortunately, this leads to the inevitable contention for SSD space. Further, the operating system's disk device drivers are not tuned for the special capabilities of SSDs, and some performance is lost. The second alternative is to use the SSD as an extended memory, in much the same manner as IBM's extended storage. Special system calls for accessing the SSD bypass the usual disk-handling code, and a 4096-byte sector can be accessed in 25 μs. The last alternative is to use the SSD as a logical device cache, i.e., as a second-level cache for multitrack chunks of files that resides between the system's in-main memory file cache and the physical disk devices. Cray engineers have observed workload speedups for their UNIX-like operating system of a factor of four over conventional disk when the cache is enabled. These results indicate that SSDs are most appropriate for containing "hot spot" data [Gawlick, 1987]. Conventional wisdom has it that 20% of the data receives 80% of the accesses, and this has been widely observed in transaction processing systems [Gawlick, 1987].

If SSDs are to be used to replace magnetic disks, then they must be made nonvolatile, and herein lies their greatest weakness. This can be achieved through battery back-up, but the technique is controversial. First, it is difficult to verify that the batteries will be fully charged when needed, i.e., when conventional power fails. Second, it is difficult to determine how long is long enough when powering the SSD with batteries. This should probably be long enough to off-load the disk's contents to magnetic media. Fortunately, low-power DRAM and wafer scale integration technology are making feasible longer battery hold times.

Another weakness is their cost. At the present time, there is more than a 10 to 20 times difference in price between the cost of a megabyte of magnetic disk memory and a megabyte of DRAM. While wafer scale integration may bring this price down in the future, for the near term SSDs will be limited to a staging or caching function.

Disk Caches

Disk caches place buffer memories between the host and the device. If disk data is likely to be re-referenced, caches can be effective in eliminating the seek and rotational latencies. Unfortunately, this effectiveness depends critically on the access behavior of the applications. Truly random access with little re-referencing cannot make effective use of disk caches. However, applications that exhibit a large degree of sequential access can use a cache to good purpose, because data can be staged into the cache before it is actually requested.

Disk caches can become even more useful if they are made nonvolatile using the battery back-up techniques described in the previous subsection (and with the same potential problems). A nonvolatile cache will allow "fast writes": the application need not wait for the write I/O to actually complete before it is notified that it has completed. For some applications environments, disk caches have the beneficial effect of reducing the number of reads and thus the number of I/O requests seen by the disks. This has the interesting side effect of increasing the percentage of writes found in the I/O mix, and some observers believe that writes may dominate I/O performance in future systems.

As already mentioned, a disk cache can also lead to better utilization of the host-to-device pathways. A device can transfer data into a cache even if the pathway is in use by another device on the same string. Thus caches are effective in avoiding rotational position sensing misses.

Disk Scheduling

The mechanical delays as seen by a set of simultaneous I/O requests can be reduced through effective disk scheduling. For example, seek times can be reduced if a *shortest-seek-time-first* scheduling algorithm is used [Smith, 1981]. That is, among the queue of pending I/O requests, the one next selected for service is the one that requires the shortest seek time from the current location of the read/write heads. The literature on disk scheduling algorithms is vast, and the effectiveness of a particular scheduling approach depends critically on the workload. It has been observed that scheduling algorithms work best when there are long queues of pending requests; unfortunately, this situation seems to occur rarely in existing systems [Smith, 1981].

Disk Arrays

An alternative to the approaches just described is to exploit parallelism by grouping together a number of physical disks and making these appear to applications as a single logical disk. This has the advantage that the bandwidth of several disks can be harnessed to service a single logical I/O request or can support multiple independent I/Os in parallel. Further, arrays can be constructed using existing, widely available disk technology, rather than the more specialized and more expensive approaches described in the previous subsection. For example, Cray offers a device called the DS-40, which appears as a single logical disk device but which is actually implemented internally as four drives. A logical track is constructed from sectors across the four disks. The DS-40 can transfer at a peak rate of 20 Mbyte/s, with a sustained transfer rate of 9.6 Mbyte/s, and thus is strictly faster than the DD-49.

Defining Terms

Arm: A mechanical assembly that positions the head to the correct track for reading or writing.

Bandwidth: The amount of data per unit time flowing between host computers and storage devices.

Cylinder: A stack of tracks at one acuator position.

Disk drive: An HDA plus all associated electronics.

Head: An electromagnet that produces switchable magnetic fields to read and record bit streams on a platter's track.

Head disk assembly (HDA): The collection of platters, heads, arms, and actuators, plus the air-tight casing, that makes up the storage device. Basically, this is everything but the electronics for controlling the drive and interfacing it to a computer system.

Latency: How long it takes to service an individual request.

Maximal areal density (MAD): The maximum number of bits that can be stored per square inch. Computed by multiplying the bits per inch in a disk track times the number of tracks per inch of media.

Platters: Metal disks covered with a magnetic material for recording information.

Rotational latency: The time it takes for the desired sector to rotate under the head position before it can be read or written.

Rotational position sensing: A storage device interrupts the I/O controller when the desired sector is under the heads.

Sector: A unit of a storage that is physically read or written at the same time.

Seek time: The time needed to position the heads to the appropriate track position containing the desired data.

Spindle: The collection of disk platters.

Track buffer: A memory buffer embedded in the disk drive. It can hold the contents of the current disk track.

Tracks: The circular recording regions on a platter.

Transfer time: The time taken to physically transfer the bytes from disk to the host.

Throughput: The number of requests for disk service per unit time.

Winchester disk: A magnetic disk in which the read/write heads fly above the recording surface on an air bearing. This is in contrast to contact recording, such as a floppy disk, in which the head and the magnetic media are actually touching.

Related Topic

36.2 Magnetic Recording

References

I. Y. Bucher and A. H. Hayes, "I/O performance measurement on Cray-1 and CDC 7600 computers," *Proceedings of the Cray Users Group Conference*, October 1980.

J. Buzen, "BEST/1 analysis of the IBM 3880-13 cached storage controller," *Proc. CMG XIII Conference*, 1982.

J. P. Buzen and A. Shum, "I/O architecture in MVS/370 and MVS/XA," *CMG Transactions*, vol. 54, pp. 19–26, Fall 1986.

J. P. Buzen and A. Shum, "A unified operational treatment of RPS reconnect delays," *Proc. 1987 Sigmetrics Conference*, Performance Evaluation Review, vol. 15, no. 1, May 1987.

Cray Research, Inc., "CRAY Y-MP Computer Systems Functional Description Manual," HR-4001, January 1988.

P. D. Frank, "Advances in head technology," presentation at Challenges in Disk Technology Short Course, Institute for Information Storage Technology, University of Santa Clara, Santa Clara, Calif., December 12–15, 1987.

D. Gawlick, Private Communication, November 1987.

J. M. Harker et al., "A quarter century of disk file innovation," *IBM Journal of Research and Development*, vol. 25, no. 5, pp. 677–689, September 1981.

G. Houtekamer, "The local disk controller," *Proc. 1985 Sigmetrics Conference*, August 1985.

N. P. Kronenberg, H. Levy, and W. D. Strecker, "VAXClusters: A closely-coupled distributed system," *ACM Trans. on Comp. Systems*, vol. 4, no. 2, pp. 130–146, May 1986.

P. Massiglia, *Digital Large System Mass Storage Handbook*, Colorado Springs, Col.: Digital Equipment Corporation, 1986.

S. Reinhardt, "A blueprint for the UNICOS operating system," *Cray Channels*, vol. 10, no. 3, pp. 20–24, Fall 1988.

A. J. Smith, "Input/output optimization and disk architectures: A survey," in *Performance and Evaluation 1*, North-Holland Publishing Company, 1981, pp. 104–117.

L. D. Stevens, "The evolution of magnetic storage," *IBM Journal of Research and Development*, vol. 25, no. 5, pp. 663–675, September 1981.

A. Vasudeva, "A case for disk array storage system," *Proc. Reliability Conference*, Santa Clara, Calif., 1988.

J. Voelcker, "Winchester disks reach for a gigabyte," *IEEE Spectrum*, pp. 64–67, February 1987.

Further Information

International Business Machines (IBM) Corporation developed the first rotating magnetic storage device in the mid-1950s and has always been an industry leader in the storage industry. In honor of the 25-year anniversary of the invention of the magnetic disk, *IBM's Journal of Research and Development* in September 1981 reviewed the development of the technology up to that time. Two particularly notable papers are

L. D. Stevens, "The evolution of magnetic storage," *IBM Journal of Research and Development*, vol. 25, no. 5, pp. 663–675, September 1981.

J. M. Harker et al., "A quarter century of disk file innovation," *IBM Journal of Research and Development*, vol. 25, no. 5, pp. 677–689, September 1981.

For a more up-to-date review of progress in the disk drive industry, see:

J. Voelker, "Winchester disks reach for a gigabyte," *IEEE Spectrum*, pp. 64–67, February 1987.

80.3 Magnetic Tape[1]

Peter A. Lee

Computers depend on memory to execute programs and to store program code and data. They also need access to stored program code and data in a **nonvolatile memory** (i.e., a form in which the information is not lost when the power is removed from the computer system). Different types of memory have been developed for different tasks. This memory can be categorized according to its price per bit, **access time**, and other parameters. Table 80.1 shows a typical hierarchy for memory which places the smallest and fastest memory at the top in level 0 and in general the largest, slowest, and cheapest at the bottom in level 4 [Ciminiera and Valenzano, 1987]. Auxiliary (secondary or mass) memory of level 4 forms the large storage capacity for program and code that are not currently required by the CPU. This is usually nonvolatile and is at a low cost per bit. Computer **magnetic tape** falls within this category and is the subject of this section.

A Brief Historical Review

Probably the first recorded storage device, developed by Schickard in 1623, used mechanical positions of cogs and gears to work a semi-automatic calculator. Then came Pascal's calculating machine based on 10 digits per wheel. In 1812 punched cards were used in weaving looms to store patterns for woven material. Since that time there have been many mechanical and, latterly, electromechanical devices developed for memory and storage.

In 1948 at Manchester University in England the cathode ray tube (Williams) and the magnetic drum were developed. These consisted of 1024 bits and 1280 bits and a magnetic drum capacity of 120K bits. Cambridge University developed the mercury delay line in 1949, which represented the first fully operational delay line memory, consisting of 576 bits per tube with a total capacity of 18K bits and a circulation time of 1.1 ms.

The first commercial computer with a magnetic tape system was introduced in 1951. The UNIVAC I had a magnetic tape system of 1.44M bits on 150 feet of tape and was capable of storing 128 characters per inch. The tape could be read at a rate of 100 ips. Optical memories are now available as very fast storage devices and will replace magnetic storage in the next few years. At present these devices are expensive although it is envisaged that optical disks with large silicon caches will be the storage arrangement of the future where computer systems utilizing CAD software and image processing can take advantage of the large storage capacities with fast access times. In the future, semiconductor memories are likely to continue their advancing trend.

Introduction

Today's microprocessors are capable of addressing up to 16 Mbytes of main memory. To take advantage of this large capacity, it is usual to have several programs residing in memory at the same time. With intelligent memory

[1]Based on P. A. Lee, "Memory subsystems," in *Digital Systems Reference Book*, B. Holdworth and G. R. Martins, Eds., Oxford: Butterworth-Heinemann, 1991, chap. 2.6. With permission.

TABLE 80.1 Memory Hierarchy

	Data	Code	MMU
Level 0	CPU register	Instruction registers	MMU registers
Level 1	Data cache	Instruction cache	MMU memory
Level 2		On-board cache	
Level 3		Main memory	
Level 4		Auxiliary memory	

Source: P. A. Lee, "Memory subsystems," in *Digital Systems Reference Book*, B. Holdsworth and G. R. Martin, Eds., Oxford: Butterworth-Heinemann, 1991, p. 2.6/3. With permission.

management units (MMUs), the programs can be swopped in and out of the main memory to the auxiliary memory when required. For the system to keep pace with this program swopping, it must have a fast auxiliary memory to write to. In the past, most auxiliary systems like magnetic tape and disks have had slow access times, and this has meant that expensive systems have evolved to cater for this requirement. Now that auxiliary memory has improved, and access times are fast and the memory cheap, computer systems have been developed that provide memory swopping with large nonvolatile storage systems. Although the basic technology has not changed over the last 20 years, new materials and different approaches have meant that a new form of auxiliary memory has been brought to the market at a very cheap cost.

Magnetic Tape

Magnetic tape currently provides the cheapest form of storage for large quantities of computer data in a nonvolatile form. The tape is arranged on a reel and has several different packaging styles. It is made from a polyester transportation layer with a deposited layer of oxide having a property similar to a ferrite material with a large hysteresis. Magnetic tape is packaged either in a cartridge, on a reel, or in a cassette. The magnetic cartridge is manufactured in several tape lengths and cartridge sizes capable of storing up to 2 G (giga) bytes of data. These can be purchased in many popular preformatted styles.

The magnetic tape reel is usually 1⁄2 inch or 1 inch wide and has lengths of 600, 1200, and 2400 feet. Most reels can store data at rates from 800 bits per inch (bpi) up to 6250 bpi. The reel-to-reel magnetic tape reader is generally bulkier and more expensive than the cartridge readers due to the complicated pneumatic drive mechanisms, but it provides a large data storage capacity with high access speeds [Wiehler, 1974]. An example of a typical magnetic tape drive with the reel-to-reel arrangement is shown in Fig. 80.16.

A cheap storage medium is the magnetic cassette. Based on the audio cassette, this uses the normal audio cassette recorder for reading and writing data via the standard Kansas City interface through a serial computer I/O line. A logic data "1" is recorded by a high frequency and a logic data "0" by a lower frequency. High-density cassettes can store up to 60 Mbytes of data on each tape and are popular with the computer games market as a cheap storage medium for program distribution.

Both reel-to-reel and cartridge tapes are generally organised by using nine separate tracks across the tape as shown in Fig. 80.17(a).

Each track has its own read and write head operated independently from other tracks [see Fig. 80.17(b)]. Tracks 1 to 8 are used for data and track nine for the parity bit. Data is written on the tape in rows of magnetized islands, using for example EBCDIC (Extended Binary Coded Decimal Interchange Code).

Each read/write head is shaped from a **ferromagnetic material** with an air gap 1 μm wide as seen in Fig. 80.18. The writing head is concerned with converting an electrical pulse into a magnetic state and can be magnetized in one of two directions. This is done by passing a current through the magnetic coil which sets up a leakage field across the 1-μm gap. When the current is reversed the field across the gap is changed, reversing the polarity of the magnetic field on the tape. The head magnetizes the passing magnetic tape recording the state of the magnetic field in the air gap. A logic 1 is recorded as a change in polarity on the tape, and a logic 0 is recorded as no change in polarity, as seen in Fig. 80.19. Reading the magnetic tape states from the tape and converting them to electrical signals is done by the read head. The bit sequences in Fig. 80.19 show the change in magnetic

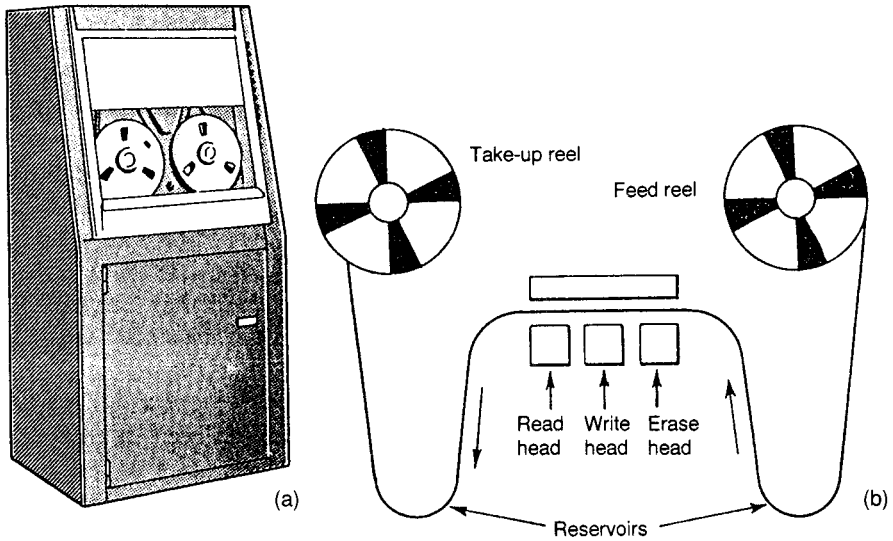


FIGURE 80.16 (a) Magnetic tape drive. (b) Magnetic tape reel arrangement. (*Source:* K. London, *Introduction to Computers*, London: Faber and Faber, 1986, p. 141. With permission.)

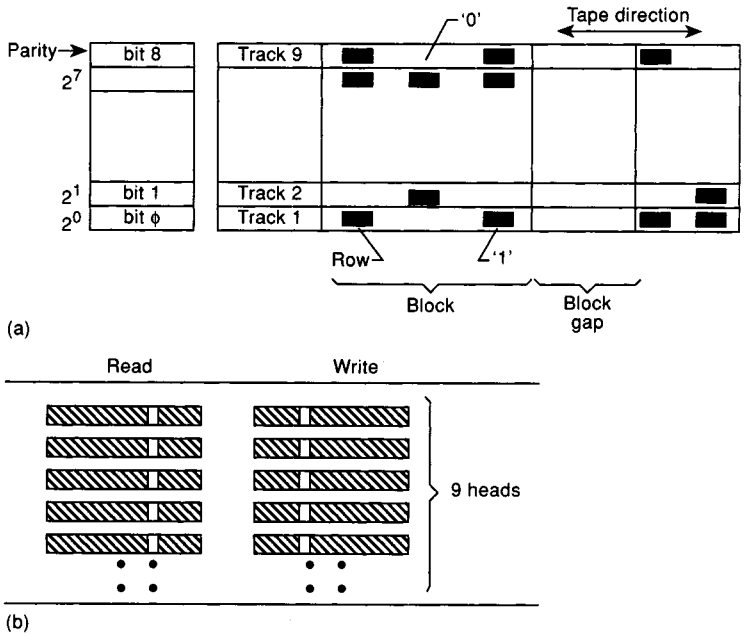


FIGURE 80.17 Magnetic tape format. (*Source:* P. A. Lee, "Memory subsystems," in *Digital Systems Reference Book,* B. Holdsworth and G. R. Martin, Eds., Oxford: Butterworth-Heinemann, 1991, p. 2.6/11. With permission.)

states on the tape. When the tape is passed over the read head, it induces a voltage into the magnetic coil which is converted to digital levels to retrieve the original data.

Tape Format

Information is stored on magnetic tape in the form of a coherent sequence of rows forming a block. This usually corresponds to a page of computer memory and is the minimum amount of data written to or read from

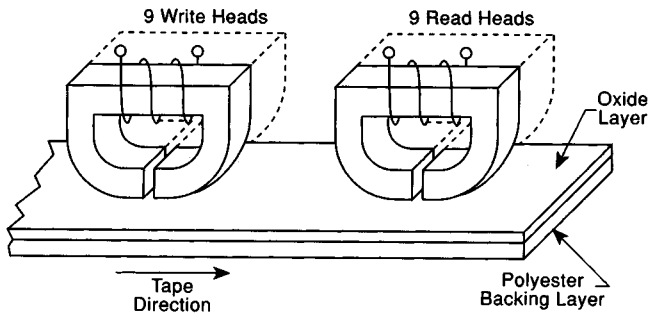


FIGURE 80.18 Read/write head layout. (*Source:* P. A. Lee, "Memory subsystems," in *Digital Systems Reference Book,* B. Holdsworth and G. R. Martin, Eds., Oxford: Butterworth-Heinemann, 1991, p. 2.6/12. With permission.)

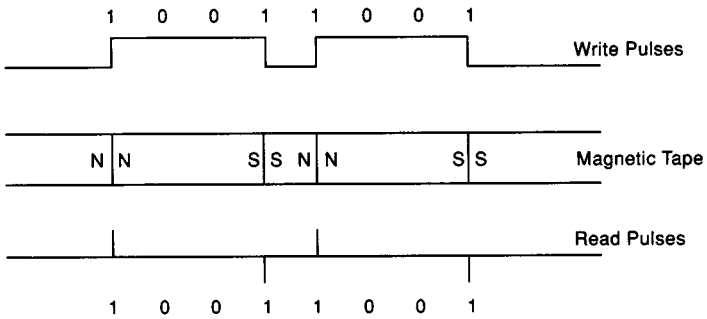


FIGURE 80.19 Write and read pulses on magnetic tape. (*Source:* P. A. Lee, "Memory subsystems," in *Digital Systems Reference Book,* B. Holdsworth and G. R. Martin, Eds., Oxford: Butterworth-Heinemann, 1991, p. 2.6/12. With permission.)

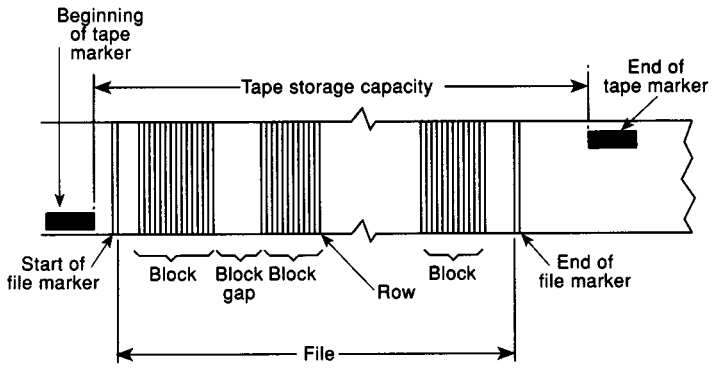


FIGURE 80.20 Magnetic tape format. (*Source:* P. A. Lee, "Memory subsystems," in *Digital Systems Reference Book,* B. Holdsworth and G. R. Martin, Eds., Oxford: Butterworth-Heinemann, 1991, p. 2.6/12. With permission.)

magnetic tape with each program statement. Each block of data is separated by a block gap which is approximately 15 mm long and has no data stored in it. This is shown in Fig. 80.20.

Block gaps are used to allow the tape to accelerate to its operational speed and for the tape to decelerate when stopping at the end of a block. Block gaps use up to 50% of the tape space available for recording, although this may be reduced by making the block sizes larger but has the disadvantage of requiring larger memory buffers to accommodate the data.

A number of blocks make up a file identified by a tape file marker which is written to the tape by the tape controller. The entire length of tape is enclosed between the beginning and end of tape markers. These normally consist of a photosensitive material that triggers sensors on the read/write heads. When a new tape is loaded, it normally advances to the beginning of a tape marker and then it is ready for access by the CPU. The end of tape marker is used to prevent the tape from running off the end of the tape spool and indicates the limit of the storage length.

Recording Modes

Several recording modes are used with the express objective of storing data at the highest density and with the greatest reliability of noncorruption of retrieved data. Two popular but contrasting modes are the *non-return-to-zero* (NRZ) and *phase encoding* (PE) modes. These are incompatible although some magnetic tape drives have detectors to sense the mode and operate in a bimodal way. The NRZ technique is shown in Fig. 80.19, where only the 1 bit is displayed by a reversal of magnetization on the tape. The magnetic polarity remains unchanged for logic 0. An external clock track is also required for this mode because a pulse is not always generated for each row of data on the tape.

The PE technique allows both the 0 and 1 states to be displayed by changes of magnetization. A 1 bit is given by a north-to-north pole on the tape, and a 0 bit is given by a south-to-south pole on the tape. PE provides approximately double the recording density and processor speed of NRZ. PE tapes carry an identification mark called a *burst*, which consists of successive magnetization changes at the beginning of track 4. This allows the tape drive to recognize the tape mode and configure itself accordingly.

Defining Terms

Access time: The cycle time for the computer store to present information to the CPU. Access times vary from less than 40 ns for level 0 register storage up to tens of seconds for magnetic tape storage.

Auxiliary (secondary, mass, or backing) storage: Computer stores which have a capacity to store enormous amounts of information in a *nonvolatile* form. This type of memory has an access time usually greater than main memory and consists of magnetic tape drives, magnetic disk stores, and optical disk stores.

Ferromagnetic material: Materials that exhibit high magnetic properties. These include metals such as cobalt, iron, and some alloys.

Magnetic tape: A polyester film sheet coated with a *ferromagnetic* powder, which is used extensively in auxiliary memory. It is produced on a reel, in a cassette, or in a cartridge transportation medium.

Nonvolatile memory: The class of computer memory that retains its stored information when the power supply is cut off. It includes magnetic tape, magnetic disks, flash memory, and most types of ROM.

Related Topic

36.2 Magnetic Recording

References

L. Ciminiera and A. Valenzano, *Advanced Microprocessor Architectures*, Reading, Mass.: Addison-Wesley, 1987.

B. Holdsworth and G. Martin, Eds, *Digital Systems Reference Book*, Oxford: Butterworth-Heinemann, 1991, pp 2.6/1–2.6/11.

R. Hyde, "Overview of memory management," *Byte,* pp. 219–225, April 1988.

J. Isailovíc, *Video Disc and Optical Memory Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

K. London, I*ntroduction to Computers*, London: Faber and Faber Press, 1986, p. 141.

M. Mano, *Computer Systems Architecture*, Englewood Cliffs, N.J.: Prentice-Hall, 1982.

R. Matick, *Computer Storage Systems & Technology*, New York: John Wiley, 1977.

A. Tanenbaum, *Structured Computer Organisation*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

G. Wiehler, *Magnetic Peripheral Data Storage*, Heydon & Son, 1974.

Further Information

The *IEEE Transactions on Magnetics* is available from the IEEE Service Center, Customer Service Department, 445 Hoes Lane, Piscataway, NJ 08855-1331; 800-678-IEEE (outside the USA: 908-981-0060). An IEEE-sponsored Conference on Magnetism and Magnetic Materials was held in December 1992. The British Tape Industry Association (BTIA) has a computer media committee, and further information on standards, etc. can be obtained from British Tape Industry Association, Carolyn House, 22-26 Dingwall Road, Croydon CR0 9XF, England. The equivalent American Association also provides information on computer tape and can be contacted at International Tape Manufacturers' Association, 505 Eighth Avenue, New York, NY 10018.

80.4 Magneto-Optical Disk Data Storage

M. Mansuripur

Since the early 1940s, magnetic recording has been the mainstay of electronic information storage worldwide. Audio tapes provided the first major application for the storage of information on magnetic media. Magnetic tape has been used extensively in consumer products such as audio tapes and video cassette recorders (VCRs); it has also found application in backup/archival storage of computer files, satellite images, medical records, etc. Large volumetric capacity and low cost are the hallmarks of tape data storage, although sequential access to the recorded information is perhaps the main drawback of this technology. Magnetic hard disk drives have been used as mass storage devices in the computer industry ever since their inception in 1957. With an areal density that has doubled roughly every other year, hard disks have been and remain the medium of choice for secondary storage in computers.[1] Another magnetic data storage device, the floppy disk, has been successful in areas where compactness, removability, and fairly rapid access to the recorded information have been of prime concern. In addition to providing backup and safe storage, inexpensive floppies with their moderate capacities (2 Mbyte on a 3.5-in. diameter platter is typical nowadays) and reasonable transfer rates have provided the crucial function of file/data transfer between isolated machines. All in all, it has been a great half-century of progress and market dominance for magnetic recording devices, which are only now beginning to face a potentially serious challenge from the technology of optical recording.

Like magnetic recording, a major application area for optical data storage systems is the secondary storage of information for computers and computerized systems. Like the high-end magnetic media, optical disks can provide recording densities in the range of 10^7 bits/cm^2 and beyond. The added advantage of optical recording is that, like floppies, these disks can be removed from the drive and stored on the shelf. Thus the functions of the hard disk (i.e., high capacity, high data transfer rate, rapid access) may be combined with those of the floppy (i.e., backup storage, removable media) in a single optical disk drive. Applications of optical recording are not confined to computer data storage. The enormously successful audio **compact disk (CD)**, which was introduced in 1983 and has since become the de facto standard of the music industry, is but one example of the tremendous potentials of the optical technology.

A strength of optical recording is that, unlike its magnetic counterpart, it can support read-only, write-once, and erasable/rewritable modes of data storage. Consider, for example, the technology of optical audio/video disks. Here the information is recorded on a master disk which is then used as a stamper to transfer the embossed patterns to a plastic substrate for rapid, accurate, and inexpensive reproduction. The same process is employed in the mass production of read-only files (CD-ROM, O-ROM) which are now being used to distribute software, catalogues, and other large databases. Or consider the write-once read-many (WORM) technology, where one can permanently store massive amounts of information on a given medium and have rapid, random access to them afterwards. The optical drive can be designed to handle read-only, WORM, and erasable media all in one unit, thus combining their useful features without sacrificing performance and ease of use or occupying too

[1] At the time of this writing, achievable densities on hard disks are in the range of 10^7 bits/cm^2. Random access to arbitrary blocks of data in these devices can take on the order of 10 ms, and individual read/write heads can transfer data at the rate of several megabits per second.

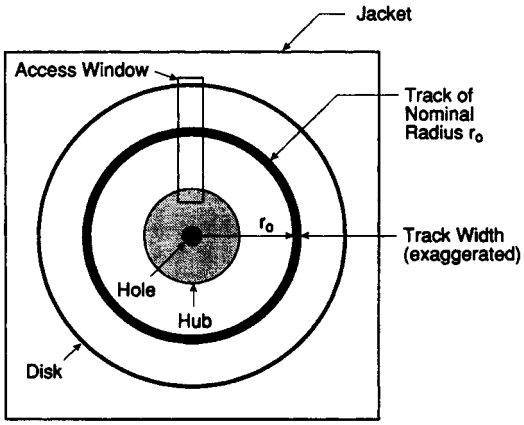


FIGURE 80.21 Physical appearance and general features of an optical disk. The read/write head gains access to the disk through a window in the jacket; the jacket itself is for protection purposes only. The hub is the mechanical interface with the drive for mounting and centering the disk on the spindle. The track shown at radius r_0 is of the concentric-ring type.

much space. What is more, the media can contain regions with prerecorded information as well as regions for read/write/erase operations, both on the same platter. These possibilities open new vistas and offer opportunities for applications that have heretofore been unthinkable; the interactive video disk is perhaps a good example of such applications.

In this article we will lay out the conceptual basis for optical data storage systems; the emphasis will be on disk technology in general and magneto-optical disk in particular. The first section is devoted to a discussion of some elementary aspects of disk data storage including the concept of track and definition of the access time. The second section describes the basic elements of the optical path and its functions; included are the properties of the semiconductor laser diode, characteristics of the beamshaping optics, and certain features of the focusing objective lens. Because of the limited depth of focus of the objective and the eccentricity of tracks, optical disk systems must have a closed-loop feedback mechanism for maintaining the focused spot on the right track. These mechanisms are described in the third and fourth sections for automatic focusing and automatic track following, respectively. The physical process of thermomagnetic recording in magneto-optic (MO) media is described next, followed by a discussion of the MO readout process in the sixth section. The final section describes the properties of the MO media.

Preliminaries and Basic Definitions

A disk, whether magnetic or optical, consists of a number of **tracks** along which the information is recorded. These tracks may be concentric rings of a certain width, W_t, as shown in Fig. 80.21. Neighboring tracks may be separated from each other by a guard band whose width we shall denote by W_g. In the least sophisticated recording scheme imaginable, marks of length Δ_0 are recorded along these tracks. Now, if each mark can be in either one of two states, present or absent, it may be associated with a binary digit, 0 or 1. When the entire disk surface of radius R is covered with such marks, its capacity C_0 will be

$$C_0 = \frac{\pi R^2}{(W_t + W_g)\Delta_0} \quad \text{bits per surface} \tag{80.2}$$

Consider the parameter values typical of current optical disk technology: $R = 67$ mm corresponding to 5.25-in. diameter platters, $\Delta_0 = 0.5$ μm which is roughly determined by the wavelength of the read/write laser diodes, and $W_t + W_g = 1$ μm for the track pitch. The disk capacity will then be around 28×10^9 bits, or 3.5 gigabytes. This is a reasonable estimate and one that is fairly close to reality, despite the many simplifying assumptions made in its derivation. In the following paragraphs we examine some of these assumptions in more detail.

The disk was assumed to be fully covered with information-carrying marks. This is generally not the case in practice. Consider a disk rotating at $\mathcal{N}$ revolutions per second (rps). For reasons to be clarified later, this rotational speed should remain constant during the disk operation. Let the electronic circuitry have a fixed clock duration T_c. Then only pulses of length T_c (or an integer multiple thereof) may be used for writing. Now, a mark written along a track of radius r, with a pulse-width equal to T_c, will have length ℓ, where

$$\ell = 2\pi \mathcal{N}\, rT_c \tag{80.3}$$

Thus for a given rotational speed $\mathcal{N}$ and a fixed clock cycle T_c, the minimum mark length ℓ is a linear function of track radius r, and ℓ decreases toward zero as r approaches zero. One must, therefore, pick a minimum usable track radius, $r_{\min}$, where the spatial extent of the recorded marks is always greater than the minimum allowed mark length, Δ_0. Equation (80.3) yields

$$r_{\min} = \frac{\Delta_0}{2\pi \mathcal{N}\ T_c} \tag{80.4}$$

One may also define a maximum usable track radius r_{max}, although for present purposes $r_{max} = R$ is a perfectly good choice. The region of the disk used for data storage is thus confined to the area between $r_{\min}$ and $r_{\max}$. The total number N of tracks in this region is given by

$$N = \frac{r_{\max} - r_{\min}}{W_t + W_g} \tag{80.5}$$

The number of marks on any given track in this scheme is independent of the track radius; in fact, the number is the same for all tracks, since the period of revolution of the disk and the clock cycle uniquely determine the total number of marks on any individual track. Multiplying the number of usable tracks N with the capacity per track, we obtain for the usable disk capacity

$$C = \frac{N}{\mathcal{N}\ T_c} \tag{80.6}$$

Replacing for N from Eq. (80.5) and for $\mathcal{N}\ T_c$ from Eq. (80.4), we find,

$$C = \frac{2\pi r_{\min}(r_{\max} - r_{\min})}{(W_t + W_g)\Delta_0} \tag{80.7}$$

If the capacity C in Eq. (80.7) is considered a function of $r_{\min}$ with the remaining parameters held constant, it is not difficult to show that maximum capacity is achieved when

$$r_{\min} = \tfrac{1}{2}\ r_{\max} \tag{80.8}$$

With this optimum $r_{\min}$, the value of C in Eq. (80.7) is only half that of C_0 in Eq. (80.2). In other words, the estimate of 3.5 gigabyte per side for 5.25-in. disks seems to have been optimistic by a factor of two.

One scheme often proposed to enhance the capacity entails the use of multiple zones, where either the rotation speed $\mathcal{N}$ or the clock period T_c is allowed to vary from one zone to the next. In general, zoning schemes can reduce the minimum usable track radius below that given by Eq. (80.8). More importantly, however, they allow tracks with larger radii to store more data than tracks with smaller radii. The capacity of the zoned disk is somewhere between C of Eq. (80.7) and C_0 of Eq. (80.2), the exact value depending on the number of zones implemented.

A fraction of the disk surface area is usually reserved for **preformat** information and cannot be used for data storage. Also, prior to recording, additional bits are generally added to the data for **error correction coding** and other housekeeping chores. These constitute a certain amount of overhead on the user data and must be allowed for in determining the capacity. A good rule of thumb is that overhead consumes approximately 20% of the raw capacity of an optical disk, although the exact number may vary among the systems in use. Substrate defects and film contaminants during the deposition process can create bad **sectors** on the disk. These are typically identified during the certification process and are marked for elimination from the sector directory. Needless to say, bad sectors must be discounted when evaluating the capacity.

Modulation codes may be used to enhance the capacity beyond what has been described so far. Modulation coding does not modify the minimum mark length of Δ_0, but frees the longer marks from the constraint of being integer multiples of Δ_0. The use of this type of code results in more efficient data storage and an effective number of bits per Δ_0 that is greater than unity. For example, the popular (2, 7) modulation code has an effective bit density of 1.5 bits per Δ_0. This or any other modulation code can increase the disk capacity beyond the estimate of Eq. (80.7).

The Concept of Track

The information on magnetic and optical disks is recorded along tracks. Typically, a track is a narrow annulus at some distance r from the disk center. The width of the annulus is denoted by W_t, while the width of the guard band, if any, between adjacent tracks is denoted by W_g. The track pitch is the center-to-center distance between neighboring tracks and is therefore equal to $W_t + W_g$. A major difference between the magnetic floppy disk, the magnetic hard disk, and the optical disk is that their respective track pitches are presently of the order of 100, 10, and 1 μm. Tracks may be fictitious entities, in the sense that no independent existence outside the pattern of recorded marks may be ascribed to them. This is the case, for example, with the audio compact disk format where prerecorded marks simply define their own tracks and help guide the laser beam during readout. In the other extreme are tracks that are physically engraved on the disk surface before any data is ever recorded. Examples of this type of track are provided by pregrooved WORM and magneto-optical disks. Figure 80.22 shows micrographs from several recorded optical disk surfaces. The tracks along which the data are written are clearly visible in these pictures.

It is generally desired to keep the read/write head stationary while the disk spins and a given track is being read from or written onto. Thus, in an ideal situation, not only should the track be perfectly circular, but also the disk must be precisely centered on the spindle axis. In practical systems, however, tracks are neither precisely circular, nor are they concentric with the spindle axis. These eccentricity problems are solved in low-performance floppy drives by making tracks wide enough to provide tolerance for misregistrations and misalignments. Thus the head moves blindly to a radius where the track center is nominally expected to be and stays put until the reading or writing is over. By making the head narrower than the track pitch, the track center is allowed to wobble around its nominal position without significantly degrading the performance during the read/write operation. This kind of wobble, however, is unacceptable in optical disk systems, which have a very narrow track, about the same size as the focused beam spot. In a typical situation arising in practice, the eccentricity of a given track may be as much as ±50 μm while the track pitch is only about 1 μm, thus requiring active track-following procedures.

One method of defining tracks on an optical disk is by means of pregrooves that are either etched, stamped, or molded onto the substrate. In **grooved media of optical storage**, the space between neighboring grooves is the so-called land [see Fig. 80.23(a)]. Data may be written in the grooves with the land acting as a guard band. Alternatively, the land regions may be used for recording while the grooves separate adjacent tracks. The groove depth is optimized for generating an optical signal sensitive to the radial position of the read/write laser beam. For the push-pull method of track-error detection the groove depth is in the neighborhood of $\lambda/8$, where λ is the wavelength of the laser beam.

In digital data storage applications, each track is divided into small segments or sectors, intended for the storage of a single block of data (typically either 512 or 1024 bytes). The physical length of a sector is thus a few millimeters. Each sector is preceded by header information such as the identity of the sector, identity of the corresponding track, synchronization marks, etc. The header information may be preformatted onto the

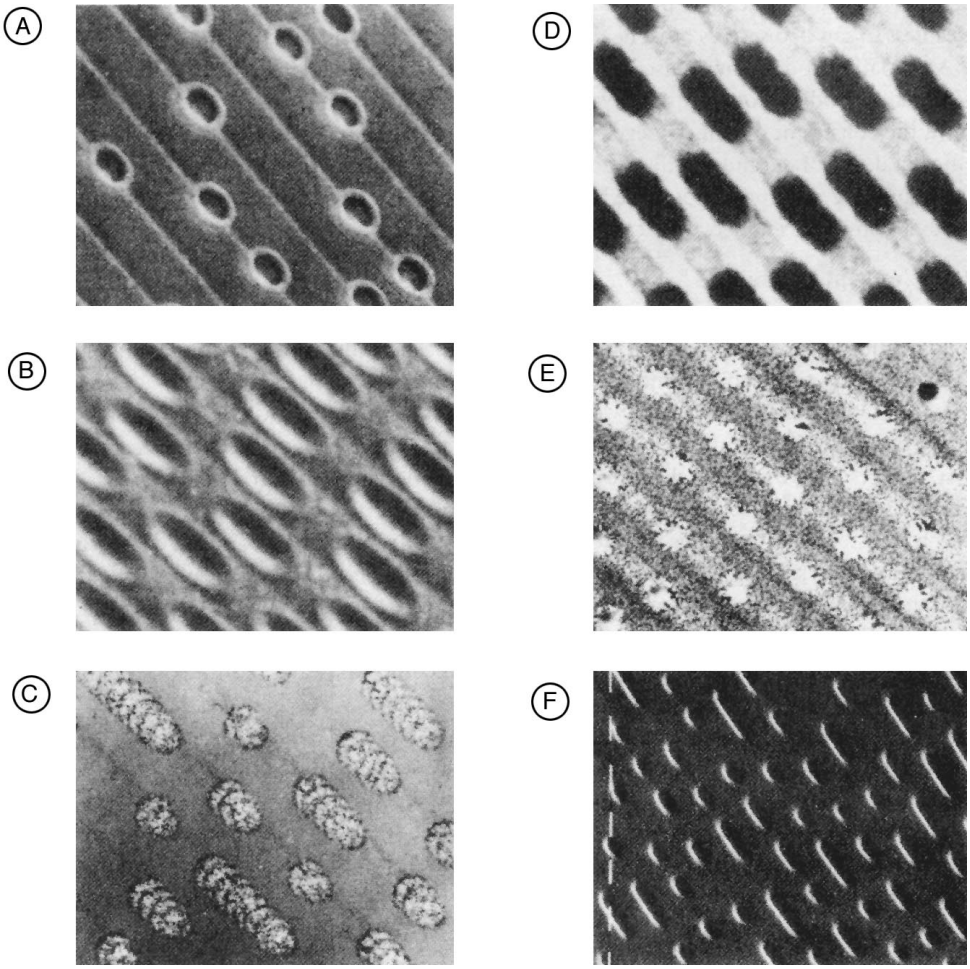


FIGURE 80.22 Micrographs of several types of optical storage media. The tracks are straight and narrow (track pitch = 1.6 μm), with an orientation angle of ≃ –45°. (A) Ablative, write-once tellurium alloy. (B) Ablative, write-once organic dye. (C) Amorphous-to-crystalline, write-once phase-change alloy GaSb. (D) Erasable, amorphous magneto-optic alloy GdTbFe. (E) Erasable, crystalline-to-amorphous phase-change tellurium alloy. (F) Read-only CD-audio, injection-molded from polycarbonate with a nickel stamper. (*Source: Ullmann's Encyclopedia of Industrial Chemistry*, 5th ed., vol. A14, Weinheim: VCH, 1989, p. 196. With permission.)

substrate, or it may be written on the storage layer prior to shipping the disk. Pregrooved tracks may be "carved" on the optical disk either as concentric rings or as a single continuous spiral. There are certain advantages to each format. A spiral track can contain a succession of sectors without interruption, whereas concentric rings may each end up with some empty space that is smaller than the required length for a sector. Also, large files may be written onto (and read from) spiral tracks without jumping to the next track, which occurs when concentric tracks are used. On the other hand, multiple-path operations such as write-and-verify or erase-and-write, which require two paths each for a given sector, or still-frame video are more conveniently handled on concentric-ring tracks.

Another track format used in practice is based on the sampled-servo concept. Here the tracks are identified by occasional marks placed permanently on the substrate at regular intervals, as shown in Fig. 80.23. Details of track following by the sampled-servo scheme will follow shortly; suffice it to say at this point that servo marks help the system identify the position of the focused spot relative to the track center. Once the position is determined it is fairly simple to steer the beam and adjust its position.

Disk Rotation Speed

When a disk rotates at a constant angular velocity ω, a track of radius r moves with the constant linear velocity $V = r\omega$. Ideally, one would like to have the same linear velocity for all the tracks, but this is impractical except in a limited number of situations. For instance, when the desired mode of access to the various tracks is sequential, such as in audio and video disk applications, it is possible to place the head in the beginning at the inner radius and move outward from the center thereafter while continuously decreasing the angular velocity. By keeping the product of r and ω constant, one can thus achieve constant linear velocity for all the tracks.[1] Sequential access mode, however, is the exception rather than the norm in data storage systems. In most applications, the tracks are accessed randomly with such rapidity that it becomes impossible to adjust the rotation speed for constant linear velocity. Under these circumstances, the angular velocity is best kept constant during the normal operation of the disk. Typical rotation speeds are 1200 and 1800 rpm for slower drives and 3600 rpm for the high data rate systems. Higher rotation rates (5000 rpm and beyond) are certainly feasible and will likely appear in future storage devices.

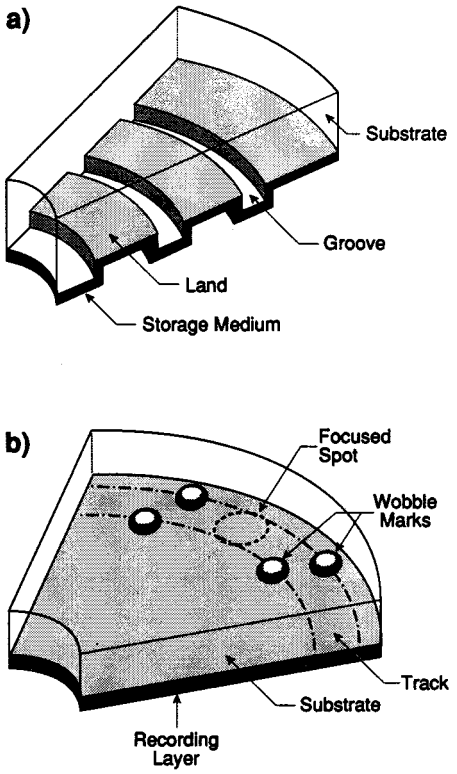


FIGURE 80.23 (a) Lands and grooves in an optical disk. The substrate is transparent, and the laser beam must pass through it before reaching the storage medium. (b) Sampled-servo marks in an optical disk. These marks which are offset from the track-center provide information regarding the position of focused spot.

Access Time

The direct-access storage device or DASD, used in computer systems for the mass storage of digital information, is a disk drive capable of storing large quantities of data and accessing blocks of this data rapidly and in arbitrary order. In read/write operations it is often necessary to move the head to new locations in search of sectors containing specific data items. Such relocations are usually time-consuming and can become the factor that limits performance in certain applications. The access time τ_a is defined as the average time spent in going from one randomly selected spot on the disk to another. τ_a can be considered the sum of a seek time, τ_s, which is the average time needed to acquire the target track, and a latency, τ_l, which is the average time spent on the target track waiting for the desired sector. Thus,

$$\tau_a = \tau_s + \tau_l \tag{80.9}$$

The latency is half the revolution period of the disk, since a randomly selected sector is, on the average, halfway along the track from the point where the head initially lands. Thus for a disk rotating at 1200 rpm $\tau_l = 25$ ms, while at 3600 rpm $\tau_l \simeq 8.3$ ms. The seek time, on the other hand, is independent of the rotation speed, but is determined by the traveling distance of the head during an average seek, as well as by the mechanism of head actuation. It can be shown that the average length of travel in a random seek is one third of the full stroke. (In our notation the full stroke is $r_{max} - r_{min}$.) In magnetic disk drives where the head/actuator assembly is relatively light-weight (a typical Winchester head weighs about 5 grams) the acceleration and deceleration periods are short, and seek times are typically around 10 ms in small drives (i.e., 5.25 and 3.5 in.). In optical disk systems,

[1]In compact disk players the linear velocity is kept constant at 1.2 m/s. The starting position of the head is at the inner radius $r_{min} = 25$ mm, where the disk spins at 460 rpm. The spiral track ends at the outer radius $r_{max} = 58$ mm, where the disk's angular velocity is 200 rpm.

on the other hand, the head, being an assembly of discrete elements, is fairly large and heavy (typical weight ≃100 grams), resulting in values of τ_s that are several times greater than those obtained in magnetic recording systems. The seek times reported for commercially available optical drives presently range from 20 ms in high-performance 3.5-in. drives to about 80 ms in larger drives. We emphasize, however, that the optical disk technology is still in its infancy; with the passage of time, the integration and miniaturization of the elements within the optical head will surely produce lightweight devices capable of achieving seek times of the order of a few milliseconds.

The Optical Path

The **optical path** begins at the light source which, in practically all laser disk systems in use today, is a semiconductor GaAs diode laser. Several unique features have made the laser diode indispensable in optical recording technology, not only for the readout of stored information but also for writing and erasure. The small size of this laser has made possible the construction of compact head assemblies, its coherence properties have enabled diffraction-limited focusing to extremely small spots, and its direct modulation capability has eliminated the need for external modulators. The laser beam is modulated by controlling the injection current; one applies pulses of variable duration to turn the laser on and off during the recording process. The pulse duration can be as short as a few nanoseconds, with rise and fall times typically less than 1 ns. Although readout can be accomplished at constant power level, i.e., in CW mode, it is customary for noise reduction purposes to modulate the laser at a high frequency (e.g., several hundred megahertz during readout).

Collimation and Beam Shaping

Since the cross-sectional area of the active region in a laser diode is only about one micrometer, diffraction effects cause the emerging beam to diverge rapidly. This phenomenon is depicted schematically in Fig. 80.24(a). In practical applications of the laser diode, the expansion of the emerging beam is arrested by a collimating lens, such as that shown in Fig. 80.24(b). If the beam happens to have aberrations (astigmatism is particularly severe in diode lasers), then the collimating lens must be designed to correct this defect as well.

In optical recording it is most desirable to have a beam with circular cross section. The need for shaping the beam arises from the special geometry of the laser cavity with its rectangular cross section. Since the emerging beam has different dimensions in the directions parallel and perpendicular to the junction, its cross section at the collimator becomes elliptical, with the initially narrow dimension expanding more rapidly to become the major axis of the ellipse. The collimating lens thus produces a beam with elliptical cross section. Circularization may be achieved by bending various rays of the beam at a prism, as shown in Fig. 80.24(c). The bending changes the beam's diameter in the plane of incidence but leaves the diameter in the perpendicular direction intact.

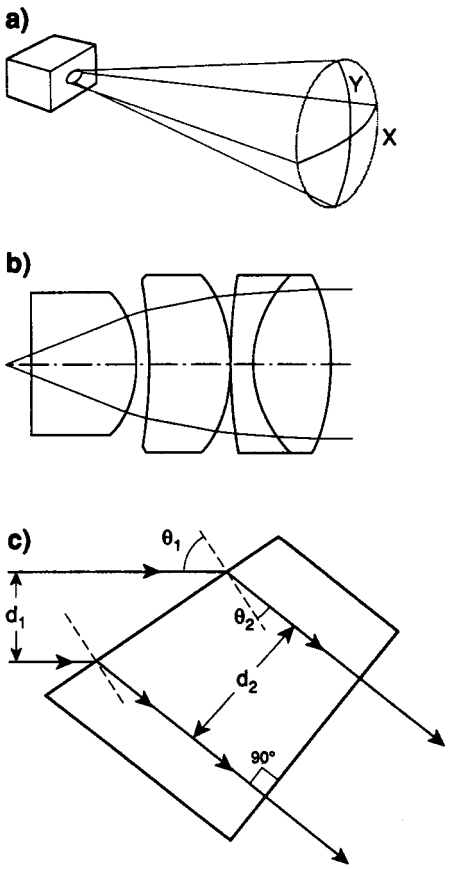


FIGURE 80.24 (a) Away from the facet, the output beam of a diode laser diverges rapidly. In general, the beam diameter along X is different from that along Y, which makes the cross section of the beam elliptical. Also, the radii of curvature R_x and R_y are not the same, thus creating a certain amount of astigmatism in the beam. (b) Multi-element collimator lens for laser diode applications. Aside from collimating, this lens also corrects astigmatic aberrations of the beam. (c) Beam shaping by deflection at a prism surface. θ_1 and θ_2 are related by the Snell's law, and the ratio d_2/d_1 is the same as $\cos\theta_2/\cos\theta_1$. Passage through the prism circularizes the elliptical cross section of the beam.

Focusing by the Objective Lens

The collimated and circularized beam of the diode laser is focused on the surface of the disk using an **objective lens**. The objective is designed to be aberration-free, so that its focused spot size is limited only by the effects of diffraction. Figure 80.25(a) shows the design of a typical objective made from spherical optics. According to the classical theory of diffraction, the diameter of the beam, d, at the objective's focal plane is given by

$$d \simeq \frac{\lambda}{NA} \qquad (80.10)$$

where λ is the wavelength of light, and NA is the numerical aperture of the objective.[1]

In optical recording it is desired to achieve the smallest possible spot, since the size of the spot is directly related to the size of marks recorded on the medium. Also, in readout, the spot size determines the resolution of the system. According to Eq. (80.10) there are two ways to achieve a small spot: first by reducing the wavelength and, second, by increasing the numerical aperture of the objective. The wavelengths currently available from GaAs lasers are in the range of 670–840 nm. It is possible to use a nonlinear optical device to double the frequency of these diode lasers, thus achieving blue light. Good efficiencies have been demonstrated by frequency doubling. Also recent developments in II–VI materials have improved the prospects for obtaining green and blue light directly from semiconductor lasers. Consequently, there is hope that in the near future optical storage systems will operate in the wavelength range of 400–500 nm. As for the numerical aperture, current practice is to use a lens with NA $\simeq$0.5–0.6. Although this value might increase slightly in the coming years, much higher numerical apertures are unlikely, since they put strict constraints on the other characteristics of the system and limit the tolerances. For instance, the working distance at high numerical aperture is relatively short, making access to the recording layer through the substrate more difficult. The smaller depth of focus of a high numerical aperture lens will make attaining/maintaining proper focus more of a problem, while the limited field of view might restrict automatic track-following procedures. A small field of view also places constraints on the possibility of read/write/erase operations involving multiple beams.

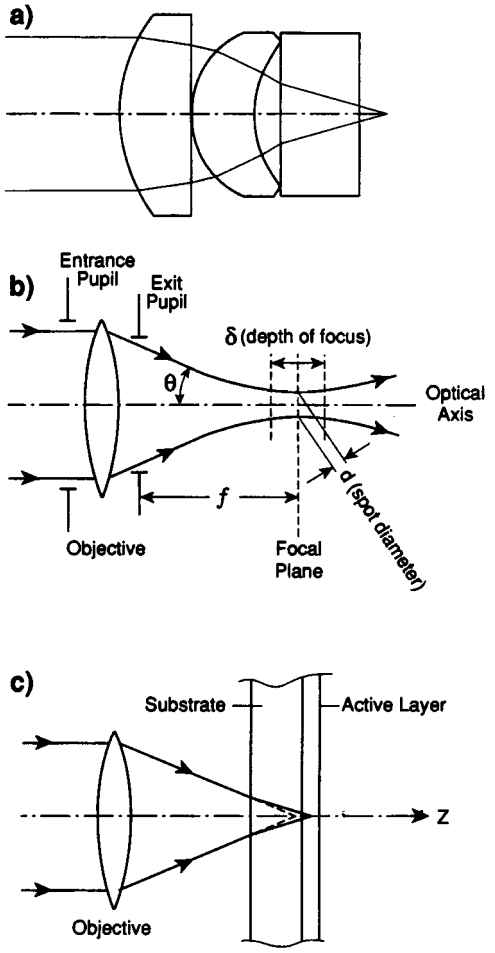


FIGURE 80.25 (a) Multi-element lens design for a high numerical aperture video disk objective. (*Source:* D. Kuntz, "Specifying laser diode optics," *Laser Focus*, March 1984. With permission.) (b) Various parameters of the objective lens. The numerical aperture is $NA = \sin\theta$. The spot diameter d and the depth of focus δ are given by Eqs. (80.10) and (80.11), respectively. (c) Focusing through the substrate can cause spherical aberration at the active layer. The problem can be corrected if the substrate is taken into account while designing the objective.

The depth of focus of a lens, δ, is the distance away from the focal plane over which tight focus can be maintained [see Fig. 80.25(b)]. According to the classical diffraction theory

[1]Numerical aperture is defined as $NA = n \sin\theta$, where n is the refractive index of the image space, and θ is the half-angle subtended by the exit pupil at the focal point. In optical recording systems the image space is air whose index is very nearly unity; thus for all practical purposes $NA = \sin\theta$.

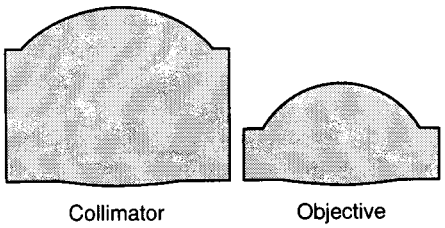


FIGURE 80.26 Molded glass aspheric lens pair for optical disk applications. These singlets can replace the multi-element spherical lenses shown in Figs. 80.24(b) and 80.25(a).

$$\delta \simeq \frac{\lambda}{NA^2} \tag{80.11}$$

Thus for a wavelength of $\lambda = 700$ nm and $NA = 0.6$, the depth of focus is about ± 1 μm. As the disk spins under the optical head at the rate of several thousand rpm, the objective lens must stay within a distance of $f \pm \delta$ from the active layer if proper focus is to be maintained. Given the conditions under which drives usually operate, it is impossible to make rigid enough mechanical systems to yield the required positioning tolerances. On the other hand, it is fairly simple to mount the objective lens in an actuator capable of adjusting its position with the aid of closed-loop feedback control. We shall discuss the technique of **automatic focusing** in the next section. For now, let us emphasize that by going to shorter wavelengths and/or larger numerical apertures (as is required for attaining higher data densities) one will have to face a much stricter regime as far as automatic focusing is concerned. Increasing the numerical aperture is particularly worrisome, since δ drops with the square of *NA*.

A source of spherical aberrations in optical disk systems is the substrate through which the light must travel to reach the active layer of the disk. Figure 80.25(c) shows the bending of the rays at the disk surface that causes the aberration. This problem can be solved by taking into account the effects of the substrate in the design of the objective, so that the lens is corrected for all aberrations including those arising at the substrate. Recent developments in molding of aspheric glass lenses have gone a long way in simplifying the lens design problem. Figure 80.26 shows a pair of molded glass aspherics designed for optical disk system applications; both the collimator and the objective are single-element lenses and are corrected for aberrations.

Automatic Focusing

We mentioned in the preceding section that since the objective has a large numerical aperture ($NA \geq 0.5$), its depth of focus δ is rather shallow ($\delta \simeq \pm 1$ μm at $\lambda = 780$ nm). During all read/write/erase operations, therefore, the disk must remain within a fraction of a micrometer from the focal plane of the objective. In practice, however, the disks are not flat and they are not always mounted rigidly parallel to the focal plane, so that movements away from focus occur a few times during each revolution. The peak-to-peak movement in and out of focus may be as much as 100 μm. Without automatic focusing of the objective along the optical axis, this runout (or disk flutter) will be detrimental to the operation of the system. In practice, the objective is mounted on a small motor (usually a voice coil) and allowed to move back and forth in order to keep its distance within an acceptable range from the disk. The spindle turns at a few thousand rpm, which is a hundred or so revolutions per second. If the disk moves in and out of focus a few times during each revolution, then the voice coil must be fast enough to follow these movements in real time; in other words, its frequency response must extend to several kilohertz.

The signal that controls the voice coil is obtained from the light reflected from the disk. There are several techniques for deriving the focus error signal, one of which is depicted in Fig. 80.27(a). In this so-called obscuration method a secondary lens is placed in the path of the reflected light, one-half of its aperture is covered, and a split detector is placed at its focal plane. When the disk is in focus, the returning beam is collimated and the secondary lens will focus the beam at the center of the split detector, giving a difference

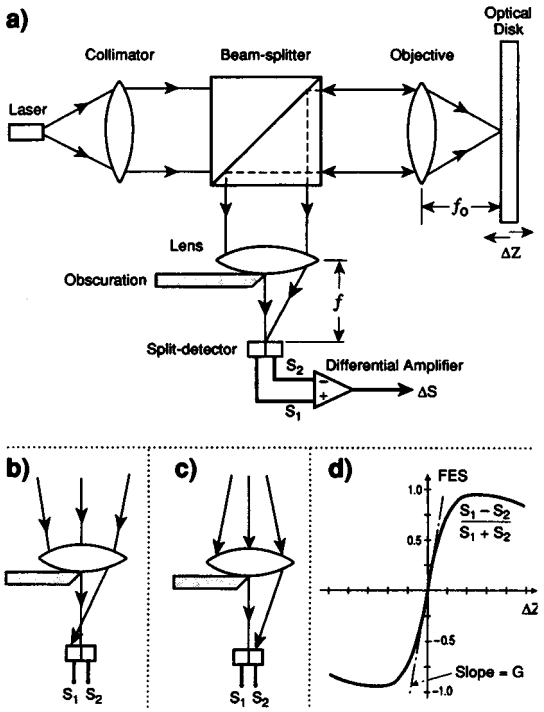


FIGURE 80.27 Focus error detection by the obscuration method. In (a) the disk is in focus, and the two halves of the split detector receive equal amounts of light. When the disk is too far from the objective (b) or too close to it (c), the balance of detector signals shifts to one side or the other. A plot of the focus error signal (FES) versus defocus is shown in (d), and its slope near the origin is identified as the FES gain, *G*.

signal ΔS equal to zero. If the disk now moves away from the objective, the returning beam will become converging, as in Fig. 80.27(b), sending all the light to detector #1. In this case ΔS will be positive and the voice coil will push the lens towards the disk. On the other hand, when the disk moves close to the objective, the returning beam becomes diverging and detector #2 receives the light [see Fig. 80.27(c)]. This results in a negative ΔS that forces the voice coil to pull back in order to return ΔS to zero. A given focus error detection scheme is generally characterized by the shape of its focus error signalΔS versus the amount of defocus Δz; one such curve is shown in Fig. 80.27(d). The slope of the focus error signal (FES) curve near the origin is of particular importance, since it determines the overall performance and stability of the servo loop.

Automatic Tracking

Consider a track at a certain radial location, say r_0, and imagine viewing this track through the access window shown in Fig. 80.21. It is through this window that the head gains access to arbitrarily selected tracks. To a viewer looking through the window, a perfectly circular track centered on the spindle axis will look stationary, irrespective of the rotation rate. However, any eccentricity will cause an apparent radial motion of the track. The peak-to-peak distance traveled by a track (as seen through the window) depends on a number of factors including centering accuracy of the hub, deformability of the substrate, mechanical vibrations, manufacturing tolerances, etc. For a typical 3.5-in. disk, for example, this peak-to-peak motion can be as much as 100 μm during one revolution. Assuming a revolution rate of 3600 rpm, the apparent velocity of the track in the radial direction will be several millimeters per second. Now, if the focused spot remains stationary while trying to read from or write to this track, it is clear that the beam will miss the track for a good fraction of every revolution cycle.

Practical solutions to the above problem are provided by **automatic tracking** techniques. Here the objective is placed in a fine actuator, typically a voice coil, which is capable of moving the necessary radial distances and

maintaining a lock on the desired track. The signal that controls the movement of this actuator is derived from the reflected light itself, which carries information about the position of the focused spot. There exist several mechanisms for extracting the track error signal (TES); all these methods require some sort of structure on the disk surface in order to identify the track. In the case of read-only disks (CD, CD-ROM, and video disk), the embossed pattern of data provides ample information for tracking purposes. In the case of write-once and erasable disks, tracking guides are "carved" on the substrate in the manufacturing process. As mentioned earlier, the two major formats for these tracking guides are pregrooves (for continuous tracking) and sampled-servo marks (for discrete tracking). A combination of the two schemes, known as continuous/composite format, is often used in practice. This scheme is depicted in Fig. 80.28 which shows a small section containing five tracks, each consisting of the tail end of a groove, synchronization marks, a mirror area used for adjusting focus/track offsets, a pair of wobble marks for sampled tracking, and header information for sector identification.

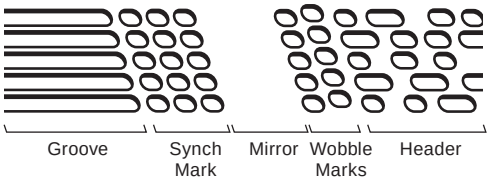


FIGURE 80.28 Servo fields in continuous/composite format contain a mirror area and offset marks for tracking

Tracking on Grooved Regions

As shown in Fig. 80.23(a), grooves are continuous depressions that are either embossed or etched or molded onto the substrate prior to deposition of the storage medium. If the data is recorded on the grooves, then the lands are not used except for providing a guard band between neighboring grooves. Conversely, the land regions may be used to record the information, in which case grooves provide the guard band. Typical track widths are about one wavelength. The guard bands are somewhat narrower than the tracks, their exact shape and dimensions depending on the beam size, required track-servo accuracy, and the acceptable levels of cross-talk between adjacent tracks. The groove depth is usually around one-eighth of one wavelength ($\lambda/8$), since this depth can be shown to give the largest TES in the push-pull method. Cross sections of the grooves may be rectangular, trapezoidal, triangular, etc.

When the focused spot is centered on track, it is diffracted symmetrically from the two edges of the track, resulting in a balanced far field pattern. As soon as the spot moves away from the center, the symmetry breaks down and the light distribution in the far field tends to shift to one side or the other. A split photodetector placed in the path of the reflected light can therefore sense the relative position of the spot and provide the appropriate feedback signal. This strategy is depicted schematically in Fig. 80.29; also shown in the figure are intensity plots at the detector plane for light reflected from various regions of the disk. Note how the intensity shifts to one side or the other depending on the direction of motion of the spot.

Sampled Tracking

Since dynamic track runout is usually a slow and gradual process, there is actually no need for continuous tracking as done on grooved media. A pair of embedded marks, offset from the track center as in Fig. 80.23(b), can provide the necessary information for correcting the relative position of the focused spot. The reflected intensity will indicate the positions of the two servo marks as two successive short pulses. If the beam happens to be on track, the two pulses will have equal magnitudes and there will be no need for correction. If, on the other hand, the beam is off-track, one of the pulses will be stronger than the other. Depending on which pulse is the stronger, the system will recognize the direction in which it has to move and will correct the error accordingly. The servo marks must appear frequently enough along the track to ensure proper track following. In a typical application, the track might be divided into groups of 18 bytes, 2 bytes dedicated as servo offset areas and 16 bytes filled with other format information or left blank for user data.

Thermomagnetic Recording Process

Recording and erasure of information on a magneto-optical disk are both achieved by the **thermomagnetic process**. The essence of thermomagnetic recording is shown in Fig. 80.30. At the ambient temperature the film

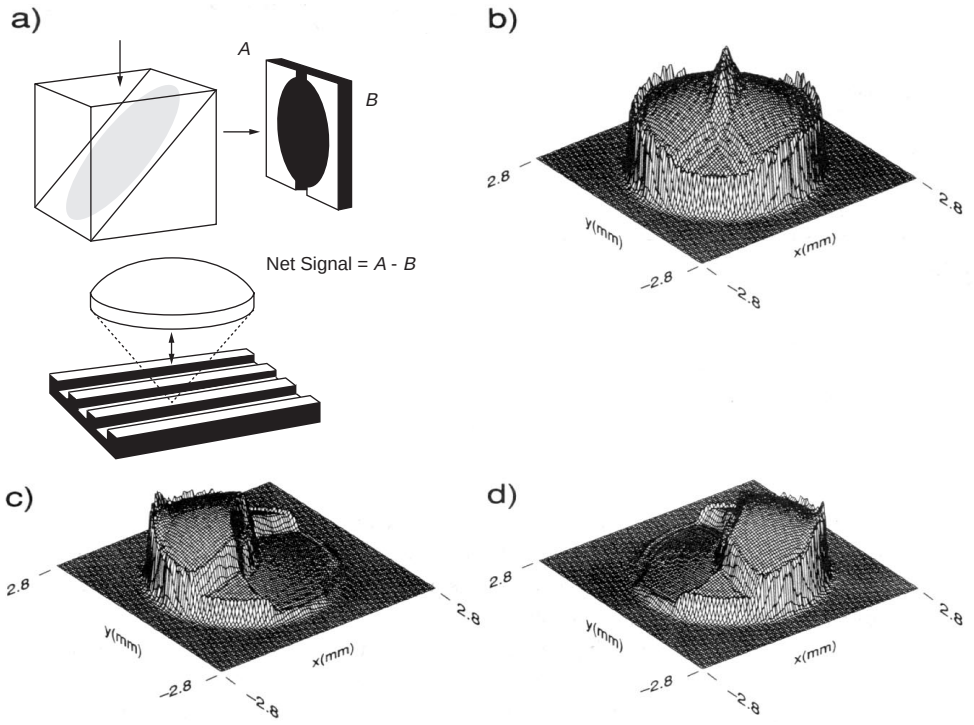


FIGURE 80.29 (a) Push-pull sensor for tracking on grooves. (b) Calculated distribution of light intensity at the detector plane when the disk is in focus and the beam is centered on track. (c) Calculated intensity distribution at the detector plane with disk in focus but the beam centered on the groove edge. (d) Same as (c) except for the spot being focused on the opposite edge of the groove.

has a high magnetic coercivity[1] and therefore does not respond to the externally applied field. When a focused beam raises the local temperature of the film, the hot spot becomes magnetically soft (i.e., its coercivity drops). As the temperature rises, coercivity drops continuously until such time as the field of the electromagnet finally overcomes the material's resistance to reversal and switches its magnetization. Turning the laser off brings the temperatures back to normal, but the reverse-magnetized domain remains frozen in the film. In a typical situation in practice, the film thickness may be around 300 Å, laser power at the disk $\simeq$10 mW, diameter of the focused spot $\simeq$1 μm, laser pulse duration $\simeq$50 ns, linear velocity of the track $\simeq$10 m/s, and the magnetic field strength $\simeq$200 gauss. The temperature may reach a peak of 500 K at the center of the spot, which is sufficient for magnetization reversal, but is not nearly high enough to melt or crystalize or in any other way modify the material's structure.

The materials of magneto-optical recording have strong perpendicular magnetic anisotropy. This type of anisotropy favors the "up" and "down" directions of magnetization over all other orientations. The disk is initialized in one of these two directions, say up, and the recording takes place when small regions are selectively reverse-magnetized by the thermomagnetic process. The resulting magnetization distribution then represents the pattern of recorded information. For instance, binary sequences may be represented by a mapping of zeros to up-magnetized regions and ones to down-magnetized regions (non-return to zero or NRZ). Alternatively, the NRZI scheme might be used, whereby transitions (up-to-down and down-to-up) are used to represent the ones in the bit-sequence.

[1]Coercivity of a magnetic medium is a measure of its resistance to magnetization reversal. For example, consider a thin film with perpendicular magnetic moment saturated in the $+Z$ direction. A magnetic field applied along $-Z$ will succeed in reversing the direction of magnetization only if the field is stronger than the coercivity of the film.

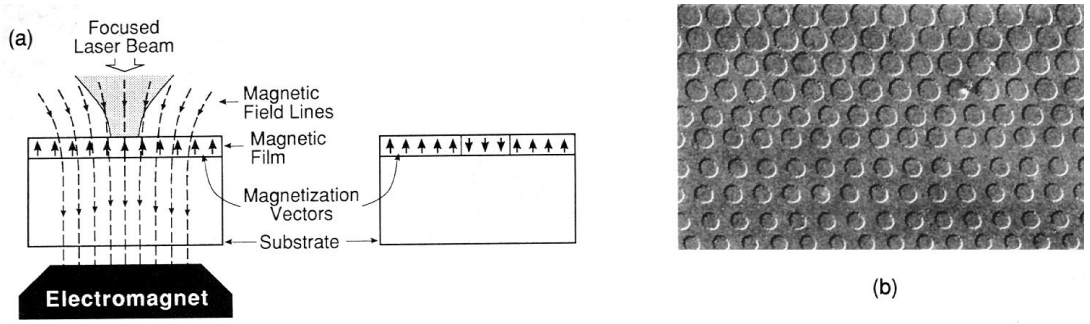


FIGURE 80.30 Thermomagnetic recording process. (a) The field of the electromagnet helps reverse the direction of magnetization in the area heated by the focused laser beam. (b) Lorentz micrograph of domains written thermomagnetically. The various tracks shown here were written at different laser powers, with power level decreasing from top to bottom. (*Source:* F. Greidanus et al., Paper 26B-5, presented at the International Symposium on Optical Memory, Kobe, Japan, September 1989. With permission.)

Recording by Laser Power Modulation (LPM)

In this traditional approach to thermomagnetic recording, the electromagnet produces a constant field, while the information signal is used to modulate the power of the laser beam. As the disk rotates under the focused spot, the on/off laser pulses create a sequence of up/down domains along the track. The Lorentz electron micrograph in Fig. 80.30(b) shows a number of domains recorded by LPM. The domains are highly stable and may be read over and over again without significant degradation. If, however, the user decides to discard a recorded block and to use the space for new data, the LPM scheme does not allow direct overwrite; the system must erase the old data during one disk revolution cycle and record the new data in a subsequent revolution cycle.

During erasure, the direction of the external field is reversed, so that up-magnetized domains in Fig. 80.30(a) now become the favored ones. Whereas writing is achieved with a modulated laser beam, in erasure the laser stays on for a relatively long period of time, erasing an entire sector. Selective erasure of individual domains is not practical, nor is it desired, since mass data storage systems generally deal with data at the level of blocks, which are recorded onto and read from individual sectors. Note that at least one revolution period elapses between the erasure of an old block and its replacement by a new block. The electromagnet therefore need not be capable of rapid switchings. (When the disk rotates at 3600 rpm, for example, there is a period of 16 ms or so between successive switchings.) This kind of slow reversal allows the magnet to be large enough to cover all the tracks simultaneously, thereby eliminating the need for a moving magnet and an actuator. It also affords a relatively large gap between the disk and the magnet, which enables the use of double-sided disks and relaxes the mechanical tolerances of the system without overburdening the magnet's driver.

The obvious disadvantage of LPM is its lack of direct overwrite capability. A more subtle concern is that it is perhaps unsuitable for the PWM (pulse width modulation) scheme of representing binary waveforms. Due to fluctuations in the laser power, spatial variations of material properties, lack of perfect focusing and track following, etc., the length of a recorded domain along the track may fluctuate in small but unpredictable ways. If the information is to be encoded in the distance between adjacent domain walls (i.e., PWM), then the LPM scheme of thermomagnetic writing may suffer from excessive domain-wall jitter. Laser power modulation works well, however, when the information is encoded in the position of domain centers (i.e., pulse position modulation or PPM). In general, PWM is superior to PPM in terms of the recording density, and, therefore, recording techniques that allow PWM are preferred.

Recording by Magnetic Field Modulation

Another method of thermomagnetic recording is based on magnetic field modulation (MFM) and is depicted schematically in Fig. 80.31(a). Here the laser power may be kept constant while the information signal is used to modulate the magnetic field. Photomicrographs of typical domain patterns recorded in the MFM scheme are shown in Fig. 80.31(b). Crescent-shaped domains are the hallmark of the field modulation technique. If one assumes (using a much simplified model) that the magnetization aligns itself with the applied field within

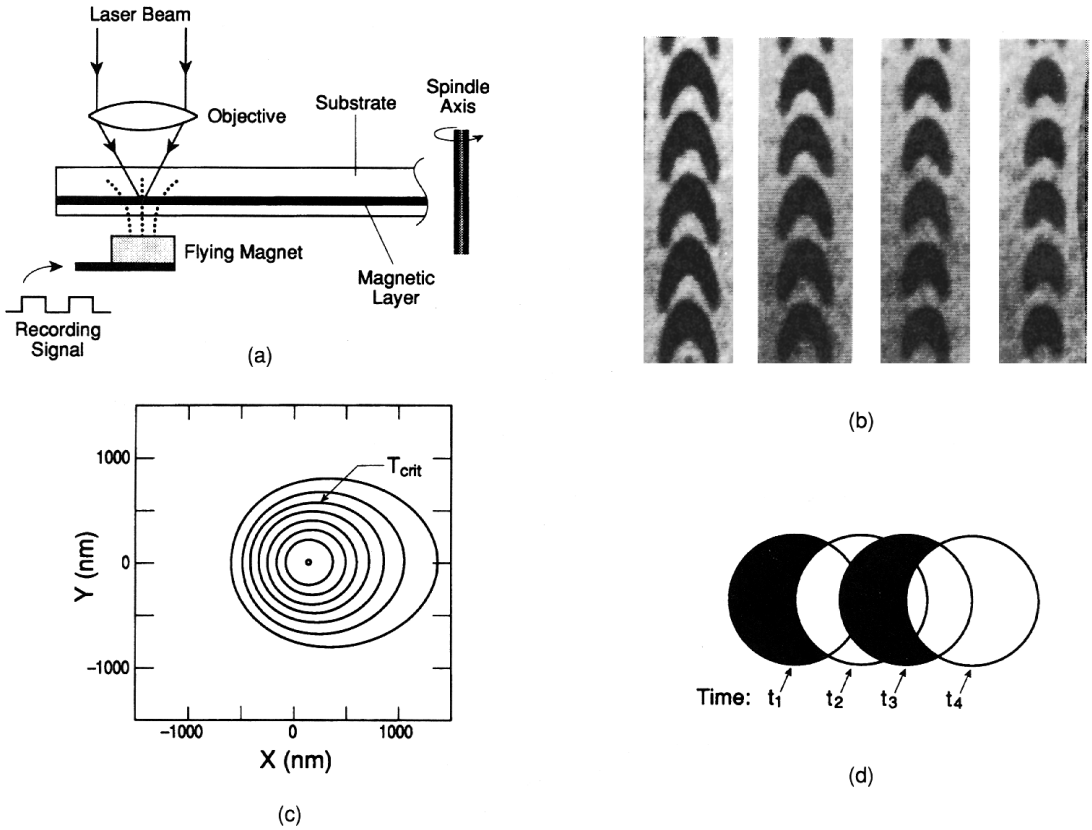


FIGURE 80.31 (a) Thermomagnetic recording by magnetic field modulation. The power of the beam is kept constant, while the magnetic field direction is switched by the data signal. (b) Polarized-light microphotograph of recorded domains. (c) Computed isotherms produced by a CW laser beam, focused on the magnetic layer of a disk. The disk moves with constant velocity under the beam. The region inside the isotherm marked as T_{crit} is above the critical temperature for writing, that is, its magnetization aligns with the direction of the applied field. (d) Magnetization within the heated region (above T_{crit}) follows the direction of the applied field, whose switchings occur at times t_i. The resulting domains are crescent-shaped.

a region whose temperature has passed a certain critical value, T_{crit}, then one can explain the crescent shape of these domains in the following way: With the laser operating in the CW mode and the disk moving at constant velocity, temperature distribution in the magnetic medium assumes a steady-state profile, such as that shown in Fig. 80.31(c). Of course, relative to the laser beam, the temperature profile is stationary, but in the frame of reference of the disk the profile moves along the track with the linear track velocity. The isotherm corresponding to T_{crit} is identified as such in the figure; within this isotherm the magnetization aligns itself with the applied field. Figure 80.31(d) shows a succession of critical isotherms along the track, each obtained at the particular instant of time when the magnetic field switches direction. From this picture it is easy to infer how the crescent-shaped domains form and also understand the relation between the waveform that controls the magnet and the resulting domain pattern.

The advantages of magnetic field modulation recording are that (1) direct overwriting is possible and (2) domain-wall positions along the track, being rather insensitive to defocus and laser power fluctuations, are fairly accurately controlled by the timing of the magnetic field switchings. On the negative side, the magnet must now be small and fly close to the disk surface, if it is to produce rapidly switched fields with a magnitude of a hundred gauss or so. Systems that utilize magnetic field modulation often fly a small electromagnet on the opposite side of the disk from the optical stylus. Since mechanical tolerances are tight, this might compromise the removability of the disk. Moreover, the requirement of close proximity between the magnet and the storage medium dictates the use of single-sided disks in practice.

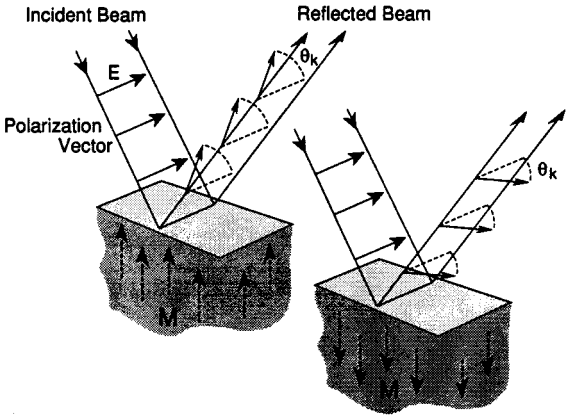


FIGURE 80.32 Schematic diagram describing the polar magneto-optical Kerr effect. Upon reflection from the surface of a perpendicularly magnetized medium, the polarization vector undergoes a rotation. The sense of rotation depends on the direction of magnetization, **M**, and switches sign when **M** is reversed.

Magneto-Optical Readout

The information recorded on a perpendicularly magnetized medium may be read with the aid of the polar **magneto-optical Kerr effect**. When linearly polarized light is normally incident on a perpendicular magnetic medium, its plane of polarization undergoes a slight rotation upon reflection. This rotation of the plane of polarization, whose sense depends on the direction of magnetization in the medium, is known as the polar Kerr effect. The schematic representation of this phenomenon in Fig. 80.32 shows that if the polarization vector suffers a counterclockwise rotation upon reflection from an up-magnetized region, then the same vector will rotate clockwise when the magnetization is down. A magneto-optical medium is characterized in terms of its reflectivity R and its Kerr rotation angle θ_k. R is a real number (between 0 and 1) that indicates the fraction of the incident power reflected back from the medium at normal incidence. θ_k is generally quoted as a positive number, but is understood to be positive or negative depending on the direction of magnetization; in MO readout, it is the sign of θ_k that carries the information about the state of magnetization, i.e., the recorded bit pattern.

The laser used for readout is usually the same as that used for recording, but its output power level is substantially reduced in order to avoid erasing (or otherwise obliterating) the previously recorded information. For instance, if the power of the write/erase beam is 20 mW, then for the read operation the beam is attenuated to about 3 or 4 mW. The same objective lens that focuses the write beam is now used to focus the read beam, creating a diffraction-limited spot for resolving the recorded marks. Whereas in writing the laser was pulsed to selectively reverse-magnetize small regions along the track, in readout it operates with constant power, i.e., in CW mode. Both up- and down-magnetized regions are read as the track passes under the focused spot. The reflected beam, which is now polarization-modulated, goes back through the objective and becomes collimated once again; its information content is subsequently decoded by polarization-sensitive optics, and the scanned pattern of magnetization is reproduced as an electronic signal.

Differential Detection

Figure 80.33 shows the differential detection system that is the basis of magneto-optical readout in practically all erasable optical storage systems in use today. The beam splitter (BS) diverts half of the reflected beam away from the laser and into the detection module.[1] The polarizing beam splitter (PBS) splits the beam into two parts, each carrying the projection of the incident polarization along one axis of the PBS, as shown in Fig. 80.33(b). The component of polarization along one of the axes goes straight through, while the component

[15]The use of an ordinary beam splitter is an inefficient way of separating the incoming and outgoing beams, since half the light is lost in each pass through the splitter. One can do much better by using a so-called "leaky" polarizing beam splitter.

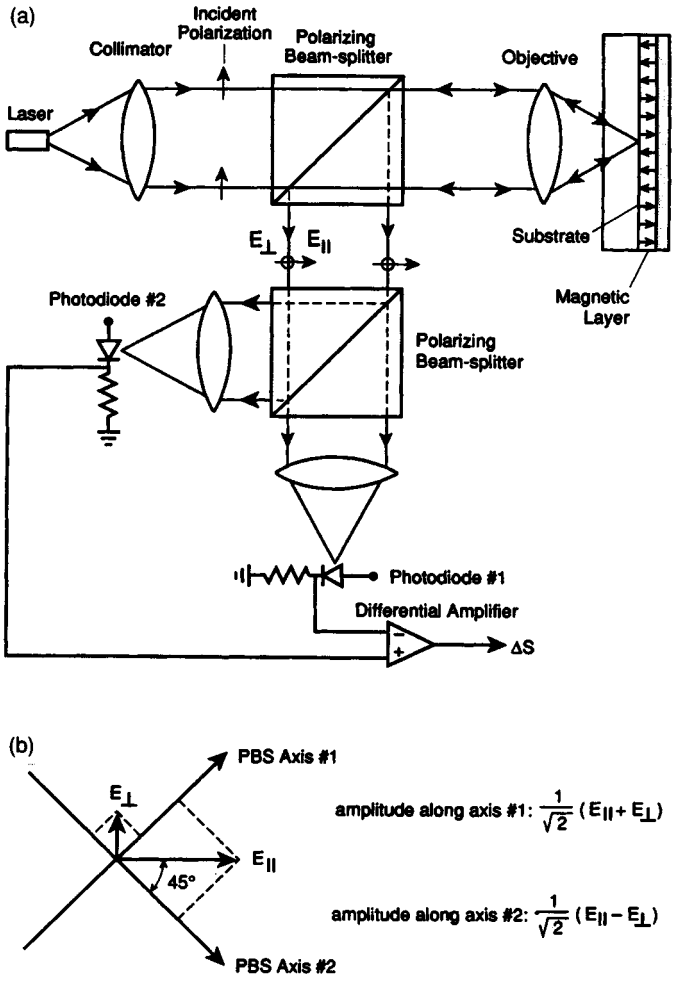


FIGURE 80.33 Differential detection scheme utilizes a polarizing beam splitter and two photodetectors in order to convert the rotation of polarization to an electronic signal. $E_{||}$ and $E_\perp$ are the reflected components of polarization; they are, respectively, parallel and perpendicular to the direction of incident polarization. The diagram in (b) shows the orientation of the PBS axes relative to the polarization vectors.

along the other axis splits off and branches to the side. The PBS is oriented such that in the absence of the Kerr effect its two branches will receive equal amounts of light. In other words, if the polarization, upon reflection from the disk, did not undergo any rotations whatsoever, then the beam entering the PBS would be polarized at 45° to the PBS axes, in which case it would split equally between the two branches. Under this condition, the two detectors generate identical signals and the differential signal ΔS will be zero. Now, if the beam returns from the disk with its polarization rotated clockwise (rotation angle = θ_k), then detector #1 will receive more light than detector #2, and the differential signal will be positive. Similarly, a counterclockwise rotation will generate a negative ΔS. Thus, as the disk rotates under the focused spot, the electronic signal ΔS reproduces the pattern of magnetization along the scanned track.

Materials of Magneto-Optical Data Storage

Amorphous rare earth transition metal alloys are presently the media of choice for erasable optical data storage applications. The general formula for the composition of the alloy may be written $(Tb_yGd_{1-y})_x(Fe_zCo_{1-z})_{1-x}$ where terbium and gadolinium are the rare earth (RE) elements, while iron and cobalt are the transition metals

(TM). In practice, the transition metals constitute roughly 80 atomic percent of the alloy (i.e., $x \simeq 0.2$). In the transition metal subnetwork, the fraction of cobalt is usually small, typically around 10%, and iron is the dominant element ($z \simeq 0.9$). Similarly, in the rare earth subnetwork Tb is the main element ($y \simeq 0.9$) while the gadolinium content is small or it may even be absent in some cases. Since the rare earth elements are highly reactive to oxygen, RE-TM films tend to have poor corrosion resistance and, therefore, require protective coatings. In multilayer disk structures, the dielectric layers that enable optimization of the medium for the best optical/thermal behavior also perform the crucial function of protecting the MO layer from the environment.

The amorphous nature of the material allows its composition to be continuously varied until a number of desirable properties are achieved. In other words, the fractions x, y, z of the various elements are not constrained by the rules of stoichiometry. Disks with very large areas can be coated uniformly with thin films of these media, and, in contrast to polycrystalline films whose grains and grain boundaries scatter the beam and cause noise, amorphous films are continuous, smooth, and substantially free from noise. The films are deposited either by sputtering from an alloy target or by co-sputtering from multiple elemental targets. In the latter case, the substrate moves under the various targets and the fraction of a given element in the alloy is determined by the time spent under each target as well as the power applied to that target. During film deposition the substrate is kept at a low temperature (usually by chilled water) in order to reduce the mobility of deposited atoms and thus inhibit crystal growth. The type of the sputtering gas (argon, krypton, xenon, etc.) and its pressure during sputtering, the bias voltage applied to the substrate, deposition rate, nature of the substarte and its pretreatment, and temperature of the substrate all can have dramatic effects on the composition and short-range order of the deposited film. A comprehensive discussion of the factors that influence film properties will take us beyond the intended scope here; the interested reader may consult the vast literature of this field for further information.

Defining Terms

Automatic focusing: The process in which the distance of the disk from the objective's focal plane is continuously monitored and fed back to the system in order to keep the disk in focus at all times.

Automatic tracking: The process in which the distance of the focused spot from the track center is continuously monitored and the information fed back to the system in order to maintain the read/write beam on track at all times.

Compact disk (CD): A plastic substrate embossed with a pattern of pits that encode audio signals in digital format. The disk is coated with a metallic layer (to enhance its reflectivity) and read in a drive (CD player) that employs a focused laser beam and monitors fluctuations of the reflected intensity in order to detect the pits.

Error correction coding (ECC): Systematic addition of redundant bits to a block of binary data, as insurance against possible read/write errors. A given error-correcting code can recover the original data from a contaminated block, provided that the number of erroneous bits is less than the maximum number allowed by that particular code.

Grooved media of optical storage: A disk embossed with grooves of either the concentric-ring type or the spiral type. If grooves are used as tracks, then the lands (i.e., regions between adjacent grooves) are the guard bands. Alternatively, lands may be used as tracks, in which case the grooves act as guard bands. In a typical grooved optical disk in use today the track width is 1.1 μm, the width of the guard band is 0.5 μm, and the groove depth is 70 nm.

Magneto-optical Kerr effect: The rotation of the plane of polarization of a linearly polarized beam of light upon reflection from the surface of a perpendicularly magnetized medium.

Objective lens: A well-corrected lens of high numerical aperture, similar to a microscope objective, used to focus the beam of light onto the surface of the storage medium. The objective also collects and recollimates the light reflected from the medium.

Optical path: Optical elements in the path of the laser beam in an optical drive. The path begins at the laser itself and contains a collimating lens, beam shaping optics, beam splitters, polarization-sensitive elements, photodetectors, and an objective lens.

Preformat: Information such as sector address, synchronization marks, servo marks, etc., embossed permanently on the optical disk substrate.

Sector: A small section of track with the capacity to store one block of user data (typical blocks are either 512 or 1024 bytes). The surface of the disk is covered with tracks, and tracks are divided into contiguous sectors.

Thermomagnetic process: The process of recording and erasure in magneto-optical media, involving local heating of the medium by a focused laser beam, followed by the formation or annihilation of a reverse-magnetized domain. The successful completion of the process usually requires an external magnetic field to assist the reversal of the magnetization.

Track: A narrow annulus or ring-like region on a disk surface, scanned by the read/write head during one revolution of the spindle; the data bits of magnetic and optical disks are stored sequentially along these tracks. The disk is covered either with concentric rings of densely packed circular tracks or with one continuous, fine-pitched spiral track.

Related Topics

42.2 Optical Fibers and Cables • 43.1 Introduction

References

A. B. Marchant, *Optical Recording*, Reading, Mass.: Addison-Wesley, 1990.

P. Hansen and H. Heitman, "Media for erasable magneto-optic recording," *IEEE Trans. Mag.*, vol. 25, pp. 4390–4404, 1989.

M. H. Kryder, "Data-storage technologies for advanced computing," *Scientific American,* vol. 257, pp. 116–125, 1987.

G. Bouwhuis, J. Braat, A. Huijser, J. Pasman, G. Van Rosmalen, and K. S. Immink, *Principles of Optical Disk Systems*, Bristol: Adam Hilger Ltd., 1985, chap. 2 and 3.

Special issue of *Applied Optics* on video disks, July 1, 1978.

E. Wolf, "Electromagnetic diffraction in optical systems. I. An integral representation of the image field," *Proc. R. Soc. Ser. A*, vol. 253, pp. 349–357, 1959.

M. Mansuripur, "Certain computational aspects of vector diffraction problems," *J. Opt. Soc. Am. A,* vol. 6, pp. 786–806, 1989.

D. O. Smith, "Magneto-optical scattering from multilayer magnetic and dielectric films," *Opt. Acta,* vol. 12, p. 13, 1965.

P. S. Pershan, "Magneto-optic effects," *J. Appl. Phys.,* vol. 38, pp. 1482–1490, 1967.

K. Egashira and R. Yamada, "Kerr effect enhancement and improvement of readout characteristics in MnBi film memory," *J. Appl. Phys.,* vol. 45, pp. 3643–3648, 1974.

H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, London: Oxford University Press, 1954.

P. Kivits, R. deBont, and P. Zalm, "Superheating of thin films for optical recording," *Appl. Phys.,* vol. 24, pp. 273–278, 1981.

M. Mansuripur, G. A. N. Connell, and J. W. Goodman, "Laser-induced local heating of multilayers," *Appl. Opt.,* vol. 21, p. 1106, 1982.

J. Heemskerk, "Noise in a video disk system: experiments with an (AlGa)As laser," *Appl. Opt.,* vol. 17, p. 2007, 1978.

A. Arimoto, M. Ojima, N. Chinone, A. Oishi, T. Gotoh, and N. Ohnuki, "Optimum conditions for the high frequency noise reduction method in optical video disk players," *Appl. Opt.,* vol. 25, p. 1398, 1986.

M. Mansuripur, G. A. N. Connell, and J. W. Goodman, "Signal and noise in magneto-optical readout," *J. Appl. Phys.,* vol. 53, p. 4485, 1982.

J. W. Beck, "Noise considerations of optical beam recording," *Appl. Opt.,* vol. 9, p. 2559, 1970.

S. Chikazumi and S. H. Charap, *Physics of Magnetism*, New York: John Wiley, 1964.

B. G. Huth, "Calculation of stable domain radii produced by thermomagnetic writing," *IBM J. Res. Dev.*, pp. 100–109, 1974.

A. P. Malozemoff and J. C. Slonczewski, *Magnetic Domain Walls in Bubble Materials*, New York: Academic Press, 1979.

A. M. Patel, "Signal and error-control coding," in *Magnetic Recording,* vol. II, C. D. Mee and E. D. Daniel, Eds. New York: McGraw-Hill, 1988.
K. A. S. Immink, "Coding methods for high-density optical recording," *Philips J. Res.,* vol. 41, pp. 410–430, 1986.
L. I. Maissel and R. Glang, Eds., *Handbook of Thin Film Technology,* New York: McGraw-Hill, 1970.
G. L. Weissler and R. W. Carlson, Eds., *Vacuum Physics and Technology,* vol. 14 of *Methods of Experimental Physics*, New York: Academic Press, 1979.
T. Suzuki, "Magneto-optic recording materials," *Mater. Res. Soc. Bull.,* pp. 42-47, Sept. 1996.
K. G. Ashar, *Magnetic Disk Drive Technology,* New York: IEEE Press, 1997.

Further Information

Proceedings of the *Optical Data Storage Conference* are published annually by SPIE, the International Society for Optical Engineering. These proceedings document the latest developments in the field of optical recording each year. Two other conferences in this field are the *International Symposium on Optical Memory* (ISOM), whose proceedings are published as a special issue of the *Japanese* J*ournal of Applied Physics*, and the *Magneto-Optical Recording International Symposium* (MORIS), whose proceedings appear in a special issue of the *Journal of the Magnetics Society of Japan.*

devices the techniques were tailored to gate networks of the type described above. The term *gate* was already in use in the forties to denote the logical elements discussed earlier.

There are very many good references on Boolean algebra and we may quote only a selected few of them. Suffice it to mention the texts by Hill and Peterson [1974], Kohavi [1978], and Hohn [1966]. These books give a sufficiently rigorous formulation of the subject, tailored to the analysis and the design of combinational networks. In addition, like most of the earlier books, Hohn's and Kohavi's texts also contain a discussion of the Boolean techniques used in connection with relay circuits. (Some of the more recent works completely omit this topic, which has been but totally overshadowed by the impressive development of electronic networks.) The reader interested in studying the relation of switching algebra to Boolean algebras in general is referred to Preparata and Yeh [1973] for an elementary introduction.

Defining Terms

Boolean algebra: The algebra of logical values enabling the logical designer to obtain expressions for digital circuits.

Boolean expressions: Expressions of logical variables constructed using the connectives *and, or,* and *not.*

Boolean functions: Common designations of binary functions of binary variables.

Combinational logic: Interconnections of memory-free digital elements.

Switching theory: The theory of digital circuits viewed as interconnections of elements whose output can switch between the logical values of 0 and 1.

Related Topic

79.2 Logic Gates (IC)

References

G. Boole, *An Investigation of the Laws of Thought,* New York: Dover Publication, 1954.
F.J. Hill and G.R. Peterson, *Introduction to Switching Theory and Logical Design,* New York: Wiley, 1974.
F.E. Hohn, *Applied Boolean Algebra,* New York: Macmillan, 1966.
Z. Kohavi, *Switching and Finite Automata Theory,* New York: McGraw-Hill, 1978.
F.P. Preparata and R.T. Yeh, *Introduction to Discrete Structures,* Reading, Mass.: Addison-Wesley, 1973.
C.E. Shannon, "A symbolic analysis of relay and switching circuits," *Trans. AIEE,* vol. 57, pp. 713–723, 1938.

81.2 Logic Circuits

Richard S. Sandige

Section 81.2 deals with two-state (high or low, 1 or 0, or true or false) logic circuits. Two-state logic circuits can be broken down into two major types of circuits: **combinational logic circuits** and **sequential logic circuits.** By definition, the external output signals of combinational logic circuits are totally dependent on the external input signals applied to the circuit. In contrast, the output signals of sequential logic circuits are dependent on all or part of the present state output signals of the circuit that are fed back as input signals to the circuit as well as any external input signals if they should exist. Sequential logic circuits can be subdivided into **synchronous** or **clock-mode** circuits and asynchronous circuits. Asynchronous circuits can be further divided into fundamental-mode circuits and pulse-mode circuits. Fig. 81.15 is the graphic classification of logic circuits.

Combinational Logic Circuits

The block diagram in Fig. 81.16 illustrates the model for combinational logic circuits. The logic elements inside the block entitled *combinational logic circuit* can be any configuration of two-state logic elements such that the output signals are totally dependent on the input signals to the circuit as indicated by the functional relationships in the figure.

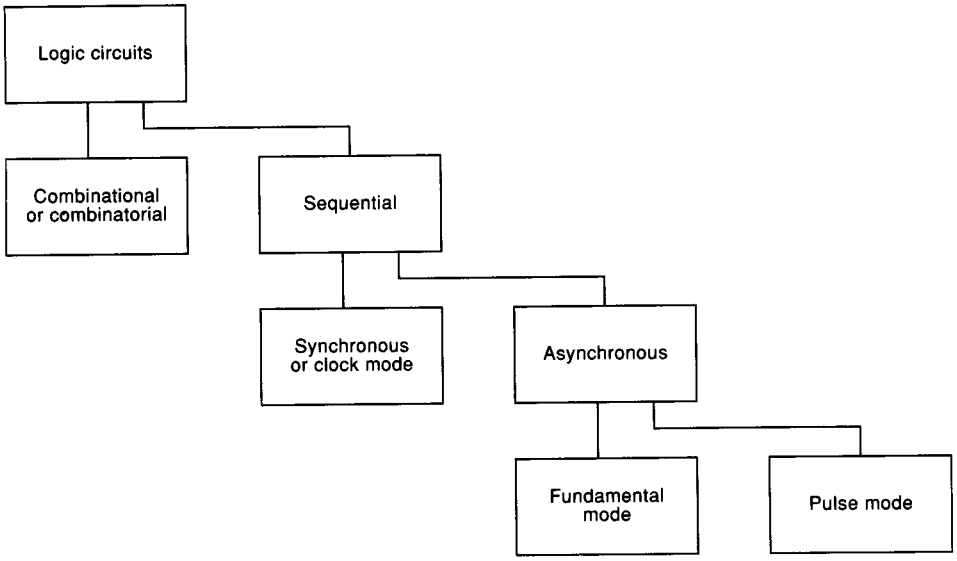


FIGURE 81.15 Graphic classification of logic circuits.

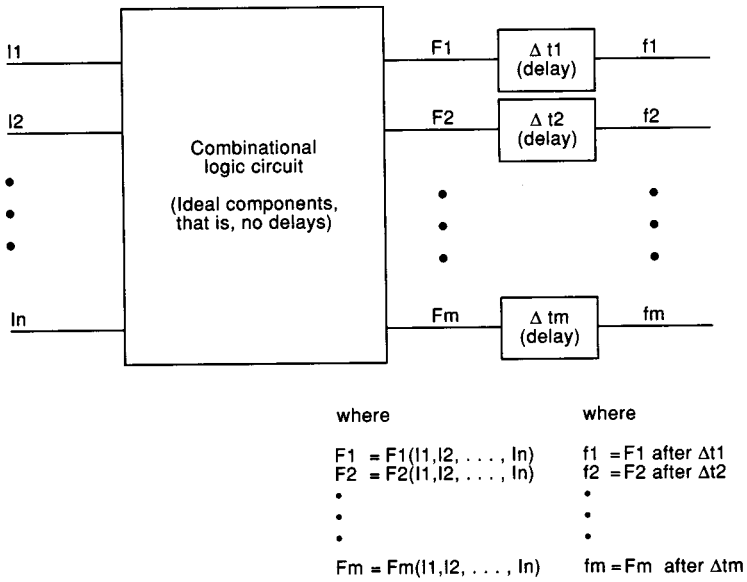


FIGURE 81.16 Block diagram model for combinational logic circuits.

The logic elements can be anything from relays with their slow on and off switching action to modern off-the-shelf integrated circuit (IC) transistor switches with their extremely fast switching action. Modern ICs exist in various technologies and circuit configurations such as transistor-transistor logic (TTL), complementary metal-oxide semiconductor (CMOS), emitter-coupled logic (ECL), and integrated injection logic (I^2L), just to name a few.

The delays in the outputs of the model in Fig. 81.16 represent lumped delays, that is, worst-case delays through the longest delay path from the inputs to each respective output of the combinational logic circuit.

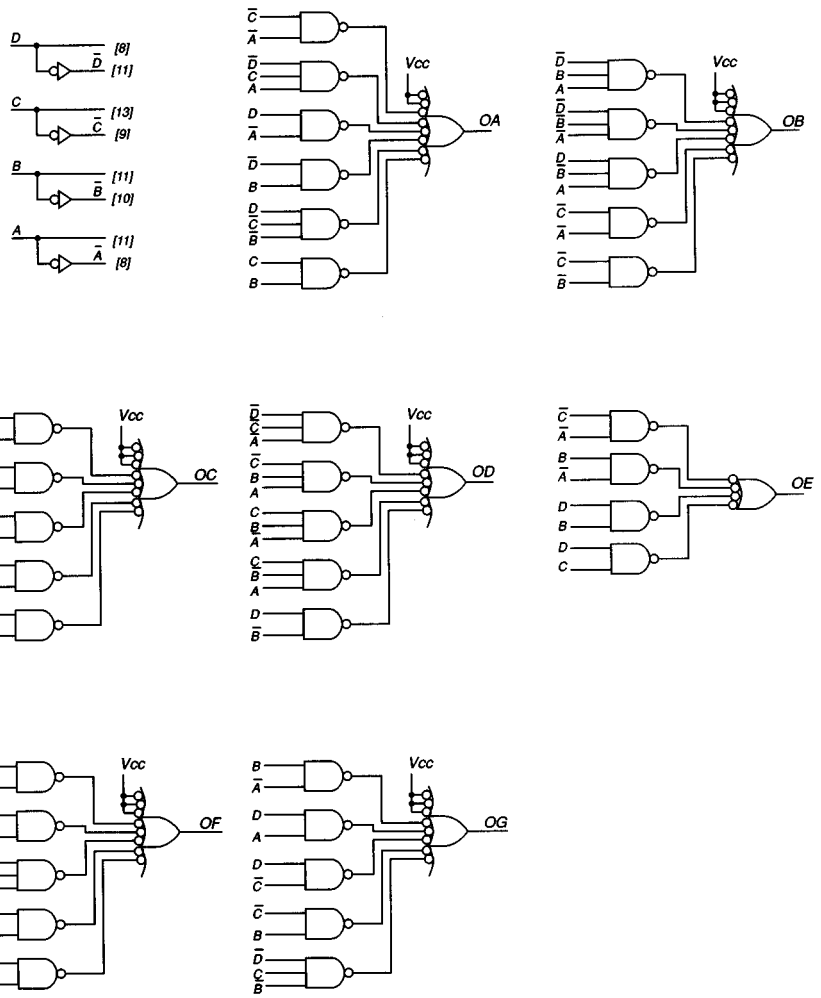


FIGURE 81.17 Gate-level logic circuit for binary to seven-segment hexadecimal character generator.

The lumped delays provide an approximate measure of circuit speed or settling time (the time it takes an output signal to become stable after the input signals have become stable).

Figure 81.17 illustrates the gate-level method (random logic method) of implementing a binary to seven-segment hexadecimal character generator suitable for driving a seven-segment common cathode LED display like the one in Fig. 81.18.

The propagation delays of logic circuits are seldom shown on logic circuit diagrams; however, these delays are inherent in each logic element and must be considered in systems designs. This gate-level combinational logic circuit converts the binary input code 0000 though 1111 represented on the signal inputs D(MSB) C B A(LSB) to the binary code on the signal outputs OA through OG. These outputs generate the hexadecimal characters 0 through F when applied to a seven-segment common cathode LED display. Each of the signal lines D, $\overline{D}$, through A, $\overline{A}$ must be capable of driving the number of gate inputs shown in the brackets (**fan-out requirement**) to both the high-level and low-level required voltages. The output equations for the circuit in Fig. 81.17 are the minimum **sum of products** (SOP) equations for the 1's of the functions OA though OG, respectively, represented by the truth table in Table 81.4.

A more efficient way (in terms of package count) to implement the same combinational logic function would be to use a **medium-scale integration** (MSI) 4- to 16-1ine decoder with gates as illustrated in Fig. 81.19. The tildes are used as in-line symbols for the logical complements of D0 through D15 as recommended by IEEE.

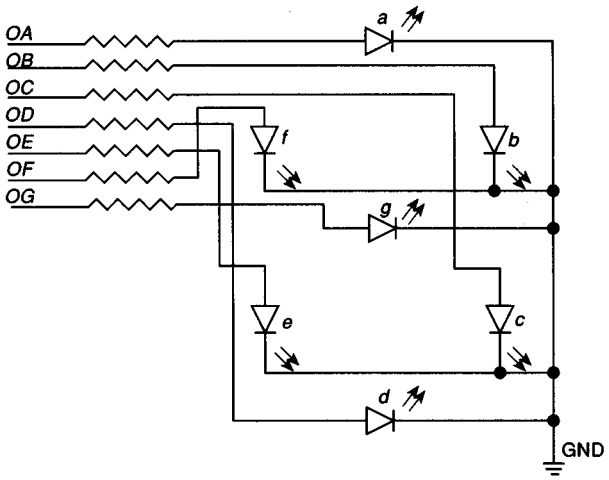


FIGURE 81.18 Seven-segment common cathode LED display.

TABLE 81.4 Truth Table for Binary to Seven-Segment Hexadecimal Character Generator

Binary Inputs				Seven-Segment Outputs							
D	*C*	*B*	*A*	*OA*	*OB*	*OC*	*OD*	*OE*	*OF*	*OG*	Displayed Characters
0	0	0	0	1	1	1	1	1	1	0	0
0	0	0	1	0	1	1	0	0	0	0	1
0	0	1	0	1	1	0	1	1	0	1	2
0	0	1	1	1	1	1	1	0	0	1	3
0	1	0	0	0	1	1	0	0	1	1	4
0	1	0	1	1	0	1	1	0	1	1	5
0	1	1	0	1	0	1	1	1	1	1	6
0	1	1	1	1	1	1	0	0	0	0	7
1	0	0	0	1	1	1	1	1	1	1	8
1	0	0	1	1	1	1	1	0	1	1	9
1	0	1	0	1	1	1	0	1	1	1	A
1	0	1	1	0	0	1	1	1	1	1	b
1	1	0	0	1	0	0	1	1	1	0	C
1	1	0	1	0	1	1	1	1	0	1	d
1	1	1	0	1	0	0	1	1	1	1	E
1	1	1	1	1	0	0	0	1	1	1	F

The decoder circuit in Fig. 81.19 requires only 8 IC packages compared to the gate-level circuit in Fig. 81.17 which requires 18 IC packages. Functionally, both circuits perform the same. The output equations for the circuit in Fig. 81.19 are the canonical or standard SOP equations for the 0's of the functions *OA* though *OG*, respectively, represented by the truth table (Table 81.4). The gates shown in Fig. 81.19 with more than four inputs are eight-input NAND gates with each unused input tied to V_{CC} via a pull-up resistor (not shown on the logic diagram).

An even more efficient way to implement the same combinational logic function would be to utilize part of a simple programmable read only memory (PROM) circuit such as the 27S19 fuse programmable PROM in Fig. 81.20. An equivalent architectural gate structure for a portion of the PROM is shown in Fig. 81.21. The X's in Fig. 81.21 represent fuses that are left intact after programming the device. The code for programming the PROM or generating the truth table of the function can be read either from the truth table (Table 81.4) or directly from each line of the circuit diagram in Fig. 81.21 (expressed in hexadecimal: 7E, 30, 6D, 79, 33, 5B, 5F, 70, 7F, 7B, 77, 1F, 4E, 3D, 4F, and 47) beginning with the first line, which represents binary input 0000,

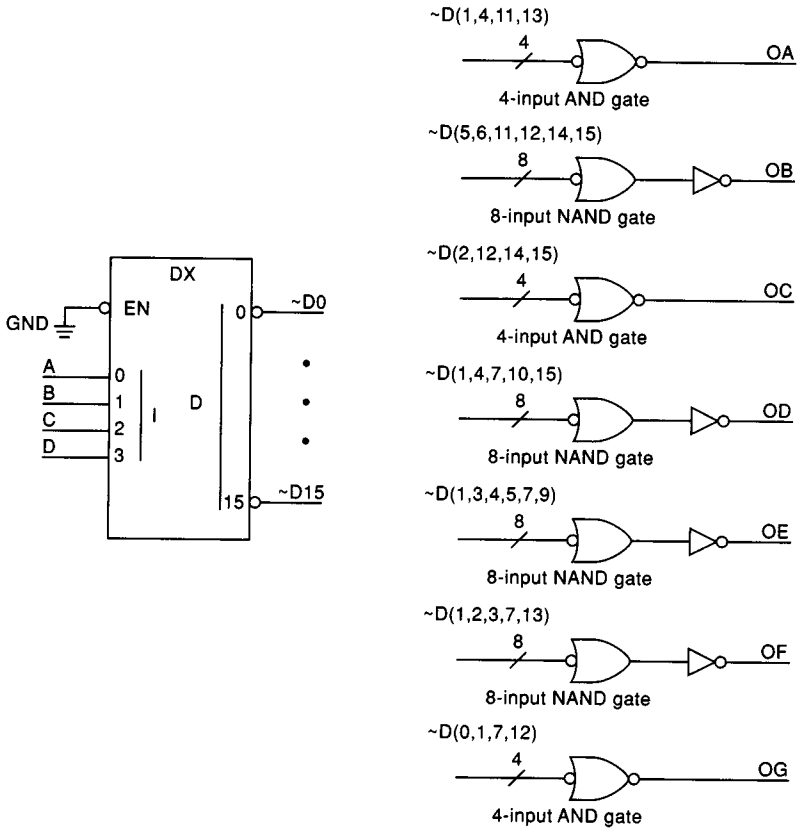


FIGURE 81.19 Decoder logic circuit for binary to seven-segment hexadecimal character generator.

down to the last line, which represents binary input 1111. The PROM solution for the combinational logic circuit is optimum since it represents a maximum efficiency by requiring only a single IC package.

Programmable logic devices (PLDs), such as PROM, programmable array logic (PAL), and programmable logic array (PLA) devices, are fast becoming the preferred devices when implementing combinational as well as sequential logic circuits. This is true because these devices (a) use less real estate on a pc board, (b) shorten design time, (c) allow design changes to be made more easily, and (d) improve reliability because of fewer connections.

Figure 81.22 shows a PAL16L8 implementation for the binary to seven-segment hexadecimal character generator that also requires just a single IC package. The fuse map for this design was obtained using the software program PLDesigner-XL. Karnaugh maps are handy tools that allow a designer to easily obtain minimum SOP equations for either the 1's or 0's of Boolean functions of up to four or five variables; however, there are a host of commercially available software programs that provide not only Boolean reduction but also equation simulation and fuse map generation for PLDs and field programmable gate arrays (FPGAs). PLDesigner-XL (trademark of Minc, Incorporated) is an example of a premier commercial software package available for logic synthesis for PLDs and FPGAs, for both combinational logic and sequential logic circuits.

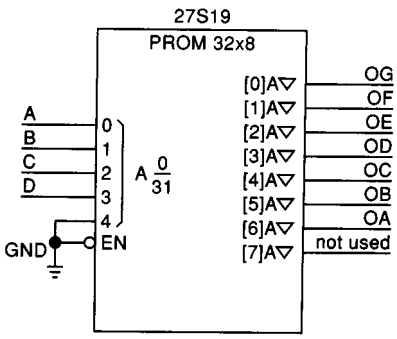


FIGURE 81.20 PROM implementation for binary to seven-segment hexadecimal character generator.

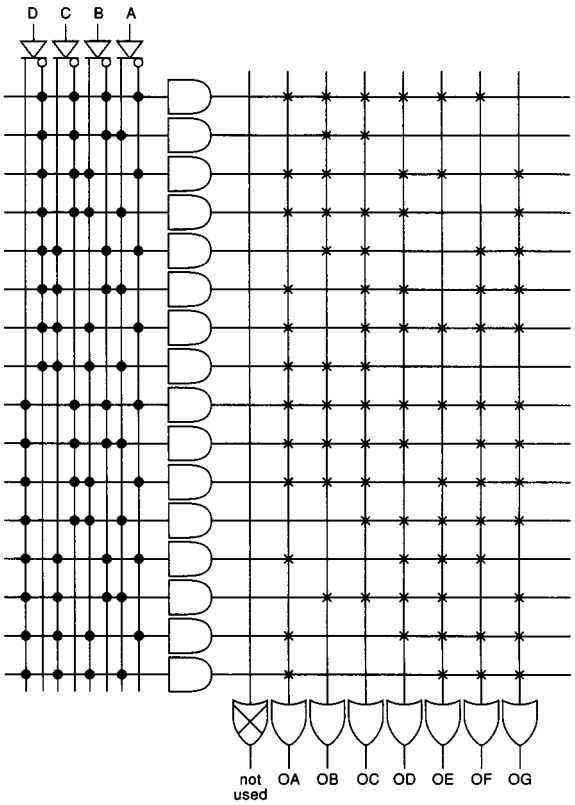


FIGURE 81.21 PROM logic circuit.

Sequential Logic Circuits

A sequential logic circuit is a circuit that has feedback such that the output signals of the circuit are functions of all or part of the present state output signals of the circuit in addition to any external input signals to the circuit. The vast majority of sequential logic circuits designed for industrial applications are synchronous or clock-mode circuits.

Synchronous Sequential Logic Circuits

Synchronous sequential logic circuits change states only at the rising or falling edge of the synchronous clock signal. To allow proper circuit operation, any external input signals to the synchronous sequential logic circuit must generate excitation inputs that occur with the proper setup time (t_{su}) and hold time (t_h) requirements relative to the designated clock edge for the memory elements being used. Synchronous or clock-mode sequential logic circuits depend on the present state of memory devices called bistables or flip-flops (asynchronous sequential logic circuits) that are driven by a system clock as illustrated by the synchronous sequential logic circuit in Fig. 81.23.

With the availability of edge-triggered *D* flip-flops and edge-triggered *J-K* flip-flops in IC packages, a designer can choose which flip-flop type to use as the memory devices in the memory section of a synchronous sequential logic circuit. Many designers prefer to design with edge-triggered *D* flip-flops rather than edge-triggered *J-K* flip-flops because *D* flip-flops are (a) more cost efficient, (b) easier to design with, and (c) more convenient since many of the available PAL devices incorporate edge-triggered *D* flip-flops in the output section of their architectures. PAL devices that contain flip-flops in their output section are referred to as registered PALs (or, in general, registered PLDs). The synchronous sequential logic circuit shown in Fig. 81.24 using edge-triggered *D* flip-flops functionally performs the same as the circuit in Fig. 81.23.

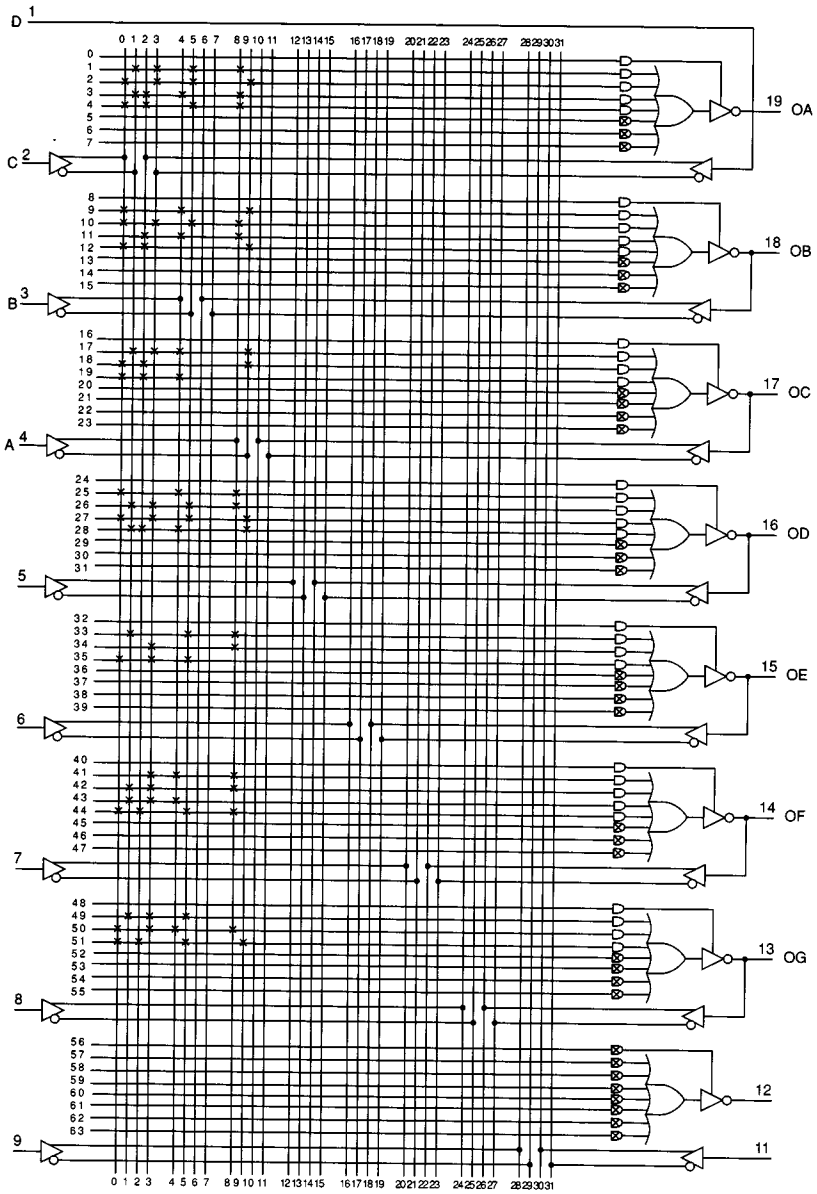


FIGURE 81.22 PAL16L8 implementation for the binary to seven-segment hexadecimal character generator. *(Source: PAL Device Data Book,* Advanced Micro Devices, Sunnyvale, Calif., 1988, p. 5–46.)

Notice that in general more combinational logic gates will be required for *D* flip-flop implementations compared to *J-K* flip-flop implementations of the same synchronous sequential function. Using a registered PAL such as a PAL16RP4A would only require one IC package to implement the circuit in Fig. 81.24. The PAL16RP4A has four edge-triggered *D* flip-flops in its output section, of which only two are required for this design.

Generally speaking, synchronous sequential logic circuits can be designed much more easily (considering design time as the criteria) than fundamental-mode asynchronous sequential logic circuits. With a system clock and edge-triggered flip-flops, a designer does not have to worry about **hazards** or **glitches** (momentary error conditions that occur at the outputs of combinational logic circuits), since outputs are allowed to become stable before the next clock edge occurs. Thus, sequential logic circuit designs allow the use of combinational hazardous circuits as well as the use of arbitrary state assignments, provided the resulting combinational logic gate count or package count is acceptable.

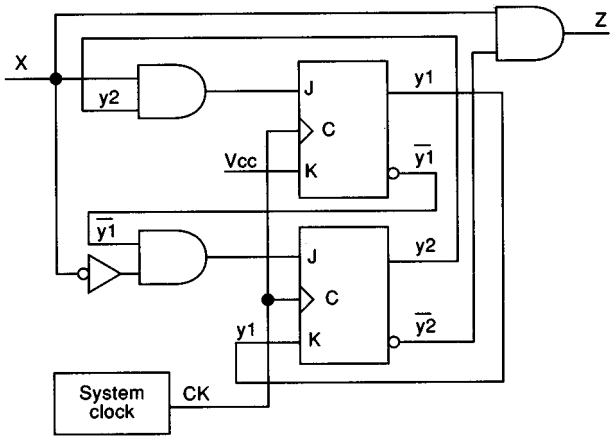


FIGURE 81.23 Synchronous sequential logic circuit using positive edge-triggered *J-K* flip-flops.

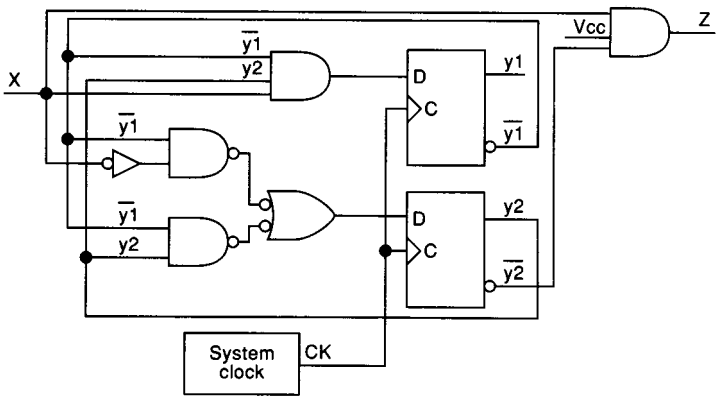


FIGURE 81.24 Synchronous sequential logic circuit using positive edge-triggered *D* flip-flops.

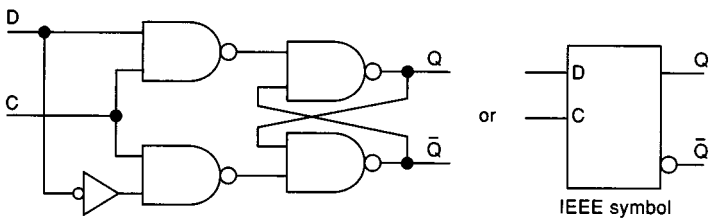


FIGURE 81.25 Fundamental-mode asynchronous sequential logic circuit.

Asynchronous Sequential Logic Circuits

Asynchronous sequential logic circuits may change states any time a single input signal occurs (either a level change for a fundamental mode circuit or a pulse for a pulse mode circuit). No other input signal change (either level change or pulse) is allowed until the circuit reaches a stable internal state. Latches and edge-triggered flip-flops are asynchronous sequential logic circuits and must be designed with care by utilizing hazard-free combinational logic circuits and race-free or critical **race-free state assignments.** Both hazards and race conditions interfere with the proper operation of asynchronous logic circuits. The gated *D* latch circuit

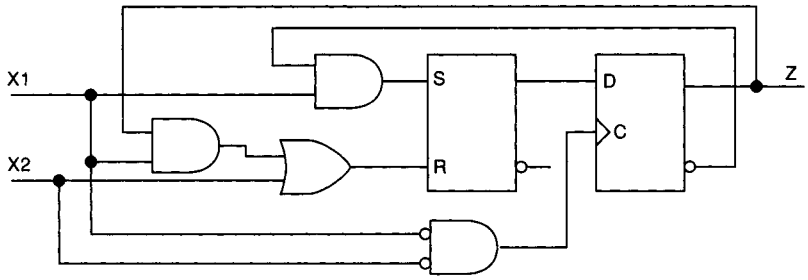


FIGURE 81.26 Double-rank pulse-mode asynchronous sequential logic circuit.

illustrated in Fig. 81.25 is an example of a fundamental-mode asynchronous sequential logic circuit that is used extensively in microprocessor systems for the temporary storage of data.

Quad, octal, 9-bit, and 10-bit transparent latches are readily available as off-the-shelf IC devices for these types of applications. For proper asynchronous circuit operation, the signal applied to the data input *D* of the fundamental-mode circuit in Fig. 81.25 must meet a minimum setup time and hold time requirement relative to the control input *C*, changing the latch to the memory mode when *C* goes to 0. This is a basic requirement for asynchronous circuits with level inputs, i.e., only one input signal is allowed to change at one time. Another restriction requires letting the circuit reach a stable state before allowing the next input signal to change.

An example of a reliable pulse-mode asynchronous sequential logic circuit is shown in Fig. 81.26. While the inputs to asynchronous fundamental-mode circuits are logic levels, the inputs to asynchronous pulse-mode circuits are pulses. Pulse-mode circuits have the restriction that the maximum pulse width of any input pulse must be sufficiently narrow such that an input pulse is no longer present when the new present state output signal becomes available. The purpose of the double-rank circuit in Fig. 81.26 is to ensure that the maximum pulse width requirement is easily met, since the output is not fed back until the input pulse is removed, i.e., goes low or goes to logic 0. The input signals to pulse-mode circuits must also meet the following restrictions: (a) only one input pulse may be applied at one time, (b) the circuit must be allowed to reach a new stable state before applying the next input pulse, and (c) the minimum pulse width of an input pulse is determined by the time it takes to change the slowest flip-flop used in the circuit to a new stable state.

Defining Terms

Asynchronous circuit: A sequential logic circuit without a system clock.

Combinational logic circuit: A circuit with external output signal(s) that are totally dependent on the external input signals applied to the circuit.

Fan-out requirement: The maximum number of loads a device output can drive and still provide dependable 1 and 0 logic levels.

Hazard or glitch: A momentary output error that occurs in a logic circuit because of input signal propagation along different delay paths in the circuit.

Hexadecimal: The name of the number system with a base or radix of 16 with the usual symbols of 0 … 9,A,B,C,D,E,F.

Medium-scale integration: A single packaged IC device with 12 to 99 gate-equivalent circuits.

Race-free state assignment: A state assignment made for asynchronous sequential logic circuits such that no more than a one-bit change occurs between each stable state transition, thus preventing possible critical races.

Sequential logic circuit: A circuit with output signals that are dependent on all or part of the present state output signals fed back as input signals as well as any external input signals if they should exist.

Sum of products (SOP): A standard form for writing a Boolean equation that contains product terms (input variables or signal names either complemented or uncomplemented ANDed together) that are logically summed (ORed together).

Synchronous or clock-mode circuit: A sequential logic circuit that is synchronized with a system clock.

Related Topic

79.2 Logic Gates (IC)

References

Advanced Micro Devices, *PAL Device Data Book,* Sunnyvale, Calif.: Advanced Micro Devices, Inc., 1988.

ANSI/IEEE Std 91-1984, *IEEE Standard Graphic Symbols for Logic Functions,* New York: The Institute of Electrical and Electronics Engineers, 1984.

ANSI/IEEE Std 991-1986, *IEEE Standard for Logic Circuit Diagrams,* New York: The Institute of Electrical and Electronics Engineers, 1986.

K.J. Breeding, *Digital Design Fundamentals,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1992.

F.J. Hill and G.R. Peterson, *Introduction to Switching Theory & Logical Design,* 3rd ed., New York: John Wiley, 1981.

M.M. Mano, *Digital Design,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1991.

E.J. McCluskey, *Logic Design Principles,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.

Minc, *PLDesigner-XL, The Next Generation in Programmable Logic Synthesis, Version 3.5, User's Guide,* Colorado Springs: Minc, Incorporated, 1996.

R.S. Sandige, *Modern Digital Design,* New York: McGraw-Hill, 1990.

Further Information

The monthly magazine *IEEE Journal on Solid-State Circuits* presents papers discussing logic circuits, for example, "Automating the Design of Asynchronous Sequential Logic Circuits," in its March 1991 issue, pp. 364–370.

The monthly magazine *IEEE Transactions on Computers* presents papers discussing logic circuits, for example, "Concurrent Logic Programming as a Hardware Descriptive Tool," in its January 1990 issue, pp. 72–88.

Also, the monthly magazine *Electronics and Wireless World* presents articles discussing logic circuits, for example, "DIY PLD," in its June 1989 issue, pp. 578–581.

81.3 Registers and Their Applications

B.R. Bannister and D.G. Whitehead

The basic building block of any **register** is the flip-flop, but, just as there are several types of flip-flop, there are many different register arrangements, and an idea of the vast range and their interrelationships is given in Fig. 81.27.

The simplest type of flip-flop is the set-reset flip-flop which can be constructed simply by cross-connecting two NAND/NOR gates. This forms an *asynchronous* flip-flop in which the set or reset signal determines both *what* the flip-flop is to do and *when* it is to operate. In fact, if a state change is required, the flip-flop begins to change state as soon as the input change is detected. This flip-flop is therefore useful as a *latch* which is used to detect when some event has occurred, and is often referred to as a *flag* since it indicates to other circuitry that the event has occurred and remains set until the controlling circuitry responds by resetting it.

Flags are widely used in digital systems to indicate a change of state and all microprocessors have a set of flags which, among other things, are used in deciding whether a program branch should or should not be made. Thus the 8086 family of microprocessors [Intel, 1989], for example, has a group of nine flags—three control flags used to control particular modes of operation of the processor and six status flags indicating whether certain conditions have resulted from the most recent arithmetic or logical instruction: zero, carry, auxiliary carry, overflow, sign and parity. For convenience, although they all act independently, these flags are grouped together into what is known as the *flag register* or *program status word register.*

Gated Registers

The more conventional meaning of register applies to a collection of identical flip-flops which are activated as a set rather than individually. They are, in general, available as four-bit or eight-bit and are used in multiples of eight bits in most cases. It is the number of flip-flops in each register that determines the width of the data

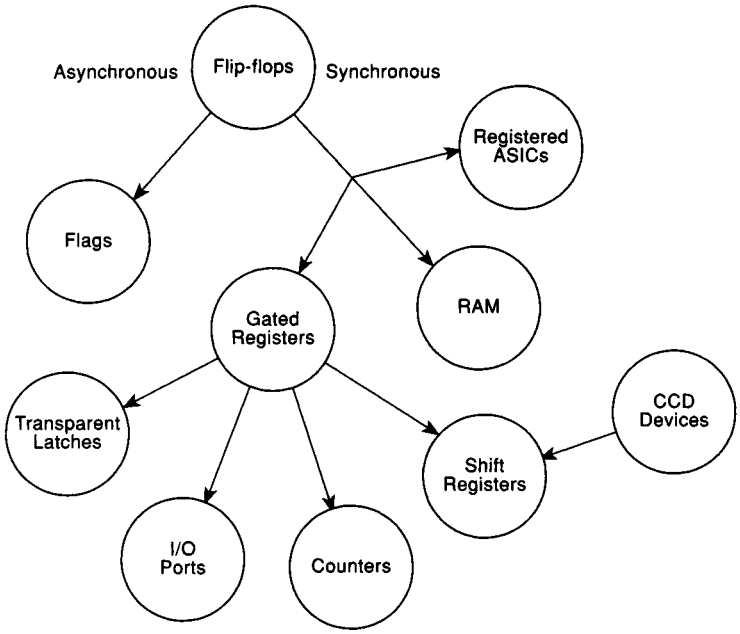


FIGURE 81.27 The register family.

bus in a microprocessor or other bus-based system and that is used to describe the microprocessor. The Z80, for instance, is said to be an eight-bit processor, indicating that the working registers are eight bits wide. The bit values in the register may represent a numerical value in standard fixed point binary, floating point, or some other coded form. Alternatively, they may indicate some logical pattern such as the settings of switches used in an industrial controller.

In order to control *when* the flip-flops set or reset we must make use of *synchronous* flip-flops, leaving the *D* or *J-K* inputs to determine *what* the flip-flop is to do logically, that is to set or reset. In all bus-organized systems it is necessary to control when the data held in the register is fed on to the output bus. This is usually achieved by means of three-state (3S) gates at the register outputs which are disabled, that is, set to their high-impedance state, until the data is required.

A multi-register, bus-structured digital system will have the *n*th bit of each register connected to bit *n* of the data bus at both the input and output of the register. In order to transfer data from one register to another, or, more strictly, to transfer a copy of the contents of one register to another register, the output gates of the source register must be enabled so that the data is fed on to the bus. This data becomes available at the inputs of all registers and is latched in under the control of the appropriate input signals. It is important in the design of the sequencing circuitry that only one set of register output gates can be enabled at any time, although the data can be latched into as many registers as required.

The signal controlling the input to the register is applied to all flip-flops simultaneously and its action depends on the type of flip-flop used. Edge-triggered flip-flops set or reset according to the value on the data inputs at the time the control signal changes. These registers are sometimes known as *staticizers*. After the few nanoseconds required for the flip-flops to settle to their new values, the register content is available at the output gating. The correct operation of the circuitry depends upon certain timing criteria being satisfied and minimum values are quoted by the manufacturers. Each is the smallest time above which the device is guaranteed to operate correctly, but in practice the device probably functions satisfactorily with smaller time intervals on at least some of the parameters. The main timing constraints occur at the inputs to the flip-flops and are illustrated in Fig. 81.28. The interval preceding the active transition of the control input is the setup time, t_{su}, during which the data signal must be held steady; t_h is the hold time and is the interval during which the data signal must be retained following the active transition of the control input; t_w is a minimum pulse width indication which applies to the control inputs such as the clock, reset, and clear. The clock pulse width is usually quoted both for the high state and for the low and is related to the maximum clocking frequency of the flip-flops used in the register.

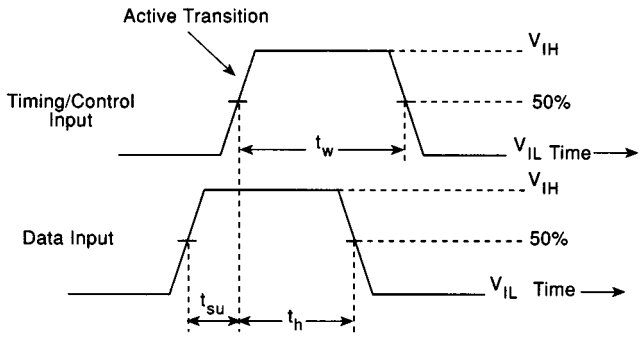


FIGURE 81.28 Control timing parameters.

An alternative flip-flop is the transparent latch, an interesting development of the simple latch. When enabled by a control signal, C, by setting C high or "1", the latch becomes a transparent section of the data path and the data value at the input simply reappears at the output. When the control signal disables the latch, that is C is low or "0", however, the last value applied to the latch is "frozen" and held until the control signal is taken high again. The 74LS373, Fig. 81.29(a), consists of eight transparent latches with a common control input labeled ENABLE. The 74LS374, Fig. 81.29(b), is a typical eight-bit register using positive edge-triggering for all flip-flops. It includes 3S output gates designed specifically for driving highly capacitive loads, such as are found in bus-organized systems, and which respond to an output control signal operating quite independently of the flip-flops. Typical minimum timing figures for the 74LS373 and 74LS374 are shown in Fig. 81.29(c) and the waveforms occurring for the two different types of register are illustrated in Fig. 81.29(d).

An extended form of transparent operation is provided in addressable latches such as the eight-bit 74LS259. As well as being able to store successive bits arriving at a single input, *D*, in the eight addressable latches, using a three-bit address, any latch can be selected for output so that the device can also act as a 1-of-8 decoder or demultiplexer. Four modes of operation are possible under the control of the *enable* and *clear* inputs. In the addressable latch mode the single-addressed latch acts transparently, with all other latches retaining their previous states. When in the memory mode all latches retain their previous states and are unaffected by address or data inputs. In the decode mode the output of the addressed latch follows the level at the *D* input, and in the clear mode all outputs are set low.

Shift Registers

There are essentially two modes of operation for a register, either *serial* or *parallel*, and those we have considered so far have operated in parallel mode. **Parallel operation** affects the entire group of bits held in the register during a single clock pulse. In **serial operation** data bits are inputted (or outputted) sequentially to (or from) the register, one bit for every clock pulse.

A register which has the facility to move the stored bits one place at a time left or right under the control of the clock pulse is called a *shift register* (Fig. 81.30).

Shift registers are normally implemented by means of *D, S-R* or *J-K* flip-flops. As an example, the 74LS165A is shown (Fig. 81.31), consisting of eight *S-R*-type flip-flops with clock, clock inhibit, and shift/load control inputs. The different functions are tabulated in Fig. 81.32.

Data presented to the eight separate inputs is loaded into the register in parallel when the shift/load input is taken low. Shifting occurs when the shift/load input is high and the clock pulse is applied, the action taking place on the low-to-high transition of the clock pulse. Registers are available which switch on the other clock edge. For example, the 74LS295A, which is a four-bit shift register with serial and parallel operating modes, carries out all data transfers and shifting operations on the high-to-low clock transition. This device also provides 3S operation. Selection of the mode of operation is carried out by suitable combinations of the MODE SELECT inputs.

Large-capacity shift registers make use of charge-coupled devices (CCD). These are MOS devices in which data bits are stored *dynamically* as charge between gate and substrate on what is effectively a distributed multi-gate

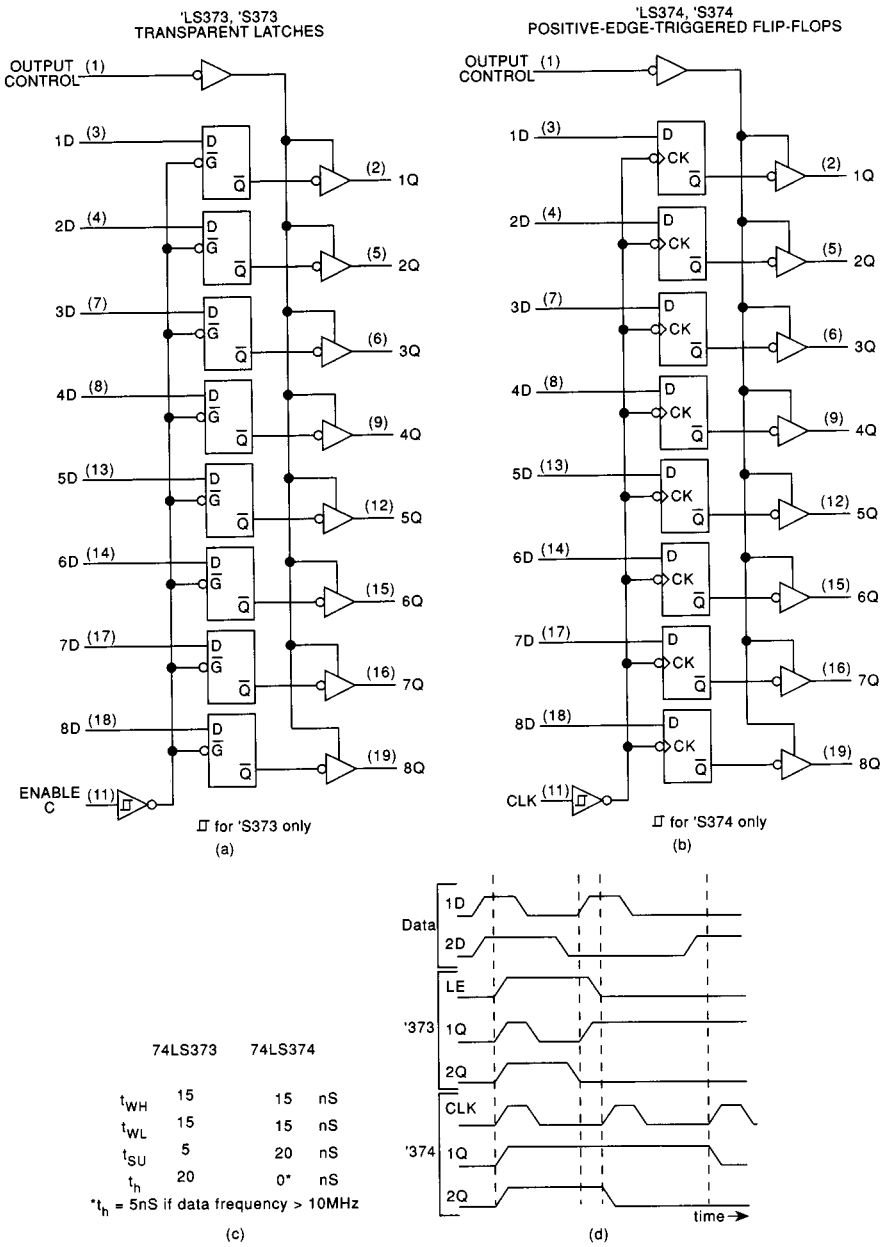


FIGURE 81.29 (a) 74LS373 and (b) 74LS374. (c) Typical minimum timing values and (d) I/O waveforms for the ′LS373/374.

MOS transistor. Consider the diagram (Fig. 81.33) depicting a section through part of an *n*-type substrate which has a series of very closely spaced gate electrodes separating the drain and source of a "stretched" MOS transistor. Using gate "G" and the first storage gate a charge packet of electrons can be introduced into the structure. The overlapping clock pulses allow this charge packet to be moved along the array. At the drain the presence of a charge under the final storage gate is detected by a change in current. Steady-state operation of this type of register is not possible since thermally generated carriers (leakage current) will ultimately cause stationary charge, held beneath a storage gate, to leak away. The result is a minimum operating *shift* frequency of around 20 kHz. The maximum length of the CCD shift register is limited by the charge transfer efficiency; each time the charge packet is transferred between storage gates a fraction is lost. The transfer efficiency also

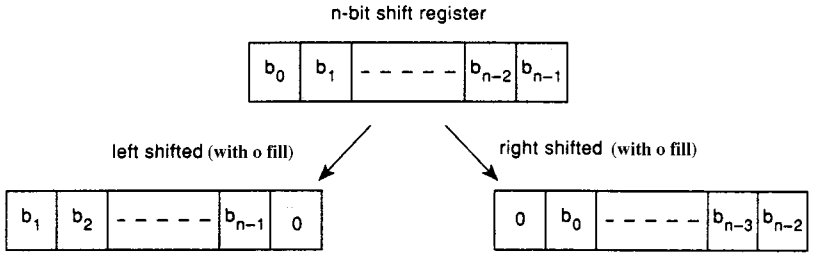


FIGURE 81.30 Shift register operation.

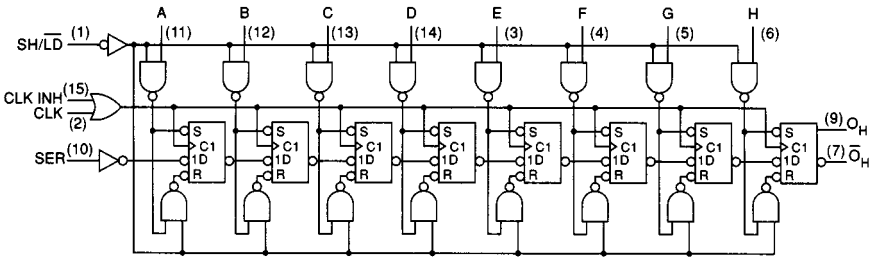


FIGURE 81.31 The 74LS165A shift register. (*Source: TTL Data Book*, Vol. 1, Texas Instruments, Inc.)

Inputs					Internal outputs		output
Shift / Load	Clock Inhibit	Clock	Serial	Parallel A . . . H	Q_A	Q_B	Q_H
L	X	X	X	a . . . h	a	b	h
H	L	L	X	X	Q_{A0}	Q_{B0}	Q_{H0}
H	L	↑	H	X	H	Q_{An}	Q_{Gn}
H	L	↑	L	X	L	Q_{An}	Q_{Gn}
H	H	X	X	X	Q_{A0}	Q_{B0}	Q_{H0}

FIGURE 81.32 Operating modes for 74LS295A.

reduces as the frequency increases, limiting the maximum shift frequency to typically 10 MHz for a 256-bit register. The CCD register is structurally a very simple device and large storage arrays are possible. Devices are available operating from two-, three- and four-phase clocks.

Transfer of data between registers is an important operation in all digital systems and sometimes the data may be modified during the transfer. For example, in an addition routine the contents of one register will be added to the contents of another, the resulting sum then being returned to one of the two registers. Parallel or serial addition could be used, but the example shown in Fig. 81.34 is that of a serial adder. The registers could each be made up of two 74LS295A devices as previously described. The data to be added is transferred to the two shifting registers, *A, B,* using parallel loading of data into the registers, carried out prior to the application of the shift clock pulses shown.

On the falling edge of each clock pulse the data is right-shifted one place. The resultant sum of the two bits plus any carry bit is, on the same clock edge, entered back into register *A*. The *D*-type flip-flop is used to delay the carry bit until the next add time: Data entered at *D* does not appear at *Q* until the falling edge of the clock pulse has occurred, and at such time it is, therefore, too late to modify the previous addition. At the end of the

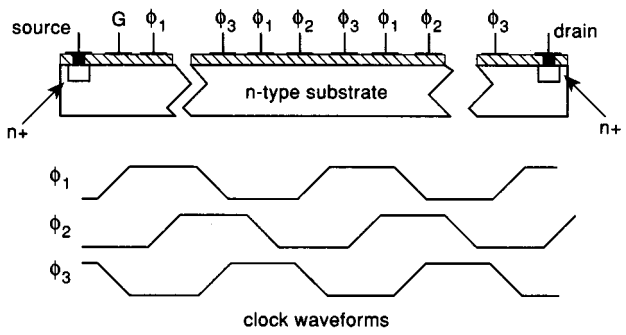


FIGURE 81.33 The CCD register.

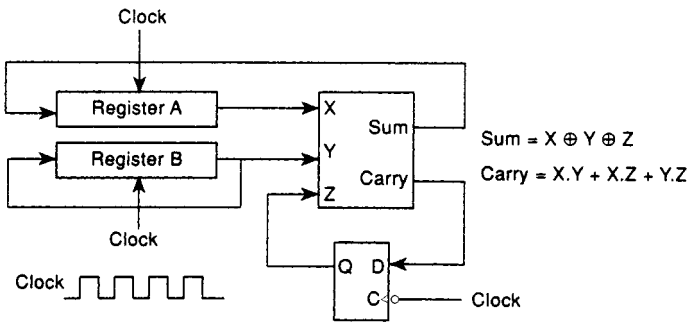


FIGURE 81.34 The serial adder using shifting registers.

addition process, when all the data bits have been shifted through the registers, register *A* contains the sum, and register *B* is unchanged.

A single shift register can be arranged to provide its own input by means of feedback circuits, and its action then becomes **autonomous**, since the only external signal required is the clock signal. There are only a finite number of states of the feedback shift register (FSR), and the output sequence from the register will, therefore, repeat with a cycle length not greater than 2^n bits, where n is the number of flip-flops in the register. This property can be used to create a counter known as a Johnson counter in which the shift register has the *J* and *K* inputs of the first stage fed directly from the Q' *and* Q outputs, respectively, of the last stage. This simple form of feedback leads to the name *twisted-ring counter,* and the result is the generation of a *creeping* or *stepping* code with $2n$ different states. This form of counter is convenient only when the count is small, as the number of flip-flops quickly becomes excessive, but is ideal for a simple decade counter. Unlike standard binary decade counters, the Johnson decade counter requires five flip-flops but no additional feedback circuitry. Gating needed to detect specific settings of the counter is also very simple [Bannister and Whitehead, 1987].

Another set of sequences is obtained if the feedback arrangements are restricted to the use of exclusive-OR, that is modulo-2 functions, and, by correct choice of function, the **linear feedback shift register (LFSR)** so formed generates a *maximal length sequence,* or *m-sequence.* A maximal length sequence has a length of $2^n - 1$ bits (the all-zeros state is not included, since the mod-2 feedback would not allow any escape from that state, so the sequence has a 0 missing) with useful properties of repeatable randomness and is, therefore, described as a *pseudorandom binary sequence* (PRBS). The number of maximal length sequences for a register of length *n*, and the feedback arrangements to achieve them, are not at all obvious, but have been worked out for a large number of cases [Messina, 1972]. A 4-bit LFSR will produce only one maximal length *sequence,* but a 10-bit register can produce 30 distinct *m*-sequences, and a 30-bit register produces no less than 8,910,000 distinct sequences!

Shift registers can be used in parallel to form a *first-in, first-out* (*FIFO*) memory. These are typically 128 × 8-bit register memories with independent input and output buses. At the input port, data is controlled by a

shift-in clock operating in conjunction with an input ready signal which indicates whether the memory is able to accept further words or is now full. The data entered is automatically shifted in parallel to the adjacent memory location if it is empty and as this continues the data words stack up at the output end of the memory. At the output port, data transfers are controlled by a shift-out clock and its associated output ready signal. The output ready signal indicates either that a data word is ready to be shifted out or that the memory is now empty. FIFOs can easily be cascaded to any desired depth and operated in parallel to give any required word length. This type of memory is widely used in controlling transfers of data between digital subsystems which operate at different clock rates and is often known as an *elastic buffer.*

Register Transfer Language

The transfer of data between registers is described using a simple notation termed the register transfer language or RTL. For data transferred from register A to register B we write: $B \leftarrow A$. The symbol $\leftarrow$ is called the *transfer operator.* Note that this statement does not indicate how many bits are to be transferred. To define the size of the register we declare the size thus: $A[8]$, $B[16]$, here defining an 8-bit register and a 16-bit register. If the action to be taken is the transfer of the most significant bit (7th bit) of register A to the least significant bit (bit 0) of register B, then we write: $B[0] \leftarrow A[7]$. Usually data is transferred by the control signal or a clock pulse. If such a signal is designated "C," then we would describe the action by C: $B \leftarrow A$.

Returning to the serial adder circuit shown earlier, we could describe the register transfers thus:

$$A[8],B[8],D[1]$$

$$C\text{: } A[7] \leftarrow A[0] \oplus B[0] \oplus D[0],\ B[7] \leftarrow B[0],\ D[0] \leftarrow \text{Carry}$$

Here in the declaration statement we refer to the D-type flip-flop as a single-bit register. Simultaneous processes are separated by a comma; sequential processes would be separated by a semicolon. The symbol "$\oplus$" is the exclusive-or (XOR) operator. Other logical operations include NOT, AND, OR. The AND operation is also called the *masking* operation because it can be used to remove (or select) specific sections of data from a register. Thus the operation $A[8] \leftarrow A[8]$ &3CH will result in the most significant two bits and the least significant two bits of the eight-bit register A being set to zero. Note that 3CH refers to the hexadecimal number 3C, i.e., 00111100 in binary. Some other terms commonly used are as follows:

$D \leftarrow A'$	transfer the complement of A to D
$A \leftarrow A + 1$	increment A
$A[8{:}15] \leftarrow B[8{:}15]$	transfer bits 8 through 15 from B to A

In order to differentiate between arithmetic and logical operations it is usual to represent OR and AND by $\vee$ and $\wedge$. Table 81.5 lists some typical RTL examples that include arithmetic, bit-by-bit logic, shift, rotate, scale and conditional operations. It is assumed that the three registers are set initially to $A = 10110$, $B = 11000$ and $C = 00001$.

Input/Output Ports

The working registers provided in microprocessors may be thought of as high-speed extensions to the memories used for storing programs and data. The random access memories (RAM) themselves are also arrays of registers, though the form of circuit used differs considerably from the more conventional register. The need to transfer data in and out of the system has led manufacturers to produce special registers which are further extensions to the internal memory and are known as **input and output ports.** One of the simplest of input/output ports is the Intel 8212 (Fig. 81.35). This has two modes of operation selected by the mode input, *MD.* With *MD* at 0 the device acts as an input port and a peripheral unit can enter data on the *DI* lines by sending a high strobe

TABLE 81.5 Typical RTL Examples

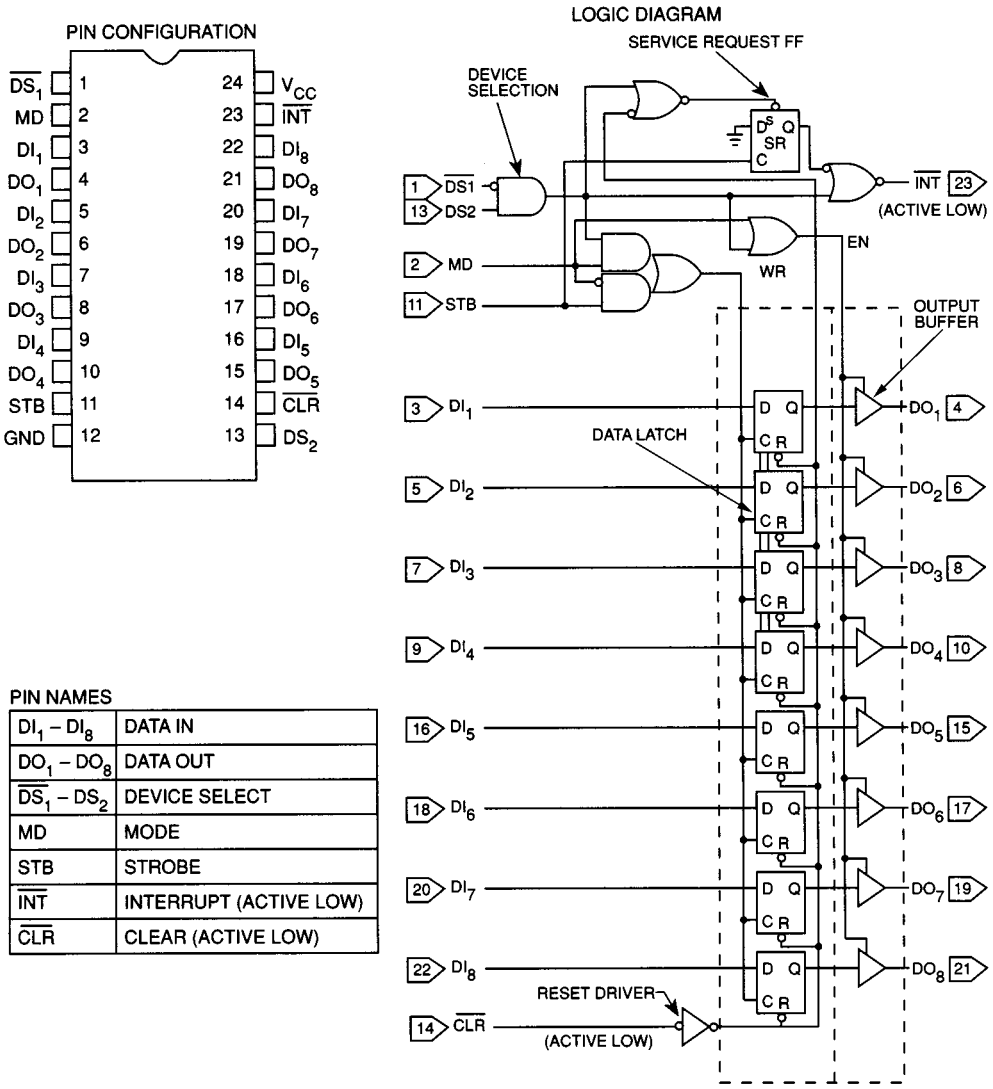
Type of Operation	Meaning	Register Bits after Operation
General		
$A_3 \leftarrow A_2$	Bit 2 of A to bit 3 of A	A = 11110
$A_3 \leftarrow B_4$	Bit 4 of B to bit 3 of A	A = 11110
$A_{1-3} \leftarrow B_{1-3}$	Bits 1 through 3 of B to bits 1 through 3 of A	A = 11000
$A_{1,4} \leftarrow B_{1,4}$	Bits 1 and 4 of B to bits 1 and 4 of A	A = 10100
$A_{1-3} \leftarrow B_z$	Groups of bit Z of B to bits 1 through 3 of A	A = 11000
Arithmetic		
$B \leftarrow 0$	Clear B	B = 00000
$A \leftarrow B_2 + C$	Sum of B and C to A	A = 11001
$A \leftarrow B - C$	Difference $B - C$ to A	A = 10111
$C \leftarrow C + 1$	Increment C by 1	C = 00010
Logic		
$A \leftarrow B \wedge C$	Bit-by-bit AND result of B and C to A	A = 00000
$A \leftarrow B \vee C_4$	OR operation result of B with bit 4 of C to A	A = 11000
$C \leftarrow \overline{C}$	Complement C	C = 11110
$B \leftarrow \overline{B} + 1$	2's complement of B	B = 01000
$B \leftarrow A \oplus C$	XOR operation result of A and C to B	B = 10111
Serial		
$B \leftarrow$ sr B	Shift right B one bit	B = 01100
$B \leftarrow$ sl B	Shift right B one bit	B = 00110
$B \leftarrow$ sr2 B	Shift right B two bits	B = 00110
$B \leftarrow$ rr B	Rotate right B one bit	B = 01100
$B \leftarrow$ rl2 B	Rotate left two bits	B = 00011
$B \leftarrow$ scr B	Scale B one bit (shift right with sign bit unchanged)	B = 11100
$B \leftarrow$ scl B	Scale B one bit (shift left with sign bit unchanged)	B = 10000
$B, C \leftarrow$ sr2 B, C	Shift right concatenated B and C two bits	B, C = 0011000000
Conditional		
IF (B_4 = 1)	If bit 4 of B is a 1, then C is cleared	C = 00000
$C \leftarrow 0$		
IF ($B \geq C$)	If B is greater than or equal to C, then B is cleared and C is set to 1	B = 00000
$B \leftarrow 0$, $C_1 \leftarrow 1$		C = 00011

Initial values: A = 10110, $B = \underbrace{11000}_{z}$ and C = 00001.

signal, STB. When the central processor is ready for the data it selects the port by setting the correct address bits on the *device select* inputs. This enables the 3S output buffers and data is routed to the processor data bus via the *DO* lines. This device also includes a service request flip-flop to generate an interrupt signal to the processor when the data is ready. In the alternative mode of use, with the mode input at 1, the device select logic routes data from the processor, now connected to the *DI* inputs, so the 8212 acts as an output port. The data is immediately available to the peripheral unit on the *DO* lines, as the 3S output buffers are permanently enabled. A more sophisticated range of input and output facilities is provided by most microprocessor manufacturers in the form of programmable input/output ports or peripheral interfaces. These are special registers with appropriate buffers and additional built-in control and status registers to facilitate proper system operation.

The majority of input/output devices use 3S bidirectional buffers which switch to the high-impedance state when not enabled. Some input/output ports, however, such as the 8051 family of microcontrollers and derivatives, are provided with *quasi-bidirectional* ports. In this construction, each port pin has an internal pull-up transistor, as shown in Fig. 81.36. For the port pin to be operative as an input the port latch must contain a 1, so that the output FET driver is turned off. (All port latches in the 8051 are set by the reset function.) Under this condition, the pin voltage is pulled high but can be taken low by an external signal when required. These inputs can, therefore, be driven in a normal way by TTL and MOS circuits and can also cope with open-collector or open-drain circuits using the pull-up transistors as load resistors.

Yet more flexible capabilities are provided in the Rockwell 6522 versatile interface adapter (VIA), which, in addition to two 8-bit bidirectional input/output ports, contains two 16-bit programmable timer/counters and



PIN NAMES

$DI_1 - DI_8$	DATA IN
$DO_1 - DO_8$	DATA OUT
$\overline{DS_1} - DS_2$	DEVICE SELECT
MD	MODE
STB	STROBE
$\overline{INT}$	INTERRUPT (ACTIVE LOW)
$\overline{CLR}$	CLEAR (ACTIVE LOW)

FIGURE 81.35 The Intel 8212. (*Source: Microprocessor Component Handbook,* Intel Corp.)

a serial data port using a serial-parallel, parallel-serial shift register. (See internal register summary in Fig. 81.37) Each line of the input/output ports, *A* and *B*, can be individually programmed as an input or an output by means of *data direction registers,* DDR*A* and DDR*B,* respectively, which "mirror" the ports. That is, a bit in the data direction register set at 1 causes the corresponding bit of the port to act as an output, with a value determined by the output latch setting. If the bit in the data direction register is set at 0, this causes the corresponding bit of the port to act as an input, and incoming data may be latched into an internal register under the control of a special peripheral control line, *CA*1 or *CB*1. Data may be written into register bits which are programmed as inputs but the output signal is unaffected. The input latching may be enabled or disabled. When reading a peripheral port with the latching disabled, the value read is the current value on the input pin; with the latching enabled, the value read is the value which existed at the input pin when the active transition occurred at the control input.

The interval timer/counters can operate in a variety of modes and are retriggerable so that an interrupt signal is generated either once, when the interval expires, or continually, when subsequent intervals expire. If an external output is required, it can be either a one-shot output or a squarewave, with the output changing level each time the interval count is completed. The main timer/counter consists of two 8-bit latches operating in

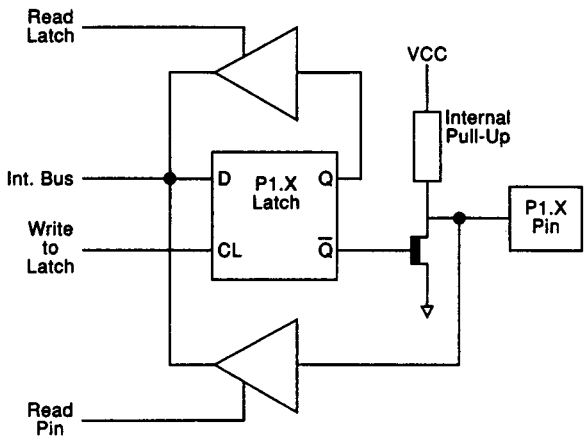


FIGURE 81.36 Intel 8051 quasi-bidirectional port.

Register Number	RS Coding				Register Desig.	Description	
	RS3	RS2	RS1	RS0		Write	Read
0	0	0	0	0	ORB/IRB	Output Register "B"	Input Register "B"
1	0	0	0	1	ORA/IRA	Output Register "A"	Input Register "A"
2	0	0	1	0	DDRB	Data Direction Register "B"	
3	0	0	1	1	DDRA	Data Direction Register "A"	
4	0	1	0	0	T1C-L	T1 Low-Order Latches	T1 Low-Order Counter
5	0	1	0	1	T1C-H	T1 High-Order Counter	
6	0	1	1	0	T1L-L	T1 Low-Order Latches	
7	0	1	1	1	T1L-H	T1 High-Order Latches	
8	1	0	0	0	T2C-L	T2 Low-Order Latches	T2 Low-Order Counter
9	1	0	0	1	T2C-H	T2 High-Order Latches	
10	1	0	1	0	SR	Shift Register	
11	1	0	1	1	ACR	Auxiliary Control Register	
12	1	1	0	0	PCR	Peripheral Control Register	
13	1	1	0	1	IFR	Interrupt Flag Register	
14	1	1	1	0	IER	Interrupt Enable Register	
15	1	1	1	1	ORA/IRA	Same as Reg 1 Except No "Handshake"	

FIGURE 81.37 Internal register summary of Rockwell 6522 VIA. (*Source: Synertek Data Book.*)

conjunction with a 16-bit counter. The latches store the data which is to be loaded into the counter and are loaded sequentially, since 16 bits are required for the counter but the data bus is only 8 bits wide. When loading the counter with a specific value, the low-order byte is actually routed to the low-order 8-bit latch. Then, when the high-order byte is supplied, it is simultaneously written into both the high-order 8-bit latch and the high-order counter byte, and the low-order byte is also transferred from the latch to the counter. Countdown then begins on the next clock pulse. This method ensures that the correct 16-bit value is loaded, but it also means that the current count value can be modified, during counting if required, by writing to the counter, or the *subsequent* counts can be modified by writing to the latches.

The shift register performs serial data transfers in and out of a given pin under the control of an internal modulo-8 counter. When all the necessary control registers and interrupt handling registers are included with the data direction registers, together with the ports and counters of direct interest to the user, the device requires a total of 16 registers. These are individually addressed using four register select pins, *RS*0–*RS*3.

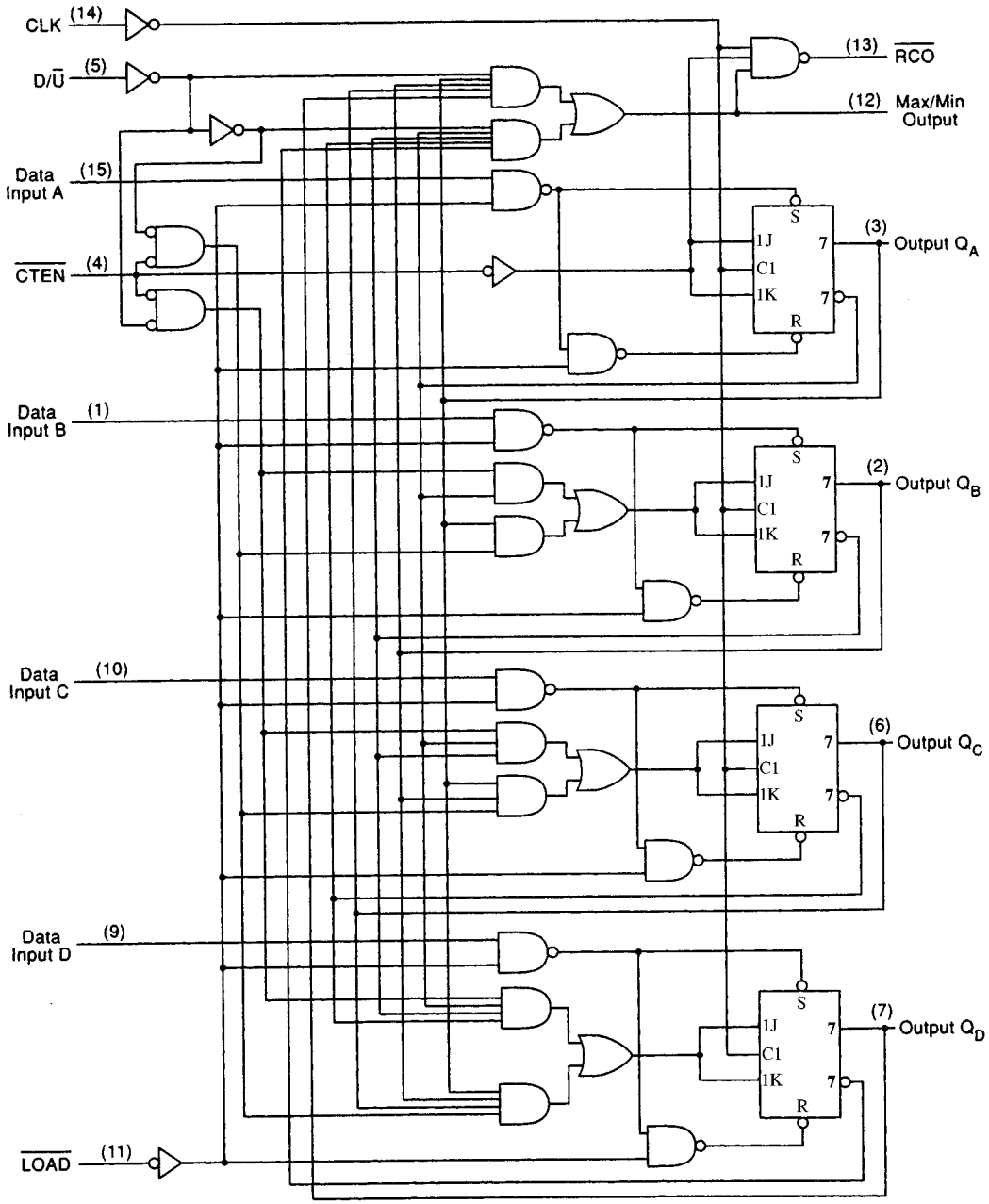


FIGURE 81.38 The 74LS191 programmable counter.

Counters

A register can be loaded with any combination by applying the correct bit pattern to the input data lines and activating the control line. As with the feedback shift registers, it is then only a small step to arrange that the register itself provides the input data by use of feedback connections and, if other circuitry is included to increment the value each time, we have a synchronous counter. The 74L5191 (Fig. 81.38) is a programmable counter which retains the facility for parallel loading of external data.

Each output may be preset to either level by entering the data at the inputs while the LOAD signal is low. The outputs change to the new values independently of the count pulses, and counting continues when pulses

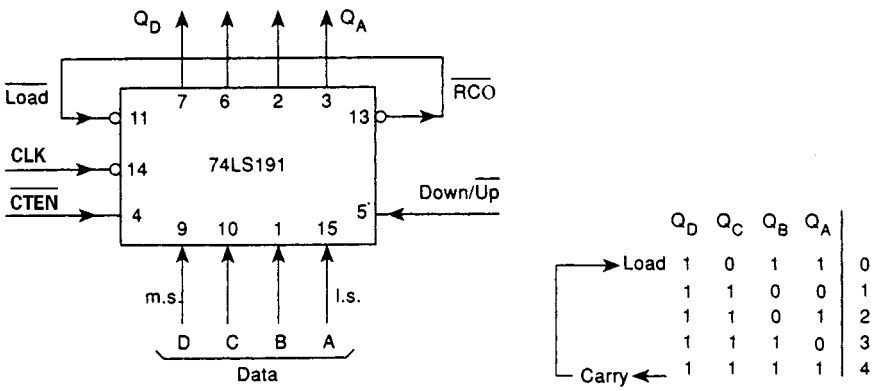


FIGURE 81.39 Programmable counter giving modulo-5.

are applied to the clock input. The master-slave flip-flops are triggered by a low-to-high transition of the clock. The "terminal count" and "ripple clock" outputs facilitate cascading of several counters. The ripple clock carry/borrow output signal, RCO, is a pulse equal in length to the clock pulse when the counter overflows or underflows, that is, when it is incremented from 1111 or decremented from 0000. By using this signal to reload the value at the data inputs we create a counter of modulus less than 16. Figure 81.39, for example, shows the arrangement to give a modulo-5 count, by reloading 1011 each time the ripple clock pulse occurs.

Registered ASICs

Developments in application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs) over recent years have provided digital system designers with a wide range of flexible devices which can be programmed for the specific job in hand. The vast majority of digital subsystems involve sequential logic to a greater or lesser extent, and the array manufactures have provided most devices with registers at either input or output and, sometimes, at both. The Altera EP512 is a good example, and its input structure is shown in Fig. 81.40. The device has eight programmable inputs in which each may be configured as any one of the following:

Synchronous *D*-type flip-flop (register)
Asynchronous *D*-type flip-flop (register)
Synchronous latch
Asynchronous latch
Flowthrough latch (transparent latch)

The *internal latch enable* (ILE) input carries the clock when synchronous mode is selected, whereas asynchronous mode control signals are generated internally. The EP512 contains 12 macrocells and the input/output architecture provides each macrocell with over 50 possible configurations. Each I/O unit can be individually configured for combinatorial or registered output, with the output polarity also programmable. Four different types of flip-flops (*D, T, J-K, S-R*) can be implemented in each I/O unit, for either synchronous or asynchronous operation.

Standard Graphic Symbols

The use of standardized graphical symbols is becoming widespread and the family of registers have their own coherent set of symbols. Two representative examples are given in Fig. 81.41. As shown, the eight-bit shift register is designated SGR8. The direction of shift is given by the arrow. The "1*D*" is part of a notation called *dependency notation.* Clock input C1 controls the inputs labeled 1D, of which only one of four is shown. The reset "*R*" and the clock are common to all units and are shown as inputs to the common block. The external reset line carries a polarity symbol which indicates that a low signal must be applied to reset the 4-bit register. For further details see the References at the end of this section.

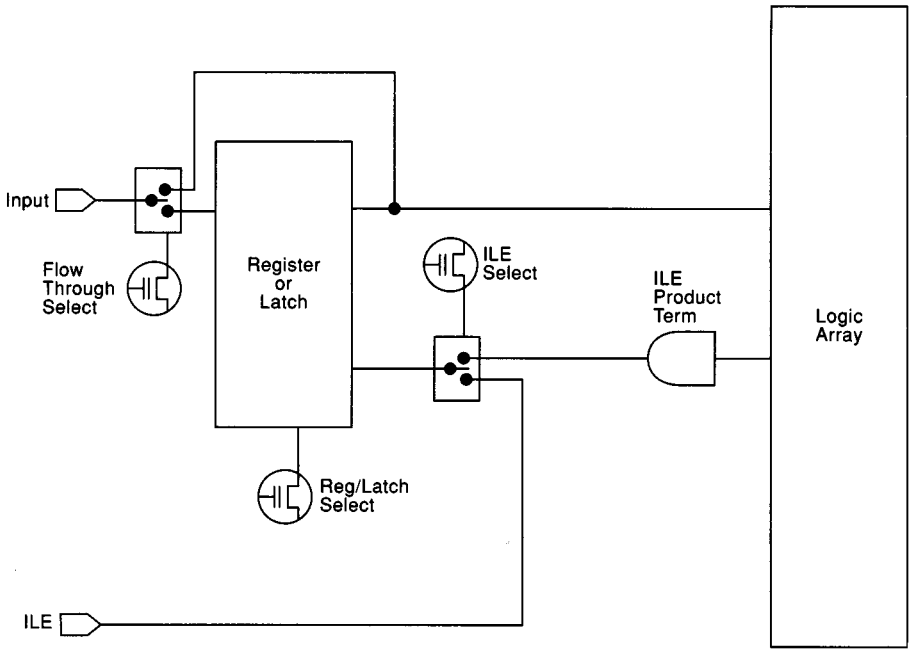


FIGURE 81.40 Input arrangements of the Altera EP512. (*Source: Altera Data Book*, 1988.)

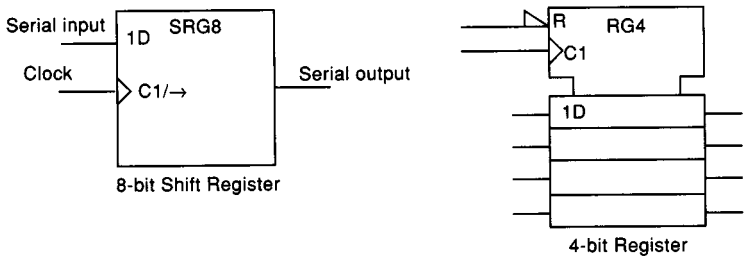


FIGURE 81.41 Standard graphic symbols. (*Source:* ANSI/IEEE Std. 91-1984.)

Defining Terms

Autonomous operation: Operation of a sequential circuit in which no external signals, other than clock signals, are applied. The necessary logic inputs are derived internally using feedback circuits.

Input/output port: A form of register designed specifically for data input-output purposes in a bus-oriented system.

Linear feedback shift register (LFSR): An autonomous feedback shift register in which the feedback function involves only exclusive-OR operations.

Parallel operation: Data bits on separate lines (often in multiples of eight) are transferred simultaneously under control of signals common to all lines.

Register: A circuit formed from several identical gated flip-flops or latches and capable of storing several bits of data.

Serial operation: Data bits on a single line are transferred sequentially under the control of a single signal.

Related Topic

79.3 Bistable Devices

References

Altera, *Data Book,* 1988.
B.R. Bannister and D.G. Whitehead, *Fundamentals of Modern Digital Systems*, London: Macmillan, 1987.
IEEE, Standard Graphic Symbols for Logic Functions, ANSI/IEEE Std. 91-1984, New York, 1984.
Intel Corporation, *8086/8088 User's Manual.*
Intel Corporation, *Microprocessor and Peripheral Handbook.*
E.L. Johnson and M.A. Karim, *Digital Design*, Boston: PWS Engineering, 1987.
I. Kampel, *A Practical Introduction to the New Logic Symbols,* Boston: Butterworth, 1985.
A. Messina, "Considerations for non-binary counter applications," *Computer Design,* vol. 11, no. 11, Nov. 1972.
Synertek, *Data Book.*
Texas Instruments, Inc., *TTL Data Book.*

Further Information

The monthly journal *IEEE Transactions on Computers* regularly has articles involving the design and application of registers and associated systems. Further information can be obtained from IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

The *IEE Proceedings-E, Computers and Digital Techniques,* published bi-monthly by the Institution of Electrical Engineers (Michael Faraday House, Six Hills Way, Stevenage, Herts. SG1 2AY UK), is also a useful source of information on the application of register devices.

81.4 Programmable Arrays

Martin Bolton

Programmable arrays or **programmable logic devices** (PLDs) are general-purpose combinational or sequential digital components whose ultimate function is determined by the designer. They leave the manufacturer in an unprogrammed state. The configurations of internal switches are fixed after the particular logic function for the PLD has been prepared and checked using a computer-aided design package appropriate for the PLD family used. PLDs belong to the family of application-specific integrated circuits (ASICs).

PLDs are manufactured today in most digital integrated circuit technologies—principally CMOS and bipolar silicon, and gallium arsenide. The programmable switches themselves can be fuses, antifuses, floating-gate MOSFETs, and RAM cells. The floating-gate devices can usually be erased and reprogrammed, while the RAM-based devices can be reconfigured dynamically. Most PLDs have to be programmed with the aid of a programmer, a unit which is able to deliver the appropriate sequences of programming pulses which configure the PLD's arrays of switches in the pattern specified by the user. Some PLDs can be programmed by sending a data stream to the device in its application environment. This is *in-system programming.*

There are today two major classes of PLDs—those based on the **programmable logic array** (PLA) and **field-programmable gate arrays** (FPGAs). PLA technology is the oldest and is now restricted to the less complex circuits. FPGAs, on the other hand, are a more recent technology and are able to implement complex systems equivalent to networks of several thousand logic gates. The first part of this section will explain the principles of PLA-based PLDs; the second will introduce the concepts of FPGAs. A final section will briefly cover the requirements of computer-aided design for programmable logic.

PLA-Structured Devices

It is possible to represent any Boolean function in the form of a *sum of products.* For example, the expressions for the outputs of a full adder can be represented as:

$$\text{SUM} = A\overline{BC} + ABC + \overline{A}\,\overline{BC} + \overline{AB}C \tag{81.14}$$

$$\text{CARRY} = AC + AB + BC \tag{81.15}$$

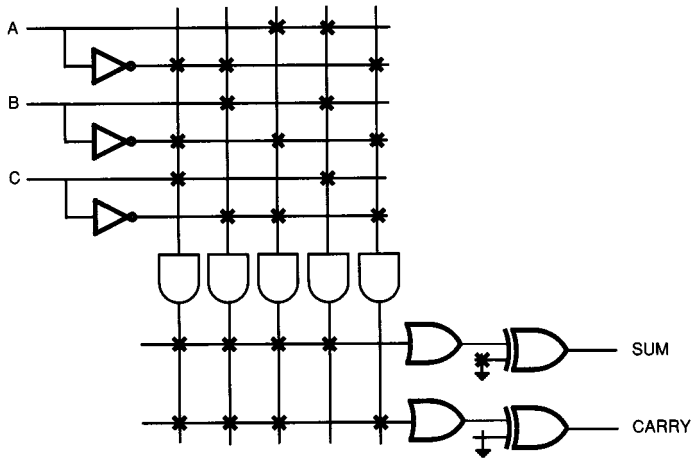


FIGURE 81.42 A full adder implemented in a PLA with programmable output inversion.

where A, B, and C are the inputs. The first equation has four *product terms,* or *products;* the second has three.

A PLA enables functions expressed in this form to be directly implemented. Each product term is generated by a gate which can be programmed to form the AND of any subset of the inputs and their complements, while subsets of products can be summed in a set of programmable OR gates. The programmable gates are constructed in the form of arrays, with the input lines being orthogonal to the product lines, which are themselves orthogonal to the output lines, as shown in Fig. 81.42.

This PLA has three inputs, five product terms, and two outputs. Each input is fed into the first array, the *AND array,* in true and complement form. This enables any product to be formed. The products are all fed into the *OR array.* A cross at an intersection indicates a connection. Notice that some products are used to contribute to both outputs. The ability to share product terms in this way is an important feature of the PLA. The product terms which make up the inputs to the OR gate for the SUM output are those given in Eq. (81.14). The equation for the CARRY output has been modified for use in this PLA which also has programmable output inversions, indicated by the two exclusive-OR gates. The modified equation for the inverse of CARRY, which now shares some products with the first function, is

$$\overline{\text{CARRY}} = A\overline{BC} + \overline{A}\overline{BC} + \overline{AB}C + \overline{ABC} \tag{81.16}$$

Only one new product is now required instead of the three which would have been required without the double negation of the CARRY function.

A PLA is able to implement any set of combinational logic functions, limited only by the number of inputs, number of outputs, and number of product terms. A memory also has this ability, but since in this type of device every combination of inputs is decoded and mapped to a unique set of outputs (via a memory word), the number of inputs handled cannot be as great as with a PLA of similar physical size, since a PLA does not decode every input combination. However, where the set of functions to be implemented would require too many products for a practical PLA, a memory, or a multiple-level network, would have to be used.

An important economy is possible by making use of the fact that for many sets of functions the OR array is not needed for the sharing of products between outputs. Figure 81.43 shows an example of such a function, a multiplexer. Such a function can be implemented by a PLA with only an AND array; the OR array has degenerated into a set of OR gates. The drawback is that the PLA is no longer universal, because a fixed allocation of products to OR gates has to be chosen. PLAs of this structure have become known as PAL devices or just *PALs.* (The term "PAL," which stands for "**programmable array logic**," is a trademark of Advanced Micro Devices, Inc.). There are many applications where the product term sharing capability of the full PLA is wasted, and a PAL-based solution is more economical. Also, because there is only one programmable array, the propagation delay will generally be smaller. Address decoding in microprocessor systems is a very common application of combinational PAL devices.

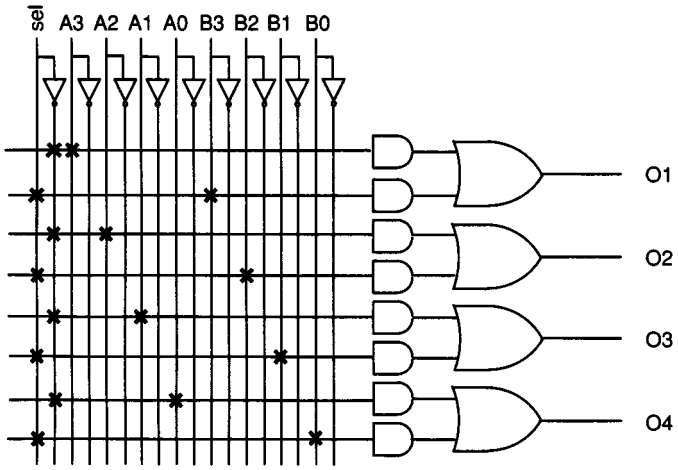


FIGURE 81.43 A four-bit multiplexer implemented in a PAL-structured programmable array.

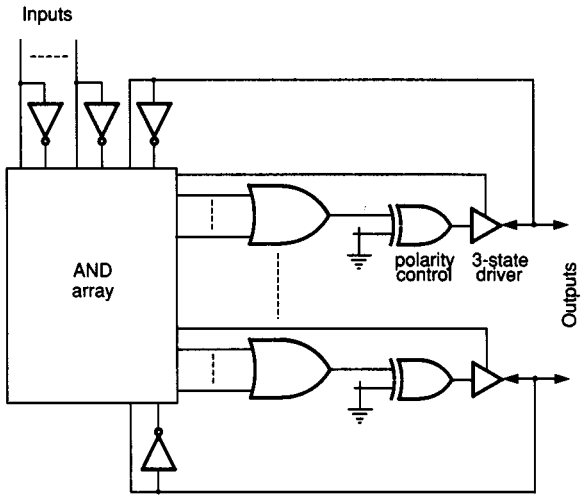


FIGURE 81.44 PAL-structured programmable array with output controls.

PAL devices are available in a wide variety of sizes. Because there is a fixed allocation of products to outputs in a simple PAL device, more different types are needed than in the case of PLAs. This limitation has lately been overcome to some extent in some devices by allowing a limited allocation of products between adjacent outputs.

The flexibility of PAL-structured PLAs can be enhanced by adding controllable three-state output drivers, as shown in Fig. 81.44. The drivers, controlled by additional product terms (or *control terms*), allow those pins to which they are connected to act as either inputs, outputs, or both. This feature allows a single device to be used in a wider range of applications than would be possible with a device with a fixed allocation of inputs and outputs. Note also the polarity control, similar to that in the PLA of Fig. 81.42.

The array structures introduced above can be extended by adding clocked flip-flops to the outputs of the arrays to create general-purpose sequential circuits. A PLA extended in this way is known as a **sequencer**, whose generic structure is shown in Fig. 81.45.

The outputs, which are produced by the OR array, feed either directly to the pins (indicated by Ob in the figure) or to flip-flop inputs (NSa, NSb, and Oa, where "NS" stands for "next state"). The flip-flop outputs are fed back either into the AND array (PSa and PSb, where "PS" stands for "present state") or to pins (PSb and Oa′). The sequencer's primary inputs to the PLA AND array are labeled I. Not every sequencer device will have

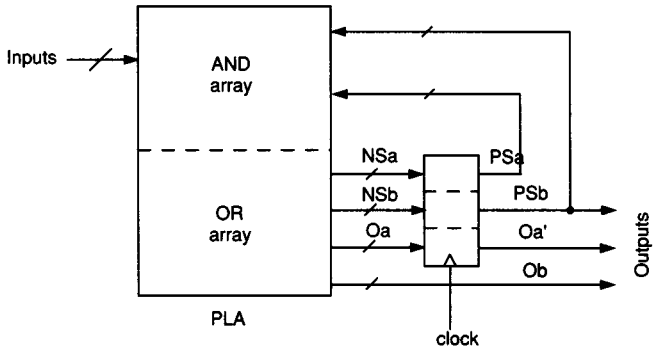


FIGURE 81.45 The structure of a sequencer.

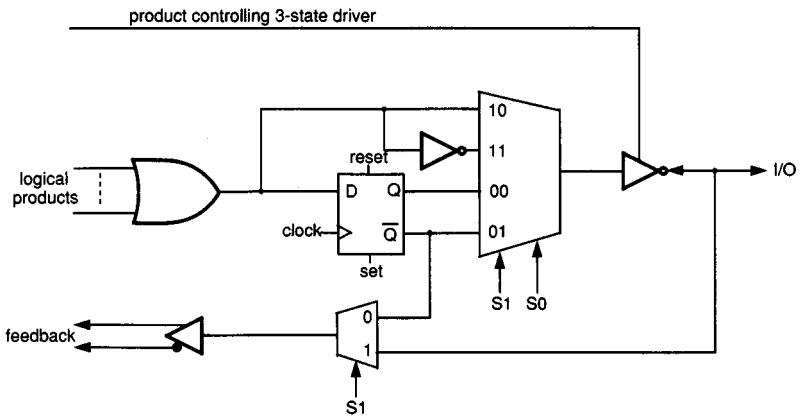


FIGURE 81.46 The output macrocell of the 22V10 PAL device.

all of these paths in its structure. Sequencers are characterized by number of inputs, number of flip-flops, and number of outputs.

A sequencer such as that shown above allows the direct implementation of a synchronous finite state machine with both Moore-type (PSb) or the Mealy-type outputs (Oa′, Ob). The transitions between states in the behavioral specification of the state machine map directly into the product terms of the sequencer.

PAL-structured arrays are also manufactured with clocked flip-flops attached to the array outputs. These are often known as *registered PALs*. State machine descriptions map less naturally into these arrays, but this fact is of less importance nowadays with the reliance on computer-aided design. Registered PALs find wide use in the data paths of digital systems, where special-purpose registers, counters, and data routing functions are needed. Some PAL devices have enhancements such as the exclusive-ORing of outputs to make them more useful in these applications.

Just as a combinational PAL device can be made more general purpose by the addition of a controlled output, a sequential one can be generalized by adding programmable *macrocells* to the array outputs. The output macrocell of the 22V10 device, the first of the *generic* PAL devices to be introduced, is shown in Fig. 81.46.

The modes of this macrocell are controllable by two dedicated bits S1 and S0, programmed into the device, and by a product term controlling the three-state driver. The four possible configurations determined by the programming of S1 and S0 are:

00: registered output, active low, fed back into array
01: registered output, active high, fed back into array
10: combinational input/output, active low
11: combinational input/output, active high

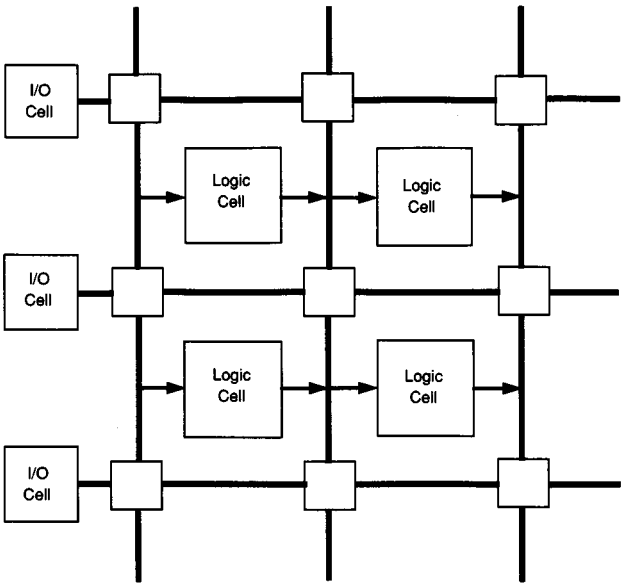


FIGURE 81.47 Interconnection structure in an FPGA.

This form of PAL device is now very widely used and has the advantage that many fewer device types are needed to handle a range of applications. The fastest devices have combinational propagation delays in the 5- to 10-ns range and maximum clock frequencies in excess of 100 MHz.

The single-array devices described so far cannot be extended in size indefinitely; it is difficult to efficiently use large arrays, and the long internal lines required cannot be driven fast enough while maintaining a reasonable power consumption. For these reasons many PLD architectures based on partitioned arrays and switching matrices have been introduced. There is not space to elaborate on these here. A number of these designs are described in Moore and Luk [1991]. However, the most important high-complexity array structure, the FPGA, is introduced in the next section.

Programmable Gate Arrays (FPGAs)

Gate arrays are semicustom devices based on an array of simple cells selected from a library surrounded by an interconnection network. In conventional gate array technology, the interconnection pattern is defined by metallization layers applied at the final stage of manufacture. FPGAs dispense with this final stage by possessing a fixed interconnection network which includes programmable crosspoint switches, as shown in Fig. 81.47. The cells, instead of being selected from a library, are generic and have programmable function. The price for doing this is a lower logic density and higher resistance interconnections, with concomitant effect on speed of operation. Nevertheless, this disadvantage is more than overcome by the short design turnaround times achievable. A major tradeoff in gate array architecture is between cell complexity and interconnection channel capacity. This remains the case with FPGAs.

Each logic cell in an FPGA is a small programmable logic block which will usually contain one or more flip-flops. Typically these cells will have up to ten inputs and a smaller number of outputs.

Computer-Aided Design for Programmable Arrays

As programmable devices become more complex, the capabilities of the computer-aided design support become more and more important. Small PLA and PAL devices can be programmed from a manually created table, but this method is error-prone and cannot be recommended. The earliest widely available design aid for programmable logic was the compiler known as PALASM. This was able to translate the functional specification of a device, expressed in the form of Boolean equations, into the device programming specification for a PAL

device. It also allowed the function of the device to be simulated. Other manufacturer-specific compilers for the simpler PLDs have been produced, but nowadays these needs can be served by one of the very capable Boolean language-based universal compilers.

Recognizing the need for design entry based on schematics, the manufacturers of the more complex PLDs and FPGAs began to offer this alternative; in fact for FPGAs this is an easier option for the design tool writer. Schematics for a PLD describe the interconnection of a collection of notional primitive cells.

Converting a language-based description into the layout of an FPGA is an application of *logic synthesis,* a technology which has evolved independently of programmable logic design tools, but which has now converged with them. However, programmable devices offer new problems to the designer of general-purpose design tools due to the great variety of architectures now existing for the complex devices. Programmable logic design tools are now becoming more integrated into mainstream computer-aided design systems and are starting to adopt the same standards, for example use of the VHDL language for specification.

Defining Terms

Field-programmable gate array (FPGA): A PLD which consists of a matrix of programmable cells embedded in a programmable routing mesh. The combined programming of the cell functions and routing network define the function of the device.

Programmable array logic (PAL) device: A PLA with no OR array, but instead a fixed set of OR gates into which are fed sets of product terms. ("PAL" is a trademark of Advanced Micro Devices, Inc.)

Programmable logic array (PLA): A PLD which consists of an AND array forming logical products of the input literals and an OR array which sums these products to form a set of output functions.

Programmable logic device (PLD): An integrated circuit which is able to implement combinational and/or sequential digital functions which are defined by the designer and programmed into the device.

Sequencer: A PLA which has a set of flip-flops for storage of outputs which can be fed back into the PLA as inputs, enabling the implementation of a finite state machine.

Related Topics

25.3 Application-Specific Integrated Circuits • 81.2 Logic Circuits

References

R.C. Alford, *Programmable Logic Designer's Guide*, Indianapolis: H.W. Sams, 1989.

J. Birkner and V. Coli, *PAL Handbook,* 2nd ed., New York: McGraw-Hill, 1981.

M.J.P. Bolton, *Digital Systems Design with Programmable Logic*, Wokingham, England: Addison-Wesley, 1990.

G. Bostock, *Programmable Logic Handbook*, London: Collins, 1987 (also published as *Programmable Logic Devices: Technology and Applications*, New York: McGraw-Hill, 1988).

J.D. Broesch, *Practical Programmable Circuits: A Guide to PLDs, State Machines and Microcontrollers*, San Diego: Academic Press, 1991.

P.K. Lala, *Digital Systems Design with Programmable Logic Devices*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

W.R. Moore and W. Luk, Eds., *FPGAs,* Abingdon, England: Abingdon EE&CS Books, 1991.

D. Pellerin and M. Holley, *Practical Design Using Programmable Logic*, Englewood Cliffs, N.J.: Prentice-Hall, 1991.

Further Information

Programmable logic is a fast-moving field, with many vendors continually introducing new devices and improved versions of existing architectures. The magazines *Electronic Design, EDN,* and *Computer Design* carry regular announcements of new programmable logic hardware and software products and often print articles illustrating applications and design methods. All of the vendors publish application notes which are the best source of device-specific design information.

81.5 Arithmetic Logic Units

Bill D. Carroll

Arithmetic logic units (ALUs) are combinational logic circuits that can perform basic arithmetic (addition or subtraction) or logical (AND, OR, NOT, etc.) operations on two m-bit operands. ALUs may be constructed from standard integrated circuits or programmable logic devices and are available as single-chip medium-scale integrated circuits. Integrated ALUs may be cascaded to form longer word lengths than are available in a single device.

This section covers the design of arithmetic and logic circuits in sufficient detail for the reader to design and implement basic arithmetic logic units and to understand the operation and utilization of commercial ALU chips. The reader wanting more details or more in-depth discussion of the subject is referred to the References and other sources given at the end of the section.

In the material that follows, it is assumed that operands are signed n-bit binary numbers with the left-most bit representing the sign (0 for positive and 1 for negative) when discussing arithmetic operations/circuits. Negative numbers will be represented in two's complement form. Recall that the two's complement of an n-bit number A is $A' + 1$ where A' represents the bit-wise complement of A. Unsigned n-bit binary numbers are assumed for logic operations/circuits.

Basic Adders and Subtracters

The basic building block for most arithmetic circuits is the **full adder**. A full adder is a logic circuit that produces the two-bit sum (S and C) of three one-bit binary numbers (X, Y, and Z). Table 81.6 shows the truth table and logic equations of a full adder. A logic symbol and a gate-level realization of a full adder are shown in Fig. 81.48.

The addition of two n-bit binary numbers ($X = x_{n-1} \ldots x_1 x_0$ and $Y = y_{n-1} \ldots y_1 y_0$) can be accomplished with n full adders cascaded as shown in Fig. 81.49. Such a circuit is called a **ripple-carry adder** since carries produced

TABLE 81.6 Full Adder Truth Table and Logic Equations

X	Y	Z	S	C	
0	0	0	0	0	
0	0	1	1	0	$S = XYZ + XY'Z' + X'YZ' + X'Y'Z$
0	1	0	1	0	
0	1	1	0	1	
1	0	0	1	0	$C = XY + XZ = YZ$
1	0	1	0	1	
1	1	0	0	1	
1	1	1	1	1	

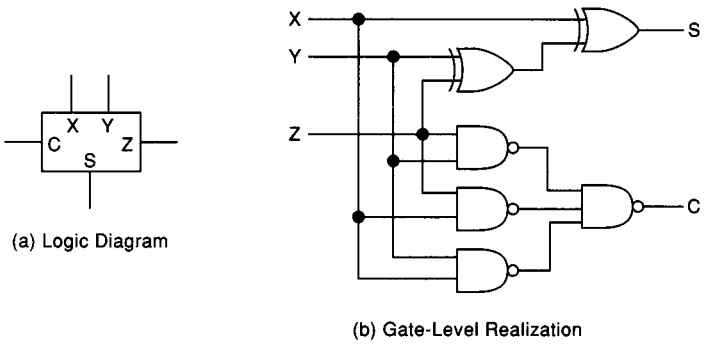


FIGURE 81.48 Full adder.

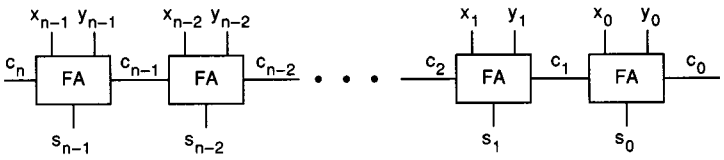


FIGURE 81.49 Ripple-carry adder for two n-bit binary numbers.

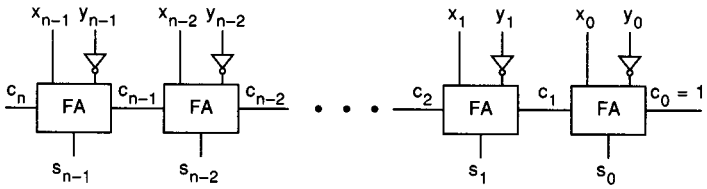


FIGURE 81.50 Two's complement subtracter.

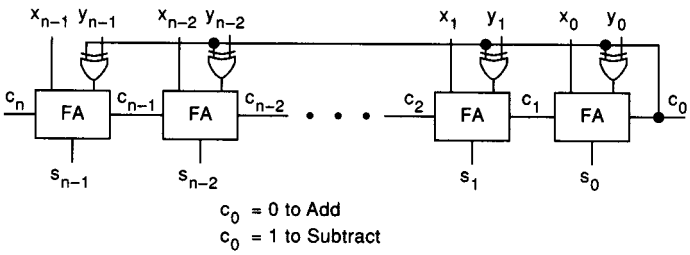


FIGURE 81.51 Two's complement adder/subtracter.

by lower-order stages must propagate or ripple through the higher-order stages before the addition operation is complete.

Ripple-carry adders are simple in both operation and structure but are slow since in the worst case ($X = 1 \ldots 11$ and $Y = 0 \ldots 01$) a carry produced in the least significant full adder must propagate through all the more significant ones. The worst-case add time, t_{add}, is given below where t_{pd} is the propagation delay introduced at each stage.

$$t_{add} = nt_{pd}$$

This assumes that all addend bits are presented to the adder simultaneously. It is important to note that in the least significant full adder, t_{pd} represents the time to compute c_1 from x_0 and y_0 and in the most significant full adder the time to compute s_{n-1} after c_{n-1} is received. In the intermediate stages, t_{pd} is the time needed to compute c_{i+1} from c_i. The propagation delay is approximately equal to the delay of a three-level logic circuit which is consistent with the realization of a full adder given in Fig. 81.48.

Subtraction can easily be performed by adding the minuend to the negative of the subtrahend. In a two's complement number system, $X - Y$ can thus be obtained by computing $X + Y' + 1$. The ripple-carry adder described above can be easily modified to perform this computation by placing inverters on the Y inputs of each full adder and by making the carry-in (c_0) equal to 1. The resulting two's complement subtracter is shown in Fig. 81.50.

A device that can perform either addition or subtraction can be built by replacing the inverters in the subtracter with exclusive-OR gates and using the carry-in (c_0) as a control signal. The resulting two's complement adder/subtracter is shown in Fig. 81.51. The device will function as a ripple-carry adder when $c_0 = 0$ and as a two's complement subtracter when $c_0 = 1$.

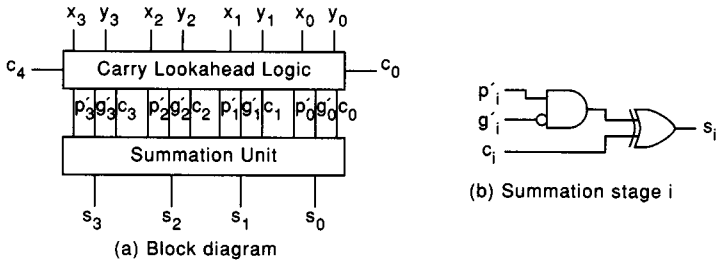


FIGURE 81.52 Carry lookahead adder.

High-Speed Adders

Several different adder designs have been developed for performing high-speed addition. These include **carry lookahead adders (CLAs)**, carry-completion adders, conditional-sum adders, and carry-select adders. Carry lookahead adders have gained wide acceptance in the design of ALUs due to the speed obtained and because they can be conveniently implemented in integrated circuit form.

This material covers only the carry lookahead approach. However, before beginning that discussion, let's briefly explain why fully parallel adders are not feasible. Addition is a combinational process so it is theoretically possible to construct a $2n$-bit "full-adder" that can be realized by a three-level combinational logic circuit and that can perform addition of two n-bit numbers in the time equal to the delay of the circuit. However, such circuits are too costly in terms of gate fan-in to be implemented for reasonable values of n. Carry lookahead is a practical and effective compromise between fully parallel adders and ripple-carry adders. The block diagram of a four-bit CLA is shown in Fig. 81.52(a).

CLAs are based on the observation that a carry-out (c_i) of the ith stage of a full adder is produced by either the propagation of the carry-in (c_{i-1}) through the ith stage or the generation of a carry in the ith stage. This can be seen in the following logic equations for c_i:

$$\begin{aligned} c_i &= x_{i-1}y_{i-1} + x_{i-1}c_{i-1} + y_{i-1}c_{i-1} \\ &= x_{i-1}y_{i-1} + (x_{i-1} + y_{i-1})c_{i-1} \\ &= g_{i-1} + p_{i-1}c_{i-1} \end{aligned}$$

where $g_i = x_i y_i$ and $p_i = x_i + y_i$ are the generate and propagate terms, respectively, for stage i for $i = 0$ to $n - 1$.

The carry equations for an n-bit adder can be derived by repeatedly applying the above equation. The following set of equations results for the $n = 4$ case:

$$c_1 = g_0 + p_0c_0$$

$$c_2 = g_1 + p_1g_0 + p_1p_0c_0$$

$$c_3 = g_2 + p_2g_1 + p_2p_1g_0 + p_2p_1p_0c_0$$

$$c_4 = g_3 + p_3g_2 + p_3p_2g_1 + p_3p_2p_1g_0 + p_3p_2p_1c_0$$

The carry equations can be realized by three-level combinational logic circuits to form the carry lookahead logic block shown in Fig. 81.52(a). The sum (s_i) bits for the ith stage of an adder can be written in terms of g_i, p_i, and c_i and generated by the logic circuit given in Fig. 81.52(b). This completes the description of the four-bit CLA.

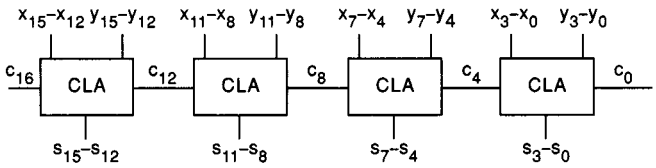


FIGURE 81.53 Cascaded carry lookahead adders.

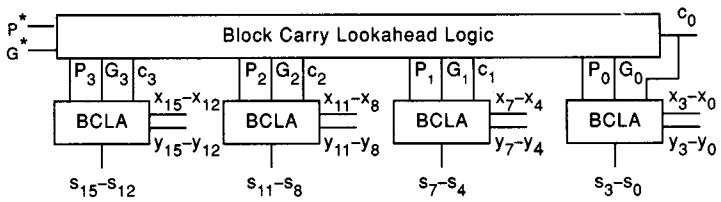


FIGURE 81.54 Block carry lookahead adders.

Now let's examine the add-time, t_{add}, for a CLA. Assume that both addends are applied to the CLA simultaneously and that $c_0 = 0$. Also, let t_{pd} represent the propagation delay of a three-level logic circuit. There are two components that contribute to the add time. First, the three-level carry lookahead logic must produce the carries. This takes t_{pd}. Then, the summation unit must produce the final values of the sum bits. This step takes a time equal to the propagation delay of the exclusive-OR gate in the summation unit which is t_{pd} since an exclusive-OR gate can be realized as a three-level combinational logic circuit. Hence, the add time for a CLA is

$$t_{add} = 2t_{pd}$$

The above result indicates that the add time of a CLA is not only much faster than a ripple-carry adder but is also independent of the length (n) of the addends. Hence, one might conclude that CLAs are the final answer to the high-speed adder problem. However, a closer look at the set of carry equations given above quickly reveals that the equations become progressively more complex in the number of product terms and literals. Therefore, fan-in constraints will eventually limit the practicality of realizing the equations in three-level logic. The actual limit is technology dependent. Standard single-chip medium-scale ALUs typically handle four-bit operands, although longer lengths are certainly feasible with today's technology.

CLAs may be cascaded to produce an adder for longer operands. Figure 81.53 shows a cascade of four 4-bit CLAs to produce a 16-bit adder. Carries are produced using carry lookahead logic within each CLA stage but must propagate between stages in a manner reminiscent of a ripple-carry adder. Hence the add time of cascaded CLAs is dependent on the number of stages in the cascade. The four-stage adder shown in Fig. 81.53 has a worst-case add time of $5t_{pd}$. In general, the add time of an m-stage cascade is $(m + 1)t_{pd}$.

The carry lookahead approach can be applied at a higher level to eliminate the propagation of carries between CLA stages or blocks. This approach uses **block carry lookahead adders** (BCLAs) and block carry lookahead (BCL) logic as shown in Fig. 81.54. A BCLA is a CLA modified to produce block carry propagate (P) and block carry generate (G) outputs instead of a carry-out. BCL logic is a combinational logic circuit that generates block carries (C_j) for each BCLA from the P and G outputs of lower-order BCLAs and c_0. Logic equations for the block carry logic can be derived by repeated application of the following equations for a typical block:

$$C_j = G_j + P_j C_{j-1}$$

where

$$G_j = [g_3 + p_3 g_2 + p_3 p_2 g_1 + p_3 p_2 p_1 g_0]_j$$

and

$$P_j = [p_3 p_2 p_1 p_0]_j$$

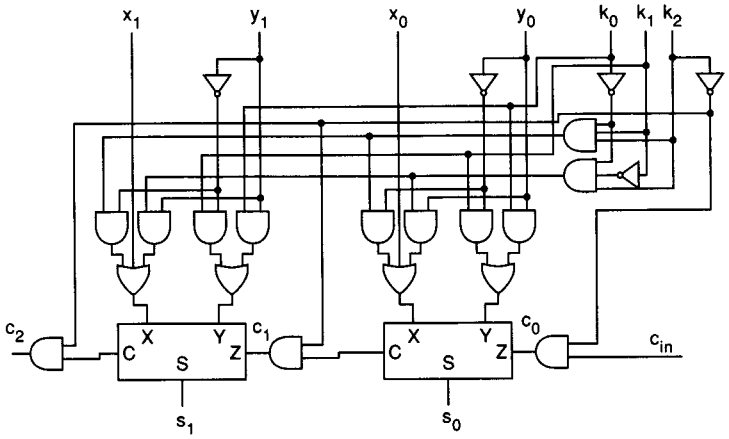


FIGURE 81.55 Multifunction ALU.

TABLE 81.7 Functions Performed by the Multifunction ALU

Control Inputs					
*k*2	*k*1	*k*0	*cin*	Result	Function
0	0	0	0	$S = X$	Transfer X
0	0	0	1	$S = X + 1$	Increment X
0	0	1	0	$S = X + Y$	Addition
0	0	1	1	$S = X + Y + 1$	Add with carry in
0	1	0	0	$S = X - Y - 1$	Subtract with borrow
0	1	0	1	$S = X - Y$	Subtraction
0	1	1	0	$S = X - 1$	Decrement X
0	1	1	1	$S = X$	Transfer X
1	0	0	. . .	$S = X$ OR Y	Logical OR
1	0	1	. . .	$S = X$ OR Y	Exclusive-OR
1	1	0	. . .	$S = X$ AND Y	Logical AND
1	1	1	. . .	$S =$ NOT X	Bit-wise complement

BCLAs and block carry logic units are available in standard medium-scale integrated circuits. Extension of the carry lookahead concept to k levels is possible in theory. However, more than three levels is usually not practical.

Multifunction Arithmetic Logic Units

Devices that can provide a variety of addition, subtraction, and logical operations can be easily designed around the adders/subtracters presented in the previous sections. The logic diagram of the first two stages of an n-bit multifunction ALU is given in Fig. 81.55. Operand inputs for the device are $X = x_{n-1} \ldots x_1 x_0$ and $Y = y_{n-1} \ldots y_1 y_0$ and the output is $S = s_{n-1} \ldots s_1 s_0$. The function performed on the operand(s) is determined by the values of the control inputs k_2, k_1, k_0, and c_{in} as shown in Table 81.7. The given realization is based on a ripple-carry adder for simplicity of presentation. However, the same design approach can be used with other adders such as carry lookahead.

Standard Integrated Circuit ALUs

The devices described above are generic in nature but are similar in function and realization to many commercially available integrated circuit products. Representative products are summarized in Table 81.8.

TABLE 81.8 Typical Integrated Circuit Arithmetic and Logic Devices

Part Number	Function	Features
74LS181	4-bit multifunction (16) ALU	BCL outputs
74LS182	Carry lookahead generator	Use with 74LS181 for BCL
74LS183	Full adder	Two per package
74LS283	4-bit binary adder	Internal CL
74LS381	4-bit multifunction (8) ALU	BCL outputs
74LS382	4-bit multifunction (8) ALU	Ripple-carry output

Defining Terms

Arithmetic logic unit (ALU): A combinational logic circuit that can perform basic arithmetic and logical operations on *n*-bit binary operands.

Block carry lookahead adder (BCLA): An adder that uses two levels of carry lookahead logic.

Carry lookahead adder (CLA): A high-speed adder that uses extra combinational logic to generate all carries in an *m*-bit block in parallel.

Full adder (FA): A combinational logic circuit that produces the two-bit sum of three one-bit binary numbers.

Ripple-carry adder (RCA): A basic *n*-bit adder that is characterized by the need for carries to propagate from lower- to higher-order stages.

Related Topic

81.1 Combinational Networks and Switching Algebra

References

J. Gosling, *Design of Arithmetic Units for Digital Computers,* New York: Springer-Verlag, 1980.

K. Hwang, *Computer Arithmetic: Principles, Architecture, and Design,* New York: John Wiley and Sons, 1979.

M.M. Mano, *Digital Logic and Computer Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1979.

V.P. Nelson, H.T. Nagle, B.D. Carroll, and J.D. Irwin, *Digital Logic Circuit Analysis and Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1995.

E.E. Swartzlander, Jr., Ed., *Computer Arithmetic, Volume I,* Los Alamitos, Calif.: IEEE Computer Society Press, 1980.

E.E. Swartzlander, Jr., Ed., *Computer Arithmetic, Volume II,* Los Alamitos, Calif.: IEEE Computer Society Press, 1990.

TTL Data Book, Texas Instruments, Inc., Dallas, Texas, 1988.

J.F. Wakerly, *Digital Design Principles and Practices,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1994.

S. Waser and M.J. Flynn, *Introduction to Arithmetic for Digital Systems Designers,* New York: CBS College Publishing, 1982.

Further Information

The reader wanting information on the theoretical aspects of computer arithmetic is referred to the *IEEE Transactions on Computers,* a monthly publication of the Institute for Electrical and Electronics Engineers, Inc., 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

More information on the specifications and applications of integrated circuits can be found in the data books and application notes published by electronics manufacturers such as Texas Instruments, Motorola, and National Semiconductor.

Discussions of multiplication, division, and floating-point arithmetic can be found in numerous textbooks on computer architecture.

Staudhammer, J., Chen, S.-L., Windley, P.J., Frenzel, J.F. "Microprocessors"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

82

Microprocessors

John Staudhammer
University of Florida and Office of Naval Research

Sue-Ling Chen
Contractor, Allied Signal

Phillip J. Windley
Brigham Young University

James F. Frenzel
University of Idaho

82.1 Practical Microprocessors
Types of Microprocessors and Microcontrollers • Software for μP/μC Systems • Packaging and Cost • Programming of μPs • Development Support • Comparison of μP/μC Chips • Trends in μP/μC Developments

82.2 Applications
Data Collection • Control • Computing

82.1 Practical Microprocessors

John Staudhammer and Sue-Ling Chen

A **microprocessor** (μP) is a semiconductor die containing the components of a computer central processor, complete with instruction processing unit, arithmetic, interrupt, and basic communication facilities. Such devices have been around since the early 1970s and have greatly benefitted from the continuing improvements in electronics. As microelectronic technology allows the **feature sizes** of components to be shrunk, more powerful μP systems are being put on single **dies**. One of the early microprocessors, the Intel 8080, is still a widely used controller; its understanding and study are recommended for all designers of microprocessor systems [see Gaonkar, 1984]. Early processors contained a few thousands of transistors, while the top-end microprocessor has over 1.5 million. The early processors were 4-bit machines, initially intended for hand-held calculator use; the largest chips today are intended for use as full central processing units of large computers. What makes these devices *micro* is that they are microelectronic devices. These devices are typically built of NMOS (N-type metal-oxide semiconductor) or CMOS (complementary metal-oxide semiconductor) circuits; the CMOS version is typically a bit more expensive and requires less power. An excellent overview of microprocessors and systems using them is found in Raffiquzzaman [1990].

Application of a microprocessor involves adding memory for program and data, and input/output circuitry, which may involve analog/digital and digital/analog converters. As feature sizes of electronic components shrink, the manufacturer of the chip has three options:

1. Reduce the chip size (the die size)—this may result in a cheaper device of the same capability.
2. Increase the processing power at the same cost.
3. Add **peripheral** circuits to the processor, thus putting on-chip devices normally added to the μP. Adding the right combination of peripherals has the greatest benefit for system cost.

A **microcontroller** (μC) is a microprocessor with peripherals on the same chip. These include various types of memory, interrupt structures, communication means, timing and data acquisition circuits.

All three chip developments listed above occur simultaneously, thus keeping a successful product line going for many years. For example, the Motorola 6800 processor was introduced over 20 years ago; today it is still available, but the manufacturer steers designers to the successor chips, the 6809 μP and the 68HC11 μC.

A μP chip communicates to its peripherals by means of three sets of lines: the bidirectional data bus, the memory address bus, and the control bus. In addition, the μP will also send/receive data on communication

lines; the number and types of these vary greatly among chips. A μC will have on the same chip a number of memory elements, timers, communication ports and buffers, counters, and analog/digital converters in addition to a processor; communication to the peripheral circuits is through ports assigned to these devices, as well as any needed external memories. A discussion of general μC systems is given in Clements [1987], while Myers and Budde [1988] present the details of a very-high-performance μC.

All processors require an external clock, typically a crystal. The processor internal circuits run at this speed, but external devices usually use a submultiple of this rate, the bus clock, which is usually 2 to 4 times slower. The advertised clock rate is usually the external clock rate, not the bus cycle speed.

There are well over a hundred different μP systems on the market. They each have peculiarities and may have some advantages. They differ by the kind of data they handle, the amount of processing they do, and the software support they enjoy. What makes μP practical is not so much its claimed prowess, typically stated in peak instruction execution capability (MIPS, millions of instructions per second), but rather its ease of use in a given application, determined to a large degree by the kind and amount of support software available from the manufacturer.

Types of Microprocessors and Microcontrollers

The yearly compendium of μP/μC chips [Markowitz, 1997] categorizes these chips by the width of the data path: 4, 8, 16, 32, and 64 bits. In addition, high-performance chips include bit/word slice chips—these are meant to implement the functions of a central processing unit for a limited number of bits (4 or 8 bits) and are meant to be concatenated for handling an entire computer word. For a discussion of these chips and their uses, see Mick and Brick [1980].

The most precious resource in a μP chip is the number of connection pins. Great efforts are made to utilize the ones used by various control lines and mode selections for the processor. Hence, a given μP chip may have four or more modes of operation: it is simply cheaper to build a flexible system, rather than several different ones.

The vast majority of μPs possess a richness of data access means (addressing modes); they support many different ways of working with memory and external data items. They are complex instruction set computers (CISC). A prime example is the Motorola 68000 with five data groupings and nine addressing modes.

Even the simplest 4-bit processor in wide use (National Semiconductor COP400) has a 10-bit address bus (and a 4-bit data bus). The device can have up to 1 K words (1024 bytes) of memory on the chip, which is often enough for a simple dedicated task. Thus, these chips often appear as single items in simple computer-controlled devices. Most of these chips are found in embedded applications, in kitchen appliances, and in toys.

By far the largest volume of μPs are 8-bit devices; they use 8-bit-wide (1 byte) data paths, but have address busses usually 16 (or more) bits. Often the data bus is time-multiplexed with the lower byte of the address bus. These types of devices can access as much as 65 K memory locations. The typical instruction execution time is 3 to 7 bus cycles. A widely used 8-bit processor is the Zilog Z-80.

The 16-bit μP chips have 16 data and address lines. The typical execution time is 2 to 5 bus cycles.

The 32-bit and 64-bit μP chips are the high end of these devices and find application in advanced personal computers, high-performance workstations, and digital controllers. They are characterized by high cost (compared to most μC chips) and extensive support circuitry. The typical execution time is 1 to 2 bus cycles. They represent the developing efforts in μP technology.

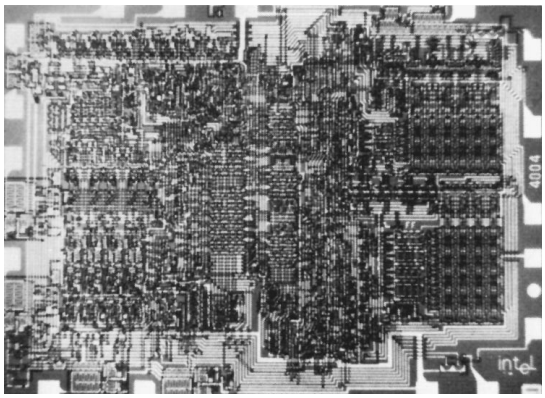
Microcontroller chips contain a μP and various items that make up a μP-based system: there is usually some random access memory (RAM) for holding volatile data, at least one kind of read only memory (ROM) for holding the control program, communication peripherals, including parallel interface(s), serial communication adapter(s), various counting and timing circuits for measuring input pulses, and analog/digital converters. Because each processor has an external clock, usually a crystal clock, timing intervals can be determined to great precision. Many procedures have been developed to take advantage of this precision: voltage-to-frequency converters are used externally to bring pulses to the μP system, which then proceeds to accurately count them. The timing and counting capabilities of μP systems gives them their ubiquitous applicability.

Software for μP/μC Systems

Manufacturers of μP/μC devices have gone to extraordinary efforts to make their devices attractive to system designers. Much software is offered for use with the chips and to support the design effort. ROM-based monitor

The Integrated Circuit

Microelectronics has been the most significant area of development in electrical technology in recent years and has had a profound effect on the course of electrical engineering. The integrated circuit was one of the major developments of its time.



The Intel 4004, the first commercial microprocessor, was originally designed for a programmable calculator. It represented the first consolidation of the arithmetic and logic functions of several chips onto a single integrated circuit. (Photo courtesy of the IEEE Center for the History of Electrical Engineering.)

The increasing complexity of electronic devices meant that even transistorized circuits could be too large and heavy, especially for aerospace applications. In addition, the reliability of such circuits was limited by the ever-increasing number of interconnections. The integrated circuit was a solution to both of these problems. They actually changed the way electrical circuits were designed. Engineers had grown accustomed to creating circuits with a minimum of active components, since transistors and diodes were relatively more expensive than resistors and capacitors. However, active components are both smaller and easier to put on a silicon chip than inactive ones. Thus, the circuits most adaptable to integration are digital circuits, with many active components performing "yes-no" logic functions. Since these are the types of circuits used in computers, it actually encouraged engineers to look for digital solutions to design problems.

The blending of microelectronics and computers has produced the microprocessor, the computer on a chip. This has made possible tremendous reductions in computer size and cost, and consequently has made computers far more available for previously undreamed-of applications. Not since the development of electric power systems began over one hundred years ago have engineers produced such a fundamental and far-reaching tool for change. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

programs are available in most chips so that normal communication tasks can be accomplished easily; programs for usual input/output tasks, data acquisition, timing, and program examples are distributed so that (relatively) error-free software is available for the designers. Most of these are distributed through public-access dial-up bulletin boards. Programming then becomes largely an adaptation of these programs to the tasks that the system is to perform. Of course, the system task analysis and program system design are still the designer's responsibilities, as is the conduct of a software validation.

The μP receives instructions, from its internal or external memory, and data, as a combination of zeros and ones; it is this machine code instruction that controls the operation of the system. The writing of such machine code is far too tedious. An English-like language, machine mnemonics, is used to describe each machine operation and this is then translated to the 0's and 1's which are stored in the memory. For example ADD 123 may stand for the machine code necessary to perform the addition of the number contained in memory location 123 to the content of the computer's arithmetic register, leaving the result in the arithmetic register. The manufacturers supply assemblers to accomplish the mnemonic language assembly into code that may be loaded into the μP memory. These assemblers typically will run on personal computers and various minicomputer systems. Programs that monitor the operation of an actual chip through the μP serial port(s) are effective development tools.

Hence, the typical software includes an assembler, a loader, and a monitor. More user-oriented software for these tasks is available from third-party vendors; many of these extra-cost software packages are highly cost effective. Interactive assemblers and debugging tools are particularly good investments.

Packaging and Cost

The simplest processors are housed in normal dual in-line packages (DIP) of 16 pins or more, spaced 0.3 inches wide. Thus, they resemble conventional transistor-transistor logic (TTL) packages. The mid-range ones are usually in large DIP packages, 0.8 inches wide, with 40 to 68 pins (2 to 4 inches long.) The high-performance ones come in multirow pin packages and may require 200 connections and may use unusual chip sockets. The high performance chips may require special cooling.

Virtually all processors are multisourced, i.e., they are available from more than one manufacturer. This is an important consideration for continued product support. The processors come in various speed grades; the higher speed ones may be 3 to 5 times as costly as the slow-speed versions.

The low-end processors, with no memory and a few communication ports, cost less than $1 in quantity. The widely used Z-80, developed about 20 years ago, costs about $1, while the high-end processors (i860, Pentium Pro) cost around $1000, in low quantities.

Programming of μPs

Most μP systems are programmed in assembly language. Each of the instructions that the processor can execute is given an English-like name (i.e., ADD for the instruction to add two numbers) and translator programs are available from the μP manufacturers to translate the English-like statements (the source code) into the machine codes (which are streams of zeros/ones that the μP interprets). Such assemblers are available from public-access bulletin boards maintained by several vendors; however, convenient assemblers for specific μP systems are usually bought from third-party vendors. Almost all μP assemblers are meant to be used on personal computers; some are available for minicomputer systems. One should check with the chip manufacturer for current software.

High-level language compilers are also available for most μPs, typically from third-party vendors. These are usually C-language compilers, but Pascal compilers also may be available. Here again, the best guide is the chip manufacturer.

The programming of μP systems usually involves carefully tailored code to control signals that interact with the processor chip. Usually the functions that are required of the processor are time-critical real-time control actions. Programs for such applications can easily become highly intricate and require detailed knowledge of all the functions of the processor chip. For a discussion of these problems, see Chapters 3 and 4 of Peatman [1988].

Development Support

Complex μP/μC systems are designed top-down: the task statements are successively refined to a set of smaller tasks, until they can be implemented in relatively short subroutines. The design approach for micro-based systems is described and illustrated in Peatman [1988].

The most important support software for a designer is programs for checking the system, both for logical flow and for cycle-by-cycle activity. Often subtle errors and data dependencies will occur; finding them may be a daunting task. Simulators are the first-level checkout tools for programs. Simulators do not use the actual hardware; rather they use a software model of it. They calculate and show the contents of all computer registers, of ports and selected memory locations, so that an effective check of the internal operations can be made. Most simulators are from third-party vendors.

The actual operation of the μP/μC system may be checked with an in-circuit emulator (ICE), an expensive but effective tool that replaces the μP or μC pin-for-pin in the actual circuit. The ICE typically uses a more powerful computer to mimic the performance of the μP/μC being developed. The ICE tracks signals (including many transients) and can be used to effectively show the behavior of the system, as well as the expected response from the μP and its associated software.

Comparison of μP/μC Chips

As with all computer devices, the advertised speeds and performance figures must be carefully interpreted. These numbers tend to measure only the performance of the manufacturer's test cases and may not be directly comparable [see Hennessey and Patterson, 1990]. These numbers may not be applicable to any one user's requirements. Since the market for μP/μC chips is highly competitive, small differences become amplified in advertising.

Unfortunately, getting a valid number for comparison purposes requires an extensive effort at benchmarking. The advertised performance figures may be taken as a guide, if the task is similar to the benchmark programs. If the processor does not pass the comparison with a comfortable margin, however, one should opt for a higher-performance version of the same chip (if much programming already has been done) or select a clearly superior other candidate.

Trends in μP/μC Developments

An entire microprocessor may be used as a building block in an application-specific integrated circuit (ASIC), available from various vendors. ASIC and VLSI design tools may be used to design such systems, tailored to specific user applications. The μP manufacturers are developing μC chips extending the use of their basic processors. For any application, the best procedure is to invite several vendors to propose alternate systems. Current chips are merely indications of devices to come.

High-performance μP systems are becoming more RISC (reduced instruction set computer) oriented. The object is to execute one instruction per clock period (versus the prevailing 3 to 7 cycles) and to obtain a more regular processor structure. The most popular μP/μC systems are CISC processors. For a user it makes little difference what the internal structure of the processor is; the availability of user support software is far more critical. The trend is to make processors simpler, thus speeding up program execution, even if some programs may have to use more program steps in replacing many "convenience" instructions.

Nevertheless, the very high end μP systems have such a wide data path (32 or 64 bits) that more than one instruction may be accessed at one time. These machines are termed very long instruction word (VLIW) machines. They are able to execute more than one instruction per clock cycle, but they will be more complex internally. For users, both the simpler RISC and the complex VLIW machines will provide increased performance.

Microelectronics is able to produce 2 million transistor chips. While they are expensive, they will replace virtually all of the peripheral chips: memory, timers, and communication channels can all be controlled from a single chip. However, the humble 8-bit chip is still be the most cost-effective workhorse of the bulk of μC applications.

Defining Terms

Die: The piece a silicon wafer containing all electronics.

Feature size: The characteristic size of electronic components on a die.

Microcontroller (μC): A microelectronic chip incorporating a μP and memory, communication, as well as other computer support functions.

Microprocessor (μP): A microelectronic chip that carries out all operations of a computer central processing unit: instruction fetch, execution, interrupt and management of address, data and control lines which are connected to the chip.

Peripheral: Device that supports the functions of the processor. The peripheral may be all electronic (a communications adapter) or may contain mechanical parts (a disk memory). To the processor a peripheral appears as electronics with timing constraints.

Related Topic

82.2 Applications

References

A. Clements, *Microprocessor System Design*, Boston: PWS Publishers, 1987.

R.S. Gaonkar, *Microprocessor Architecture, Programming, and Applications with the 8085/8080A,* 2nd ed., Columbus, Ohio: Merrill Publishing Company, 1984.

J.L. Hennessey and D.A. Patterson, *Computer Architecture: A Quantitative Approach,* Palo Alto, Calif.: Morgan Kaufmann Publishers, 1990.

M.C. Markowitz, "EDN's 24th annual μP/μC chip directory," *EDN Magazine*, September 25, 1997.

J.R. Mick and J. Brick, *Bit-Slice Microprocessor Design*, New York: McGraw-Hill, 1980.

G.J. Myers and D.L. Budde, *The 80960 Microprocessor Architecture*, New York: Wiley Interscience, 1988.

J.B. Peatman, *Design with Microcontrollers*, New York: McGraw-Hill, 1988.

M. Rafiquzzaman, *Microprocessors and Microcomputer-Based System Design*, Boca Raton, Fla.: CRC Press, 1990.

Further Information

The magazine *EDN* runs an annual microprocessor/microcontroller review, typically late in the year (i.e., September 25, 1997). These are handy compendia of characteristics.

The IEEE magazine *Micro* presents detailed articles on device development and applications of microprocessors and systems which embed them. At mid-year the magazine carries a set of articles based on the Hot Chips Symposium, presenting developments in μP and chip technology for high-performance workstations and systems.

Specifics of various microprocessor and microsystem chips are found in the respective manufacturer's reference literature; for any design one must become familiar with the applicable manual and design notes. Even for a modest chip the reference manual may run 300 pages; in addition, the manufacturer's free-of-charge support software is of comparable size. For example the Motorola 68HC11 reference manual is 512 pages, and there is over 1 megabyte of design support software available from the manufacturer. The Intel i860 data book is 150 pages, hardware and programmer reference manuals are over 300 pages, and the support software is several megabytes.

82.2 Applications

Phillip J. Windley and James F. Frenzel

Microprocessors are cheap, small, and consume little power. In addition, in recent years their performance has increased at a greater rate than the performance of larger computers. These factors have led to an explosion in the application of microprocessors. A short section could never do justice to every application; therefore, we will view representative applications in three broad areas:

- Data collection, where microprocessors are used to monitor sensors and either record the collected information or communicate the information to some other computer.
- Control, where microprocessors have largely replaced analog electronics for controlling everything from manufacturing robots to home appliances.
- Computing, where microprocessors have transformed the concept of computer and made parallel processing possible.

Admittedly, these categories are not strictly disjoint. They do, however, represent the most pervasive uses for microprocessors at an abstract level.

Data Collection

In data collection the microprocessor-based system serves primarily as a low-cost data recorder. Basic functions include the polling of sensors, acceptance of data, data storage, and data transmission or display. Additional features might include preprocessing of the raw data. Such a classification spans a broad range of applications, from automotive diagnostics to space-born monitoring stations.

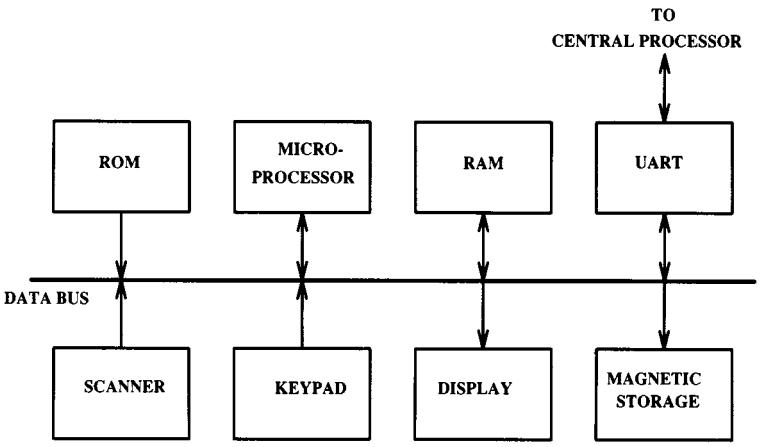


FIGURE 82.1 Point-of-sale terminal system.

Microprocessors are well suited as the controller for such tasks because of their cost and flexibility. Sufficient numbers of processors may be used to allow real-time data acquisition. Because the microprocessor is programmable, sensors may be added, removed, or rearranged without major system impact. Finally, because the microprocessor is a computational device, calculations may be performed on the recorded data to produce useful information, such as calculating speed from distance and time. In the next section we will examine the components of one such system, the retail point-of-sale terminal [Hordeski, 1984].

Point-of-Sale Terminal

The function of a point-of-sale (POS) terminal is characteristic of the applications under the category of data collection. The microprocessor is not being used for intensive computations, nor for controlling a complex process, but rather to collect data, perform some processing, and then pass the results on to a central collector. The cost and flexibility of the microprocessor make it an excellent choice over special-purpose hardware.

System Components. In addition to the microprocessor and storage capability, the typical retail terminal has one or more input devices for entering prices (e.g., keyboard, bar code scanner) and one or more output devices for displaying totals (e.g., paper tape, display). Often these terminals are part of a large network of terminals and may support additional features beyond totaling purchases such as automated inventory control and credit checking. A complete system is shown in Fig. 82.1, including magnetic tape for storing transactions and a **universal asynchronous receiver/transmitter (UART)** for communication with a central processing facility. Because of the high unit volume, it is desirable to keep the cost and complexity low. Typically, each terminal will have limited storage capability, relying on a central processor for maintaining store inventory and credit checks. In order to reduce communication traffic with the central processor, however, each terminal generally has in storage the current price for all items.

Universal Product Code. The use of the Universal Product Code (UPC) has enabled the development of intelligent POS terminals which can "read" the UPC symbol and determine the identity of the item. The UPC symbol consists of ten decimal digits, split into two fields of five digits each. Each digit is encoded using a 7-bit binary number, represented by a group of 7 dark (binary 1) and light (binary 0) bars. The five left-hand digits are encoded using odd parity and the right-hand digits are encoded using even parity. This allows correct recognition of the symbol, independent of its orientation.

For groceries, the first five decimal digits identify the manufacturer and the second group of five digits identify the specific product. There are additional codes in use as well, such as the National Drug Code. By using a microprocessor-based system, a POS terminal can be quickly reconfigured to recognize a different code (or multiple codes) through a simple software change.

Operation. A typical sale might involve the following steps. The clerk inquires whether the sale is to be a cash purchase or charged to an account. If the latter, the clerk enters the necessary information and the terminal

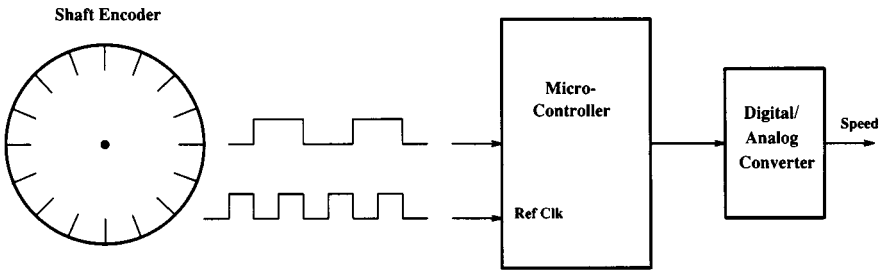


FIGURE 82.2 Digital tachometer.

transmits a request to a central processor, inquiring as to the available credit. In the interim, items are entered, either through the bar code scanner or the keypad, and the price and running total are displayed. The identity of the items purchased is also stored for later transmission to the central processor responsible for inventory control. Finally, the terminal checks the available credit against the total and records the transaction for later transmission to the central processor.

Digital Tachometer

Another example of using a microprocessor for data collection is the implementation of a digital tachometer [Bonert, 1989]. The microprocessor samples the output of a shaft **encoder** and compares it with a reference signal to determine the rotational speed. The calculated value is passed to a digital-to-analog converter to generate an analog speed signal. The system is shown in Fig. 82.2.

Speed Evaluation Methods. Various methods may be used to evaluate the speed value, all of which involve some combination of pulse counting and time measurement. The constant elapsed time (CET) method provides a good compromise between measurement accuracy and response time. The CET method records the number of encoder pulses observed during a fixed time interval. The rotational speed, n, is then given by

$$n = C_p/(C_t m/T_c)$$

where C_p is the number of encoder pulses, C_t is the number of clock pulses, m is the number of encoder marks per turn, and T_c is the clock pulse period.

Implementation. Rather than continuously stopping and resetting external counters, it is possible to take advantage of features often found in modern **microcontrollers**, microprocessors containing additional interface circuitry. Microcontrollers often contain counters, timers, and **capture registers**. Capture registers allow the storing of timer or counter values triggered by an external signal. At the start of evaluation, the rising edge of the next encoder pulse triggers the capture of the timer count and the pulse count. After a minimum evaluation time has elapsed, the next encoder pulse again triggers the capture of the current counter values. The rotational speed can then be computed using the difference between the captured values. A flowchart of the algorithm is shown in Fig. 82.3.

Performance. Using an encoder with 1024 marks per revolution, a 2-MHz reference clock, and an evaluation period of 2.3 ms resulted in a measurable speed range of 25.5–4883 rpm. The maximum relative error was 0.123%, induced primarily by the encoder tolerance [Bonert, 1989].

Control

Microprocessors are ubiquitous in control applications. While some custom analog controllers are still built, the advantages of cost and flexibility inherent in microprocessors make them a natural choice. The advantages of microprocessors are particularly obvious in mass-produced goods where time-to-market can be a significant driving force.

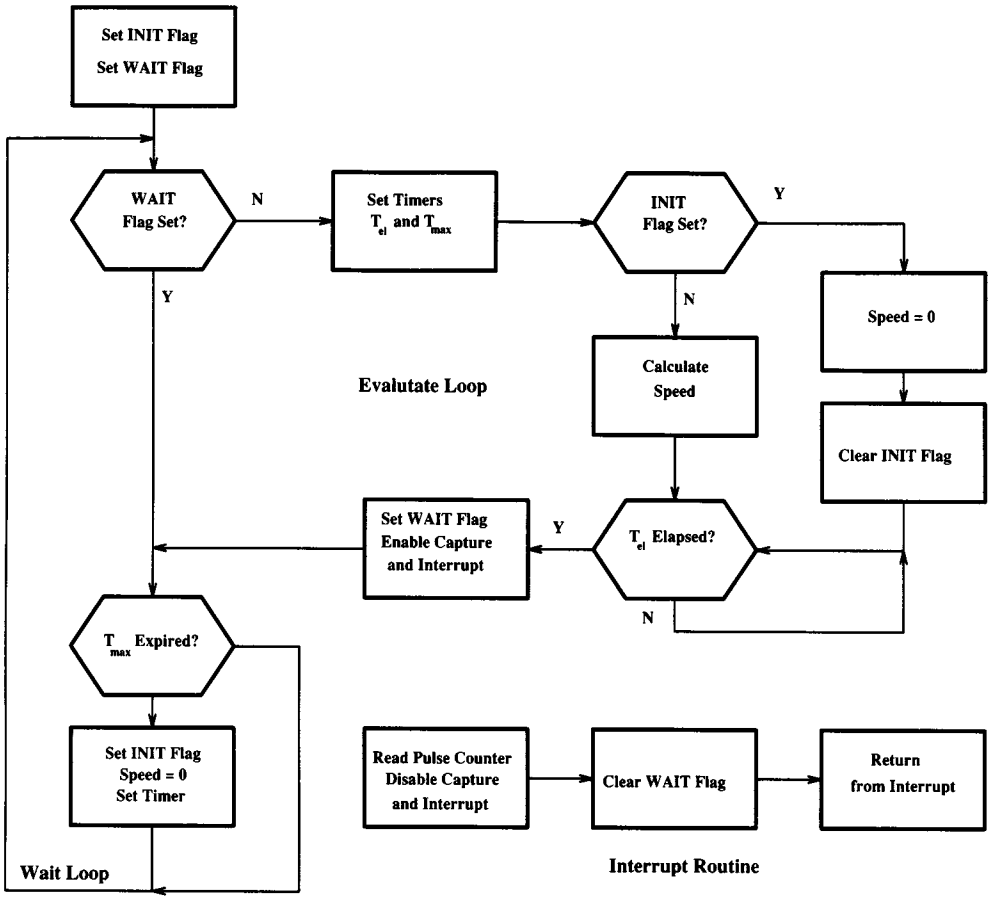


FIGURE 82.3 Tachometer program. (*Source:* Bonert, 1989.)

Microcontrollers

Microprocessors designed especially for use in control applications are called *microcontrollers*. Typically, the major difference between a microcontroller and a standard microprocessor is the presence of scratchpad RAM, input and output ports, timers, and even analog-to-digital (**A/D**) and digital-to-analog (**D/A**) converters on-chip.

Figure 82.4 shows a simplified microcontroller architecture. The process to be controlled is monitored by means of sensors. The outputs from the sensors are fed to A/D converters which convert the analog signals from the sensors to digital signals appropriate for use in the microprocessor. The microprocessor reads the digital signal from the A/D converter and uses it for input to a control program stored in the microprocessor memory. The program produces digital outputs which are fed to D/A converters. The analog outputs from the D/A converters (which are typically low power) are fed to amplifiers, and the amplified signal is used to control actuators that affect the process being controlled.

Control Applications

Consumer Electronics. A survey of the typical home will show numerous microprocessors where 10 years ago, there were none. Microprocessors are used for controlling VCRs, TVs, stereo equipment, microwave ovens, sprinkler systems, telephone equipment, heating systems, and virtually every other appliance using electricity.

Manufacturing. Microprocessors have found numerous applications in manufacturing. Perhaps none is better known than the robot. Microprocessor technology has made the modern robot possible. Robot arms used in

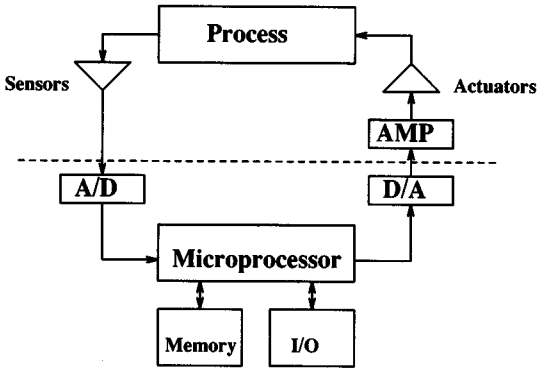


FIGURE 82.4 Typical microcontroller design.

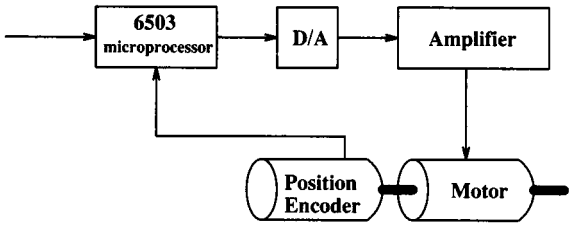


FIGURE 82.5 Microprocessor-controlled servomechanism from a PUMA 560 robot.

manufacturing typically have five or six joints. Current practice is to treat each joint in the robot arm as a separate servomechanism with its own control system. For example, the PUMA 560 robot arm, manufactured by Unimation, has six rotating joints. Each joint is controlled by an individual microcontroller system. Another computer calculates paths and sends individual joint motion information to the six joint servomechanisms [Fu et al., 1987].

The servomechanism system shown in Fig. 82.5 consists of an 8-bit Rockwell 6503 microprocessor, a D/A converter, an amplifier, a joint motor, and an encoder. The 6503 microprocessor receives joint position information from the supervisory computer every 28 ms. The microprocessor calculates the joint error information by comparing the current position to the desired joint position using the PID (proportional-integral-derivative) control method. The error is converted to an analog signal by the D/A converter and amplified before going to the joint motor. The encoder is connected to the motor shaft and provides a digital signal to the microprocessor.

The microprocessor performs the following functions:

1. Receives the desired joint position from the supervisory computer every 28 ms
2. Reads the position signal from the encoder every 0.875 ms
3. Calculates the error every 0.875 ms
4. Sends the error to the D/A converter

The microprocessor calculates joint error and sends the correction signal to the joint motor 32 times for every joint position received from the supervisory computer.

Transportation. Microprocessors are used for control applications in every facet of the transportation industry. Microprocessors are used to control the operation of the vehicles themselves such as controlling engines, air surfaces in aircraft, antilock brakes in automobiles, and rudders in ships. Microprocessors are also used in wide-area applications such as traffic control.

In controllers for motor traffic, the microprocessor has replaced hardwired logic and analog systems to provide systems which are much more capable and typically more reliable [Hordeski, 1984]. A typical traffic light controller is shown in Fig. 82.6. The microprocessor provides the CPU, memory, and I/O ports. The system

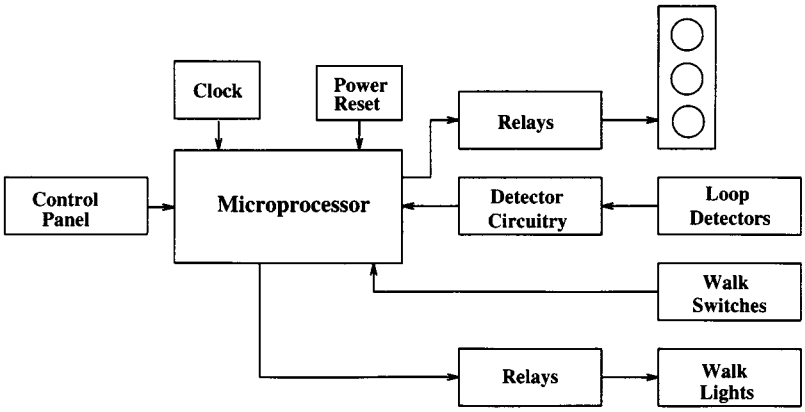


FIGURE 82.6 Traffic control system. (*Source:* M. Hordeski, *Microprocessors in Industry*, New York: Van Nostrand Reinhold, 1984, p. 398. With permission.)

includes a real-time clock for timing external events and a power-fail restart unit which restarts the system after a power failure (including restoring volatile data). The system monitors traffic at the intersection through the use of loop detectors and controls the traffic by changing the traffic lights. Other components of the system monitor and control pedestrian traffic and provide an interface to the system for human operators.

The loop detectors are paired coils of wire placed under the pavement. The impedance of the loop detectors changes in response to the presence of a car on the roadway. The change in impedance changes the frequency of an RC oscillator, which is converted to a digital signal reported to the microprocessor. Loop detectors can be used to monitor the presence of a car at a traffic light, the length of a line of cars, and the speed of traffic.

The function of the traffic controller is to optimize traffic flow. For example, during busy periods of the day, the goal may be to optimize flow through an intersection. Another goal may be to ensure that traffic flows smoothly in certain directions to effectively feed larger roads. Traffic lights can be synchronized to provide a highway through a busy network of roads by ensuring that a car that enters the roadway and maintains a recommended speed can travel along the entire length without stopping at a traffic light. On the other hand, during periods of low use, such as night and early morning, the system may monitor for the presence of a car at an intersection and immediately switch the light to let it pass.

Microprocessors offer advantages in traffic control situations in addition to optimized traffic flow. When properly designed, the system can provide a certain degree of fault tolerance. When a loop detector is giving a faulty value, the system can be programmed to ignore its value and use values from adjoining lanes. An error report can be forwarded to a central traffic facility and after repairs are made, the loop brought automatically on-line. The system can also monitor feedback information from the traffic light to ensure that the lights are actually lit. When a problem is detected, the system can enter an emergency mode and report the problem.

Social Issues

The explosive growth in the use of microprocessors in control applications has caused discussion about the utility and safety of such devices.

An issue many people can identify with is feature overload. The advent of cheap microprocessors has turned design upside-down. Designers can add additional features for very little additional increased manufacturing cost. Competition spurs even more features until even the simplest of consumer items come with thick instruction manuals. Naturally, consumers become frustrated with features that are difficult to use.

Perhaps more important are the safety hazards that may be engendered by replacing analog control systems with digital control systems. Most analog systems are based on physical properties with continuous behavior. Digital systems, on the other hand, are discrete and are thus much more prone to problems where small errors can result in large changes in behavior due to the digital representation of value; a single bit change can result in a large change in magnitude. Digital control systems are becoming more and more prevalent in systems controlling aircraft, automobiles, nuclear power plants, and other safety-critical systems. Engineers who design

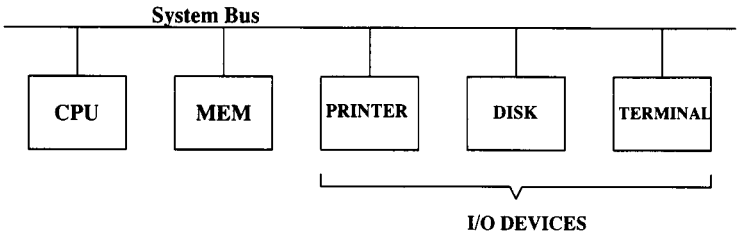


FIGURE 82.7 Major components of a simple microcomputer.

the systems and officials charged with ensuring their safety are still coming to grips with the implications of this trend. New techniques for analyzing computer system designs for errors are being developed which promise to alleviate some of these concerns [Windley, 1995].

Computing

While microprocessors have been put to a plethora of interesting special-purpose uses such as data collection and control, perhaps the most visible use of microprocessors has been in the area of general-purpose computing.

Microcomputers

The advent of microprocessors has resulted in a personal computer on virtually every desktop. Even the slowest of these computers rival the performance of the largest computers available 15 years ago.

Figure 82.7 shows the major hardware components of a simple microcomputer. The central processing unit (CPU) is the execution engine of the microcomputer and is most often a microprocessor. One popular family of microprocessors used as the CPU in microcomputers is manufactured by Intel. These chips, with names such as the 8088, 80286, 80386, 80486, and Pentium are used in microcomputers such as the IBM personal computer. Another important family of microprocessors is the Motorola Power Pc series, which is used in microcomputers manufactured by Apple Computer [Matloff, 1992].

In addition to the CPU, there are a number of other components in a microcomputer. General-purpose memory is not typically part of the microprocessor but must be added as a separate component. In simple microcomputers, the memory may be directly attached to the microprocessor. In more complex designs, the memory is attached to the microprocessor by a system bus that allows system components other than the microprocessor to access memory as well. In addition, the memory may have its own controller, called a memory management unit.

Other components in the system include input/output (I/O) interfaces to devices such as printers, terminals, disks, mice, and so on. The common feature of all of these devices is that they interface the microprocessor to the outside world. All of the components in the microcomputer are connected together by a system bus. The bus is a set of parallel wires that carry information from one component to another.

Multiprocessing

The desire for greatly increased computer performance has fueled research in using microprocessors as the computing engines in multiprocessors which would achieve performance gains over single-processor computers through the use of numerous low-cost microprocessors.

There are numerous multiprocessor architectures. An example architecture that is well suited to using large numbers of microprocessors is the hypercube. The hypercube architecture was originally developed by Charles Seitz and others at California Institute of Technology in the early 1980s. The hypercube depends on using large numbers of commodity microprocessors, each with private memory, in a hypercube network [Bell, 1989].

In a hypercube network, N microprocessors are arranged in an n-dimensional cube, where $N = 2^n$. Each processor is connected to n other processors and the longest communications path from any processor to any other is n links. For example, a three-dimensional hypercube contains eight processors and is arranged as a standard cube, where the nodes are the processors and the edges of the cube are the communication paths.

Figure 82.8 shows a four-dimensional hypercube represented as a tesseract. A four-dimensional hypercube has 16 processors, each is connected to 4 other processors, and the longest path between any two processors (shown in bold in Fig. 82.8) is 4. Thus, doubling the number of processors results in a unit increase in the communications path length. This logarithmic relationship results in the great advantage of the hypercube: it scales well. A system with 1024 processors has a maximum communications path length of just 10.

There are several manufacturers of hypercube systems including NCUBE and Intel. Most of these systems have between 32 and 1024 processors. NCUBE has a hypercube architecture with 8192 nodes operating at 2.4 megaflops each.

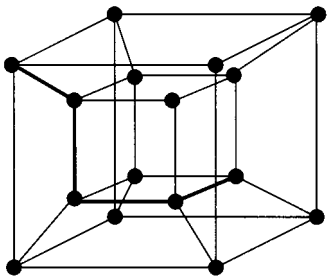


FIGURE 82.8 A four-dimensional hypercube.

Digital Signal Processing

Digital signal processing (DSP) may be considered a specific example belonging to the category of computation. Specialized microprocessors are finding widespread application in many areas of digital signal processing such as telecommunications, speech processing, medical imaging, and radar [Aliphas and Feldman, 1987]. These microprocessors are designed for very high data rates and contain specialized circuitry to accelerate computations that are specific to signal processing. Figure 82.9 illustrates the architectural differences between digital signal processors and conventional microprocessors.

Architectural Features. A common task among most signal processing algorithms is the summation of multiple products. The most notable distinction between general-purpose microprocessors and digital signal processors is the existence of a high-speed multiplier-accumulator [Allen, 1985]. This circuitry can complete a multiply-add operation in one cycle, as opposed to roughly 25 cycles for a conventional microprocessor. Traditionally, only fixed-point arithmetic was available, but newer DSP chips provide floating-point arithmetic with 32 bits of precision.

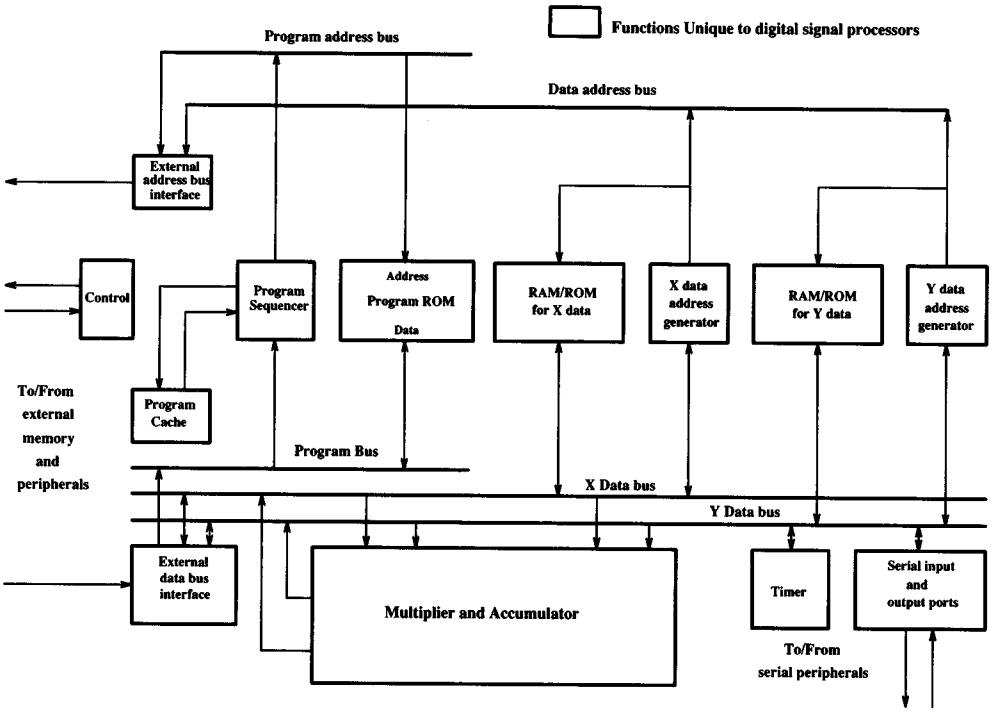


FIGURE 82.9 Digital signal processor architecture. (*Source:* Aliphas and Feldman, 1987.)

The second most noticeable feature on DSP chips is the existence of multiple data buses and memories. Many chips have two data memories, each with a data bus, allowing the simultaneous fetch of two operands for the multiply-accumulate operation. Furthermore, most chips use the Harvard architecture, characterized by separate program and data memories, so that instructions and data can be fetched simultaneously. Others use a modified Harvard architecture, where data can be stored in slower, cheaper program memory and moved to the faster data memory as needed.

Finally, DSP chips typically have separate arithmetic-logic units (ALU) for data arithmetic and address calculations. This serves two purposes: (1) data calculations can proceed unhindered by address calculations, maintaining a high throughput, and (2) each unit can be specialized for its particular task. For example, the data ALU may have additional circuitry to support saturation arithmetic, whereas the ALU used for address calculations may provide indexing, auto-increment, or even bit-reversal, an operation required for the fast Fourier transform (FFT).

Dedicated digital signal processors offer an excellent alternative or supplement to general-purpose microprocessors for signal processing applications. As a slave to a conventional processor, the DSP chip is freed from communicating with peripherals, increasing throughput. For additional performance, DSP chips may be operated in a multiprocessor configuration, controlled by a central processor. Such an arrangement would be appropriate for applications such as phased-array radar, where the volume of data and uniformity of the calculations lend themselves to distributed processing.

Defining Terms

A/D: Analog to digital. Usually a device that changes an analog signal to a digital signal of corresponding magnitude.

Capture registers: Internal registers which, triggered by a specified internal or external signal, store or "capture" the contents of an internal timer or counter.

D/A: Digital to analog. Usually a device that changes a digital signal to an analog signal of corresponding magnitude.

Encoder: A sensor that directly creates a digital signal for use in a control application. An example is a shaft encoder that turns an angular shaft position into a digital signal.

Interrupts: Special hardware on a computer that suspends the executing program so that another procedure can be run to service an external device.

Microcontroller: A special-purpose microprocessor with scratchpad RAM, input and output ports, timers, and even analog to digital (A/D) and digital-to-analog (D/A) converters on-chip used in control applications.

Universal asynchronous receiver/transmitter (UART): Circuitry (often a separate module), which provides all of the interface functions necessary for a microprocessor to communicate with a serial device.

Related Topic

82.1 Practical Microprocessors

References

A. Aliphas and J. Feldman, "The versatility of digital signal processing chips," *IEEE Spectrum*, vol. 24, no. 6, pp. 40–45, June 1987.

J. Allen, "Computer architecture for digital signal processing," *Proceedings of the IEEE*, vol. 73, no. 5, pp. 852–873, May 1985.

G. Bell, "The future of high performance computers in science and engineering," *Communications of the ACM*, 32(9), pp. 1091–1099, September 1989.

R. Bonert, "Design of a high performance digital tachometer with a microcontroller," *IEEE Transactions on Instrumentation and Measurement*, vol. 38, no. 6, pp. 1104–1108, December 1989.

K.S. Fu, R.C. Gonzalez, and C.S.G. Lee, *Robotics: Control, Sensing, Vision, and Intelligence*, New York: McGraw-Hill, 1987.

M. Hordeski, *Microprocessors in Industry*, New York: Van Nostrand Reinhold, 1984.
N. S. Matloff, *IBM Microcomputer Architecture and Assembly Language*, Englewood Cliffs, N.J.: Prentice-Hall, 1992.
P. J. Windley, "Formal modeling and verification of microprocessors," *IEEE Trans. on Computers,* vol. 44, no. 1, January 1995.

Further Information

Byte magazine is a good resource for entry-level articles on microprocessor applications. For subscriptions contact: BYTE, One Phoenix Mill Lane, Petersborough, NH 03458.

The Institute of Electrical and Electronics Engineers (IEEE) publishes several magazines and journals that frequently contain articles concerning microprocessor applications. *IEEE Micro* is a bimonthly magazine which addresses the design and use of microprocessors and minicomputers. *IEEE Computer* is a monthly magazine covering all aspects of computing. Three pertinent journals published bimonthly by the IEEE are *Transactions on Industry Applications, Transactions on Industrial Electronics*, and *Transactions on Instrumentation and Measurement*. The address for the IEEE Service Center is 445 Hoes Lane, Piscataway, NJ 08855.

Morris, J.E., Martin, A., Weber, L.F. "Displays"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

83

Displays

83.1 Light-Emitting Diodes
Semiconductor Device Principles • Semiconductor Materials • Device Efficiency • Interfacing

83.2 Liquid-Crystal Displays
Principle of Operation • Interfacing

83.3 The Cathode Ray Tube
Monochrome CRTs • Color CRTs • Contrast and Brightness • Measurements on CRTs • Projection Screen

83.4 Color Plasma Displays
Introduction • Color Plasma Display Markets • Color Plasma Display Attributes • Gas Discharge Physics • Current Limiting for Plasma Displays • ac Plasma Displays • Color Plasma Display Devices • Gray Scale

James E. Morris
State University of New York at Binghamton

André Martin
Hughes Display Products

Larry F. Weber
Plasmaco, subsidiary of Matsushita

83.1 Light-Emitting Diodes

James E. Morris

The light-emitting diode (LED) has found a multitude of roles as the field of optoelectronics has bloomed. Infrared devices are used in conjunction with spectrally matched phototransistors in optoisolation couplers, hand-held remote controllers, interruptive, reflective and fiber-optic sensing techniques, etc. Visible spectrum applications include simple status indicators and dynamic power level bar graphs on a stereo or tape deck. This section will concentrate on digital display applications of visible output devices.

Semiconductor Device Principles

The operation of an LED is based on the recombination of electrons and holes in a semiconductor. As an electron carrier in the conduction band recombines with a hole in the valence band, it loses energy ΔE equal to the bandgap E_g with the emission of a photon of frequency

$$\upsilon = c/\lambda = \Delta E/h \tag{83.1}$$

where λ is the radiation wavelength and h is Planck's constant.

The incidence of recombination under equilibrium conditions is insufficient for practical applications but can be enhanced by increasing the minority carrier density. In an LED, this is accomplished by forward biasing the diode, the injected minority carriers recombining with the majority carriers within a few diffusion lengths of the junction edge. Figure 83.1 illustrates the process. The potential barrier eV_o is reduced by forward bias eV, leading to net forward current and the minority carrier distributions shown on either side of the depletion layer. As the carriers diffuse away from the junction edges, these distributions decay exponentially because of recombination with the majority carriers. Each recombination event shown on either side of the junction gives off a photon. This process is called **injection electroluminescence.**

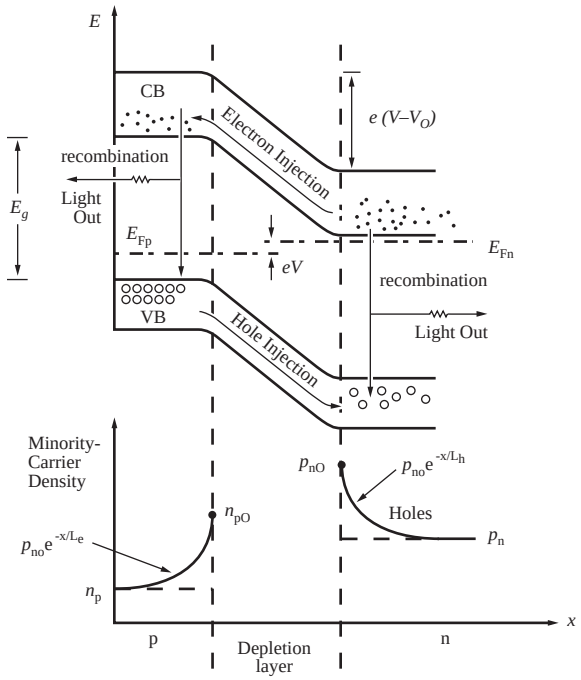


FIGURE 83.1 Light emission due to radiative recombination of injected carriers in a forward-biased *pn* junction.

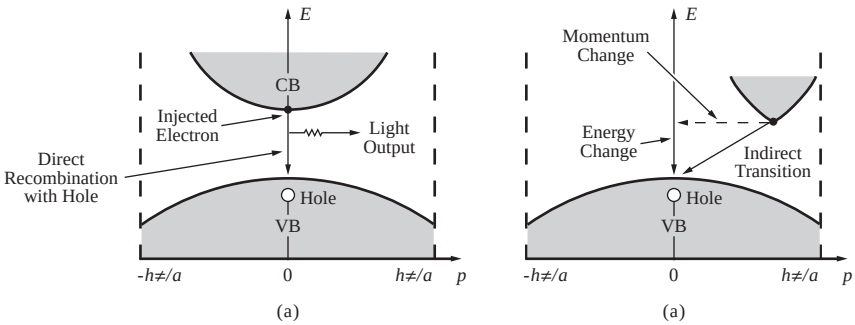


FIGURE 83.2 (a) Interband recombination in a direct-bandgap semiconductor; (b) recombination in an indirect-gap semiconductor also involves a momentum change.

Equation (83.1) implies that the radiation emitted will be monochromatic, but in practice $\Delta E > E_g$, and there is a spectral distribution corresponding to the energy distributions of the carriers in the conduction and valence bands.

Semiconductor Materials

Silicon is the most common material used in current semiconductor technologies, but it is not at all suitable for an LED. The reason is that silicon has an indirect bandgap, and a direct bandgap is required for process efficiency. Direct and indirect bandgaps are compared in Fig. 83.2, where carrier energy is plotted versus momentum for both cases. The photon momentum

$$p = h\lambda = h\upsilon/c \tag{83.2}$$

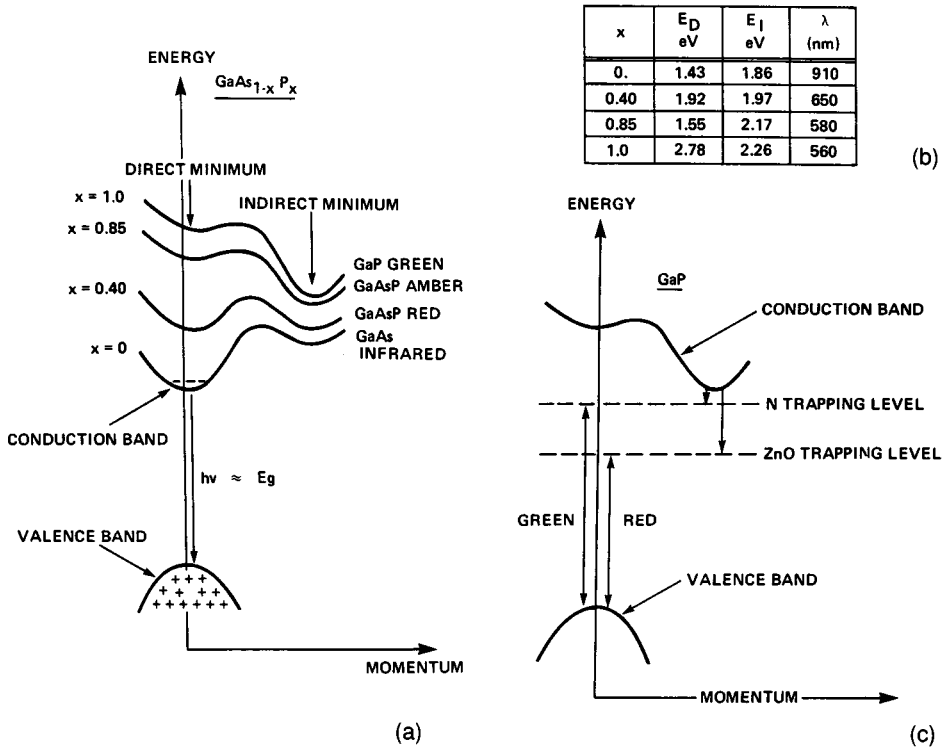


FIGURE 83.3 (a) Plot of momentum versus bandgap energy, and (b) corresponding semiconductor parameters for various compounds of the GaAs/GaP system; (c) plot of momentum versus bandgap energy for indirect GaP materials showing special trapping levels. (*Source:* S. Gage et al., *Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981, pp. 1.3–4. With permission.)

(where c is the velocity of light) is very small, and conservation of momentum can be readily accommodated by small deviations from the vertical transition shown in Fig. 83.2(a). For the indirect case illustrated in Fig. 83.2(b), the energy change ΔE defines the photon energy and momentum, again according to Eqs. (83.1) and (83.2), but conservation of momentum additionally requires that the much greater electron momentum on the order of $h/2a$ be accounted for. For lattice dimensions, a, on the order of 10^{-10} m and wavelengths, λ, on the order of 10^{-6} m, it is clearly not possible for both conservation criteria to be met without the participation of a third body, i.e., a phonon. The two consequences of this result are that the indirect transition is inefficient (in that it must transfer momentum and hence thermal energy to the lattice) and less likely to occur than the direct transition (because of the requirement for all three particles to simultaneously meet the energy and momentum conditions). Indirect bandgaps therefore lead to long diffusion lengths and recombination times, which produce good transistors but poor LEDs.

The most common direct-bandgap semiconductor is GaAs, but the photon wavelength calculated for $E_g = E_D = 1.43$ eV as listed in Fig. 83.3(b) is in the infrared. Such a material may be ideal for communications and sensory optoelectronic applications but is unsuitable for display purposes. The bandgap may be adjusted, however, by the substitution of phosphorus for arsenic in the lattice as shown in Fig. 83.3(a). The color range listed corresponds to the range of LED colors commonly available: red, yellow, and green. The direct and indirect bandgaps, E_D and E_I, of $GaAs_{1-x}P_x$ vary with x as

$$E_D = 1.441 + 1.091x + 0.210x^2 \tag{83.3}$$

and

$$E_I = 1.977 + 0.144x + 0.211x^2 \tag{83.4}$$

[Wang, 1989], enabling one to design the material to produce the required LED color.

Note the continuous transition from the direct GaAs to the indirect GaP. The materials have an indirect bandgap for $x > 0.4$ and have the same problems as light emitters as silicon. The efficiency of an indirect-gap emitter can be greatly enhanced by the introduction of appropriate impurity recombination centers, as shown in Fig. 83.3(c). In the process shown, an injected minority carrier electron (in *p*-type material) is first trapped by the localized impurity (which is itself electrically neutral but which introduces a local potential to the lattice which attracts electrons). The center is then negatively charged and attracts a hole to complete the recombination process, which produces the photon. The recombination center solves the momentum transfer problem, because the trapped electron is localized to the impurity lattice site and has a momentum range according to the Heisenberg Uncertainty Principle of

$$\Delta p \sim h/2\pi a \tag{83.5}$$

that is, sufficient to include the processes shown in the diagram at $p \sim 0$. In the cases used as examples, a nitrogen atom substitutes for a phosphorus, or a zinc–oxygen pair substitutes for adjacent gallium–phosphorus atoms in the $GaAs_{1-x}P_x$ lattice.

The $GaAs_{1-x}P_x$ system is well established, but can only produce wavelengths defined by the range of energy gap widths, i.e., down to green. Blue LEDs require higher band-gap materials:

(a) SiC technology is well developed for high temperature semiconductor applications, but it has an indirect band gap, so its emission efficiency is very poor [Pierret, 1996].

(b) GaN (and In/Al GaN alloys) is a direct band gap material system producing successful blue and blue-green devices [Jiles, 1994; Nakamura, 1995; Pierret, 1996].

(c) II-IV compounds such as ZnS and ZnSe possess direct band gaps in the 1.5–3.6eV range, offering the possibility of full spectrum LEDs within the single materials system [Jiles, 1994].

Device Efficiency

In considering LED efficiencies, it is convenient to consider the emission process to consist of three distinct steps: (a) excitation, (b) recombination, and (c) extraction. These will be discussed with reference to Fig. 83.4.

(a) Photons created by minority electron recombination on the *p*-type side of the junction are more likely to be successfully emitted from the surface of the device, for the structure shown in Fig. 83.4(a) and (b) if the *p*-type region is a thin surface layer. For a given total LED current, I, made up of electron, hole, and space-charge region recombination components, I_n, I_p, and I_r, respectively, the electron injection efficiency (which provides the excitation) is

$$\gamma_n = I_n/(I_n + I_p + I_r) \tag{83.6}$$

In principle, all the physical processes described above apply equally to both electrons and holes. However, the electron mobility, μ_n, is greater than that of a hole, μ_p, and since

$$I_n/I_p = N_d\,\mu_n/N_a\,\mu_p \tag{83.7}$$

(where N_d, N_a are *n*-type donor and *p*-type acceptor doping densities, respectively) greater γ_n is attainable for a given doping ratio than hole injection efficiency, γ_p. Consequently, LEDs are usually p-n^+ diodes constructed as in Fig. 83.4, with the *p*-layer at the surface.

(b) Some of the recombinations undergone by the excess electron distribution, Δn, in the *p*-type region will lead to radiation of the photon desired, but others will not, because of the existence of doping and various impurity levels in the bandgap. The total recombination rate, R, can be written in terms of the radiative and nonradiative rates, R_r and R_{nr}, as

$$R = R_r + R_{nr} \tag{83.8}$$

where

$$R_r = \Delta n/\tau_r, \qquad R_{nr} = \Delta n/\tau_{nr}, \qquad R = \Delta n/\tau \tag{83.9}$$

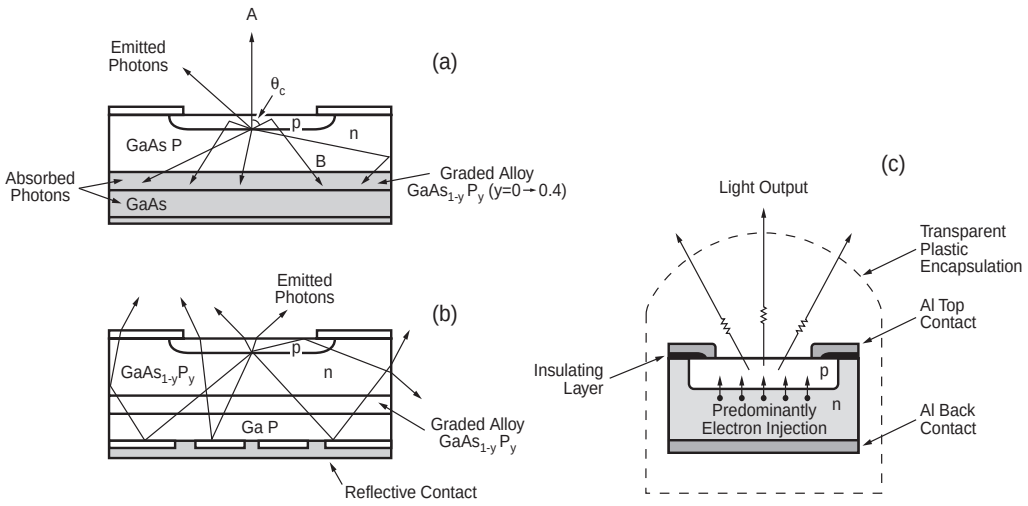


FIGURE 83.4 Effect of (a) opaque substrate, (b) transparent substrate, and (c) encapsulation on photons emitted at the *pn* junction.

and where τ_r and τ_{nr} are the minority carrier lifetimes associated with the radiative and nonradiative recombination processes, and τ is the effective lifetime. The radiative efficiency is defined as

$$\eta = R_r/(R_r + R_{nr}) = \tau/\tau_r \tag{83.10}$$

and the **internal quantum efficiency** is

$$\eta_i = \eta\gamma \tag{83.11}$$

(c) It is clear from Fig. 83.4 that many of the photons generated on either side of the junction will pass through sufficient bulk semiconductor to be reabsorbed. In fact the photon energy may be ideally suited to reabsorption if it exceeds the semiconductor direct bandgap. It is obvious, then, why GaAs is opaque and GaP transparent to photons from Ga(As:P) junctions. Clearly, a greater efficiency might be expected from the transparent substrate with reflecting contact [Fig. 83.4(b)].

The photon must strike the LED surface at an angle less than the critical angle for total internal reflection, θ_c, where

$$\sin\theta_c = n_{\text{ext}}/n_{\text{LED}} = 1/n \tag{83.12}$$

and n_{ext}, n_{LED} are the external and internal refractive indices, respectively. For air, $n_{\text{ext}} = 1$, but critical angle loss can be reduced by encapsulating the device in an epoxy lens cap [Fig. 83.4(c)] to increase both $n_{\text{ext}} > 1$ and the angle of incidence at the air interface.

Even within angles less than θ_c, there is Fresnel loss, with transmission ratio

$$T = 4n/(1 + n)^2 \tag{83.13}$$

The total **external quantum efficiency** is then the fraction of photons emitted [Neamen, 1992], given by [Yang, 1988]

$$\eta_e = 1/(1 + \alpha v_o/AT) \tag{83.14}$$

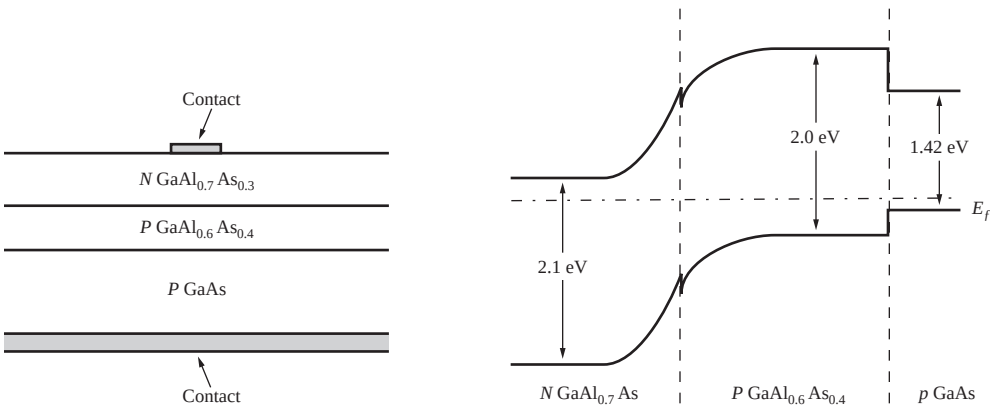


FIGURE 83.5 A GaAlAs heterojunction LED: (a) cross-sectional diagram; (b) energy-band diagram.

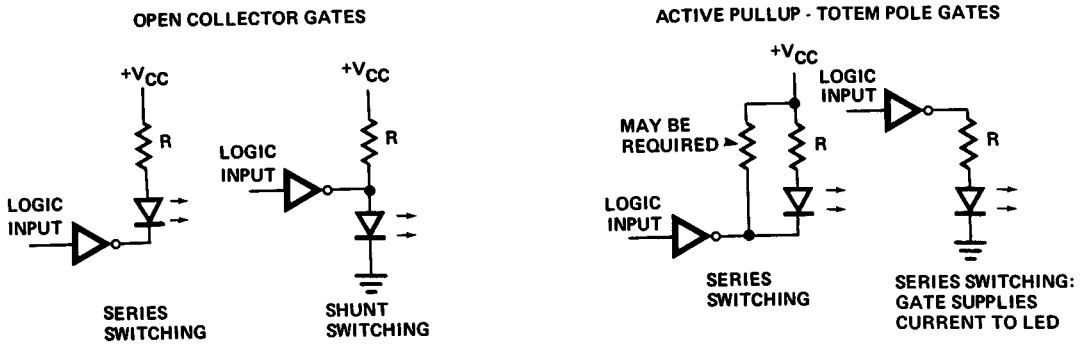


FIGURE 83.6 Digital logic can interface directly to LED lamps. (*Source:* S. Gage et al., *Optoelectronics/Fiber-Optics Applications Manual*, 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981, p. 2.20. With permission.)

where α is the average absorption coefficient, v_o is the LED volume, and A is the emitting area.

In considering LED effectiveness for display purposes, one must also include radiation wavelength in relation to the spectral response of the human eye [Sze, 1985]. Although the GaP green LED is intrinsically less efficient than the GaAsP red LED, the eye compensates for the deficiency with a greater sensitivity to green.

More recently developed heterojunction LEDs (Fig. 83.5) offer two mechanisms to improve LED efficiencies [Yang, 1988]. The electron injection efficiency can be enhanced, but, in addition, absorption losses through the wider 2.1-eV bandgap *n*-type layer are essentially eliminated for photons emitted by recombination in the lower 2.0-eV bandgap *p*-type region.

Interfacing

In circuit design applications, the LED may be treated much as a regular diode, but with a much greater forward voltage, V_P. Since one usually seeks maximum brightness from the device, it is usually conducting heavily and V_F approaches the contact potential. As one moves from GaAs to GaP [Fig. 83.3(a)], V_F varies from about 1.5 to around 2.0 V. The variation in V_F with temperature (at constant current) follows similar rules as apply to conventional diodes, but radiant power and wavelengths also change [Gage et al., 1981].

Single LEDs are commonly driven by logic gates, perhaps as status indicators, and some of the simplest interface circuits are shown in Fig. 83.6. In many cases, the gate output will not be able to source or sink sufficient current for visibility, and an amplifier will be required, as in Fig. 83.7. Bar graph displays are commonly

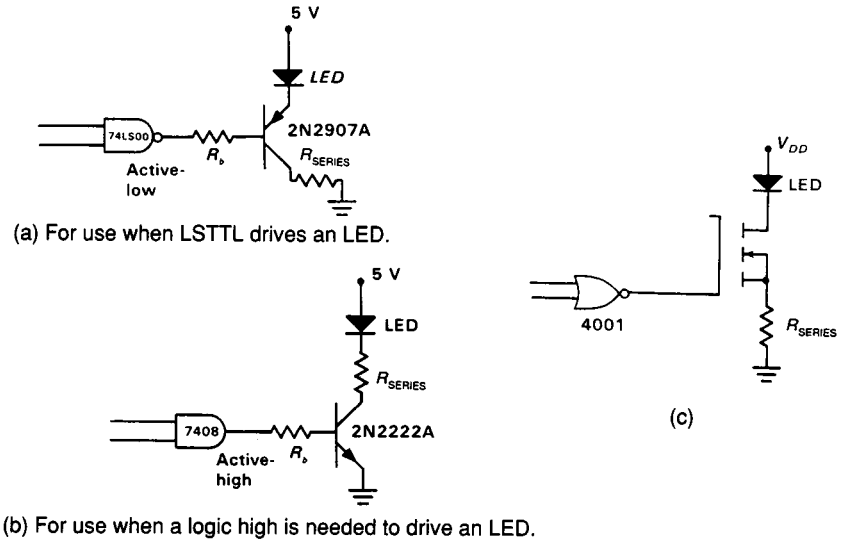


FIGURE 83.7 LED interfacing for (a) low-power transistor-transistor logic, (b) logic high drive, and (c) CMOS. (*Source:* M. Forbes and B.B. Brey, *Digital Electronics,* Indianapolis: Bobbs-Merrill, 1985, p. 242. With permission.)

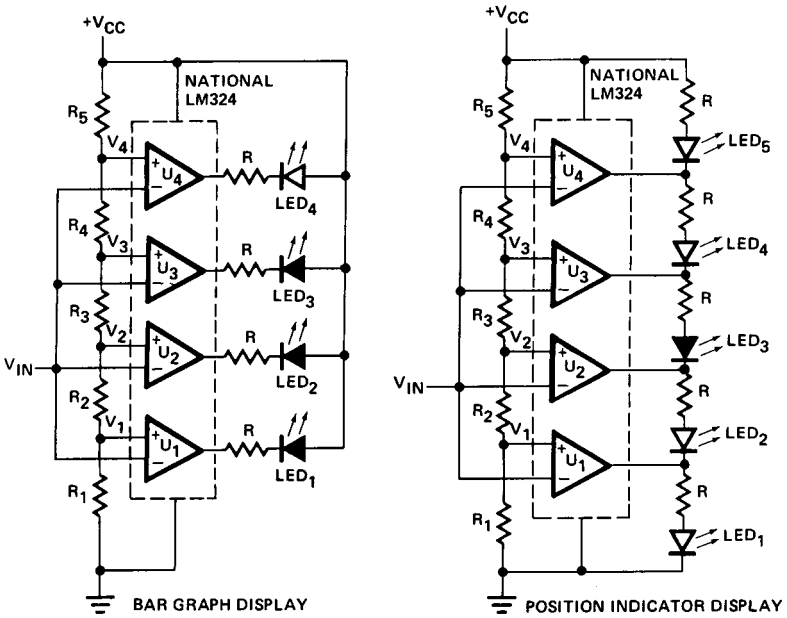


FIGURE 83.8 Operational amplifiers or voltage comparators used to decode an analog signal into a bar graph or position indicator display. (*Source:* S. Gage et al., *Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981, p. 23.3. With permission.)

used to indicate signal level on audio equipment, with a modification of the position indicator seen in Fig. 83.8 to guide fine tuning. Matrix LED arrays can be used for flexible, high-density panel displays [Fig. 83.9(a)] and are conventionally controlled by row or column strobing [Fig. 83.9(b)] controlled by a microprocessor interface.

Multiple LEDs are commonly packaged together in a single integrated device, organized in one of the standard display fonts [Fig. 83.10(a)], with decoding often included within the package [Fig. 83.10(b)]. The 7-segment

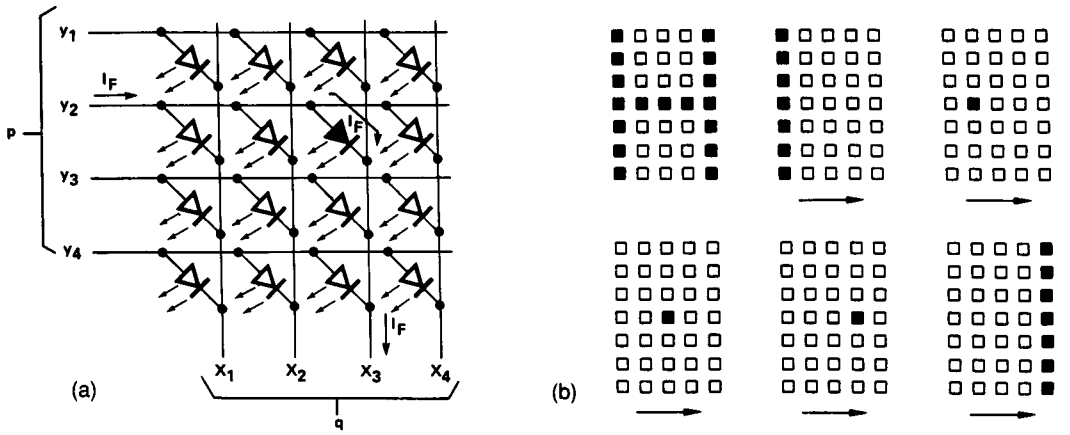


FIGURE 83.9 Matrix displays. (a) One LED will be turned on by applying the proper signal to one *x* axis and one *y* axis. (b) Character generation using column strobe methods. (*Source:* S. Gage et al., *Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981, pp. 2.25, 5.44. With permission.)

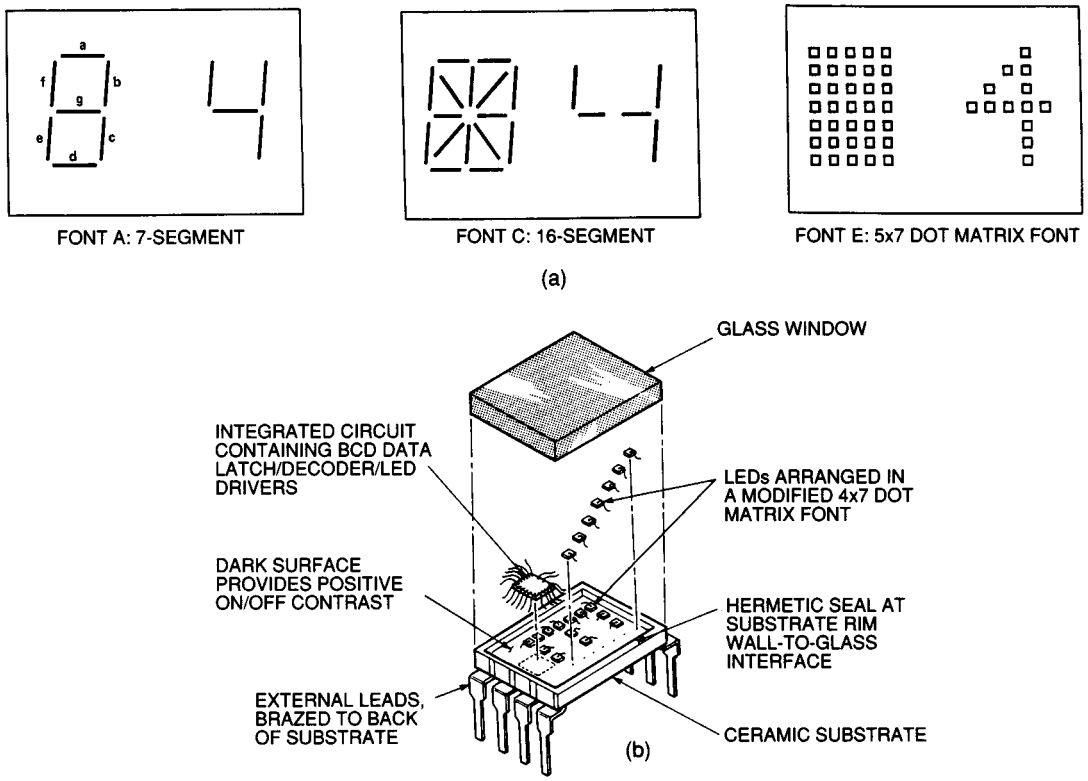


FIGURE 83.10 (a) Display fonts used in LED displays. (b) Construction features of a hermetic LED display. (*Source:* S. Gage et al., *Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981, pp. 5.3, 5.6. With permission.)

display is adequate for hexadecimal applications, but the 16-segment is required for alphanumerics. To limit pin-out requirements, the LEDs of a single package are connected in either the common anode or common cathode configuration [Fig. 83.11(a)], with multiple display digits multiplexed as illustrated in Fig. 83.11(b).

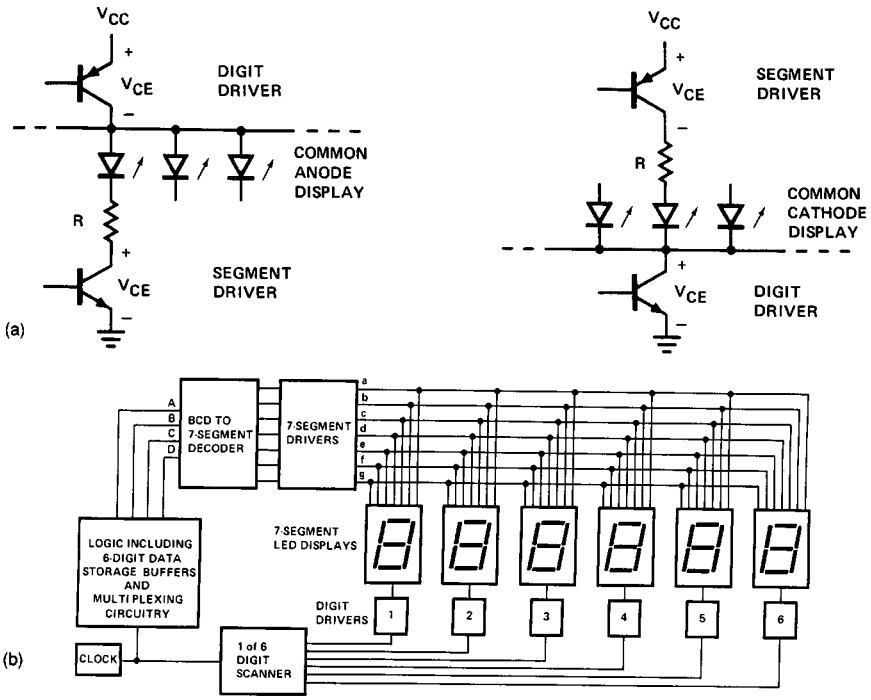


FIGURE 83.11 (a) Generalized drive circuits for strobed operation. (b) Block diagram of a strobed (multiplexed) six-digit LED display. (*Source:* S. Gage et al., *Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981, pp. 5.25, 5.23. With permission.)

Defining Terms

External quantum efficiency: The proportion of the photons emitted from the *pn* junction that escape the device structure (but sometimes alternatively defined as $\eta_i\eta_e$).

Injection electroluminescence: *Electroluminescence* is the general term for optical emission resulting from the passage of electric current; *injection electroluminescence* refers to the case where the mechanism involves the injection of carriers across a *pn* junction.

Internal quantum efficiency: The product of injection efficiency and radiative efficiency corresponds to the ratio of power radiated from the junction to electrical power supplied.

Related Topic

22.1 Physical Properties

References

J. Allison, *Electronic Engineering Semiconductors and Devices,* 2nd ed., London: McGraw-Hill, 1990.

M. Forbes and B. B. Brey, *Digital Electronics,* Indianapolis: Bobbs-Merrill, 1990.

S. Gage, D. Evans, M. Hodapp, H. Sorensen, R. Jamison, and R. Krause, *Optoelectronics/Fiber-Optics Applications Manual,* 2nd ed., New York: Hewlett-Packard/McGraw-Hill, 1981.

D. Jiles, *Introduction to the Electronic Properties of Materials,* London: Chapman & Hall, 1994.

S. Nakamura, "A bright future for blue/green LEDs," *IEEE Circuits & Devices,* 11(3), 19–23, 1995.

D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles,* Boston: Irwin, 1992.

R. F. Pierret, *Semiconductor Device Fundamentals,* New York: Addison-Wesley, 1996.

S. M. Sze, *Semiconductor Devices: Physics and Technology,* New York: Wiley, 1985.

S. Wang, *Fundamentals of Semiconductor Theory and Device Physics,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

E. S. Yang, *Microelectronic Devices,* New York: McGraw-Hill, 1988.

Further Information

More extensive semiconductor device treatments of the LED are contained in *Semiconductor Devices and Integrated Circuits* by A. G. Milnes (Van Nostrand Reinhold, New York) and in *Introduction to Optical Electronics* by K. A. Jones (Harper and Row, New York). E. Uiga provides more interfacing and design detail for the LED as a circuit element in optoelectronics (Prentice-Hall, Englewood Cliffs, NJ, 1995). Wang [1989] considers second-order effects extensively. In *Semiconductor Optoelectronics* by J. Singh (McGraw-Hill, New York, 1996), the emphasis is on communications applications, but the temperature dependence and frequency response issues covered there are also relevant to displays.

Chapter 2 of Gage et al. [1981] contains detailed information on the optical and thermal design constraints on the LED package and on LED back-lit display systems. Chapter 6 considers filtering and other techniques for the contrast enhancement required for direct sunlight viewing.

Professional society magazines are good sources of up-to-date information at the non-specialist level, especially the occasional special issues devoted to topic reviews. *IEEE Spectrum* is a good example, as is the *IEEE Circuits & Devices* magazine.

83.2 Liquid-Crystal Displays

James E. Morris

In a low-power CMOS digital system, the dissipation of a light-emitting diode (LED) or other comparable display technology can dominate the total system's power requirements. In such circumstances the low-power dissipation advantage of CMOS technology can be completely lost. This is the situation in which liquid-crystal display (LCD) technology must be used. The LED (or other active system, such as a plasma or vacuum fluorescent display) emits optical power supplied (comparatively inefficiently) by the system battery or other source. The passive LCD is fundamentally different in that the optical power is supplied externally (by sunlight or room lighting typically) and the system source need supply only the relatively minute amount of power (microwatts per square centimeter) required to change the device's reflective optical properties.

Principle of Operation

Materials classed as liquid crystals are typically liquid at high temperatures and solid at low temperatures, but in the intermediate temperature range they display characteristics of both. Although there are many different types of liquid crystals used, we will concentrate here on the use of **nematic** crystals in **twisted nematic** devices, the most common by far.

The essential feature of a liquid crystal is the long rod-like molecule. In a nematic crystal, the molecules align as shown in Fig. 83.12. If the container surface is microscopically grooved, the interface molecules will be aligned by the grooves and intermolecular forces will maintain that orientation across the liquid crystal [Fig. 83.12(a)]. The molecules will align in an electric field, and beyond a critical value, the field may be sufficient to overcome the alignment with the grooves [Fig. 83.12(b)]. (In practice, the transition is not so abrupt, and groove alignment persists at the interface itself [Fig. 83.13].)

The process of alignment in the electric field is the result of the anisotropic dielectric constant characteristic of liquid crystals. For the electric field parallel to the molecular alignment, $\varepsilon_r = \varepsilon_{\parallel}$, and for a perpendicular field, $\varepsilon_r = \varepsilon_{\perp}$. In a "positive" liquid crystal, $\varepsilon_{\parallel} > \varepsilon_{\perp}$, and the molecules align parallel to the field as described above in order to minimize the system's potential energy.

The principle of the twisted nematic cell is illustrated in Fig. 83.14. The confining plates, typically 10 μm apart, are grooved orthogonally, forcing the molecular orientation to spiral through 90 degrees [Fig. 83.14(a)]. In the LCD, two polarizers and a mirror are added as shown in Fig. 83.14(b). Incident ambient light is polarized and enters the liquid-crystal cell with the plane of polarization parallel to the molecular orientation. As the light traverses the cell, the plane of polarization is rotated by the twist in the liquid crystal, so that it reaches the opposite face with a polarization 90 degrees to the original direction, but parallel now to the direction of the second polarizer, through which it may therefore pass. The light is then reflected from the mirror and passes back through the cell, reversing the prior sequence.

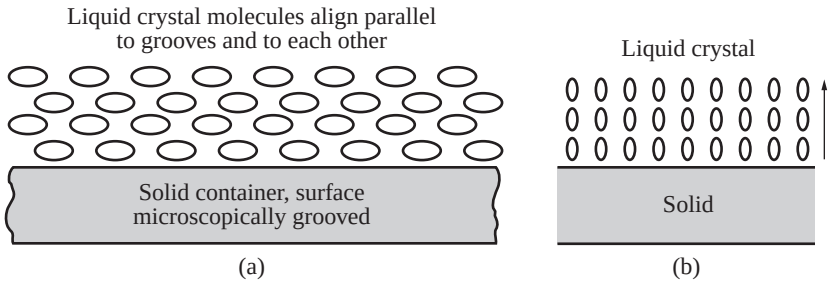


FIGURE 83.12 Liquid-crystal/grooved interface: (a) with no field applied, and (b) with an electric field ε > a critical value.

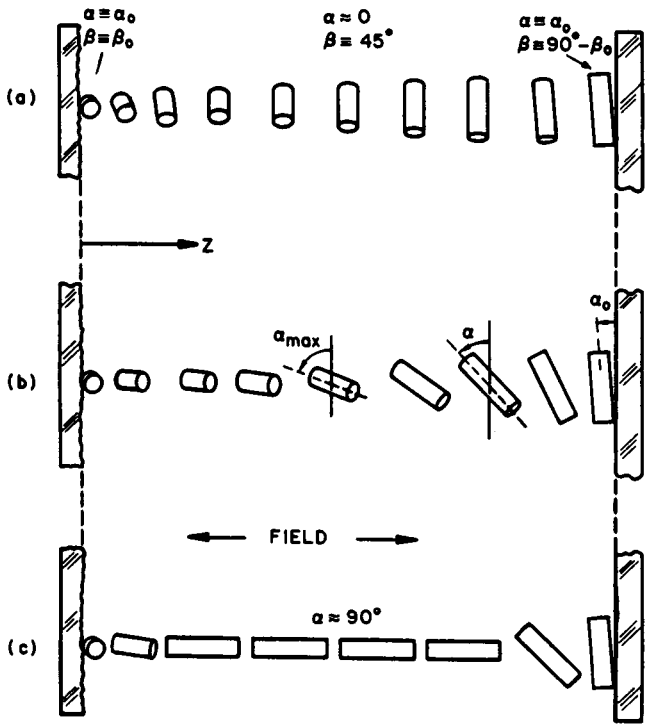


FIGURE 83.13 Diagram of the orientation of the liquid-crystal axis in a cell (a) with no applied field, (b) with about twice the critical field, and (c) with several times the critical field. Note slight permanent tilt (α_0) and turn (β_0) at the surfaces. (*Source:* G. Baur, in *The Physics and Chemistry of Liquid Crystal Devices,* G.J. Sprokel, Ed., New York: Plenum, 1980, p. 62. With permission.)

When an electric field (greater than the critical field) is applied between the transparent electrodes, usually conductive **indium–tin oxide (ITO)** thin films, the 90-degree twist in the crystal is destroyed as the molecules align parallel to the field, so that the rotation of the light's plane of polarization cannot be sustained. Consequently, the crossed polarizers effectively block reflection of the incident light from the backing mirror, and the surface appears to be dark, with excellent contrast to the light gray color of the device in the reflecting mode. The contrast ratio can be further enhanced by the use of the *super twisted nematic* crystal, where the molecular orientation is rotated through 270 degrees rather than 90 degrees.

Transmission LCDs function very similarly to the devices just described, but without the mirror, which is replaced by a powered backlighting source. Obviously, the low-power advantage of the passive device is lost in

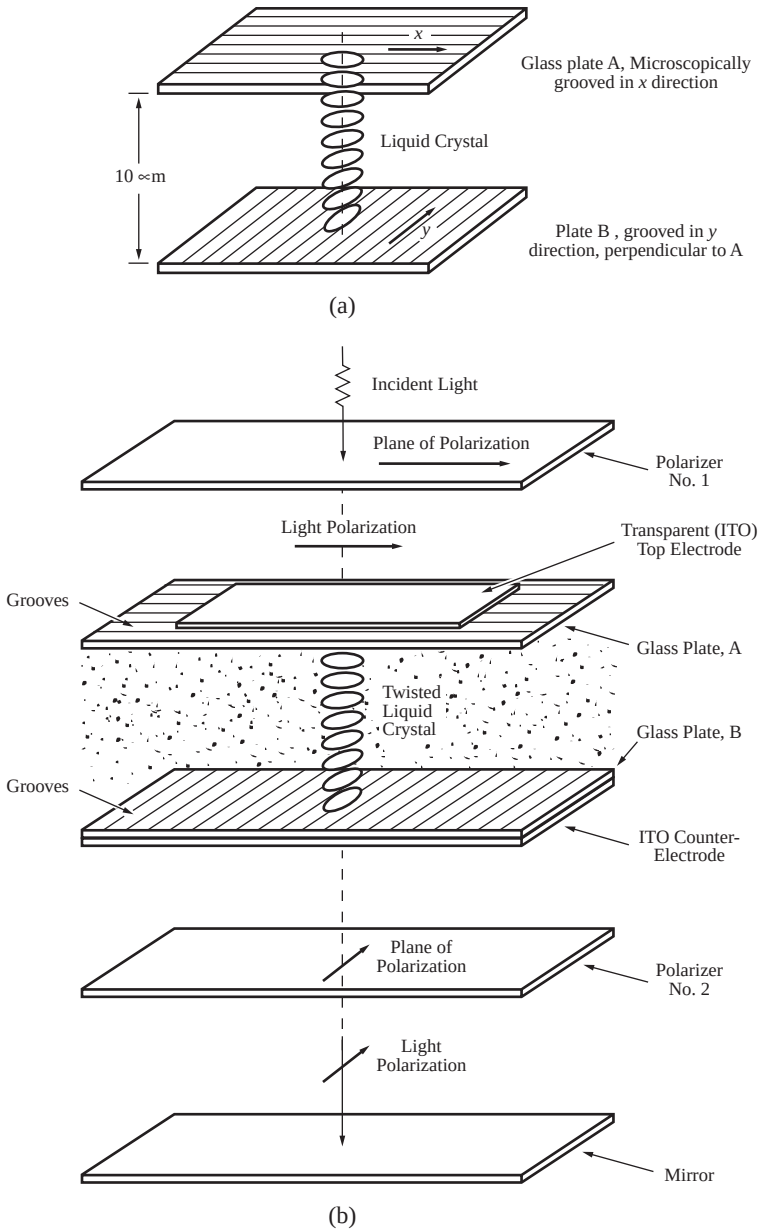


FIGURE 83.14 (a) Twisted nematic cell, $\varepsilon = 0$. (b) Liquid-crystal display element.

this active alternative, but monochromatic backlighting does provide one means of constructing displays with varied background colors.

Another form of color display is provided by **cholesteric** crystals. The three main types of liquid crystals, nematic, cholesteric, and smectic, are distinguished by the different types of molecular ordering they display. In the cholesteric crystal, the direction of molecular alignment rotates in each successive parallel plane (Fig. 83.15). The spatial period of the rotation, p, is called the pitch, and Bragg reflections occur when the wavelength of incident light meets the condition

$$\lambda = p/n \tag{83.15}$$

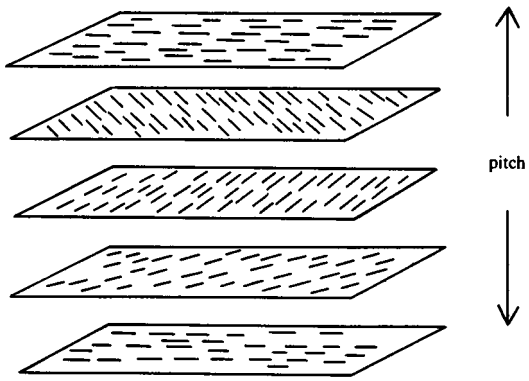


FIGURE 83.15 Cholesteric ordering: a large number of planes of nematic ordering are formed where the directors rotate as we move along a direction perpendicular to the planes. (*Source:* J. Wilson and J.F.B. Hawkes, *Optoelectronics: An Introduction,* London: Prentice-Hall, 1989, p. 145. With permission.)

where n is an integer. The liquid crystal can thus appear to be colored in incident white light. In practice, the color is strongly temperature dependent and the effect is more appropriate to temperature-sensing applications than to digital displays.

In practice, the color pixels of a large area display will be organized in the traditional television RGB format, with the colors defined by external filters or internal dyes [Braithwaite and Weaver, 1995]. The dye molecules align with the LCD molecules, and absorb correctly polarized light.

There is current interest in the development of liquid-crystal color switches where an electrical control signal would be able to change the device color, whether from white to monochromatic or continuously through the spectrum.

Uiga [1995] discusses some LCD problems, in particular the slow response times and the variation of effective critical voltages with limited viewing angles, and the temperature dependences of both.

Interfacing

LCDs can be organized in all the ways available to competing technologies, e.g., LEDs (see Section 83.1), including seven-segment, alphanumeric, and dot matrix. The LCD differs from LED displays, where each pixel or segment must be a separate device, because the LCD segment or pixel areas are defined by transparent electrodes separated from a common overlapping backplane by a single liquid crystal [Fig. 83.16(a)]. In a large matrix array, it may take a significant period to scan all pixels, and the simple addressing scheme of Fig. 83.16(b) may lead to noticeable flicker. The high off resistance of the MOSFETs of Fig. 83.16(c) can reduce this problem by increasing the discharge time to hold the LCD on after the address pulse has gone. The MOSFETs in this **active matrix** technology are implemented in practice in the form of polysilicon or hydrogenated amorphous silicon (a-Si:H) thin film transistors (TFT) [Shur, 1990; Braithwaite and Weaver, 1990].

The interfacing requirements, which are otherwise similar in multiplexing techniques, etc., are complicated by the requirement for zero net dc bias across the cell in order to avoid electrochemical degradation of the material. LCDs require ac drive signals, and square waves of frequency between 25 Hz and 1 kHz are typically used [Wilson and Hawkes, 1989]. A square wave is applied to the backplane, with in-phase and antiphase signals to the counter electrode determining whether the given pixel or segment is on or off. In practice, the state is determined by the root-mean-square (rms) value of the differential voltage applied.

Figure 83.18 illustrates the additional complexity that would be required by even a simple multiplexed addressing system. The backplane and segment drivers might correspond to rows and columns of a dot matrix, as implied in the diagram, or the backplanes may identify specific characters of an alphanumeric display. Calculating the rms values of the difference voltages shown gives 0.42 V_{tc} for the *on* pixels and 0.24 V_{tc} for *off*, from which V_{tc} can be calculated for reliable operation if the critical voltage is known for the LCD to be used.

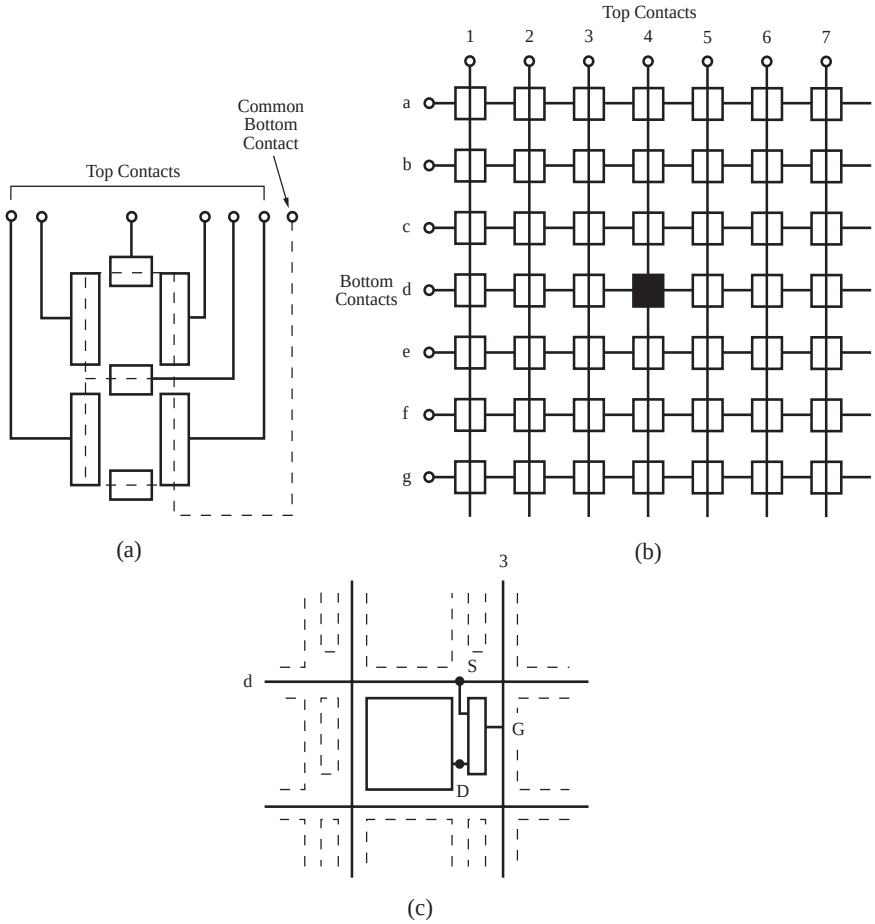


FIGURE 83.16 LCD addressing: (a) simple (seven-segment) addressing; (b) matrix addressing; and (c) matrix addressing with MOSFETs.

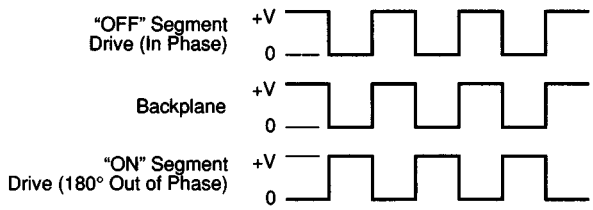


FIGURE 83.17 Drive signals from a direct connect LCD driver. (*Source:* R. Lutz, Application Note 350, in *Interface Databook,* Santa Clara, Calif.: National Semiconductor Corporation, 1990, p. 4–109. With permission.)

Defining Terms

Active matrix: Each pixel in a high density display matrix, such as for flat-screen television, requires its own active (switching element) driver (e.g., a TFT).

Cholesteric: In the cholesteric liquid crystal, successive layers of aligned molecules are rotated naturally.

Indium–tin oxide (ITO): A mixture of the semiconducting oxides SnO_2 and In_2O_3; the most common transparent conductor.

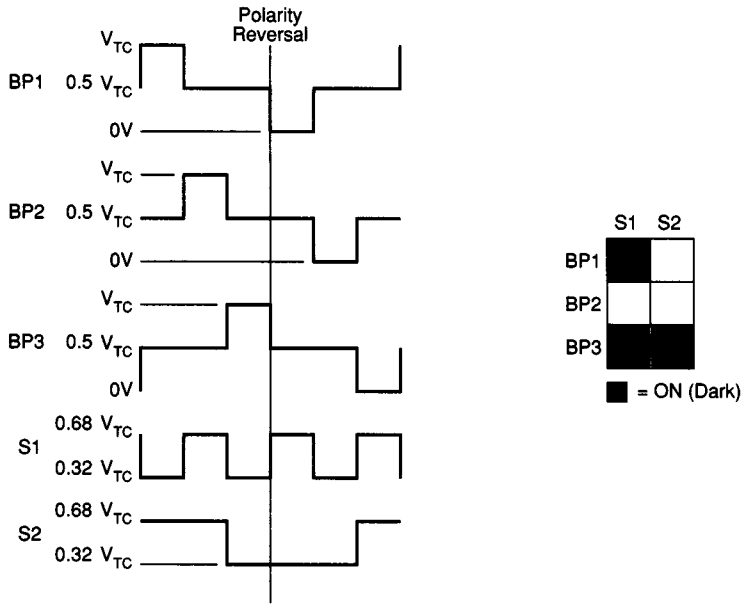


FIGURE 83.18 Example of backplane and segment patterns. (*Source:* R. Lutz, Application Note 350, in *Interface Databook,* Santa Clara, Calif.: National Semiconductor Corporation, 1990, p. 4–109. With permission.)

Nematic: The type of liquid crystal in which the molecular chains align; such alignment can be controlled across the liquid crystal if it can be constrained at the boundaries.

Twisted nematic: The alignments of the nematic planes are rotated through 90 degrees across the crystal by constraining alignments to be orthogonal at the boundaries.

Related Topics

22.1 Physical Properties • 83.1 Light-Emitting Diodes

References

J. Allison, *Electronic Engineering Semiconductors and Devices*, 2nd ed., London: McGraw-Hill, 1990.

G. Baur, "Optical characteristics of liquid crystal displays," in The Physics and Chemistry of Liquid Crystal Devices, G. J. Sprokel, Ed., New York: Plenum, 1980.

N. Braithwaite and G. Weaver, Eds., *Electronic Materials,* Milton Keynes: The Open University/Butterworths, 1990.

R. Lutz, "Designing an LCD dot matrix display interface, application note 350," in *Interface Databook*, Santa Clara, Calif.: National Semiconductor Corporation, 1990.

M. Shur, *Physics of Semiconductor Devices*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

E. Uiga, *Optoelectronics,* Englewood Cliffs, N.J.: Prentice-Hall, 1995.

J. Wilson and J. F. B. Hawkes, *Optoelectronics: An Introduction*, London: Prentice-Hall, 1989.

Further Information

Nematic liquid-crystal molecules typically incorporate two separated benzene rings in a complex chain molecule [Wilson and Hawkes, 1989]. The organic chemistry of liquid-crystal compounds will lie outside the interests of most readers but is briefly reviewed in "Liquid Crystal Materials for Display Devices," by J. A. Castellano and K. J. Harrison in *The Physics and Chemistry of Liquid Crystal Devices,* edited by G. J. Sprokel [Plenum, 1980].

One technique used in liquid-crystal color switches requires the use of electrically controlled birefringence. This topic is covered at an elementary level by Wilson and Hawkes [1989].

An interesting historical perspective on the development of LCD technology is provided by the extensive reviews of 150 patents in the field contained in *Liquid Crystal Devices,* edited by T. Kallard (*State of the Art Review,* Vol. 7, Optosonic Press, New York, 1973). The book also contains a bibliography of more than 1100 entries.

The various professional societies' magazines are excellent sources of material for recent developments in this field (and others). These publications regularly devote a special issue to research developments in a single field, at a level intended for the non-specialist. A good example in the LCD area is provided by two articles on TFT silicon for active matrix displays contained in the *Materials for Flat-Panel Displays* issue of the *MRS Bulletin,* 21(3), March 1996 (Materials Research Society), which cover the transition from a-Si:H to polysilicon, to the prospects for single crystals.

83.3 The Cathode Ray Tube

André Martin

The **cathode ray tube** (CRT) is the element which, in a display, converts an electrical signal into visual information using an electron beam adequately intensity modulated and deflected to impinge on a cathodoluminescent screen surface, in a glass envelope under vacuum.

Because of the extensive growth of electronic communication since the Second World War, information is very often presented on CRTs, mainly when the information content exceeds 100,000 picture elements (pixels).

Monochrome CRTs are widely used for computer terminals, radars, oscilloscopes, projection systems, etc., while high resolution color CRTs are preferred for imaging from computers, such as multimedia, CAD-CAM systems and digital image processors [Keller, 1991].

In 1991, the worldwide market for CRT monitors was 37 million units [Stanford Resources, 1992a], of which 20 million units were high-resolution color monitors. These 37 million do not take into account the oscilloscope, radar, projection, and other special-purpose tubes that would add a few hundred thousand to the figures. In 1996, high resolution color CRT monitors in current use are in excess of 200 million units and mroe than 50 million monitors are being produced.

The importance of the CRT in the display world can be explained by two key factors:

- The CRT is using a single serial data input to generate a picture.
- The CRT is a very efficient light-emitting display.

For example, a typical high-resolution 20-inch diagonal color monitor requires 70 W from the main ac 50/60-Hz power supply and generates a picture visible in any office environment with a luminous efficiency of 6 to 8 lm/W. These two factors, high luminous efficiency and convenience of addressing, make the CRT difficult to replace by any other type of display for large image contents.

We will describe first the monochrome CRT and then discuss the color CRTs.

Monochrome CRTs

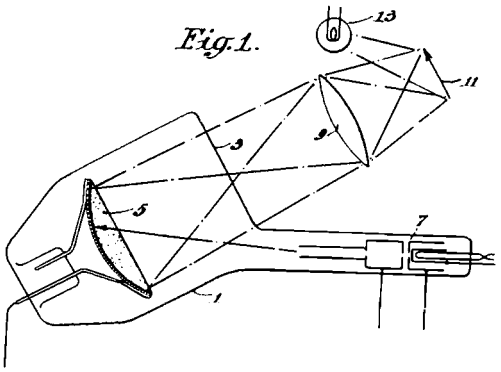
General

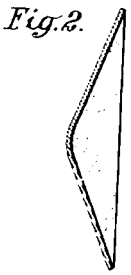
The monochrome CRT [Martin, 1986] is composed of:

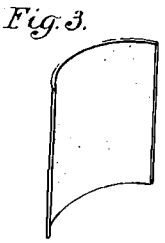
- A glass envelope with the necessary glass-to-metal seals for anode and electron gun connections. This envelope is under vacuum (about 10^{-7} mm Hg).
- A cathodoluminescent screen deposited on the faceplate, usually aluminized to improve brightness and obtain good screen potential uniformity.
- An electron gun using a hot cathode to emit electrons that are accelerated toward the screen and deflected either by electrostatic plates or by an electromagnetic deflection coil.
- Various outside and inside conductive coatings to normalize the potentials.

The Electron Gun

A typical electron gun is composed of (Fig. 83.19):







Cathode Ray Tube

Vladimir K. Zworykin
Patented December 6, 1938
#2,139,296

An excerpt from Zworykin's patent application:

I claim as my invention:

1. In combustion, a cathode ray image transmitting tube provided with an electrode in the form of a mosaic photosensitive target and with means for developing a ray of electrons and directing the ray at a surface of said target for scanning the same, a lens system for focusing an optical image of an object on the scanning portion of the surface of the target, such portion of said surface being concave toward the lens system with the curvature corresponding substantially to that of an imaginary spherical surface on which the lens system is able to focus the optical image sharply.

A spherical (or possibly hyperbolic) photosensitive cathode helped correct for the distortion inherent in aiming the electron beam over a relatively wide dispersion to scan the whole "screen". The basic principles behind Zworykin's CRT, first demonstrated in 1929, have been used in hundreds of millions of radar screens, television tubes, and now computer monitors.

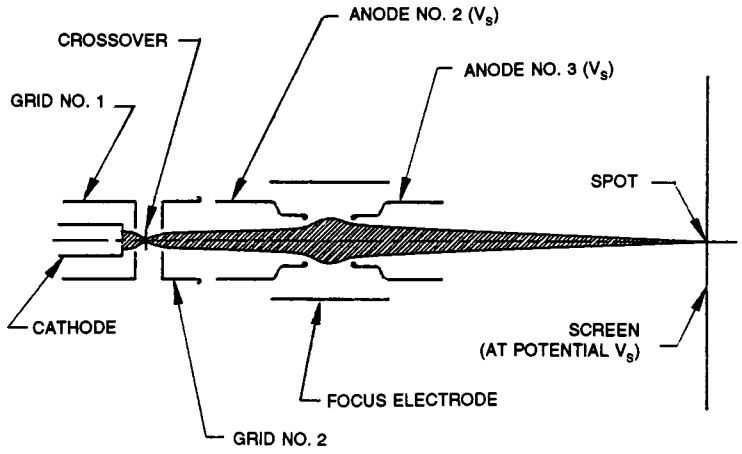


FIGURE 83.19 A typical electron gun with unipotential lens structure.

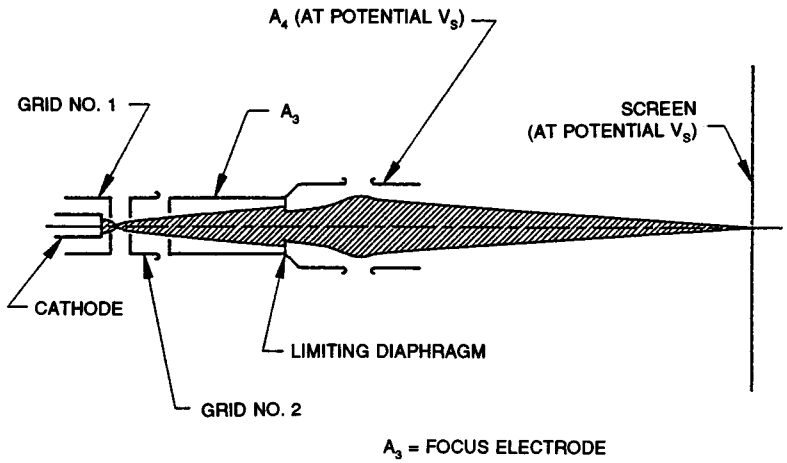


FIGURE 83.20 A typical electron gun with bipotential lens structure.

- A hot cathode that emits electrons when heated to ≈800°C and a suitable potential is applied to the adjacent electrodes G_1 and G_2.
- An apertured grid No. 1, also called G_1, which is maintained negative with respect to the cathode and whose potential controls the flow of electrons from the cathode.
- An apertured grid No. 2 placed close to G_1 (usually a few thousandths of an inch) and set at a positive voltage of a few hundred volts with respect to the cathode. This G_2 attracts the electrons controlled by the G_1 aperture potential and shapes the beam.
- An anode composed of metal cylinders to accelerate the electron beam toward the focus electrode and a final anode to further accelerate the beam toward the screen, where it focuses into a spot.

Figure 83.19 describes the unipotential lens focus structure, also called an EINZEL lens design. Another structure, widely used in modern CRTs, is the bipotential lens focus structure represented in Fig. 83.20. The bipotential structure is theoretically a better performer than the unipotential focus structure, because the lenses have less curvature of the line forces and less spherical aberration.

When optimum resolution is required, electromagnetic focus is used instead of the electrostatic focus systems described previously. Because this magnetic focus lens just bends the electron trajectories without changing the electron's speed, and because of its large diameter, the spherical aberration is reduced and the spot size is optimized.

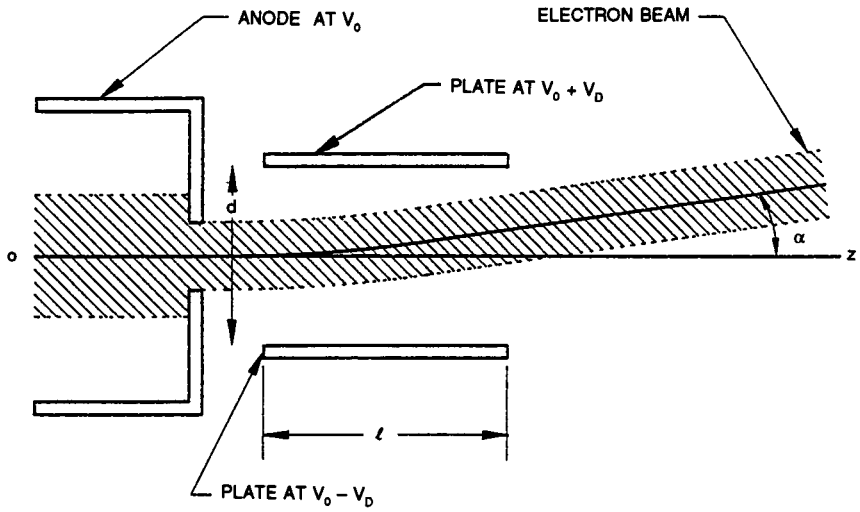


FIGURE 83.21 Principle of electrostatic deflection.

The Cathode

The cathode used on most CRTs is the oxide cathode, which consists typically of a heated nickel substrate coated with barium, strontium, and calcium oxides. This cathode works at a temperature of ≈800°C and provides a dc emission density up to 0.2 amp/cm^2 and 2 amp/cm^2 peak current density. When higher current densities are required, a tungsten-impregnated cathode (porous tungsten matrix heated at ≈1000°C and impregnated with barium and calcium aluminates) can be used. The impregnated cathode operates at higher temperature than the oxide cathode and requires sophisticated materials and techniques for its processing. The impregnated cathode, also known as a dispenser cathode, is commonly used on projection tubes and on other tubes where high beam currents are required, typically above 1.5 mA.

The Electrostatic Deflection System

An electrostatic deflection system (Fig. 83.21) consists of two sets of metal plates of length l symmetrically located with respect to the electron beam axis at a distance d of each other. At the anode outlet aperture (at potential V_0), the beam enters the deflection plates whose potentials are, respectively, $V_0 + V_D$, $V_0 - V_D$. The deflection angle α at the exit of plates is such that

$$\tan \alpha = \frac{1}{2} \frac{V_D}{V_0} \frac{l}{d}$$

In order to increase the deflection sensitivity, plates are often flared to have an optimum contour. High-frequency deflection systems incorporate delay lines to match the electron beam speed in the deflection zone with the signal propagation speed in the delay line.

The Electromagnetic Deflection System

An electromagnetic deflection coil is composed of two perpendicular windings generating electromagnetic fields perpendicular to the trajectory of the electron beam in the vertical and horizontal planes. Figure 83.22 shows the principle of electromagnetic deflection where a field of length l is applied perpendicularly to the electron beam accelerated at V_B. The beam, assuming the field intensity is uniform and of length l, is deflected onto a circular path of radius r. The corresponding angle of deflection is θ such as:

$$\sin \theta = \frac{Nil}{2.68D\sqrt{V_B}}$$

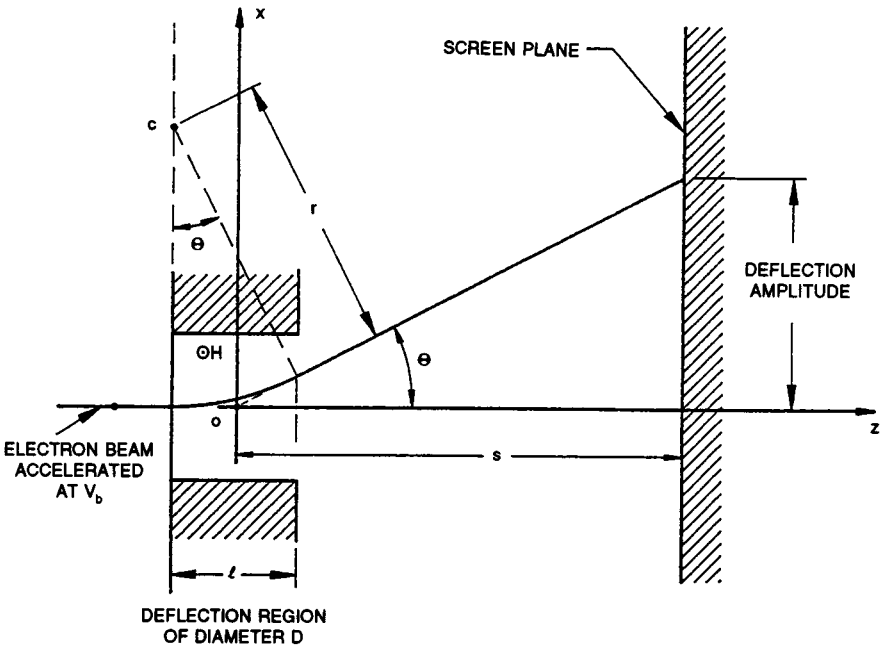


FIGURE 83.22 Principle of electromagnetic deflection.

where *Ni* is the number of ampere turns generating the magnetic field, *D* the diameter of the cylindrical winding generating the field, l is its length, and V_B the accelerating voltage expressed in volts. From the above, the deflection angle is conversely proportional to the coil diameter. As the coil diameter is limited by the tube neck diameter, it is preferable to use a small neck diameter to increase sensitivity to reduce deflection power. Because a small neck diameter cannot accommodate the large electrostatic focusing lenses required to reduce spherical aberration and spot size, a compromise must then be found between spot size and deflection power to achieve the best CRT performance when electrostatic focus is required.

The Screen

A cathodoluminescent material is characterized by several parameters:

- Color, usually expressed by its spectral distribution curve and also measured by color coordinates *xy* or $u'v'$
- Temporal characteristics, such as decay time, usually measured at 10% of initial excitation

These parameters, as well as the chemical composition of the luminescent materials, are listed in the E.I.A. Publication TEP 116 C [Publ. TEP].

Other parameters such as luminous efficiency (expressed in lumens/watt) or energy conversion efficiency (expressed in watts/watt) are also necessary and have to be required from manufacturers.

Color CRTs

The Shadow-Mask CRT

Color CRTs are widely used in commercial television and for computer displays. The shadow-mask CRT [Morrell et al., 1974] uses three electron beams deflected by one deflection coil. The beams traverse a perforated metal mask before impinging on the selected luminescent screen material, which is usually made of stripes or dots of red, blue, and green phosphors. The arrangement of the electron optics and of the deflection system is such that the three electron beams converge on the screen after passing through the shadow-mask, each beam impinging on one color, red, blue, or green, only.

Shadow-mask tubes use a mechanical selection of colors. The thin perforated steel or invar shadow-mask is welded onto a metallic frame suspended by supports in the tube glass faceplate. This structure is sensitive to shock and vibration, which may affect the position of the mask in the faceplate and the registration of the beam on the appropriate phosphor dots. These types of CRTs need specialized damping when ruggedization is required. Another type of shadow-mask CRT, the flat tension mask (FTM), is much more resistant to shock and vibration because of the thin shadow-mask foil tension sealed to the face plate. Suspended weight is minimal and the CRT can withstand high shock and vibration levels [Taki et al., 1996]. The shadow-mask tube is by far the most widespread tube for computer and high-resolution monitor displays.

Two other types of color CRTs are practically immune to shock and vibration. These are the beam index tube and the penetration tube.

The Beam Index CRT

In the beam index tube, the RGB striped screen is intermixed with indexing stripes of a UV emitting luminescent material with very fast decay time. When excited by the electron beam, the index stripes emit a light pulse that is detected by photosensors located on the transparent bulb. The signals are then digitally processed, permitting video signals to be fired at the correct position of the electron beam on the screen.

Beam index CRTs are basically used for avionic applications. A typical 150×150 mm^2 (6×6 in.) CRT can display images with a white brightness around 3400 cd/m^2 (1000 fL).

The Penetration CRT

With the penetration CRT, operation is based on the variation in depth of penetration of an electron beam in successive layers of different luminescent materials, typically red and green emitting.

At low voltage, such as 9 kV, the first layer of luminescent material is excited and a red color is obtained. At high voltage, such as 18 kV, the electrons penetrate the red layer without losing much of their energy and excite the green layer to produce a slightly desaturated green. Typically, only four colors can be produced.

These tubes can be built in any size and can also use a variety of short and long persistence luminescent materials to achieve variable persistence. This is convenient for radar or other specific limited color applications.

The Beam Matrix Flat CRT [Tully, 1994]

This flat panel CRT combines electron beams provided by a hot cathode, like a conventional CRT, and beam modulation and addressing thanks to a matrix of conductive metallic rows and columns. The electron beam generated impinges on a fluorescent screen identical to that of a standard CRT.

Such tubes are, as of 1996, still in the development stage.

Another type of flat panel using an electron beam is the field emission device (FED); these FEDs are in fact using electron beams generated in a very high vacuum by field emission, then these beams are modulated and addressed by sets of matrixed conductive electrodes, arranged in rows and columns, and impinged on a fluorescent screen [Kumar et al., 1994]. These FEDs are in the early stages of development and samples of 10-in. diagonal color panels have been demonstrated in 1996, during the SID conference.

Although the FEDs are not exactly corresponding to the definition of a CRT, they are emissive devices using an electron beam to illuminate a fluorescent screen and hence can be assimilated to the family of CRTs.

Contrast and Brightness

Contrast ratio is defined as the ratio of the luminance L_1 of the picture element to the luminance L_2 of the background as follows:

$$C_R = \frac{L_1}{L_2}$$

Brightness contrast, important to the observer, is defined as the ratio of the luminance of the picture element plus background $L_1 + L_2$ to the background luminance L_2

$$C = \frac{L_1 + L_2}{L_2} = 1 + C_R$$

In order to improve contrast, the usual technique is to use an absorbing faceplate or spectrally matched filter to absorb the light emitted by the luminescent screen once and the ambient incident light twice. In addition, antireflection coatings and antiglare treatments can be applied to the CRT face to reduce the reflections from the front surface. Depending upon the ambient lighting conditions, a compromise must usually be found between light output required and the contrast need.

Measurements on CRTs

Line Width and Modulation Transfer Function

The emitted phosphor spot of a CRT usually presents a Gaussian energy distribution. The measurement is usually performed at the 50% height of the Gaussian curve and is called $L_{0.5}$. It can be done also at the σ point of the Gaussian curve and called $L_{0.6}$, with $L_{0.5} = 1.175\ L_{0.6} = 1.175\ L\sigma$.

Spot size is related to modulation transfer function (MTF) [E.I.A., 1986] by

$$L_{0.5} = \sqrt{\frac{1}{3.56(N)^2}\ L\frac{1}{MTF}} = \frac{0.53}{N}\sqrt{L\frac{1}{MTF}}$$

where MTF is fractional, L is the neperian logarithm, N is the number of cycles/millimeter, and $L_{0.5}$ is expressed in millimeters, and can be related to $L\sigma$ by the formula

$$L_{0.5} = 1.175L\sigma$$

The spot image is normally measured using a microscope with a suitable detector such as a fiber-optic probe coupled to a photomultiplier or a CCD array.

Another method widely used is the shrinking **raster** technique. After a raster of n horizontal lines is scanned on the CRT screen, the vertical size of the screen is reduced until the line structure disappears and produces a uniform luminance to the observer. The number of raster lines is then divided into the raster height:

$$L_{SR} = \frac{\text{compressed raster height}}{n}$$

Comparing L_{SR} to $L\sigma$, authors find values of L_{SR} between 1.17 and 1.23 $L\sigma$.

Brightness

Brightness is measured as area brightness (raster luminance) L_R or peak line brightness L_p. Peak line brightness is an inverse function of the writing speed, a direct function of the refresh rate of the displayed line and, of course, of the beam current. Raster luminance L_R is related to the raster emitting surface (S), to the beam current (I), to the screen efficiency (ε), transmission (T) of the faceplate, and to the screen voltage (V_s) by the formula

$$L_R = \frac{V_S I}{\pi S} T\varepsilon$$

where ε is lumens/watt, I is amperes, S is in square meters, T is fractional, and $\pi = 3.1416$. There is no convenient relation between peak line brightness and raster brightness.

Illumination

Illumination is usually measured by illuminance in lux where: 1 lux = 1 lumen/m^2. The S.I. units are the candela/square meter or nit for luminance and the lux for illuminance. They are related to the commonly used footlambert (fL) and footcandle (fC) by the relations:

$$1\ \text{fL} = \frac{1}{\pi}\ \text{cd/ft}^2 = 3.42\ \text{cd/m}^2$$
$$1\ \text{fC} = 1\ \text{lm/ft}^2 = 10.8\ \text{lux}$$
$$1\ \text{lux} = 1\ \text{lm/m}^2$$

The notion of illumination has to be differentiated from the notion of brightness. The footcandle is often used for illuminance and is produced approximately by a source having a luminance of 1 fL (when the reflectance of the surface is 1), which makes the contrast calculation easy. However, these units must not be confused.

Projection Screen

Very often a CRT is used to project an image onto a screen. For a lambertian screen, which rediffuses the incident light in all directions, we can write

$$B = L\frac{T}{4n^2(m+1)^2}$$

where B is the luminance off the projection screen, L is the luminance of the CRT, f is the focal length of the lens used, m is the magnification factor, d is lens diameter, n is aperture, f/d, and T is the transmission of the optics.

The maximum **flux** F (lumens) from the CRT can be expressed by

$$F = \pi\ LS$$

where L is the CRT screen brightness (candelas/square meters) and S the emitting screen surface (square meters), $\pi = 3.1416$.

Conclusion

The worldwide market for 1996 CRTs is at a yearly level of 50 million units compared to 38 million units in 1991 [Stanford Resources, 1992]. High resolution color CRTs are the mainstream of this growth because there is progressive symbiosis between multimedia and high resolution color television with the upcoming of digital television and the development of the Internet. The growth of this market is presently foreseen by the electron tube community up to 100 million units yearly. It is now obvious that the competitive technologies have still to progress to be able to replace the CRT in many applications and that this progress is much slower than the sometimes optimistic predictions of the scientific and industrial community.

It is also important to note that the Electronics Industries Association is monitoring several standard committees for CRTs and that proposals for standards are submitted by the industry and by laboratories such as the National Information Display Laboratory (see Further Information). These standards are following the evolution of measurement techniques. The adoption of these standards by the industry will make it possible for display users to choose their equipment according to criteria uniformly approved around the world.

The CRT is widely used aboard commercial and military aircraft for cockpit displays, aboard surface ships and submarines, and aboard military and commercial vehicles. Because of its wide range of capabilities, versatility of use, brightness and contrast, image quality, and efficiency, the qualities of the CRT largely offset

its bulkiness and weight for most applications. CRTs are still dominating high-resolution applications ranging from 12-mm helmet-mounted displays to 1000-mm diagonal 2000×2000 pixels color monitors and, according to P. Brody [1980], will continue to dominate the market for a few decades.

Defining Terms

Cathode ray tube: A vacuum tube which uses *cathode rays* to generate a picture on a fluorescent screen. These cathode rays are in fact the electron beam deflected and modulated, which impinges on a phosphor screen to generate a picture according to a repetitive pattern refreshed at a frequency usually between 25 and 72 Hz. The term cathode rays stems from the discovery by Plücker [1858] and Hittorf [1869] of a blue glow present at some spots on a glass tube when studying high-voltage discharge in a low-pressure gas. Crookes [1879] showed that these cathode rays were deflected by a magnetic field and were in fact electrons emitted by the negative electrode. The more poetic term *cathode rays* has been kept instead of electron beam for the cathode ray tube.

Contrast: For a cathode ray tube, contrast is the evaluation of the visibility of the picture presented on the phosphor screen in a given ambient lighting. Contrast is usually measured by the contrast ratio, which is the ratio of the luminance of the picture element under evaluation to the background luminance.

Flux: Also called *radiant flux*, the radiant power emitted by a source. It can be expressed in watts for a radiometric source or in lumens when the source spectral energy distribution is between 400 and 700 nm.

Illumination: The effect of a visible radiation flux received on a given surface. Illumination is measured by the illuminance, which is the luminous flux received by surface unit, usually expressed in lux. One lux equals 1 lumen/m^2.

Raster: Also called *television raster;* it is developed by a moving spot of light generated by an electron beam scanning a CRT phosphor screen in a predetermined and repetitive pattern. A picture is generated by modulating the beam intensity, hence the spot light output, when scanning the screen surface. Usually, horizontal lines are generated scanning in a left-to-right sequence and developed top-to-bottom of the image surface.

Related Topic

83.4 Color Plasma Displays

References

T.P. Brody, "When—if ever—will the CRT be replaced by a flat display panel?" *Microelectronics Journal,* vol. 11, pp. 5–9, 1980.

W. Crookes, "On the illumination of lines of molecular pressure and the trajectory of molecules," *Philos. Trans. R. Soc. London*, vol. 170, pp. 135–164, 1879.

E.I.A. J.T 20 Committee—Meeting #59 (1986)—Test Method—Measurement of M.T.F. for Monochrome CRTs by Fourier Transform.

W. Hittorf, "Über die Elektricitätsleitung der Base," *Ann. Phys.* (*Leipzig*) [*2*], vol. 136, pp. 1–31, 1869.

P. Keller, *The Cathode Ray Tube, Technology, History, and Applications,* New York: Palisades Institute for Research Services, 1991.

N. Kumar, H. Schmidt et al., "Development of nano-crystalline diamond based field emission displays", SID 1994 Symposium Digest, Conference 6.1.

A. Martin, *Cathode Ray Tubes for Industrial and Military Applications,* vol. 67, New York: Academic Press, 1986, pp. 183–328.

A.M. Morrell, H.B. Law, E.G. Ramberg, and E.W. Herold, *Color Television Picture Tubes,* New York: Academic Press, 1974.

J. Plücker, "Über die Einwirkung der Magneten auf die elektrischen Entladungen in verdüunten Gasen," *Ann. Phys.* (*Leipzig*) [*2*], vol. 103, pp. 88–106, 1858.

Publication TEP 116C, Washington, D.C.: Electronic Industries Association.

Stanford Resources—Monitor Market Trends—1991, Menlo Park, Calif.: Information Associates, 1992a, 552 pp.
Stanford Resources—Electronic Display World, San Jose, Calif.: Stanford Resources, Inc., vol. 12, no. 3, p. 4, 1992b.
A. Taki, N. Arimoto, T. Okamoto, and Y. Ueda, "Development of 17-inch pure flat color monitor tube", SID 1996 Symposium Digest, Conference 38.4.
P. Tully, "Matsushita's color flat panel", *Information Display*, 6, 9–11, 1994.

Further Information

Electronic Industries Association (E.I.A.)
2001 I Street, N.W.
Washington, D.C. 20006

E.I.A. prints a wide range of literature on electronics components, computers and industrial electronics, communications and services. E.I.A. also administers many committees for standardization, safety, etc., for cathode ray tubes and other components. E.I.A. offers a complete Electronics Technology Curriculum and technical training books, tapes, etc.

National Information Display Laboratory
David Sarnoff Research Center–CN 8619
Princeton, NJ 08543-8619

The NIDL's strategic objective is to provide direct support to government users while promoting the development and commercialization of advanced soft copy technologies. The NIDL participates in the elaboration of standards for soft copy, for example, for high-resolution monitors, in close relation with E.I.A.

83.4 Color Plasma Displays

Larry F. Weber

Introduction

The last few years have seen an explosive growth in manufacturing capacity and interest in full color **plasma** displays. This is fueled by the realization that plasma displays can fulfill the long sought after goal of consumer-affordable hang-on-the-wall flat-panel television displays with diagonals in the range of 20 to 60 in. Color plasma displays operate on the same physical principle as fluorescent lamps. A gas discharge generates ultraviolet light which excites a phosphor layer that fluoresces visible light. Differing phosphors are used for the red, green, and blue primaries and a full color moving image is obtained by modulating each primary color sub-pixel to one of typically 256 intensity levels at 60 times a second.

One quantitative measurement of industrial activity is the list of major corporate efforts on the three distinct structures shown in Fig. 83.23. Each of these companies is now manufacturing or has demonstrated 40-in. or larger color plasma panels. Many of these companies are looking to plasma displays as the next display device opportunity that will follow the success model of the active matrix liquid crystal displays (AMLCDs).

Color Plasma Display Markets

The plasma display manufacturers have adopted the strategy of a strong attack on the greater than 40-in. diagonal NTSC television and high definition television (HDTV) markets. Display diagonals smaller than 20 in. are specifically avoided in this strategy. Plasma displays have found their proper place in the market by evading the fierce competition from the smaller diagonal LCDs and CRTs since both of these technologies have difficulty with 40- to 60-in. diagonals. In this diagonal range, plasma displays will compete primarily with projection displays.

While projection systems have recently shown a very high level of achievement they are especially vulnerable to plasma because of their limited viewing angle and bulk. It is clear that projection systems have not found

Color Plasma Display Structures

AC Single Substrate
- Fujitsu
- Mitsubishi
- NEC
- Pioneer
- Plasmaco

AC Double Substrate
- Photonics
- Thomson

DC
- Matsushita
- NHK

FIGURE 83.23 Major corporate efforts on fundamental structures.

TABLE 83.1 Color Plasma Display Attributes

1.	Diagonals of 20 to 60 in.
2.	Full 16 million colors
3.	Very strong non-linearity
4.	Inherent memory
5.	Long lifetime
6.	Very wide viewing angle
7.	Instant update time
8.	Good luminance and luminous efficiency
9.	CRT-like manufacturing model
10.	Tolerant to shock, vibration, and temperature extremes
11.	Reasonable impedance characteristics
12.	Precise digital grey scale
13.	CRT-like color gamut

much success in diagonal ranges where there is a wide viewing angle CRT alternative. Thus, a wide angle and thin plasma display alternative at a competitive price will easily dominate this large diagonal market. This market is quite large because of the expectation that many HDTV viewers will prefer the wide screen theater-like effect of the high resolution image. The challenge to the plasma displays will be to achieve prices competitive to those of projection systems.

The potential market for large plasma displays is enormous. The most optimistic market projections show annual world-wide sales of about 10 million plasma displays in the year 2002. While this projection shows a very rapid growth, it is still only 5% of the slightly less than 200 million television sets that will be sold that year. More conservative projections still show $5 billion world-wide sales for plasma displays in 2002.

Color Plasma Display Attributes

Table 83.1 shows some of the attributes of color plasma displays which make them successful. The following reviews each attribute.

1. The electrical characteristics of the gas discharge allow plasma displays to be made with diagonals in the 20- to 60-in. range. Such large diagonals are facilitated by very strong non-linearity and inherent **memory** of the discharge, as discussed in Items 3 and 4 below, which present no practical limitations to the number of lines that can be multiplexed. Also, the high impedance characteristic (covered below in Item 11) coupled with the ability to use highly conductive opaque electrodes, greatly reduces electrode loss limitations to size. Monochrome plasma displays have been sold with sizes as great as 60-in. diagonal,

having over 4 million pixels. By 1996 full color plasma displays had been demonstrated with 46-in. diagonals and 4 million sub-pixels.

2. The all-digital gray scale technique used in color plasma displays allows each primary sub-pixel to display 256 or more intensity levels. This allows full 24-bit color or 16 million colors. These 256 intensity levels are not the limit of the plasma displays but rather a convenient design point for software and system compatibility.
3. The plasma display has a very large non-linearity due to the electrical characteristic of the gas discharge used in all plasma displays. This is an electrical non-linearity, meaning that below a certain threshold voltage, the gas discharge will emit no light. Of course, above that threshold voltage the gas discharge fires and emits a desired color. Very sharp non-linearity allows plasma displays to be multiplexed without limit which makes very large plasma displays practical. This is demonstrated by a number of recently developed $1280 \times 1024 \times 3$ sub-pixel color plasma displays. This is a considerable advantage when compared to other display technologies such as the liquid crystal. The liquid crystal display does not have a very good non-linearity and, therefore, some other non-linear element, such as a thin film transistor, is sometimes added in series with each liquid crystal element to increase the display non-linearity. Of course, this greatly complicates and adds cost to this active matrix liquid crystal.
4. Most color plasma displays have inherent memory which is stored directly in the glass plasma panel. Memory is very desirable for flat panel displays because it allows the display to be very bright even for very large sizes. This is because a display with memory has a pixel duty cycle of one. Displays without memory have a pixel duty cycle of one divided by the number of scanned lines. Thus, as the non-memory displays get bigger and the number of scanned lines increases, the duty cycle and, therefore, the brightness of the display decrease. An additional value of memory is the elimination of flicker because the pixels are on all of the time.
5. The lifetime of color plasma displays can be long. Full color plasma display products have been delivered with specified lifetimes to half luminance of 10,000 to 30,000 h. While this is comparable to some CRT products, there is considerable effort to extend this lifetime further. The failure mode is usually a slow degradation in the phosphor that gradually decreases the display luminance. If all of the pixels are aged uniformly, with perhaps a randomly moving television image, the display will still be usable after the specified lifetime but at reduced luminance. However, displays used for computer images require a much tougher life specification because images such as icons may be left on the same screen location for long periods and burn in an image. These problems are very similar to the phosphor degradation observed on a CRT.
6. One of the major advantages of all plasma displays over liquid crystals is the very wide viewing angle. Plasma displays can even get brighter when viewed off axis and, therefore, have the widest viewing angle of any display technology.
7. Gas discharges switch in microseconds and so plasma displays can be updated instantly. Speed is especially important for the very popular mouse and cursor operations where the cursor would disappear when moving on a liquid crystal display. Full motion television images are not a challenge for plasma displays.
8. The luminance and **luminous efficiency** of color plasma displays are good. Displays having 450 candelas per meter squared at 1 lumen per watt have been demonstrated. Some other display technologies such as ac electroluminescence have a higher material luminous efficiency but this performance must be tempered by the fact that the plasma panel has 1000 times less electrical capacitance than the ac electroluminescent devices. The plasma panel frequently takes less power than the EL devices when the switching loss of the EL panel is considered. This favors plasma panels for larger numbers of scanned lines and favors EL panels for smaller numbers of scanned lines. The crossover point is somewhere in the region of a few hundred scanned lines.
9. Color plasma displays are manufactured in a plant that has considerable commonalty with CRT manufacturing plants. This contrasts sharply with the semiconductor-like manufacturing plant of AMLCDs. Therefore, the plasma display plant will cost much less than the AMLCD plant.
10. The structure can withstand very high levels of shock and vibration when properly mounted. Military plasma displays have been designed for in excess of 150 Gs of shock. Plasma displays can easily operate at both high and low temperature extremes. **ac plasma displays** have a temperature limit dependent

almost solely on the drive circuit characteristics. **dc plasma displays**, which use mercury, should not be operated for long periods at low temperatures without an external heater. All recently introduced color dc plasma displays do not have mercury and do not have this limitation.

11. Plasma displays have a high input impedance characteristic that makes them easy to drive. The dielectric constant of the gas is equal to one, which means that plasma displays have virtually the lowest possible electrode capacitance. This is 1000 times smaller than electroluminescent displays and about 100 times smaller than liquid crystal displays. This translates to lower current requirements and, therefore, smaller drive circuit silicon area for the plasma displays. While plasma displays do require 100-volt address drivers, it is frequently easier to design high voltage circuits than high current circuits. Also, 40-in. and larger displays can be designed with little power dissipation in panel electrodes.
12. The gray scale technique used in color plasma displays is 100% digital, which allows design of an all digital image system having reduced noise and increased stability in the color representations. This will become more important as signal sources with very high quality digital signals, such as those from digital video disks and HDTV, become widely available.
13. The color gamut of the available plasma display phosphors is very good. While the color coordinates do not yet exactly match those used in the CRT, future process adjustments are expected to produce the desired close match.

Gas Discharge Physics

A brief account of gas discharge physics will be covered below. A more detailed discussion of this material is presented in Weber [1985].

Figure 83.24 shows the important reactions that occur in a gas discharge for the monochrome gas mixture of neon and argon. The reactions in the gas volume include ionization (I), excitation (E), metastable generation (M), and Penning ionization (P). The three surface reactions that occur at the cathode cause ejection of electrons from the cathode by a bombarding neon ion, a neon **metastable atom** or by a high energy photon. The most important volume reaction is ionization (I), which can cause the generation of an avalanche in the gas volume as shown in Fig. 83.24. This avalanche is started by an electron near the cathode and as it grows toward the anode, it generates a large number of electron-ion pairs. The number of electron-ion pairs increases with increasing applied voltage across the gas. Ions, photons, or metastable atoms that are transported to the cathode can then eject electrons with a cathode surface-dependent probability and these ejected electrons will initiate further avalanches. These mechanisms act as a positive feedback system that becomes unstable when the loop gain is greater than 1. The onset of the unstable condition is defined as the gas firing voltage. Above this firing voltage the discharge current will continue to grow without bounds if the initial avalanche is primed with at least a single electron.

Figure 83.25 shows the I-V characteristic of a typical gas discharge found in plasma displays. Note that the current is plotted on a log scale over nine orders of magnitude. The most striking feature is the very strong non-linearity at the firing voltage, which is a major attribute of gas discharges that allows matrix addressing. When the discharge current has sufficient magnitude, space charge distortion sets in and the characteristic achieves a negative resistance region. Most plasma displays operate near the junction of the normal and the abnormal glow regions of the characteristic.

One critical aspect of gas discharges is the requirement for external priming as shown in the lower part of Fig. 83.25. The avalanche process shown in Fig. 83.24 needs at least one electron to start the discharge growth. Without this first electron, the discharge will not start at any voltage. Priming electrons can come from a number of different active particles created either by a prior discharge or by neighboring discharging pixels. Active particles include free electrons, free ions, metastable atoms, and ultraviolet photons.

Figure 83.26 shows the characteristics of the glow discharge commonly found in operating plasma displays. The light comes from two luminous regions: the negative glow and the positive column. All plasma displays on the market today use light from the negative glow but a few research displays have used the light from the positive column. These regions are caused by the space charge distribution of the electrons and ions that distort the electric field and voltage distribution.

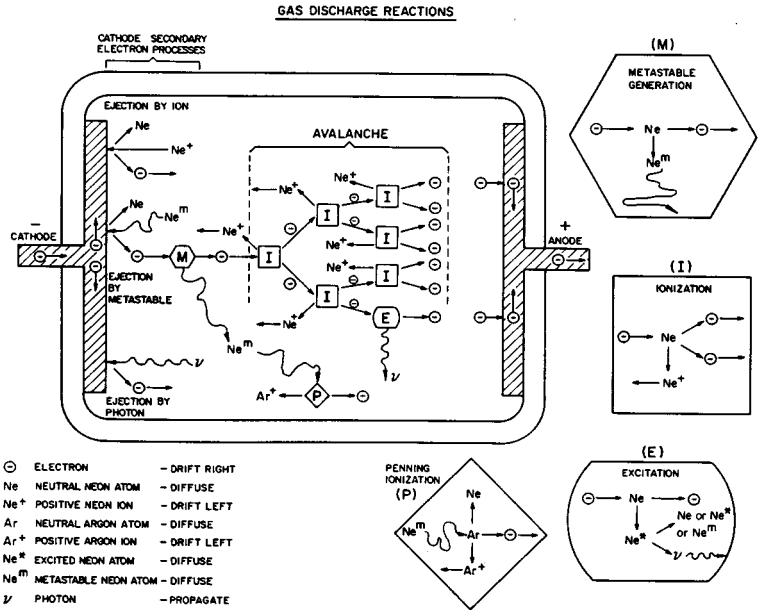


FIGURE 83.24 Model of important gas discharge reactions.

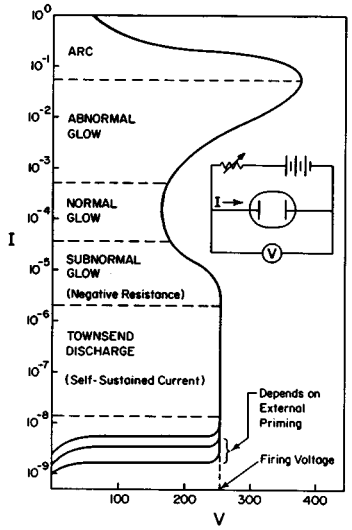


FIGURE 83.25 The I-V characteristic of a gas discharge.

Current Limiting for Plasma Displays

To avoid a catastrophic arc, the current in a gas discharge must be limited by some means. There are a number of ways of accomplishing this, but only two, shown in Fig. 83.27, have achieved commercial success. dc plasma displays use a resistor, a semiconductor current source, or a short applied voltage pulse to limit the current and have the electrodes in intimate contact with the gas discharge. ac plasma displays limit the current with an internal glass dielectric that couples the electrodes capacitively to the gas discharge.

Most of the commercially successful monochrome dc displays have the resistors or current sources connected to a display electrode external to the panel which allows only one discharge to be ignited along that electrode at any one time. This works well for scanned displays. Multiple discharges and dc memory require placing the

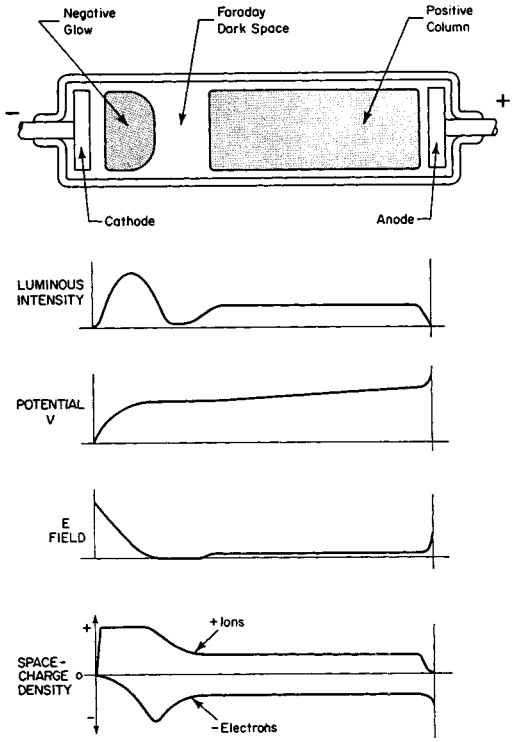


FIGURE 83.26 Luminous regions of a gas discharge.

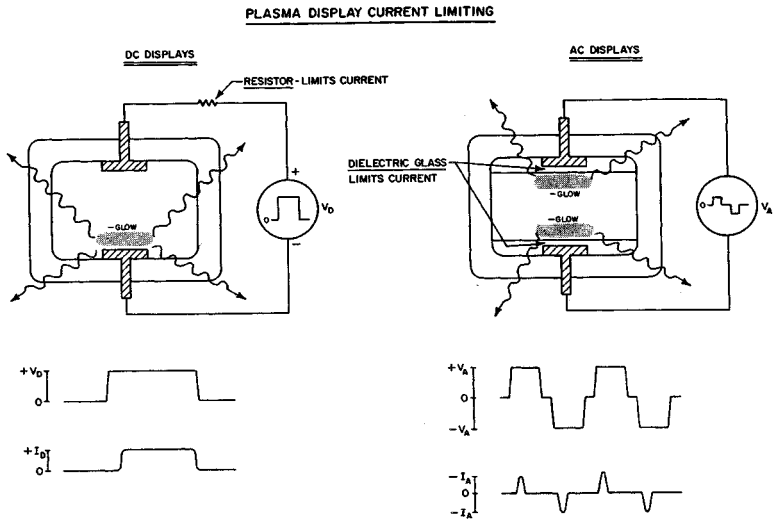


FIGURE 83.27 The two current limiting techniques used in plasma display products.

resistor internal to the panel in series with each pixel. Recent materials and process advances have allowed practical dc color displays with memory to be made having a resistor per sub-pixel.

The ac displays can achieve memory and the necessary current limiting with a simple dielectric layer that forms a capacitor in series with each pixel. When a voltage pulse is applied to an ac panel, the discharge deposits

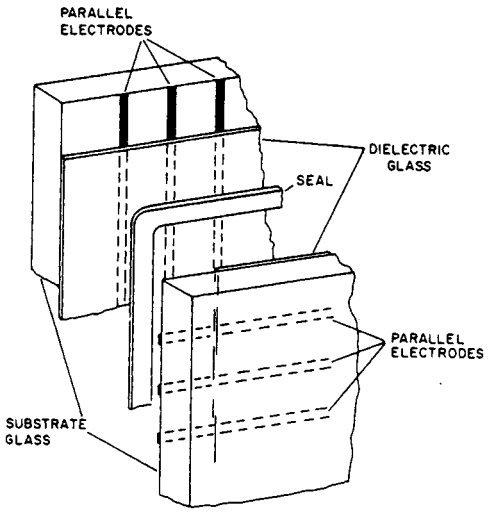


FIGURE 83.28 Monochrome ac plasma display structure.

a charge on the wall that reduces the voltage across the gas. After a short time, the discharge will extinguish and the light output will end until the applied voltage reverses polarity and a new discharge pulse occurs. This wall charge allows the ac plasma displays to operate in a memory mode, which greatly increases the brightness of large displays.

ac Plasma Displays

Figure 83.23 shows that currently, color plasma displays are dominated by the ac technology, so it is worth examining the ac monochrome structure shown in Fig. 83.28 [Criscimagna and Pleshko, 1980]. These panels are made by depositing thin film electrodes on the front and back substrates and then covering those electrodes with a thin dielectric glass. Recall from Fig. 83.27 that this dielectric glass makes a capacitor that is used to limit the discharge current. This dielectric also is used to store the charge that gives these panels inherent memory. The two substrates are then sealed together around the perimeter and filled with neon gas. The ac panels have a very simple structure that allows the pixels to be isolated simply by the action of electric fields.

The ac plasma panels require the inner surface of the dielectric that is in contact with the gas to have a special coating of magnesium oxide. This MgO layer is necessary for the panel to have low operating voltages and long life. Being a refractory oxide, MgO sputters away at a very low rate and it is also well known for its high secondary electron emission.

The largest plasma display product ever manufactured (prior to 1997) uses the Fig. 83.28 structure and has a diagonal of 60 in. with a 2048×2048 array of more than 4 million pixels [Wedding et. al., 1987]. This display operates at a very high update rate so that it will work with a standard NTSC video source. The memory feature allows this display to have the same luminance as the smaller page sized displays.

ac displays require that an ac signal, called the sustain voltage, be applied during operation as shown on the right side of Fig. 83.27. A typical sustain frequency is 50 kHz. Figure 83.29 shows the details of this operation for a pixel in both the on and off states. When a pixel is discharging, charge collects on the dielectric glass walls and influences the voltage across the gas. The component of voltage due to this charge is called the wall voltage. When a pixel is on, the wall voltage changes for each polarity reversal of the sustain voltage. This change in wall voltage coincides with a pulse of light due to the gas discharge. When the pixel is off, there are no light pulses, and the wall voltage remains at a zero level.

Pixel addressing is achieved through a partial discharge by introducing an address pulse timed between the sustain pulses. A write pulse causes the wall voltage to transit from zero volts to the final equilibrium wall voltage level. Likewise, an erase pulse causes the wall voltage to return to zero.

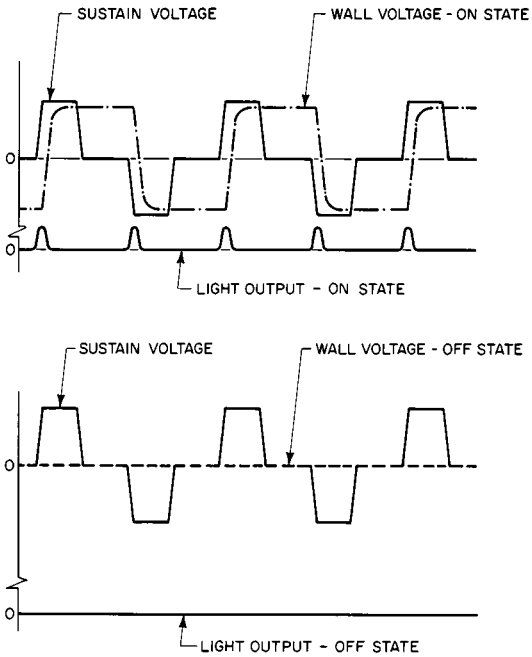


FIGURE 83.29 Sustain voltage, wall voltage, and light output for ac plasma pixels in the on and off states.

Color Plasma Display Devices

Color is achieved by placing phosphors in the plasma panel and then exciting those phosphors with the ultraviolet light of the gas discharge. This is the same principle as used in the fluorescent lamp. Xenon is the active UV generating gas which provides atomic resonance radiation at 147 nm and a molecular band centered at 173 nm. Neon or helium buffer gases are always mixed with the xenon.

Figure 83.23 indicates the companies who are making a serious effort at developing the three dominant full color structures. The basic concepts of the two ac structures are shown in Fig. 83.30. The double substrate structure is very similar to the monochrome ac structure shown in Fig. 83.28. The single substrate structure separates the discharge cathode areas from the phosphor by applying the sustain voltage only to the lower electrodes while the phosphor is on the top. The single substrate approach promises longer phosphor life because it is not directly sputtered by the energetic ions that are directed toward the cathodes.

The structure for the single substrate ac plasma displays is shown in Fig. 83.31 [Shinoda et al., 1993]. Note that the front and back substrates each have simple one dimensional features. Since the two substrate structures are positioned orthogonally, there is no critical alignment between the two substrates because the pixels will automatically occur wherever the orthogonal electrodes intersect. This allows for straightforward manufacture of large panels.

The phosphors are placed on the rear substrate of the panel in Fig. 83.31 and are excited by the ultraviolet light generated by the electrodes on the top substrate. In this case, both sets of ac sustain electrodes are on the upper substrate. The ac voltage is applied to these electrodes in the normal way and the fringing fields from these electrodes reach into the gas and create a discharge. Note that the structure in Fig. 83.31 has glass barrier rib separators between each sub-pixel. This is necessary to reduce cross-talk between the different colors that will reduce the color purity. These barrier ribs do not transmit the 147-nm or 173-nm radiation generated by the xenon gas used in color plasma displays. The phosphors are placed on all walls of the sub-pixel channel except for the front plate which has the phosphor damaging **sputtering** activity at the cathodes. This nearly complete phosphor coverage of the walls maximizes luminance while minimizing sputtering damage.

Other important features of the structure in Fig. 83.31 are the address electrodes buried beneath the phosphors of the rear substrate. These are the column electrodes that are selectively pulsed depending on the input image data. While these address operations do create discharge activity that could potentially sputter damage

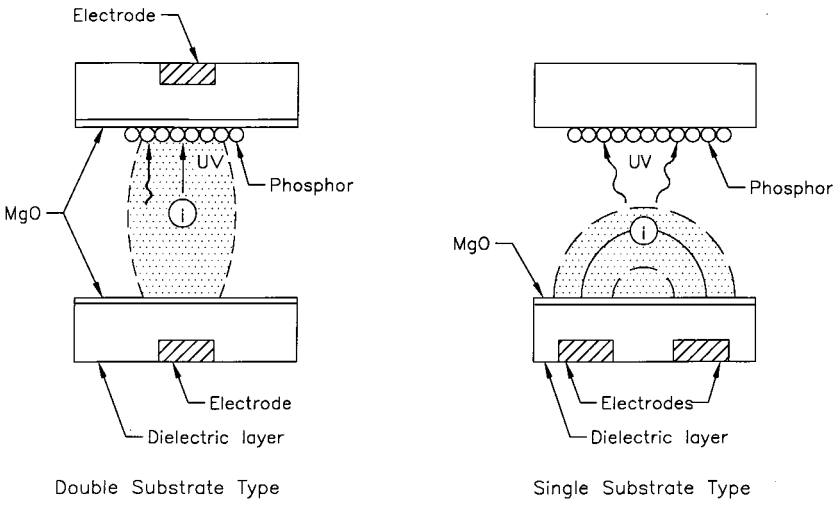


FIGURE 83.30 Two major structural designs of color ac plasma displays.

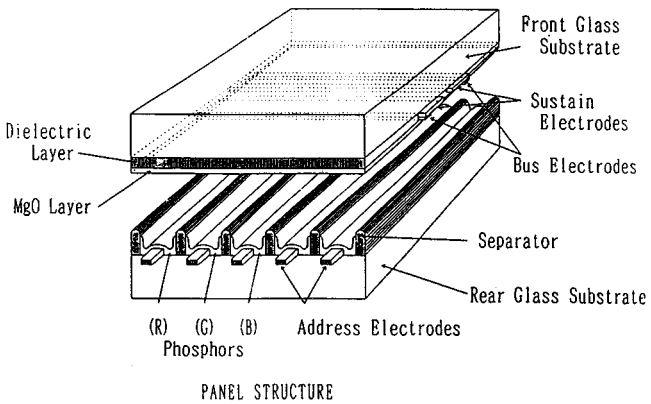


FIGURE 83.31 Structure of single substrate color ac panel.

the phosphor, the address pulse frequency is orders of magnitude lower than the sustain frequency and so the amount of address damage is minimal.

The sustain electrodes shown in Fig. 83.31 are made of a conductive transparent material such as tin oxide. Unfortunately, the resistance is orders of magnitude too high. To correct this problem, narrow bus electrodes of high conductivity materials, such as silver or chrome-copper-chrome, are placed over the tin oxide to reduce the electrode resistance to values on the order of 100 ohms.

The ac double substrate device shown on the left of Fig. 83.30 has a structure very similar to the monochrome device of Fig. 83.28 [Doyeux and Deschamps, 1995]. The major differences are the introduction of glass barrier ribs to maintain color purity and the introduction of color phosphors. The phosphors are placed on one substrate and are carefully positioned to avoid the location directly over the electrode since the sputtering action over the electrode will cause significant phosphor degradation.

Figure 83.32 shows the color dc plasma display structure [Koike et al., 1995]. The major difference between the dc and ac structures is the placement of resistors in series with each sub-pixel to limit the discharge current for the dc case as shown in Fig. 83.27. Another difference is the requirement for the dc device to have barrier ribs on all four edges of each sub-pixel. This is needed to prevent the dc discharge from spreading to neighboring sub-pixels in addition to the color purity issue discussed above for ac panels. Unlike the single substrate ac structure shown in Fig. 83.31, the alternate color phosphors are patterned along each row and each column in

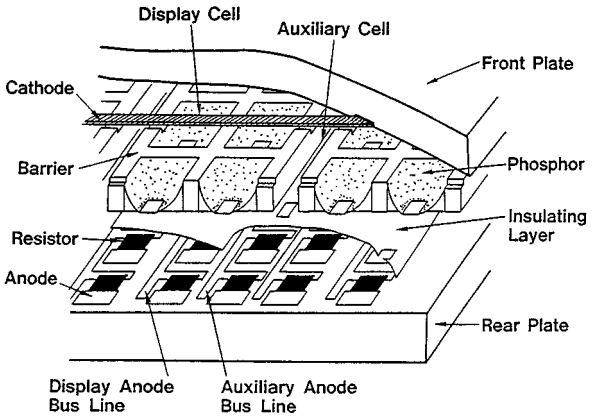


FIGURE 83.32 Structure of color dc panel.

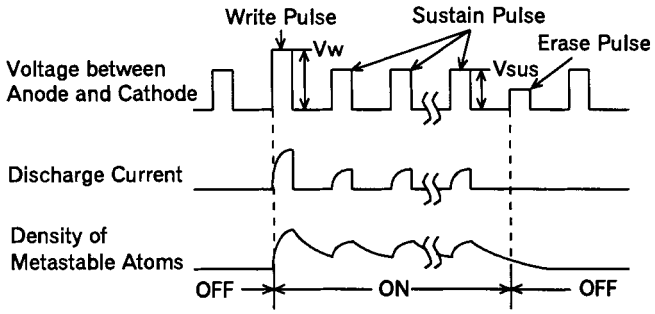


FIGURE 83.33 Pulsed memory mode for dc plasma panels.

a two-dimensional structure. This requires careful alignment between the front and the rear substrates. Priming is achieved by means of the auxiliary cells and anodes placed between the sub-pixels that generate a discharge that is not visible to the viewer.

Inherent memory is achieved for the dc plasma structure of Fig. 83.32 with the pulsed memory mode waveforms shown in Fig. 83.33 [Takano et al., 1994]. This operates on the principle that if there is metastable priming from the preceding discharge pulse, then the discharge will build up in a sufficiently short time to fully mature during the very short electrode voltage pulse. The electrode voltage pulse is adjusted to be sufficiently short to inhibit a sequence of discharges if there has not been an initiating higher amplitude address pulse. The pulse memory mode was invented over 25 years ago [Holz, 1972], and is the most widely studied technique for achieving memory in dc plasma displays. It was just recently introduced to commercial production.

Gray Scale

The ac or dc memory displays cannot use pulse intensity or pulse width modulation for gray scale because the pixels in memory mode are either on or off and such pulse perturbations would, in many cases, have the undesirable effect of changing the state of the pixel. Instead these memory displays achieve gray scale by modulating the percentage of time that the pixel is on in a given frame. This means that the pixels must be addressed multiple times per frame. In the sequence shown in Fig. 83.34 for 256 intensity levels, each frame is divided into eight sub-fields and each sub-field consists of an address period and a sustain period [Yoshikawa, 1992]. During a given address period, address pulses are applied to all pixels in the panel according to the sub-field image data. Each of the eight sub-fields has a sustain period with a different number of sustain cycles which emits an amount of light proportional to the number of sustain cycles. If each data bit of a given pixel

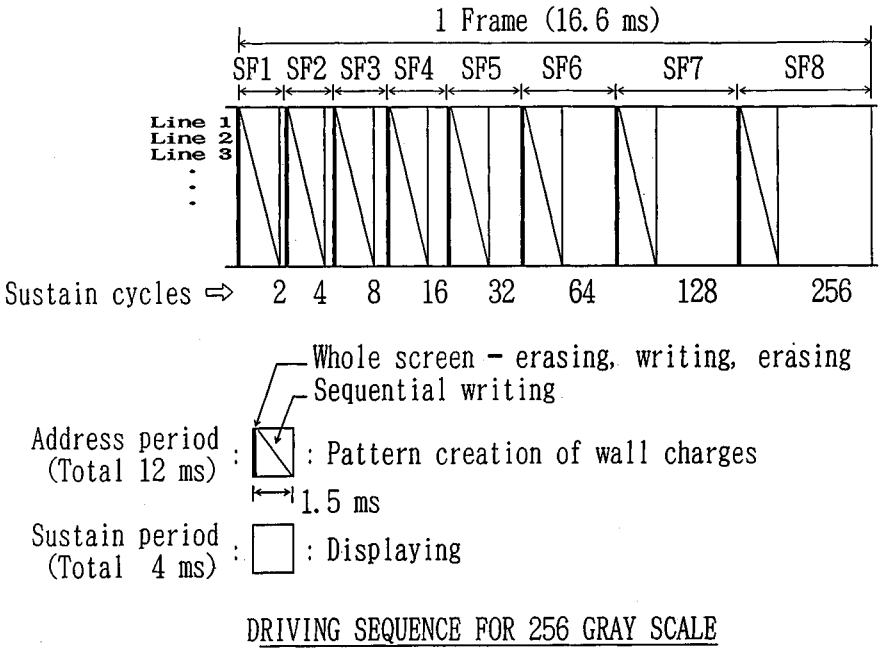


FIGURE 83.34 Addressing sequence for 256 intensity gray scale in ac memory plasma displays.

intensity word is allowed to control one of the sub-fields, then the total number of sustain cycles (and light) per frame will be proportional to the 8-bit intensity value. This results in a precise inherently digital gray scale.

Defining Terms

ac plasma displays: These employ an internal capacitive dielectric layer to limit the gas discharge current.

dc plasma displays: These employ an internal or external resistor to limit the gas discharge current.

Luminous efficiency: The measure of the display output light luminance for a given input power, usually measured in lumens per watt, which is equivalent to the nit.

Memory: The property of a display pixel that allows it to remain stable in an initially established state of luminance. Memory gives a display high luminance and absence of flicker.

Metastable atom: An atom in a temporary but long lived excited state in which photon emission is forbidden by electrodynamic theory. Metastables can give up their energy by ionizing other atoms or through wall collisions.

Plasma: The fourth state of matter comprised of positive ions and negative electrons of equal and sufficiently high density to nearly cancel out any applied electric field. Not to be confused with blood plasma.

Sputtering: The physical process whereby an ion with kinetic energy in the gas collides with a solid surface ejecting an atom from the solid into the gas.

Related Topic

83.3 The Cathode Ray Tube

References

T. N. Criscimagna and P. Pleshko, "ac Plasma Display," in *Topics in Applied Physics,* Vol. 40 Display Devices, Berlin: Springer-Verlag, 1980, pp. 91–150.

H. Doyeux and J. Deschamps, "A high-resolution 19-in. 1024 × 768 Color ac PDP," *SID Intl. Symp.,* Orlando, pp. 811–814, 1995.

G. E. Holz, "Pulsed gas discharge display with memory," *SID Intl. Symp.,* San Francisco, pp. 36–37, 1972.
J. Koike et al., "Long-life, high luminance 40-in. color DC PDP for HDTV," *Intl. Display Res. Conf.,* Hamamatsu, Japan, pp. 943–944, 1995.
T. Shinoda et al., "Development of technologies for large-area color ac plasma displays," *SID Intl. Symp.,* Seattle, pp. 161–164, 1993.
Y. Takano et al., "A 40-in. dc-PDP with new pulse-memory drive scheme," *SID Intl. Symp.,* San Jose, pp. 731–734, 1994.
L. F. Weber, "Color plasma displays," *SID Seminar Lecture Notes,* Vol. 1, San Diego, pp. M-6/1–41, 1996.
L. F. Weber, "Plasma displays," in *Flat-Panel Displays and CRTs,* L. E. Tannas Jr., Ed., New York: Van Nostrand Reinhold, 1985, pp. 332–414.
D. K. Wedding, P. S. Friedman, T. J. Soper, T. D. Holloway, and C. D. Reuter, "A 1.5 m diagonal ac gas discharge display," *SID Intl. Symp.,* New Orleans, pp. 96–99, 1987.
K. Yoshikawa et al., "A full color ac plasma display with 256 gray scale," *Intl. Display Res. Conf.,* Hiroshima, pp. 605–608, 1992.

Further Information

A more detailed account of the material presented in this section and presentations of other display technologies are provided in Weber [1985] and Weber [1996].

The Society for Information Display (SID) annual International Symposium publishes a digest of technical papers which is the best source for new display developments. Tutorial material can be found in the annual SID Seminar Lecture Notes. More research-oriented papers can be found in the technical digest of the International Display Research Conference which rotates annually among Europe, Japan, and North America. In addition, SID publishes the quarterly *Journal of the SID,* which contains more detailed archival versions of selected papers from the conferences. These materials can be obtained from the SID at 1526 Brookhollow Drive, Suite 82, Santa Ana, CA 92705-5421, or see the web site http://www.display.org/sid.

Kurumbalapitiya, D., Hoole, S.R.H. “Data Acquisition”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

84

Data Acquisition

Dhammika Kurumbalapitiya
Harvey Mudd College

S. Ratnajeevan H. Hoole
Harvey Mudd College

84.1 Introduction
84.2 The Analog and Digital Signal Interface
84.3 Analog Signal Conditioning
84.4 Sample-and-Hold and A/D Techniques in Data Acquisition
84.5 The Communication Interface of a Data Acquisition System
84.6 Data Recording
84.7 Software Aspects

84.1 Introduction

Data acquisition includes everything from gathering data, to transporting it, to storing it. The term *data acquisition* is described as the "phase of data handling that begins with sensing of variables and ends with a magnetic recording of raw data, may include a complete telemetering link" (McGraw-Hill, *Dictionary of Scientific and Technical Terms,* Second Edition, 1978). Here, the term *variables* refers to those physical quantities that are associated with a natural or artificial process. A data acquisition phase involves a real-time computing environment where the computer must be keyed to the time scale of the process. Figure 84.1 gives a simplified block diagram of a data acquisition system current in the early 1990s.

The path the data travels through the system is called the data acquisition channel. Data are first captured and subsequently translated into usable signals using transducers. In this discussion, usable signals are assumed to be electrical voltages, either unipolar (that is, single ended, with a common ground so that we need just one lead wire to carry the signal) or bipolar (that is, common mode, with the signal carried by a wire pair, so that the reference of the rest of the system is not part of the output). These voltages can be either analog or digital, depending on the nature of the measurand (the quantity being captured). When there is more than one analog input, they are subsequently sent to an analog **multiplexer** (MUX). Both the analog and the digital signals are then conditioned using signal conditioners. There are two additional steps for those conditioned analog signals. First they must be sampled (see Chapter 73.4) and next converted to digital data. This conversion is done by **analog-to-digital converters** (ADC) (see Chapter 32).

Once the analog-to-digital conversion is done, the rest of the steps have to deal with digital data only. The calendar/clock block shown in Fig 84.1 is used to add the time-of-date information, an important parameter of a real-time processing environment, into the half-processed data. The digital processor performs the overall system control tasks using a software program, which is usually called system software. These control tasks also include display, printer, data recorder, and communication interface management. A well-regulated **power supply unit** (PSU) and a stable clock are essential components in many data acquisition systems. There are systems where massive amounts of data points are produced within a very short period of time, and they are equipped with *on-board memory* so that a considerable amount of data points can be stored locally. Data are transmitted to the host computer once the local storage has reached its full capacity. Historically, data acquisition evolved in modular form, until monolithic silicon came along and reduced the size of the modules.

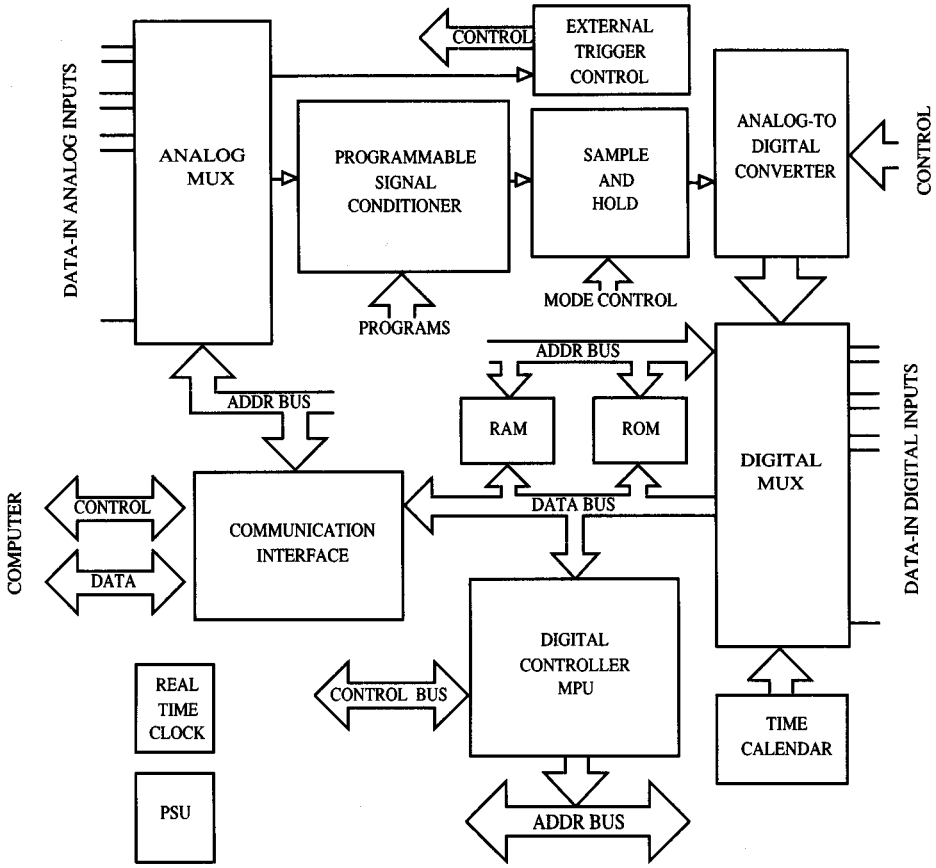


FIGURE 84.1 The block diagram of a data acquisition system.

The analysis and design of data acquisition systems is a discipline that has roots in the following subject areas: signal theory, transducers, analog signal processing, noise, sampling theory, quantizing and encoding theory, analog-to-digital conversion theory, analog and digital electronics, data communication, and systems engineering. Cost, accuracy, bit resolution, speed of operation, on-board memory, power consumption, stability of operation under various operating conditions, number of input channels and their ranges, on-board space, supply voltage requirements, compatibility with existing bus interfaces, and the types of data recording instruments involved are some of the prime factors that must be considered when designing or buying a data acquisition system. Data acquisition systems are involved in a wide range of applications, such as machine control, robot control, medical and analytical instrumentation, vibration analysis, spectral analysis, correlation analysis, transient analysis, digital audio and video, seismic analysis, test equipment, machine monitoring, and environmental monitoring.

84.2 The Analog and Digital Signal Interface

The data acquisition system must be designed to match the process being measured as well as the end-user requirements. The nature of the process is mainly characterized by its speed and number of measuring points, whereas the end-user requirement is mainly the flexibility in control. Certain processes require data acquisition with no interruption where computers are used in controlling. On the other hand, there are cases where the acquisition starts at a certain instance and continues for a definite period. In this case the acquisition cycle is repeated in a periodic manner, and it can be controlled manually or by software. Controllers access the process via the analog and digital interface submodules, which are sometimes called analog and digital front ends.

Many applications require information capturing from more than one channel. The use of the analog MUX in Fig. 84.1 is to cater to multiple analog inputs. A detailed diagram of this input circuitry is shown in Fig. 84.2

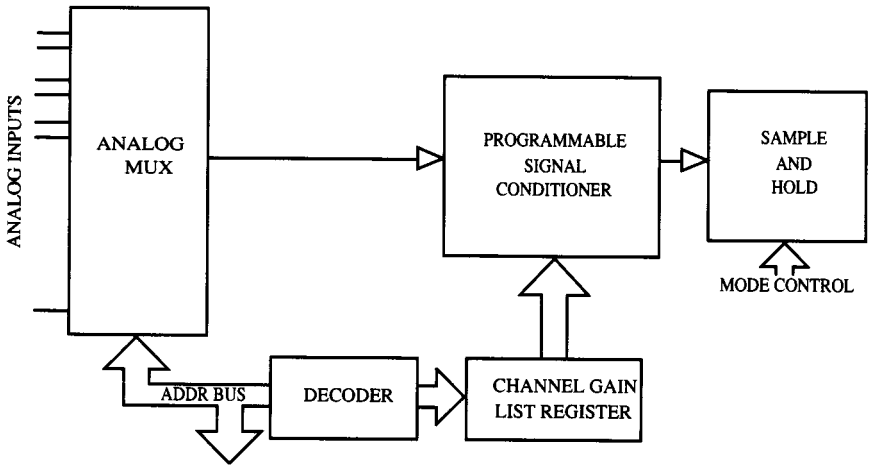


FIGURE 84.2 Analog input circuitry—the analog front end.

and the functional description is as follows. When the MUX is addressed to select an input, say, $x_i(t)$, the same address will be decoded by the decoding logic to generate another address, which is used in addressing the programmable register. The programmable register contains further information regarding how to handle $x_i(t)$. The outcome of the register is then used in subsequent tuning of the signal conditioner. Complex programmable control tasks might include automatic gain selection for each channel, and hence the contents of this register are known as the channel gain list. The MUX address generator could be programmed in many ways, and one simple way is to scan the input channels in a cyclic fashion where the address can be generated by means of a binary counter. Microprocessors are also used in addressing MUXs in applications where complex channel selection tasks are involved. Multiplexers are available in integrated circuit form, though relay MUXs are widely used because they minimize errors due to cross talk and bias currents. Relay MUX modules are usually designed as plugged-in units and can be connected according to the requirements.

There are applications where the data acquisition cycle is triggered by the process itself. In this case an analog or digital trigger signal is sent to the unit by the process, and a separate external trigger interface circuitry is supplied. The internal controller assumes its duties once it has been triggered. It takes a finite time to settle the signal $x_i(t)$ through the MUX up to the signal conditioner once it is addressed. Therefore, it is possible to process $x_{i-1}(t)$ during the selection time of $x_i(t)$ for greater speeds. This function is known as pipelining and will be illustrated in Section 84.3.

In some data acquisition applications the data acquisition module is a plugged-in card in a computer, which is installed far away from the process. In such cases, transducers—the process sensing elements—are connected to the data acquisition module using transmission lines or a radio link. In the latter case a complete demodulating unit is required at the input. When transmission lines are used in the interconnection, care must be taken to minimize electromagnetic interference since transmission lines pick up noise easily. In the case of a single-ended transducer output configuration, a single wire is adequate for the signal transmission, but a common ground must be established between the two ends as given in Fig. 84.3(a). For the transducers that have common mode outputs, a shielded twisted pair of wires will carry the signal. In this case, the shield, the transducer's encasing chassis, and the data acquisition module's reference may be connected to the same ground as shown in Fig. 84.3(c). In high-speed applications the transmission line impedance should be matched with the output impedance of the transducer in order to prevent reflected traveling waves. If the transducer output is not strong enough to transmit for a long distance, then it is best to amplify it before transmission.

Transducers that produce digital outputs may be first connected to Schmitt trigger circuits for pulse shaping purposes, and this can be considered as a form of digital signal conditioning. This becomes an essential requirement when such inputs are connected through long transmission lines where the line capacitance significantly affects the rising and falling edges of the incoming wave. Opto-isolators are sometimes used in coupling when the voltage levels of the two sides of the transducer and the input circuit of the data acquisition unit do not match each other. Special kinds of connectors are designed and widely used in interconnecting

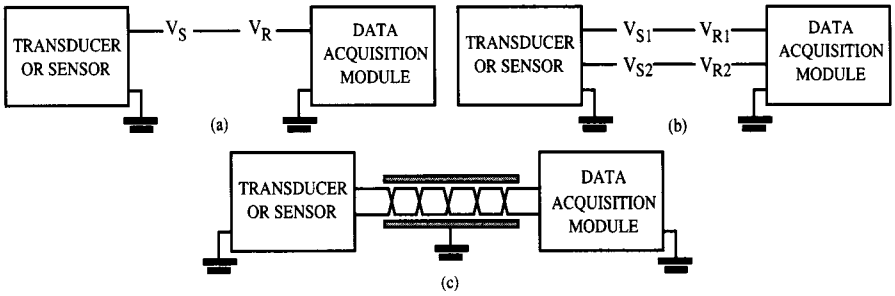


FIGURE 84.3 (a) Connecting transducers to the data acquisition unit, (b) single-ended (unipolar) output, and (c) common-mode (bipolar) output.

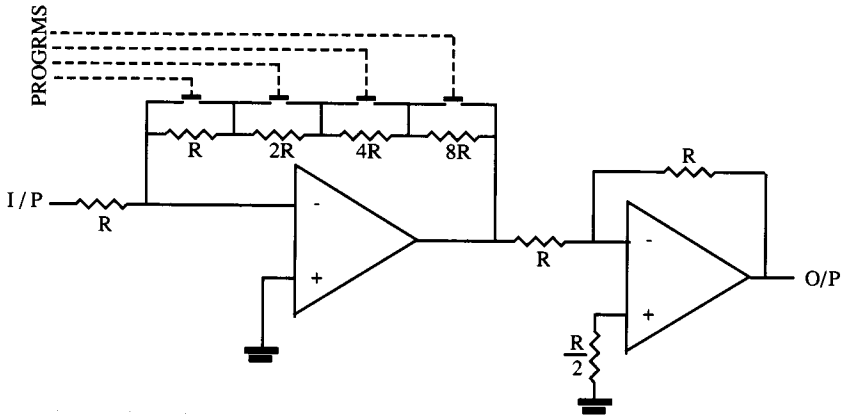


FIGURE 84.4 Programmable gain instrumentation amplifier.

transmission lines and data acquisition equipment in order to screen the signals from noise. Analog and digital signal grounds should be kept separate where possible to prevent digital signals from flowing in the analog ground circuit and including spurious analog signal noise.

84.3 Analog Signal Conditioning

The objective of an analog signal conditioner is to increase the quality of the transducer output to a desired level before analog-to-digital conversion. A signal conditioner mainly consist of a preamplifier, which is either an instrumentation amplifier or an operational amplifier and/or a filter. Coupling more and more circuits to the data acquisition channel has to be done taking great care that these signal conditioning circuits do not add more noise or unstable behavior to the data acquisition channel. General purpose signal conditioner modules are commercially available for applications. Some details were given in the previous section about programmable signal conditioners and the discussion is continued here.

Figure 84.4 shows an instrumentation amplifier with programmable gain where the programs are stored in the channel-gain list. The reason for having such sophistication is to match transducer outputs with the maximum allowable input range of the ADC. This is very important in improving accuracy in cases where transducer output voltage ranges are much smaller than the full-scale input range of an ADC, as is usually the case. Indeed, this is equally true for signals that are larger than the full-scale range, and in such cases the amplifier functions as an attenuator. Furthermore, the instrumentation amplifier converts a bipolar voltage signal into a unipolar voltage with respect to the system ground. This action will reduce a major control task as far as the ADC is concerned; that is, the ADC is always sent unipolar voltages, and hence it is possible to maintain unchanged the mode control input which toggles the ADC between the unipolar and bipolar modes of an ADC.

Values of the **signal-to-noise ratio**

$$\text{SNR} = \left[\frac{\text{RMS signal}}{\text{RMS noise}}\right]^2 \tag{84.1}$$

at the input and the output of the instrumentation amplifier are related to its **common-mode rejection ratio** (CMRR) given by

$$\text{CMRR} = \sqrt{\frac{\text{SNR}_{\text{output}}}{\text{SNR}_{\text{input}}}} \tag{84.2}$$

Hence, higher values of $\text{SNR}_{\text{output}}$ indicate low noise power. Therefore, instrumentation amplifiers are designed to have very high CMRR figures. The existence of noise will result in an error in the ADC output. The allowable error is normally expressed as a fraction of the **least significant bit** (LSB) of the code such as $\pm(1/X)$LSB. The amount of error voltage (V_{error}) corresponding to this figure can be found considering the bit resolution (N) and the ADC's maximum analog input voltage (V_{max}) as given in

$$V_{\text{error}} = \pm\left[\frac{V_{\text{max}}}{2^N - 1} \times \frac{1}{X}\right] \text{ volts} \tag{84.3}$$

Other specifications of amplifiers include the temperature dependence of the input offset voltage (V_{offset}, μV/°C) and the current (I_{offset}, pA/°C) associated with the operational amplifiers in use. High slew rate (V/μs) amplifiers are recommended in high-speed applications. Generally, the higher the bandwidth, the better the performance.

Cascading a filter with the preamplifier will result in better performance by eliminating noise. Active filters are commonly used because of their compact design, but passive filters are still in use. The cut-off frequency, f_c, is one of the important performance indices of a filter that has to be designed to match the channel's requirements. The value f_c is a function of the preamplifier bandwidth, its output SNR, and the output SNR of the filter.

84.4 Sample-and-Hold and A/D Techniques in Data Acquisition

Sample-and-hold systems are primarily used to maintain a constant magnitude representing the input, across the input of the ADC throughout a precisely known period of time. Such systems are called **sample-and-hold amplifiers** (SHA), and their characteristics are crucial to the overall system accuracy and reliability of digital data. The SHA is not an essential item in applications where the analog input does not vary more than $\pm(1/2)$LSB of voltage. As the name indicates, a SHA operates in two different modes, which are digitally controlled. In the sampling mode it acts as an input voltage follower, where, once it is triggered into its hold mode, it should ideally retain the signal voltage level at the time of the trigger. When it is brought back into the sampling mode, it instantly assumes the voltage level at the input.

Figure 84.5 shows the simplified circuit diagram of a monolithic sampling-and-hold circuit and the associated switching waveforms. The differential amplifiers function as input and output buffers, and the capacitor acts as the storage mechanism. When the mode control switch is at its *on* position, the two buffers are connected in series and the capacitor follows the input with minimum time delay, if it is small. Now, if the mode control is switched *off*, the feedback loop is interrupted, and the capacitor ideally retains its terminal voltage until the next sampling signal occurs. Leakage and bias currents usually cause the capacitor to discharge and or charge in the hold mode and the fluctuation of the hold voltage is called *droop*, which could be minimized by having a

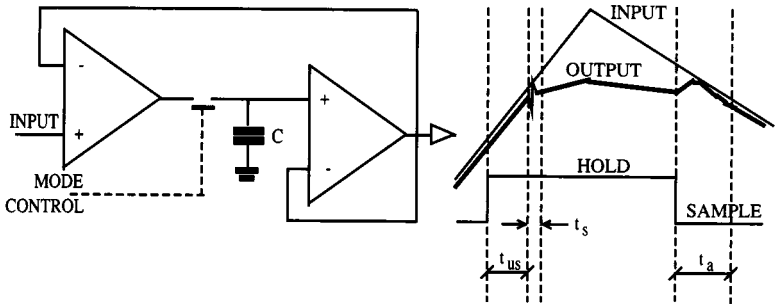


FIGURE 84.5 Sample-and-hold circuit diagram and switching waveforms.

large capacitor. Therefore, the capacitance has to be selected such that the circuit performs well in both modes. Several time intervals are defined relative to the switching waveform of SHAs. The *acquisition time* (t_a) is the time taken by the device to reach its final value after the sample command has been given. The *setting time* (t_s) is the time taken to settle the output. The *aperture uncertainty* or *aperture jitter* (t_{us}) is the range of variation of the aperture time. It is important to note here that the sampling techniques have a well-formulated theoretical background.

ADCs perform a key function in the data acquisition process. The application of various ADC technologies in a data acquisition system depends mainly on the cost, bit resolution, and speed. Successive approximation types are more common at high resolution at moderate speeds (<1 MHz). This kind of ADC offers the best trade-offs among bit resolution, accuracy, speed, and cost. Flash converters, on the other hand, are best suited for high-speed applications. Integrating-type converters are suitable for high-resolution and -accuracy applications.

Many techniques have been developed in coupling sample-hold circuits and ADCs in data acquisition systems because no single ADC or sampling technology is able to satisfy the ever increasing requirements of data acquisition applications. Figure 84.6 illustrates the various sampling and ADC configurations used in practice. It can be seen that the sampling frequencies are increased because of pipelining, parallelism, or concurrent architecture. The increase in the sampling frequency improves the bandwidth, improving in turn the SNR in the channel.

84.5 The Communication Interface of a Data Acquisition System

The communication interface is the module through which the acquired data are sent as well as other control tasks are established between the data acquisition module and the host computer (Fig. 84.1). There are basically two different ways of establishing a data link between the two. One way is to use interrupts and the other is through **direct memory access** (DMA). In the case of an interrupt-driven mode, an interrupt-request signal is sent to the computer. Upon receiving it, the computer will first finish the execution of the current instruction, suspend the next, and then send an interrupt-acknowledge signal asking the module to send data. The operation is asynchronous since the sender sends data when it wants to do so. Getting the computer ready to receive data is known as handshaking. In the case of a DMA transfer, the DMA controller is given the starting address of the memory location where the data have to be written. The DMA controller asks the computer to freeze its operations until it has finished writing data directly into the memory. The operation does not need any waiting time and therefore it is fast.

Data acquisition systems are usually designed to couple with existing computer systems, and many computer systems provide standard bus architecture, allowing users to connect various peripherals that are compatible with its bus. Data acquisition systems are computer peripherals that follow the above description. Since ADCs produce parallel data, many data acquisition systems provide outputs compatible with parallel instrument buses such as the IEEE-488 (HP-IB or GPIB) or the VMEbus. Personal computer-based data acquisition boards must have communication interfaces compatible with the computer bus in order to share resources. The RS-232 standard communication interfaces are widely used in serial data transfer. Communication interfaces for data acquisition systems are normally designed to satisfy the electrical, mechanical, and protocol standards of the interface bus. Electrical standards include power supply requirements, methods of supply, the data transfer rate (baud rate), the width of the address, and the line terminating impedance. Mechanical requirements are the

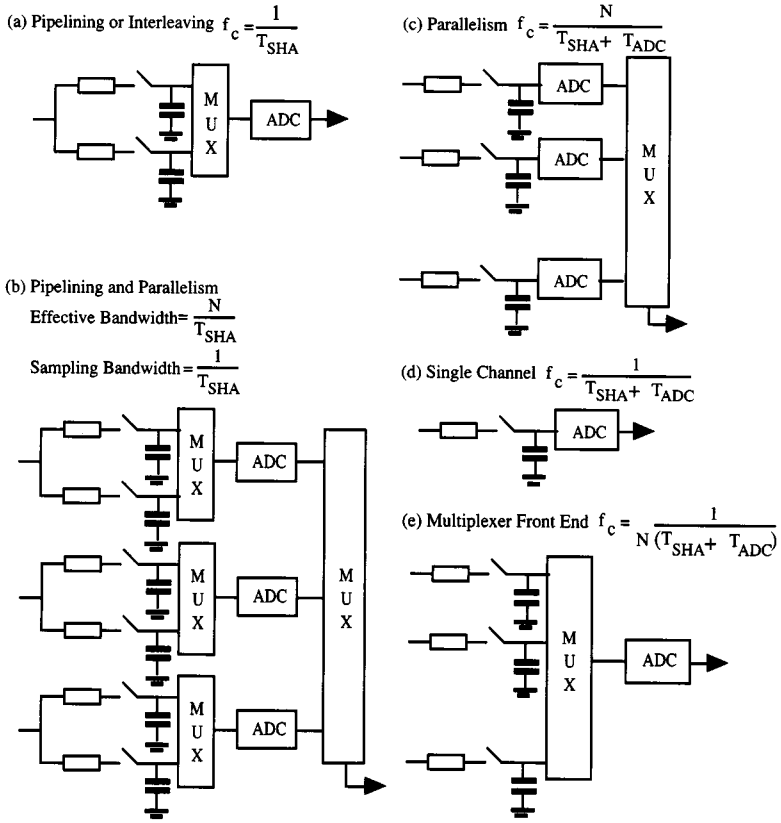


FIGURE 84.6 Coupling techniques for SHA and ADC systems.

type, size, and the pin assignments of the connectors. The data transfer protocol determines the procedure of data transfer between the two systems. A definition of the timing and input–output philosophy—whether the transfer is in synchronous, asynchronous, or quasi-synchronous mode and how errors are detected and handled—are important factors to be considered.

84.6 Data Recording

It is important to provide storage media to cater to large streams of data being produced. Data acquisition systems use graph paper, paper tapes, magnetic tapes, magnetic floppy disks, hard disks, or any combination of these as their data recorders. Paper and magnetic tape storage schemes are known as sequential access storage, whereas disk storage is called direct access storage. Tapes are cost-effective media compared to disk drives and are still in wide use. In many laboratory situations it will be much more cost effective to network a number of systems to a single, high-capacity hard drive, which acts as a file server. This adoption of digital recording provides the ultimate in signal-to-noise ratio, accuracy of signal waveform and freedom from tape transfer flutter. Data storage capacity, access time, transfer rate, and error rate are some of the performance indices that are associated with these devices.

84.7 Software Aspects

So far the discussion has been mainly on the hardware side of the data acquisition system. The other most important part is the software system associated with a data acquisition system, which can generally be divided into two—the system software and the user-interface program. Both must be designed properly in order to achieve the maximum use of the system. The system software is mainly written in assembly language with many

lines of code, whereas the user interface is built using a high-level software development tool. One main part of system software is written to handle the input–output (I/O) operations. The use of assembly language results in the fast execution of I/O commands. The I/O software has to deal with how the basic input–output programming tasks such as interrupt and DMA handling are done. The other aspects of system software are to perform the internal control tasks such as providing trigger pulses for the ADC and SHA, addressing the input multiplexer, the accessing and editing of the channel-gain list, transferring data into the on-board memory, and the addition of the clock/calendar information into data. Multitasking software programs are best suited for many data acquisition systems because it may be necessary to read data from the data acquisition module and display and print it at the same time. Menu-driven user interfaces are common and have a variety of functions built into them.

Defining Terms

Analog-to-digital converter (ADC): A device that converts analog input voltage signals into digital form.

Common-mode rejection ratio (CMRR): A measure of quality of an amplifier with differential inputs and defined as the ratio between the common-mode gain and the differential gain.

Direct memory access (DMA): The process of sending data from an external device into the computer memory with no involvement of the computer's central processing unit.

Least significant bit (LSB): The 2^0th bit in a digital word.

Multiplexer (MUX): A combinational logic device with many input channels and usually just one output. The function performed by the device is connecting one and only one input channel at a time to the output. The required input channel is selected by sending the channel address to the MUX.

Power supply unit (PSU): The one that generates the necessary voltage levels required by a system.

Sample-and-hold amplifier (SHA): A unity gain amplifier with a mode control switch where the input of the amplifier is connected to a time-varying voltage signal. A trigger pulse at the mode control switch causes it to read the input at the instance of the trigger and maintain that value until the next trigger pulse.

Signal-to-noise ratio (SNR): The ratio between the signal power and the noise power at a point in the signal traveling path.

Related Topics

32.1 D/A and A/D Circuits • 69.2 Radio • 70.1 Coding • 80.1 Integrated Circuits (RAM, ROM)

References

For further reading consult the following texts, which were used along with the authors' experience and other sources as a basis for this article:

Analog Devices, *Data Conversion Handbook,* Analog Devices, Inc., 1989/90.

R. Annino and R. Driver, *Scientific and Engineering Applications with Personal Computers*, New York: Wiley Interscience, 1986.

D. L. Feucht, *Handbook of Analog Circuit Design*, San Diego: Academic Press, 1990.

D. G. Fink and D. Christiansen (eds.), *Electronics Engineers' Handbook*, 3rd ed., New York: McGraw-Hill, 1989.

P. M. Garrett, *Analog Systems for Microprocessor and Minicomputers*, Reston, Va.: Reston Publishing Company, 1978.

P. Holloway, "Technology focus interview," *Electronic Engineering*, December 1990.

F. Jorgensen, *The Complete Handbook of Magnetic Recording*, 4th ed., Blue Ridge Summit, Penn.: Tab Books, 1995.

F. F. Mazda, *Electronic Instruments and Measurement Techniques*, New York: Cambridge University Press, 1987.

D. A. Mellichamp (ed.), *Real-Time Computing With Applications to Data Acquisition and Control*, New York: Van Nostrand Reinhold, 1983.

M. Tatkow and J. Turner, "New techniques for high-speed data acquisition," *Electronic Engineering*, September 1990.

Further Information

To probe further in the subject area, refer to the *Data Acquisition Handbook,* published by Data Translation, Marlboro, Mass., 1990, and the *Data Acquisition Handbook,* published by Rector Press, 1995.

Serra, M., Dervisoglu, B.I. "Testing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

85

Testing

Micaela Serra
University of Victoria

Bulent I. Dervisoglu
Hewlett-Packard Company

85.1 Digital IC Testing
Taxonomy of Testing • Fault Models • Test Pattern Generation • Output Response Analysis
85.2 Design for Test
The Testability Problem • Design for Testability • Future for Design for Test

85.1 Digital IC Testing

Micaela Serra

In this section we give an overview of digital testing techniques with appropriate reference to material containing all details of the methodology and algorithms. First, we present a general introduction of terminology and a taxonomy of testing methods. Next, we present a definition of fault models, and finally we discuss the main approaches for test pattern generation and data compaction, respectively.

Taxonomy of Testing

The evaluation of the reliability and quality of a digital IC is commonly called *testing*, yet it comprises distinct phases that are mostly kept separate both in the research community and in industrial practice.

1. *Verification* is the initial phase in which the first prototype chips are "tested" to ensure that they match their functional specification, that is, to verify the correctness of the design. Verification checks that all design rules are adhered to, from layout to electrical parameters; more generally, this type of functional testing checks that the circuit: (a) implements what it is supposed to do *and* (b) does not do what it is not supposed to do. Both conditions are necessary. This type of evaluation is done at the design stage and uses a variety of techniques, including logic verification with the use of hardware description languages, full functional simulation, and generation of functional test vectors. We do not discuss verification techniques here.
2. *Testing* correctly refers to the phase when one must ensure that only defect-free production chips are packaged and shipped and detect faults arising from manufacturing and/or wear-out. Testing methods must (a) be fast enough to be applied to large amounts of chips during production, (b) take into consideration whether the industry concerned has access to large expensive external tester machines, and (c) consider whether the implementation of **built-in self-test (BIST)** proves to be advantageous. In BIST, the circuit is designed to include its own self-testing extra circuitry and thus can signal directly, during testing, its possible failure status. Of course, this involves a certain amount of overhead in area, and trade-offs must be considered. The development of appropriate testing algorithms and their tool support can require a large amount of engineering effort, but one must note that it may need to be done only once per design. The speed of application of the algorithm (applied to many copies of the chips) can be of more importance.

3. *Parametric testing* is done to ensure that components meet design specification for delays, voltages, power, etc. Lately much attention has been given to **I_{DDq} testing,** a parametric technique for CMOS testing. I_{DDq} testing monitors the current I_{DD} that a circuit draws when it is in a quiescent state. It is used to detect faults such as bridging faults, transistor stuck-open faults, or gate oxide leaks, which increase the normally low I_{DD} [Jacomino et al., 1989].

The density of circuitry continues to increase, while the number of I/O pins remains small. This causes a serious escalation of complexity, and testing is becoming one of the major costs to industry (estimated up to 30%). ICs should be tested before and after packaging, after mounting on a board, and periodically during operation. Different methods may be necessary for each case. Thus by testing we imply the means by which some qualities or attributes are determined to be fault-free or faulty. The main purpose of testing is the detection of malfunctions (Go/NoGo test), and only subsequently one may be interested in the actual location of the malfunction; this is called *fault diagnosis* or *fault location.*

Most testing techniques are designed to be applied to combinational circuits only. While this may appear a strong restriction, in practice it is a realistic assumption based on the idea of designing a sequential circuit by partitioning the memory elements from the control functionality such that the circuit can be reconfigured as combinational at testing time. This general approach is one of the methods in *design for testability* (DFT) (see Section 85.2). DFT encompasses any design strategy aimed at enhancing the testability of a circuit. In particular, scan design is the best-known implementation for separating the latches from the combinational gates such that some of the latches can also be reconfigured and used as either tester units or as input generator units (essential for built-in testing).

Figure 85.1(a) shows the general division for algorithms in testing. *Test pattern generation* implies a fair amount of work in generating an appropriate subset of all input combinations, such that a desired percentage of faults is activated and observed at the outputs. *Output response analysis* encompasses methods which capture only the output stream, with appropriate transformations, with the assumption that the circuit is stimulated by either an exhaustive or a random set of input combinations. Both methodologies are introduced below.

Moreover a further division can be seen between *on-line* and *off-line* methods [see Fig. 85.1(b)]. In the former, each output word from the circuit is tested during normal operation. In the latter, the circuit must suspend normal operation and enter a "test mode," at which time the appropriate method of testing is applied. While **off-line testing** can be executed either through external testing (a tester machine external to the circuitry) or through the use of BIST, **on-line testing** (also called *concurrent checking*) usually implies that the circuit contains some coding scheme which has been previously embedded in the design of the circuitry.

If many defects are present during the manufacturing process, the manufacturing yield is lowered, and testing becomes of paramount importance. Some estimation can be given about the relationship between manufacturing yield, effectiveness of testing and defect level remaining after test [Williams, 1986]. Let Y denote the yield, where Y is some value between 1 (100% defect-free production) and 0 (all circuits faulty after testing).

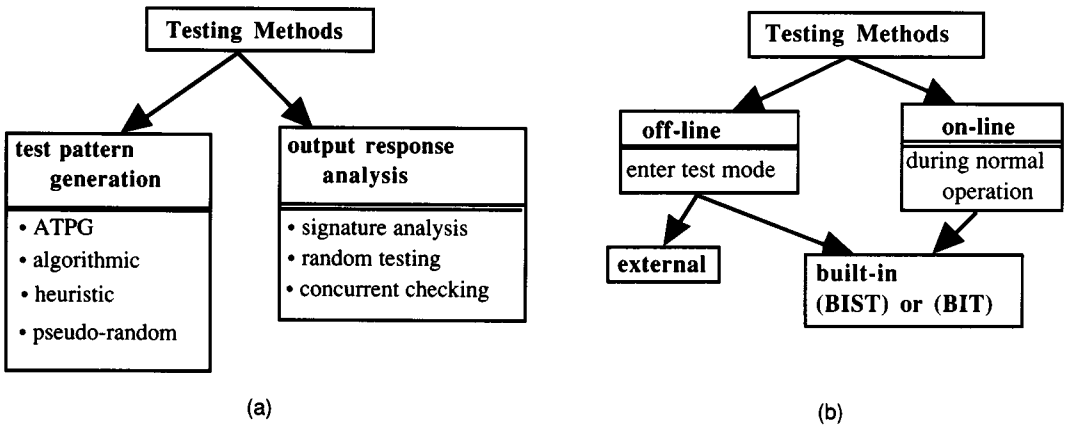


FIGURE 85.1 Taxonomy of testing methods. (a) Test pattern generation; (b) on-line and off-line methods.

Let *FC* be the **fault coverage**, calculated as the percentage of detected faults over the total number of detectable modeled faults (see below for fault models). The value of *FC* ranges from 1 (all possible faults detected) to 0 (no testing done). We are interested in the final defect level (*DL*), after test, defined as the probability of shipping a defective product. It has been shown that tests with high fault coverage (for certain fault models, see below) also have high defect coverage. The empirical equation is

$$DL = (1 - Y^{1-FC})\ 100\%$$

Plotting this equation gives interesting and practical results. Table 85.1 shows only a few examples of some practical values of *Y* and *FC*. The main conclusion to be drawn is that a very high fault coverage must be achieved to obtain any acceptable defect level value, and manufacturing yield must be continually improved to maintain reliability of shipped products.

TABLE 85.1 Examples of Defect Levels

Y	*FC*	*DL*
0.15	0.90	0.18
0.25	0.00	0.75
0.25	0.90	0.15

Fault Models

At the defect level, an enormous number of different failures could be present, and it is totally infeasible to analyze them as such. Thus failures are grouped together with regards to their logical fault effect on the functionality of the circuit, and this leads to the construction of logical fault models as the basis for testing algorithms [Abramovici et al., 1992]. More precisely, a *fault* denotes the physical failure mechanism, the *fault effect* denotes the logical effect of a fault on a signal-carrying net, and an *error* is defined as the condition (or state) of a system containing a fault (deviation from correct state). Faults can be further divided into classes, as shown in Fig. 85.2. Here we discuss only *permanent* faults, that is, faults in existence long enough to be observed at test time, as opposed to *temporary* faults (transient or intermittent), which appear and disappear in short intervals of time, or *delay* faults, which affect the operating speed of the circuit. Moreover we do not discuss **sequential faults**, which cause a combinational circuit to behave like a sequential one, as they are mainly restricted to certain technologies (e.g., CMOS).

The most commonly used fault model is that of a **stuck-at fault**, which is modeled by having a line segment stuck at logic 0 or 1 (stuck-at 1 or stuck-at 0). One may consider single or multiple stuck-at faults and Fig. 85.3 shows an example for a simple circuit. The fault-free function is shown as *F*, while the faulty functions, under

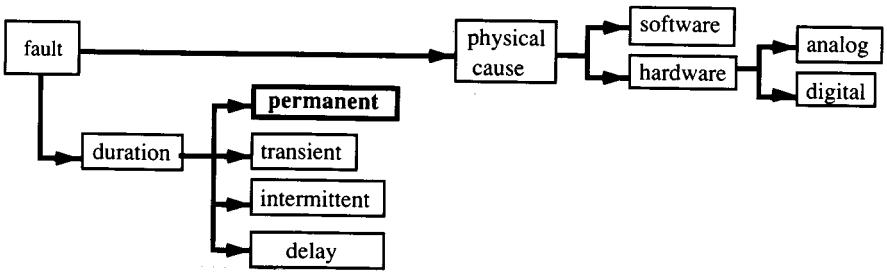


FIGURE 85.2 Fault characteristics.

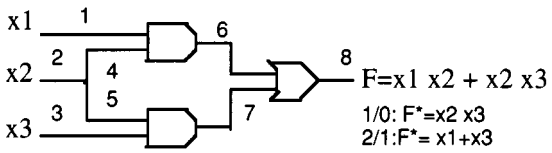


FIGURE 85.3 Single stuck-at fault example.

a	b	z	a/0	a/1	b/0	b/1	z/0	z/1
0	0	1	1	1	1	1	[0]	1
0	1	1	1	[0]	1	1	[0]	1
1	0	1	1	1	1	[0]	[0]	1
1	1	0	[1]	0	[1]	0	0	[1]
possible tests			{11}	{01}	{11}	{10}	{00 01 10}	{11}

=> Minimal Test Set T: {(01), (10), (11) }

FIGURE 85.4 Test set example.

the occurrence of the single stuck-at faults of either line 1 stuck-at 0 (1/0) or of line 2 stuck-at 1 (2/1), are shown as F^*.

Bridging faults occur when two or more lines are shorted together. There are two main problems in the analysis of bridging faults: (1) the theoretical number of possible such faults is extremely high and (2) the operational effect is of a wired logic AND or OR, depending on technology, and it can even have different effects in complex CMOS gates.

CMOS stuck-open faults have been examined recently, as they cannot be modeled from the more classical fault models and are restricted to the CMOS technology. They occur when the path through one of the *p*-channel or one of the *n*-channel transistors becomes an open circuit. The main difficulty in detecting this type of fault is that it changes the combinational behavior of a cell into a sequential one. Thus the logical effect is to retain, on a given line, the previous value, introducing a memory state. To detect such a fault, one must apply two stimuli: the first to set a line at a certain value and the second to try and change that value. This, of course, increases the complexity of fault detection.

Test Pattern Generation

Test pattern generation is the process of generating a (minimal) set of input patterns to stimulate the inputs of a circuit such that detectable faults can be exercised (if present) [Abramovici et al., 1992]. The process can be divided in two distinct phases: (1) derivation of a test and (2) application of a test. For (1), one must first select appropriate models for the circuit (gate or transistor level) and for faults; one must construct the test such that the output signal from a faulty circuit is different from that of a good circuit. This can be computationally very expensive, but one must remember that the process is done only once at the end of the design stage. The generation of a test set can be obtained either by manual methods, by algorithmic methods (with or without heuristics), or by pseudo-random methods. On the other hand, for (2), a test is subsequently applied many times to each IC and thus must be efficient both in space (storage requirements for the patterns) and in time. Often such a set is not minimal, as near minimality may be sufficient. The main considerations in evaluating a test set are the time to construct a minimal test set; the size of the test pattern generator, i.e., the software or hardware module used to stimulate the circuit under test; the size of the test set itself; the time to load the test patterns; and the equipment required (if external) or the BIST overhead.

Most algorithmic test pattern generators are based on the concept of sensitized paths. Given a line in a circuit, one wants to find a *sensitized* path to take a possible error all the way to an observable output. For example, to sensitize a path that goes through one input of an AND gate, one must set all other inputs of the gate to logic 1 to permit the sensitized signal to carry through. Figure 85.4 summarizes the underlying principles of trying to construct a test set. Each column shows the expected output for each input combination of a NAND gate. Columns 3 to 8 show the output under the presence of a stuck-at fault as per label. The output bits that permit detection of the corresponding fault are shown in a square, and thus at the bottom the minimal test set is listed, comprising the minimal number of distinct patterns necessary to detect all single stuck-at faults.

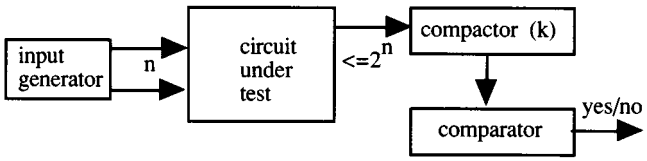


FIGURE 85.5 Data compaction testing.

The best-known algorithms are the D-algorithm (precursor to all), PODEM, and FAN [Abramovici, 1992]. Three steps can be identified in most automatic test pattern generation (ATPG) programs: (1) listing the signals on the inputs of a gate controlling the line on which a fault should be detected, (2) determining the primary input conditions necessary to obtain these signals (back propagation) and sensitizing the path to the primary outputs such that the signals and fault can be observed, and (3) repeating this procedure until all detectable faults in a given fault set have been covered. PODEM and FAN introduce powerful heuristics to speed the three steps by aiding in the sequential selection of faults to be examined and by cutting the amount of back and forward propagation necessary.

Notwithstanding heuristics, algorithmic test pattern generation is very computationally expensive and can encounter numerous difficulties, especially in certain types of networks. Newer alternatives are based on **pseudo-random pattern generation** [Bardell et al., 1987] and **fault simulation**. In this strategy, a large set of patterns is generated pseudo-randomly with the aid of an inexpensive (hardware or software) generator. Typical choices for these are linear feedback shift registers and linear cellular automata registers (see below). The pseudo-random set is used to stimulate a circuit, and, using a fault simulator, one can evaluate the number of faults that are covered by this set. An algorithmic test pattern generator is then applied to find coverage for the remaining faults (hopefully, a small number), and the pseudo-random set is thus augmented. The disadvantages are that the resulting set is very large and fault simulation is also computationally expensive. However, this method presents an alternative for circuits where the application of deterministic algorithms for all faults is infeasible.

Output Response Analysis

Especially when designing a circuit including some BIST, one must decide how to check the correctness of the circuit's responses [Bardell et al., 1987]. It is infeasible to store on-chip all expected responses, and thus a common solution is to reduce the circuit responses to relatively short sequences: this process is called *data compaction* and the short, compacted resulting sequence is called a *signature*. The normal configuration for data compaction testing is shown in Fig. 85.5. The n-input circuit is stimulated by an input pattern generator (pseudo-random or exhaustive if $n < 20$); the resulting output vector(s), of length up to $2''$, is compacted to a very short signature of length $k << 2''$ (usually k is around 16 to 32 bits). The signature is then compared to a known good value. The main advantages of this method are that (1) the testing can be done at circuit speed, (2) there is no need to generate test patterns, and (3) the testing circuitry involves a very small area, especially if the circuit has been designed using scan techniques (see Section 85.2). The issues revolve around designing very efficient input generators and compactors.

The main disadvantage of this method is the possibility of **aliasing**. When the short signature is formed, a loss of information occurs, and it can be the case that a faulty circuit produces the same signature of a fault-free circuit, thus remaining undetected. The design method for data compaction aims at minimizing the probability of aliasing. Using the compactors explained below, the probability of aliasing has been theoretically proven to be 2^{-k}, where k is the length of the compactor (and thus the length of the signature). It is important to note that (1) the result is asymptotically independent of the size and complexity of the circuit under test; (2) for $k = 16$, the probability of aliasing is only about 10^{-6} and thus quite acceptable; and (3) the empirical results show that in practice this method is even more effective. Most of all, this is the chosen methodology when BIST is required for its effectiveness, speed, and small area overhead.

A secondary issue in data compaction is in the determination of the expected "good" signature. The best way is to use fault-free simulation for both the circuit and the compactor, and then the appropriate comparator can be built as part of the testing circuitry [Bardell et al., 1987; Abramovici, 1992].

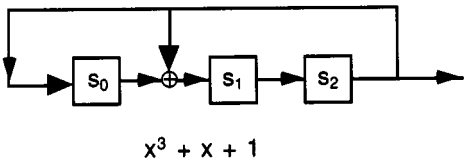


FIGURE 85.6 Autonomous LFSR.

The most important issue is in the choice of a compactor. Although no "perfect" compactor can be found, several have been shown to be very effective. Several compaction techniques have been researched: *counting techniques*, as in one's count, syndrome testing, transition count, and Walsh spectra coefficients; and *signature analysis techniques* based on linear feedback shift registers (**LFSRs**) or linear cellular automata registers (LCARs). Only these latter ones are discussed here. LFSRs and LCARs are also the preferred implementation for the input pattern generators.

LFSRs as Pseudo-Random Pattern Generators

An autonomous LFSR is a clocked synchronous shift register augmented with appropriate feedback taps and receiving no external input [Bardell et al., 1987; Abramovici, 1992]. It is an example of a general linear finite state machine, where the memory cells are simple *D* flip-flops and the next state operations are implemented by EXOR gates only. Figure 85.6 shows an example of an autonomous LFSR of length $k = 3$. An LFSR of length k can be described by a polynomial with binary coefficients of degree k, where the nonzero coefficients of the polynomial denote the positions of the respective feedback taps. In Fig. 85.6, the high-order coefficient for x^3 is 1, and thus there is a feedback tap from the rightmost cell s_2; the coefficient for x^2 is 0, and thus no feedback tap exists after cell s_1; however, taps are present from cell s_0 and to the leftmost stage since x and x^0 have nonzero coefficients. Since this is an autonomous LFSR, there is no external input to the leftmost cell.

The state of the LFSR is denoted by the binary state of its cells. In Fig. 85.6, the next state of each cell is determined by the implementation given by its polynomial and can be summarized as follows: $s_0^+ = s_2$, $s_1^+ = s_0 \oplus s_2$, $s_2^+ = s_1$, where the s_i^+ denotes the next state of cell s_i at each clock cycle. If the LFSR is initialized in a nonzero state, it cycles through a sequence of states and eventually comes back to the initial state, following the functionality of the next-state rules implemented by its polynomial description. An LFSR that goes through all possible $2^k - 1$ nonzero states is said to be described by a *primitive* polynomial (see theory of Galois fields for the definition of primitive), and such polynomials can be found from tables [Bardell et al., 1987].

By connecting the output of each cell to an input of a circuit under test, the LFSR implements an ideal input generator, as it is inexpensive in its implementation and it provides the stimuli in pseudo-random order for either exhaustive or pseudo-exhaustive testing.

LFSRs as Signature Analyzer

If the leftmost cell of an LFSR is connected to an external input, as shown in Fig. 85.7, the LFSR can be used as a compactor [Bardell et al., 1987; Abramovici, 1992]. In general, the underlying operation of the LFSR is to compute polynomial division over a finite field, and the theoretical analysis of the effectiveness of **signature analysis** is based on this functionality. The polynomial describing the LFSR implementation is seen to be the divisor polynomial. The binary input stream can be seen to represent the coefficients (high order first) of a dividend polynomial. For example, if the input stream is 1001011 (bits are input left to right in time), the dividend polynomial is $x^6 + x^3 + x + 1$. After seven clock cycles for all the input bits to have entered the LFSR, the binary output stream exiting from the right denotes the quotient polynomial, while the last state of the cells in the LFSR denotes the remainder polynomial.

In the process of computing a signature for testing the circuit, the input stream to the LFSR used as a compactor is the output stream from the circuit under test. At the end of the testing cycles, only the last state of the LFSR is examined and considered to be the compacted signature of the circuit. In most real cases, circuits have many outputs, and the LFSR is converted into a multiple-input shift register (**MISR**). A MISR is constructed by adding EXOR gates to the input of some or all the flip-flop cells; the outputs of the circuit are then fed through these gates into the compactor. The probability of aliasing for a MISR is the same as that of an LFSR;

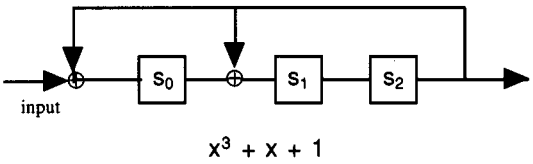


FIGURE 85.7 LFSR for signature analysis.

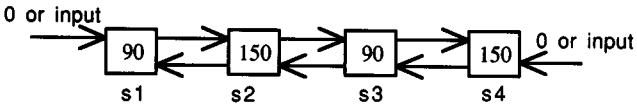


FIGURE 85.8 LCAR for signature analysis.

however, some errors are missed due to cancellation. This is the case when an error in one output at time t is canceled by the EXOR operation with the error in another output at time $t+1$. Given an equally likely probability of errors occurring, the probability of error cancellation has been shown to be 2^{1-m-N}, where m is the number of outputs compacted and N is the length of the output streams.

Given that the normal length of signatures used varies between $k = 16$ and $k = 32$, the probability of aliasing is minimal and considered acceptable in practice. In MISR, the length of the compactor also depends on the number of outputs tested. If the number of outputs is greater than the length of the MISR, algorithms or heuristics exist for combining outputs with EXOR trees before feeding them to the compactor. If the number of outputs is much smaller, various choices can be evaluated. The amount of aliasing that actually occurs in a particular circuit can be computed by full fault simulation, that is, by injecting each possible fault into a simulated circuit and computing the resulting signature. Changes in aliasing can be achieved by changing the polynomial used to define the compactor. It has been shown that primitive polynomials, essential for the generation of exhaustive input generators (see above), also possess better aliasing characteristics.

Data Compaction with Linear Cellular Automata Registers

LCARs are one-dimensional arrays composed of two types of cells: rule 150 and rule 90 cells [Cattell et al., 1996]. Each cell is composed of a flip-flop that saves the current state of the cell and an EXOR gate used to compute the next state of the cell. A rule 150 cell computes its next state as the EXOR of its present state and of the states of its two (left and right) neighbors. A rule 90 cell computes its next state as the EXOR of the states of its two neighbors only. As can be seen in Fig. 85.8, all connections in an LCAR are near-neighbor connections, thus saving routing area and delays (common for long LFSRs).

Up to two inputs can be trivially connected to an LCAR, or it can be easily converted to accept multiple inputs fed through the cell rules. There are some advantages of using LCARs instead of LFSRs: first, the localization of all connections, and second, and most importantly, it has been shown that LCARs are much "better" pseudo-random pattern generators when used in autonomous mode, as they do not show the correlation of bits due to the shifting of the LFSRs. Finally, the better pattern distribution provided by LCARs as input stimuli has been shown to provide better detection for delay faults and open faults, normally very difficult to test.

As for LFSRs, LCARs are fully described by a characteristic polynomial, and through it any linear finite state machine can be built either as an LFSR or as an LCAR. It is, however, more difficult, given a polynomial, to derive the corresponding LCAR, and tables are now used. The main disadvantage of LCARs is in the area overhead incurred by the extra EXOR gates necessary for the implementation of the cell rules. This is offset by their better performance. The corresponding multiple-output compactor is called a MICA.

Summary

Accessibility to internal dense circuitry is becoming a greater problem, and thus it is essential that a designer consider how the IC will be tested and incorporate structures in the design. Formal DFT techniques are

concerned with providing access points for testing (see *controllability* and *observability* in Section 85.2). As test pattern generation becomes even more prohibitive, probabilistic solutions based on compaction and using fault simulation are more widespread, especially if they are supported by DFT techniques and they can avoid the major expense of dedicated external testers. However, any technique chosen must be incorporated within the framework of a powerful CAD system providing semiautomatic analysis and feedback, such that the *rule of ten* can be kept under control: if one does not find a failure at a particular stage, then detection at the next stage will cost 10 times as much!

Defining Terms

Aliasing: Whenever the faulty output produces the same signature as a fault-free output.
Built-in self-test (BIST): The inclusion of on-chip circuitry to provide testing.
Fault coverage: The fraction of possible failures that the test technique can detect.
Fault simulation: An empirical method used to determine how faults affect the operation of the circuit and also how much testing is required to obtain the desired fault coverage.
***I_{DDq}* testing:** *A parametric technique to monitor the current* I_{DD} that a circuit draws when it is in a quiescent state. It is used to detect faults which increase the normally low I_{DD}.
LFSR: A shift register formed by *D* flip-flops and EXOR gates, chained together, with a synchronous clock, used either as input pattern generator or as signature analyzer.
MISR: Multiple-input LFSR.
Off-line testing: Testing process carried out while the tested circuit is not in use.
On-line testing: Concurrent testing to detect errors while circuit is in operation.
Pseudo-random pattern generator: Generates a binary sequence of patterns where the bits appear to be random in the local sense (1 and 0 are equally likely), but they are repeatable (hence only pseudo-random).
Random testing: The process of testing using a set of pseudo-randomly generated patterns.
Sequential fault: A fault that causes a combinational circuit to behave like a sequential one.
Signature analysis: A test where the responses of a device over time are compacted into a characteristic value called a signature, which is then compared to a known good one.
Stuck-at fault: A fault model represented by a signal stuck at a fixed logic value (0 or 1).
Test pattern (test vector): Input vector such that the faulty output is different from the fault-free output (the fault is stimulated and detected).

Related Topic

23.2 Testing

References

M. Abramovici, M.A. Breuer and A.D. Friedman, *Digital Systems Testing and Testable Design,* Rockville, Md.: IEEE Press, 1992.

P.H. Bardell, W.H. McAnney, and J. Savir, *Built-In Test for VLSI: Pseudorandom Techniques,* New York: John Wiley and Sons, 1987.

K. Cattell and J.C. Muzio, "Synthesis of one-dimensional linear hybrid cellular automata," *IEEE Trans. Computer Aided Design,* vol. 15, no. 3, pp. 325–335, 1996.

N.H.E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design,* Addison-Wesley, 1993.

T.W. Williams (Ed.), *VLSI Testing*, Amsterdam: North-Holland, 1986.

Further Information

The author would like to recommend reading the book by Abramovici et al. [1992] that, at the present time, gives the most comprehensive view of testing methods and design for testability. More information on deterministic pattern generation can also be found in *Fault Tolerant Computing,* edited by D.K. Pradhan, and for

newer approaches of random testing the book by Bardell et al. contains basic information. The latest state-of-the-art research is to be found mainly in proceedings of the IEEE International Test Conference.

85.2 Design for Test

Bulent I. Dervisoglu

Testing of electronic circuits, which has long been pursued as an activity that follows the design and manufacture of (at least) the prototype product, has currently become a topic of up-front investigation and commitment. Today, it is not uncommon to list the *design for testability* (DFT) features of a product among the so-called *functional* requirements in the definition of a new product to be developed. Just how such a major transformation has occurred can be understood by examining the testability problems faced by manufacturing organizations and considering their impact on time to market (TTM).

The Testability Problem

The primary objective of testing digital circuits at chip, board, or system level is to detect the presence of hardware failures induced by faults in the manufacturing processes or by operating stress or wearout mechanisms. Furthermore, during manufacturing, a secondary but equally important objective is to accurately determine which component or physical element (e.g., connecting wire) is faulty so that quick diagnosis/repair of the product becomes possible. These objectives are necessary due to imperfections in the manufacturing processes used in building digital electronic components/systems. All digital circuits must undergo appropriate level testing to avoid shipping faulty components/systems to the customer. Analog circuits may have minimum and maximum allowable input signal values (e.g., input voltage) as well as infinitely many values in between these that the component has to be able to respond to. Testing of analog circuits is often achieved by checking the circuit response at the specified upper and lower bounds as well as observing/quantifying the change of the output response with varying input signal values. On the other hand, the behavior of a digital system is characterized by discrete (as opposed to continuous) responses to discrete operating state/input signal permutations such that testing of digital circuits may be achieved by checking their behavior under every operating mode and input signal permutation. In principle this approach is valid. However, in practice, most digital circuits are too complex to be tested using such a brute force technique. Instead, test methods have been developed to test digital circuits using only a fraction of all possible test conditions without sacrificing test coverage. Here, *test coverage* is used to refer to the ratio of faults that can be detected to all faults which are taken into consideration, expressed as a percentage. At the present time the most popular *fault model* is the so-called *stuck-at* fault model that refers to individual nets being considered to be fault-free (i.e., *good network*) or considered to be permanently stuck at either one of the logic 1 or logic 0 values. For example, if the *device under test* (DUT) contains several components (or building blocks), where the sum of all input and output terminals (*nodes*) of the components is k, there are said to be $2k$ possible stuck-at faults, corresponding to each of the circuit nodes being permanently stuck at one of the two possible logic states. In general, a larger number of possible stuck-at faults leads to increased difficulty of testing the digital circuit.

For the purpose of *test pattern* (i.e., input stimulus) generation it is often assumed that the *circuit under test* (CUT) is either fault-free or it contains only one node which is permanently stuck at a particular logic state. Thus, the most widely used fault model is the so-called *single stuck-at fault* model. Using this model each fault is tested by applying a specific test pattern that, in a good circuit, drives the particular node to the logic state which has the opposite value from the state of the fault assumed to be present in the faulty circuit. For example, to test if node v is stuck at logic state x (denoted by v/x or v-x), a test pattern must be used that would cause node v to be driven to the opposite of logic state x if the circuit is not faulty. Thus, the test pattern attempts to show that node v is not stuck at x by driving the node to a value other than x, which for a two-valued digital circuit must be the opposite of x (denoted by $\sim x$). This leads to the requirement that to detect any stuck-at fault v/x, it is necessary to be able to control the logic value at node v so that it can be set to $\sim v$. If the signal value at node v can be observed directly by connecting it to a test equipment, the particular fault v/x can be detected readily. However, in most cases, node v may be an *internal* node, which is inaccessible for direct

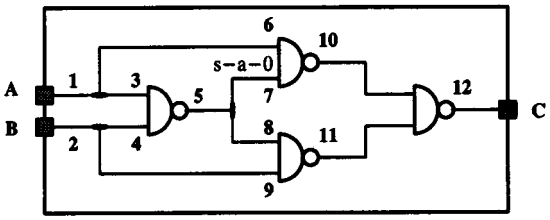


FIGURE 85.9 Example logic circuit with internal node 7 stuck at 0 (7/0).

TABLE 85.2 Test Pattern for Node 7/0 for the Circuit in Fig. 85.9

A	*B*	1	2	3	4	5	6	7	8	9	10	11	12	C
1	0	1	0	1	0	1	1	1	1	0	0	1	1	1 good circuit
1	0	1	0	1	0	1	1	**0**	1	0	**1**	1	**0**	**0** with fault 7/0

observation from outside the component package. In that case, it is necessary to create a condition where the value of the signal on an externally observable node, say node *t*, will be different for each of the two possible values that node *v* can take on, that is, node *t* shall be driven to logic state *y* or $\sim y$ depending upon whether node *v* is at logic state *x* or $\sim x$, respectively. Note that *x* and *y* may represent the same or different logic states.

The external pins of a component are the only means of applying the stimuli and observing the behavior of that component. During testing, a test pattern is used as the stimulus to detect the presence of a particular fault by causing at least one output pin of the component to take on a different value depending upon whether the targeted fault is present or not. Thus, a test pattern is used for *controlling* the circuit's nodes so that the presence of a fault on a circuit node can be *observed* on at least one of the circuit's external pins. Solving the dual problems of *controllability* and *observability* is the primary objective of all test methods. The *logic-to-pin ratio* of a digital circuit is a relative measure of the ratio of possible faults in the circuit to the number of signal pins (i.e., not including the constant power/ground pins) of that component. A large-value logic-to-pin ratio implies that logic states of a large number of circuit nodes must be controlled using a small number of external pins. As a result, conflicting requirements for controllability and observability become harder to satisfy, and the circuit is considered to be more difficult to test.

Consider Fig. 85.9, which depicts a single (hypothetical) *integrated circuit* (IC) component and shows its internal circuitry which uses four NAND gates. The nodes of the circuit are numbered 1 through 12 and the external pins of the component are labeled *A, B,* and *C*. To detect if node 7 is stuck at logic 0 (i.e., 7/0), a test pattern must be found that sets node 7 (and hence, node 5) to the logic 1 state. This can be achieved by setting either or both of the external pins *A* and *B* to the logic 0 state. Furthermore, to observe (or deduce) the value of node 7 at the only externally visible circuit pin, *C*, it is necessary to create a condition where the logic state of node 12 becomes dependent on the value of node 7. The only path from node 7 to node 12 passes through node 10, and since node 10 is the output of a NAND gate the second input to that gate (i.e., node 6) must be set to the logic 1 state by setting input pin *A* to the logic 1 state. Therefore, the only possible test pattern for 7/0 is $A = 1$ and $B = 0$. At this point, we must still continue the analysis to see if indeed node 12 will reflect the value of node 7. With input terminals *A* and *B* set to logic 1 and logic 0, respectively, node 9 will be set to logic 0, which causes node 11 to become logic 1. With these settings, the value at node 12 will be determined by the value at node 10 and the test pattern is valid. Table 85.2 shows the values of all circuit nodes when this test pattern is applied to the circuit of Fig. 85.9.

It should be evident from the simple example of a *combinational circuit* described above that test pattern generation for digital circuits can be very difficult and involved. The problem becomes much more complex when dealing with *sequential circuits,* where the *internal state variables* (i.e., bistable memory storage elements such as latches and flip-flops) must be treated as *pseudo-inputs* and *pseudo-outputs* that must be controlled and observed using the external pins of the component. In this case test patterns become *test sequences* that must be applied in precise order, and outputs must be observed only at prescribed times. Thus, the testing of sequential

circuits is much harder to achieve compared to the testing of combinational circuits. Computer programs, called automatic test pattern generation (ATPG) programs, have been developed for generating test patterns for combinational or sequential circuits. By far, the generation of test patterns for combinational circuits is better understood and automated than doing the same for sequential circuits.

Before discussing the various techniques that may be used to improve testability of digital circuits, it is necessary to mention the related problem of determining test effectiveness. A typical digital system contains a very large number of possible stuck-at faults. This and the logical complexity of the circuits make it unacceptable to "guess" how effective the test patterns (or the diagnostic program) will be in detecting all possible faults. This problem is often approached in a formal manner by using a class of test tool called a *fault simulator* program. A fault simulator uses the given set of test patterns to simulate the given circuit first when there are no faults assumed present (i.e., good circuit simulation). Next, the circuit is simulated with the same set of test patterns, but this time the effects of each possible stuck-at fault are considered one at a time. For a given test pattern, and given stuck-at-type fault, if the output of the good circuit simulation differs from the output obtained during fault simulation, then the given fault will be detected by the given test pattern. This way, it is possible to determine the percentage of all possible stuck-at faults that may be present in a digital circuit which will be covered by the given set of test patterns.

Most ATPG programs operate by picking a possible fault from among the possible faults, generating a specific test pattern that covers it, simulating the logic circuit with the newly generated test pattern to determine which other faults are incidentally covered by the same pattern, and continuing the process until all faults have been considered. Of the two related processes of *test pattern generation* and *fault simulation*, the latter is by far the more time-consuming one.

A different approach is taken in some testability analysis tools whereby rather than determining which faults are covered by a given test pattern, the analysis program assigns a numeric value to indicate the degree of difficulty of controlling and observing the digital circuit's nodes. This analysis, which can be done much more quickly compared to performing fault simulation, should be done prior to attempting to generate the test patterns for a circuit so that time will not be spent unnecessarily on digital circuits which are likely to present difficulties for the ATPG/fault-simulation process to deal with.

Design for Testability

Low-cost/high-volume manufacturing requires that product testability be considered up front since a product which is inherently hard to test will cost both time and money to achieve a desired level of quality. There are many steps that can be taken to improve the testability of digital circuits and systems. The following subsections describe some of the techniques that can be used.

Ad-Hoc Techniques [Abramovici et al., 1990; Bardell et al., 1978]

Circuit/System Reset Requirements. A simple and straightforward mechanism for resetting a digital circuit to a known state is an essential requirement for testability. It should be noted that the requirement is not only for having the reset function provided but further that it should be simple to execute. For example, applying a defined sequence of external signals to a circuit which must be synchronized with a free-running clock signal would not be considered a simple reset mechanism. Instead, keeping an external signal at some logic value for a minimum duration is a much more desirable approach. It is very desirable that the reset function be asynchronous (i.e., not require system clock pulses to execute) since during power-up a circuit may need to be reset even before free-running clock pulses can be started.

Clock Control Requirements. Another very important requirement for implementing DFT is the ability to control the clocking of the internal logic of the digital circuit. If the external clock signal is gated with some other signals such that it is necessary to determine how to set these other signals to their required values to allow the externally applied clock pulse to reach the internal flip-flop clock terminals, then the ATPG program has another level of constraints to resolve in generating the test patterns. Furthermore, some of these additional requirements may pose difficulties in satisfying them during component and/or system testing. Most ATPG programs assume that once the test pattern has been applied to the pins of the component, the system's response to that pattern can be captured by applying an external clock pulse which enables the internal flip-flops to

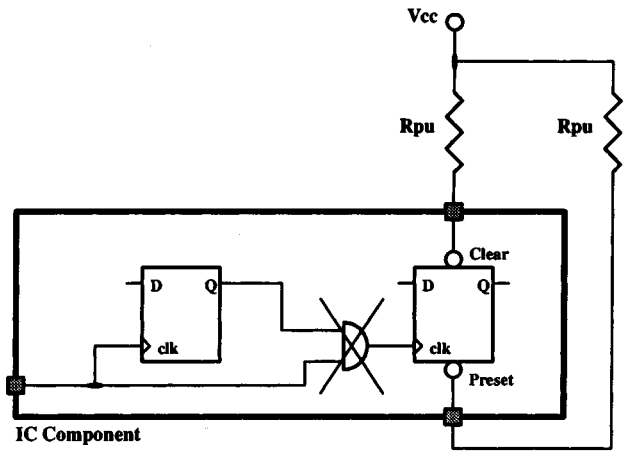


FIGURE 85.10 Using pull-up resistor to tie off unused preset/clear inputs of flip-flops.

respond to the test pattern. Thus, the ATPG programs assume that the internal flip-flop clock inputs are controlled directly from an external pin of the component. This very desirable characteristic is often expressed by stating that *externally applied clock pulses are not allowed to be gated by other signals before these reach the clock terminals of the internal flip-flops.* A side benefit of this design rule is that it prevents glitches (i.e., undesirable pulses) which might be generated at the flip-flop clock terminals due to changing the other inputs to the clock gating circuit while the clock pulse is present.

Managing "Unused" Inputs of Components. When designing digital systems from existing components there may be inputs of those components that, for the current implementation, are not needed. For example, if a two-input AND gate is needed to implement a logic circuit on a printed circuit board, it may be possible to use one of the unused three-input AND gate elements from an IC package already present on that board. In this case, the unused third input of that AND gate must be connected to the logic 1 level in order that a three-input AND function may be implemented using the other two inputs to that gate. Thus, the unused input to the AND gate may be connected directly to the V_{cc} (i.e, power supply) signal. Similarly, if a flip-flop contains unused *preset* or *clear* terminals, these may be tied off to their respective deasserted states. In many cases printed circuit boards are tested using an *in-circuit tester* which uses a *bed-of-nails* test fixture to make physical contact with selected nets on the board so that their values can be observed or controlled by the tester. For the in-circuit tester to control the value of a net it has to backdrive the output of the component which normally drives that net. Since IC components have limited output drive capabilities, the in-circuit tester can overcome the electrical drive from that component and can force that net to a value opposite the value which the driving IC is trying to achieve. By keeping such backdriving conditions to last only a very short period, damage to the opposing IC component is prevented. However, if the net is driven not by an IC but directly from the V_{cc} or ground (*Gnd*) signals, then the in-circuit tester may not be able to overcome their drive. Furthermore, back-driving the V_{cc} or *Gnd* levels would prevent the other IC components from being able to perform their normal functions. Instead, if the logic signals to such unused terminals are applied using *pull-up* or *pull-down* resistors when connecting these to the V_{cc} or *Gnd* levels, respectively, these signals may be controlled by the in-circuit tester. For example, this way it becomes possible to set/reset a flip-flop value by using the normally "unused" preset/clear terminal of that flip-flop. Note that if the flip-flop contains both a preset and a clear input which are unused, these must be pulled up (or pulled down) through separate resistors so that each can be controlled by the in-circuit tester independent of the other. This is illustrated in Fig. 85.10.

Synchronous versus Asynchronous Design Style. More than any other issue, discussions concerning synchronous versus asynchronous design style create the most disagreements concerning design for testability. Many logic designers who are experienced in using SSI and MSI IC chips have adapted a design style where synchronous (e.g., clocked) and asynchronous (e.g., self-timed) designs are freely mixed together. Using clocked flip-flops

with asynchronous preset/clear inputs is a typical example of this design style. Similarly, building latches out of, say, cross-coupled NAND gates and using these as state variables in implementing finite-state machines used to be a very common technique. However, concerns about system initialization and pattern generation have made this style undesirable for implementing DFT. Indeed, most of the so-called *structured* design styles described below make it a requirement that all internal storage elements be constructed from clocked flip-flops, and feedback loops in combinational circuits are broken with the insertion of such flip-flops, along the feedback paths. Asynchronous circuits suffer from combinational circuit hazards that are glitches created as a result of delay differences along circuit paths. Some hazards may be prevented by constraining the manner (i.e., sequence) in which circuit inputs are allowed to be changed. Whereas such constraints may be met during regular system operation, often test pattern generation algorithms cannot take such constraints into account. Therefore, asynchronous logic may create severe problems during testing.

Avoiding Redundant Logic. Technically speaking, redundancy is the only reason why a given stuck-at fault might not be detectable by any test. For example, if an INVERTER function is implemented by tying both inputs of a two-input NAND gate together, then a stuck-at 1 fault on either one of the inputs becomes undetectable since the output signal can still be determined correctly by the remaining nonfaulty input signal. This creates two problems. First, conventional ATPG programs might spend a lot of time trying to generate a test pattern for such a fault before they declare the fault untestable. Second, the presence of an undetectable fault can cause a detectable fault to become undetectable (it may also cause an undetectable fault to become detectable). For example, consider a parity checking circuit in which an existing stuck-at fault may cause the wrong parity to be generated, and the existence of a second fault may correct the parity and hence hide both failures. The remedy for these situations is to try to avoid redundancy in the first place, and when this is not possible provide additional circuit modes where the redundant circuits might be isolated. Alternately (or in addition) it may be useful to provide additional test points, as described below.

Providing Test Points. A test point is an input or output signal to control or observe intermediate signals in a logic circuit. For example, if triple redundancy has been used to implement a fault-tolerant circuit, additional output signals might be provided so that signal values from the identical functional units become individually observable, improving the testability of the overall circuit. Similarly, control signals might be provided so that, during testing, outputs from some functional units may be forced into certain states which allow easier observation of the outputs from other circuits. Recommended sites for inserting test points include redundant nets, nets with large fan-outs, preset and clear inputs of flip-flops, nets that carry system clock signals, (at least some of the) inputs to logic circuit gates with large number of inputs (i.e., large fan-in), data and/or address lines of bus lines, as well as intermediate points in cascaded circuits (such as long ripple counters, shift registers).

Logic Partitioning. Traditionally logic partitioning has been used as a strategy when the circuit is too large/complex for the test generation tools to handle. Thus, its objective is to reduce the number of circuit nodes that must be considered jointly in order to generate test patterns. The partitioning process identifies the *logic cones*, which are sections of logic receiving inputs from multiple input sources and generating a single output. Thus, a digital circuit would be broken into as many individual logic cones as there are individually observable output signals. Obviously, the logic cones may (and often do) overlap with each other since they share common input signals or intermediate signals generated from inside one partition and used in another partition. This is illustrated in Fig. 85.11(a), where two overlapping cones of logic are shown. Here, logic cones O_1 and O_2 contain primary inputs I_1, I_2, I_3, I_4 and I_3, I_4, I_5, I_6, respectively. When either partition is dependent on more inputs than what the ATPG tools or the tester can accommodate, it is possible to insert an additional gate, controlled by a tester input in order to test each partition independently of the other. This is illustrated in Fig. 85.11(b), where an additional input pin I_t has been added such that with I_t set to logic 0 by the tester, it is possible to test either partition without requiring to control shared inputs I_3 or I_4. Logic partitioning has become more important as a result of increased use of *pseudo-exhaustive testing* (to be described later).

Testing Embedded Memory Blocks. A major testability problem arises when a regular-structure memory block such as random-access memory (RAM) or read-only memory (ROM) is embedded into a logic circuit. This creates three problems:

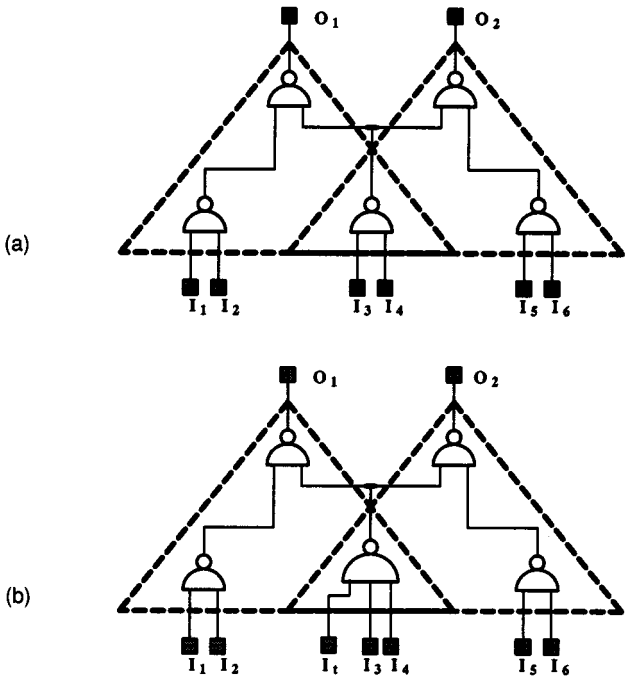


FIGURE 85.11 (a) Logic partitioning with overlapping logic cones. (b) Adding an additional test point to reduce dependence on primary inputs.

1. Testing logic that is downstream from the RAM block (i.e., output of RAM block drives the downstream logic) is difficult since this requires setting the test pattern at the RAM outputs. This problem is usually solved by providing a bypass mode where data inputs to the RAM (or ROM) block are channeled directly to the RAM (or ROM) outputs without (or in addition to) being stored inside the RAM block. This way the RAM data outputs can be controlled by controlling the data inputs as desired.
2. Testing logic that is upstream from the RAM block (i.e., outputs from logic circuit are captured by the RAM block) is difficult since the observation point is the RAM block. That is, it is necessary to access the RAM block in order to observe the test results. This problem might be solved by improving the observability of the RAM inputs and/or making the RAM outputs more easily observable as well as providing the *bypass* capability. This way, inputs to the RAM might be bypassed directly to the RAM outputs where they may be observed. This may require adding an *observe-only* register to capture the RAM outputs.
3. Testing of the RAM block itself is difficult since controlling its inputs and observing its outputs require manipulating the upstream and downstream logic circuit blocks, which may be difficult to achieve. Solution to this problem involves providing adequate control of the RAM block inputs (data, address, and read/write control) as well as providing observability of the RAM outputs. In effect, the embedded RAM block can be made testable as if it was a stand-alone block where established memory test algorithms can be applied [Breuer and Friedman, 1976].

Figure 85.12 illustrates how to improve testability of an embedded RAM structure.

Structured Techniques

An alternate approach to improving the testability of digital circuits is to carry out the circuit design by following certain rules that, by construction, assure high testability of the resulting circuits. Since the main problem in achieving testability of a digital circuit is achieving adequate controllability/observability of its internal nodes, structured DFT approaches [Bardell and McAnney, 1978] follow strict design rules that are aimed at achieving this goal. Furthermore, most structured DFT approaches require/recommend additional design rules aimed at preventing incorrect circuit operation as a result of signal races and hazards.

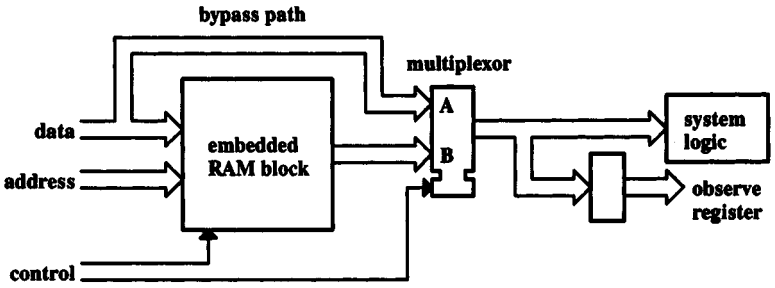


FIGURE 85.12 Providing testability in a design containing an embedded memory block.

Level-Sensitive Scan Design (LSSD). Level-sensitive **scan design** [Eichelberger and Williams, 1978] imposes strict rules on clock signal usage and allows implementing sequential behavior to be implemented only using the shift-register latch (SRL). In the first place, by not allowing any feedback involving combinational circuit elements alone, the LSSD approach prevents timing failures that might be present in purely asynchronous designs. Furthermore, rigid clocking rules are stated in order to prevent SRL data inputs from changing while the clock pulse(s) is (are) transitioning. Hence, the digital circuit is separated into two sections: (1) a robust (i.e., level-sensitive) multi-input/multi-output combinational circuit and (2) a set of SRL elements with which sequential behavior is implemented. In addition to their normal system interconnections each SRL is also connected to its two neighboring SRLs to form a shift-register structure. The serial shift input and shift output signals are labeled *scan-in* and *scan-out,* respectively, and treated as primary input/output terminals. Figure 85.13

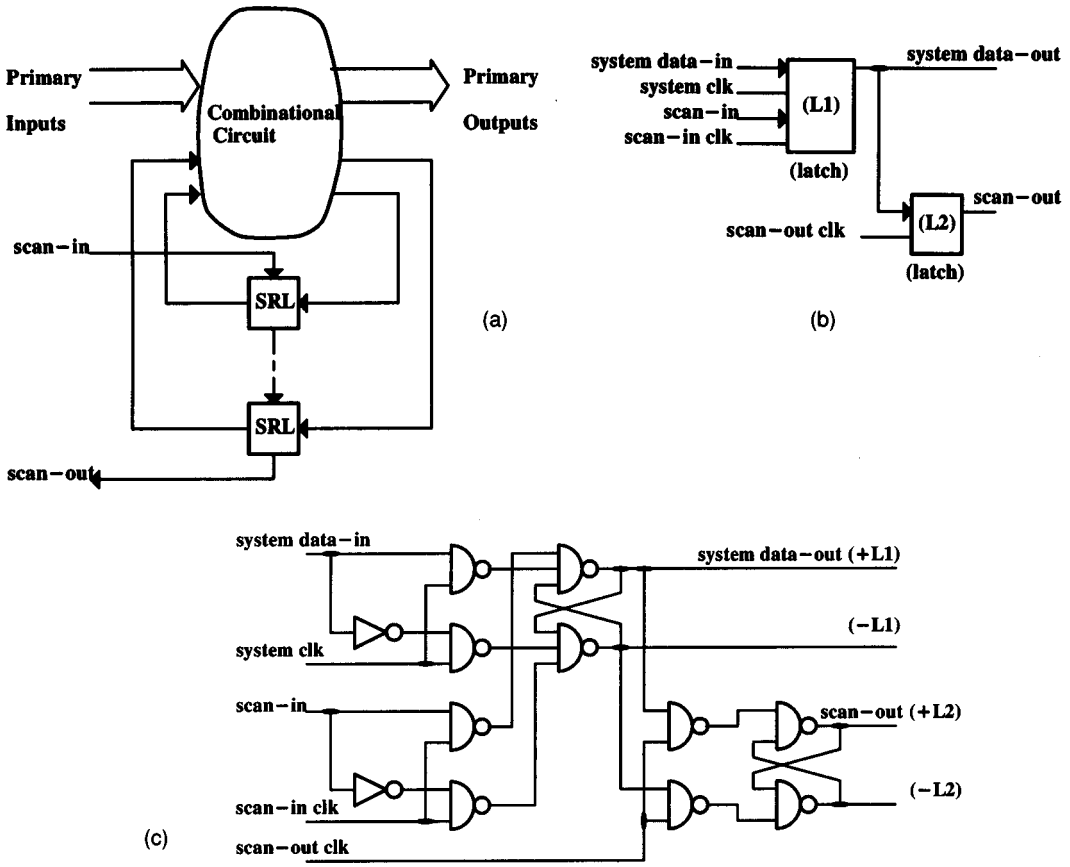


FIGURE 85.13 (a) LSSD circuit model. (b) SRL block diagram. (c) SRL logic diagram.

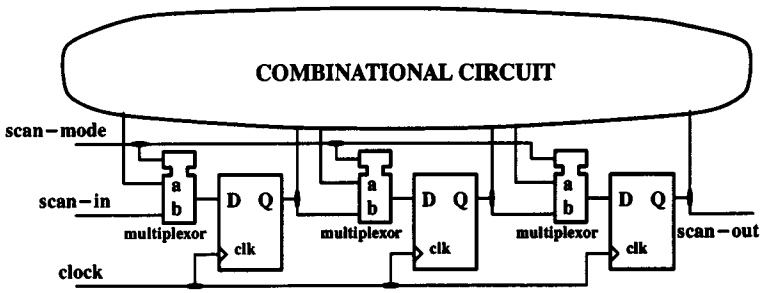


FIGURE 85.14 Model of a digital circuit with scan path.

shows an LSSD circuit model and the general form of an SRL. The significance of the shift-register (often referred to as the *scan-register*) structure is that, during testing, it allows each SRL's value to be individually controllable and observable by shifting (i.e., scanning) a serial vector into/out of the scan register. Hence, the SRLs can be treated as *pseudo-input/output terminals,* and the testing of the digital circuit is reduced to that of a combinational circuit only. Figure 85.13(a) shows an LSSD circuit model, and the general form of an SRL is given in Fig. 85.13(b). A possible gate-level circuit implementation of an SRL is shown in Fig. 85.13(c).

Among the most important LSSD design rules are the following:

1. All internal storage is implemented using SRLs. Each SRL operates such that the L1 latch accepts one or the other of the system data-in or the scan-in data values depending upon whether the system clk or the scan-in clk clock pulse is applied, respectively. The L2 latch accepts the L1 latch value when the scan-out clk clock pulse is applied. The L1 and L2 latches are stable (i.e., cannot change) when the clocks are off.
2. The SRL clocks system clk, scan-in clk, and scan-out clk must be controlled from primary circuit terminals and must be operated in nonoverlapping fashion. This eliminates dependency on minimum circuit delay and assures hazard-free (i.e, level-sensitive) operation.
3. System data-out from SRL_1 may feed the system data-in terminal of SRL_2 only if the system clk which feeds SRL_1 does not overlap with the system clk which feeds SRL_2. This rule prevents the data input to a latch from changing while its clock signal is transitioning.
4. All SRLs are interconnected into one or multiple shift registers by connecting the scan-out terminal from one SRL to the scan-in terminal of the next one in series. If multiple shift registers are implemented, each must be capable of being shifted simultaneously with the others and must have its own scan-in and scan-out primary terminals.

Scan Path. The *scan-path* [Funatsu et al., 1975] approach can be seen as a generalization of the LSSD approach since it follows the same principles but uses standard *D*-type flip-flops as the storage elements instead of the SRLs. The scannable flip-flops can be implemented using dual-ported latches (similar to the L1 latch in the SRL) or using a multiplexor to select between the scan-in and system data-in signals to feed the *D* input of a standard *D*-type flip-flop, as shown in Fig. 85.14.

Scan/Set Logic. Scan/set [Stewart, 1977] is another form of implementing scan technology whereby the sequential circuit structure is separated from its accompanying scan/set register. This is illustrated in Fig. 85.15. A variation on this scheme is the so-called shadow-register concept that has been implemented in some off-the-shelf IC components [AMDI, 1987].

Random-Access Scan. Random-access scan [Ando, 1980] uses a technique akin to addressing locations in a memory (e.g., RAM) block in order to make the states of all storage elements controllable and observable from primary input/output terminals. Using this approach, each storage element is made individually addressable (i.e., accessible) so that in order to control and/or observe the value of an individual storage element it is not necessary to shift in/shift out all other storage elements as well. Figure 85.16(a) shows the general model of a digital circuit employing the random-access scan approach. A possible gate-level circuit implementation of an addressable latch is given in Fig. 85.16(b).

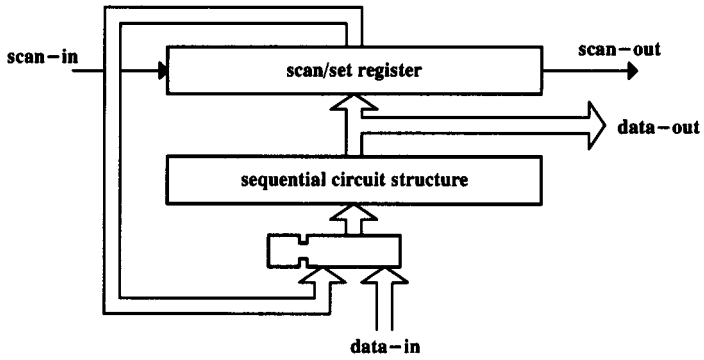


FIGURE 85.15 Generic scan/set circuit design.

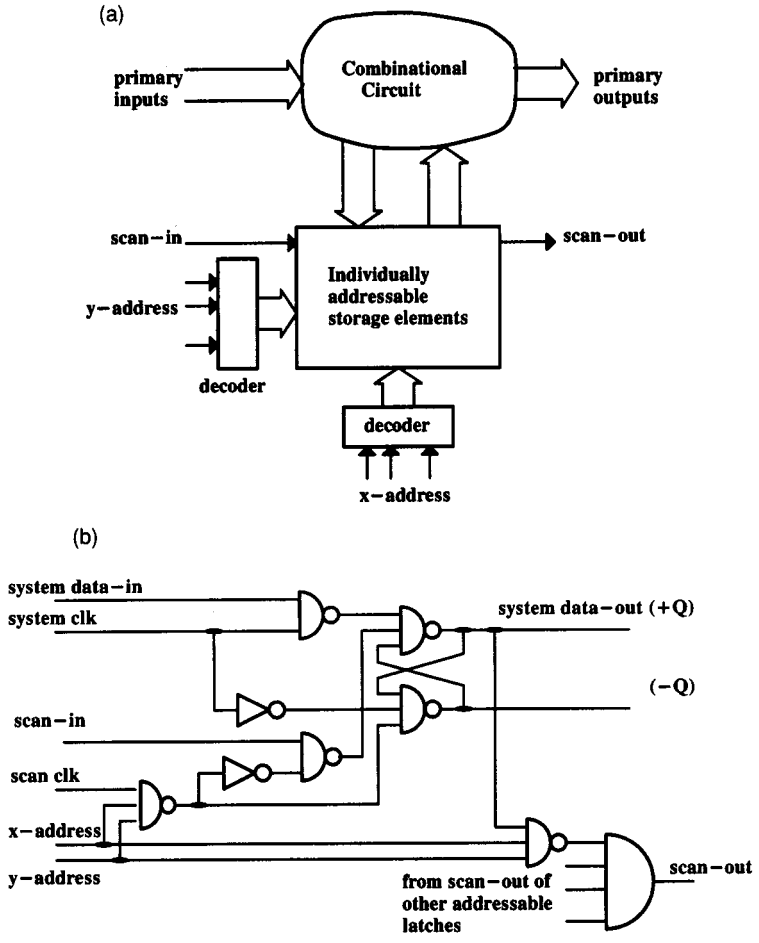


FIGURE 85.16 (a) General model for digital circuit implementing random-access scan. (b) Logic diagram for addressable latch.

Using this approach, each storage element in the circuit is given a unique *x*/*y* address and the decoded address signals are connected to the *x*/*y* address inputs of the latches. As seen in the circuit of Fig. 85.16(b), each latch can then be individually written into using the *scan-in* terminal or its output can be observed using the *scan-out* terminal, provided that the pair of *x*/*y* address lines connected to the current latch are both asserted (i.e.,

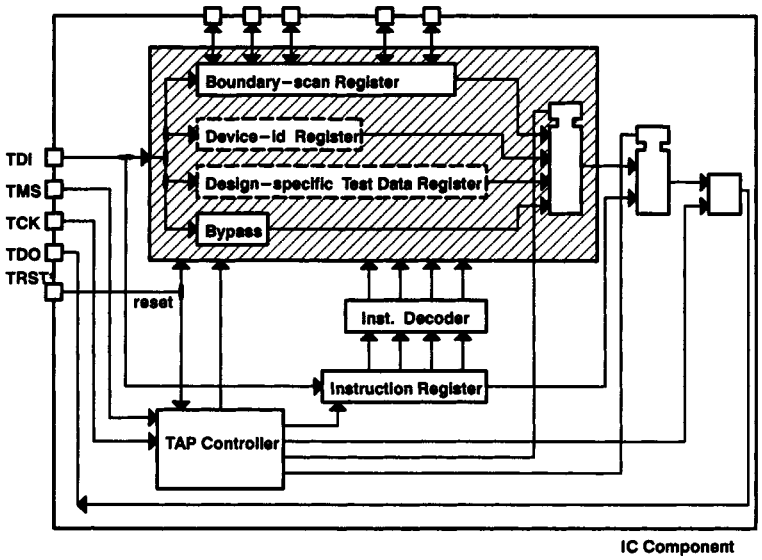


FIGURE 85.17 Architecture of IEEE 1149.1 boundary-scan standard.

set to logic 1). Furthermore, whereas it is also necessary to apply the *scan-in clk* in order to write into the latch, no clock is necessary to observe the latch output. This is a convenient feature that allows the latch values to be selectively observable even while the regular system operations are being executed. The *scan-out* values from the individual latches are combined together into a single AND gate and brought out to a primary output terminal of the circuit. This arrangement works since for any given address only one of the addressable latches will be selected and the scan-out from all other latches will be forced to the logic 1 state. On the other hand, a disadvantage of this approach is that before addressing each latch its proper address must first be applied to the circuit.

Boundary Scan. Unlike the other scan-based techniques described above, **boundary scan** [IEEE, 1990] is intended primarily for testing the board-level interconnections among the IC components on a printed circuit board (PCB). In effect, boundary scan is a special form of scan path that is implemented around every I/O pin of an IC component in order to provide controllability and observability of the I/O pin values during testing. Test control signals provided by an on-chip controller are used to disable the boundary-scan cells during regular system operation so that signal values can flow in/out of the IC component without interference from the test circuits. During testing, *output* pin values can be controlled using values preloaded into the boundary-scan register. Similarly, signal values received on the *input* pins can be captured into the boundary-scan register and subsequently shifted out to be observed on an external tester.

Boundary scan has become an important tool in achieving design for testability following the adoption of the IEEE 1149.1 Test Access Port and Boundary-Scan Architecture in 1990. The IEEE 1149.1 Standard defines a mandatory four-pin (plus an optional fifth pin) test access port (TAP) for providing the interface between the IC component and a digital tester. TAP signals comprise test data input (TDI), test data output (TDO), test clock (TCK), and test mode select (TMS) plus an optional asynchronous tap reset (TRST*) signal. The overall IEEE 1149.1 test architecture (see Fig. 85.17) includes:

- The TAP
- The TAP controller
- The instruction register (IR)
- A group of mandatory and optional test data registers (TDRs)

The TAP controller is characterized by a 16-state finite-state machine (FSM) whose behavior is defined by the IEEE 1149.1 Standard. State transitions of the TAP FSM are controlled by the TMS input line and the dedicated test clock, TCK. Figure 85.18 shows the state-transition diagram for the TAP FSM.

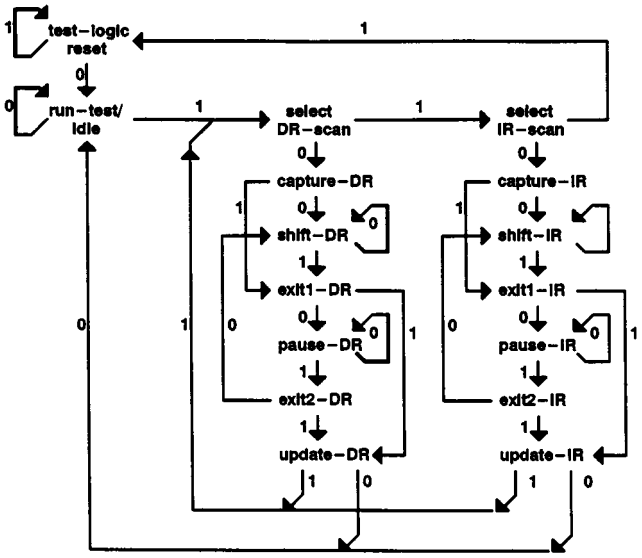


FIGURE 85.18 State-transition diagram for the TAP FSM.

A most important test data register defined by the IEEE 1149.1 Standard is the boundary-scan register that has individual cells associated with each I/O pin of the IC component. Mandatory and permissible features of the boundary-scan register cells are defined by the standard. In addition, a special single-bit register called the BYPASS register has been provided to furnish a more efficient way to shift data through IC components when multiple ICs are chained together by connecting the TDO output from one component to the TDI input of another.

Another mandatory feature of the IEEE 1149.1 Standard is the instruction register and an associated list of mandatory/permissible instructions that govern the behavior of the IC component during testing. The three mandatory instructions are called SAMPLE/PRELOAD, BYPASS, and EXTEST. SAMPLE allows taking a snapshot of the normal operation of the IC, whereas PRELOAD is used for shifting the captured values out while new values are loaded into the boundary-scan register. BYPASS allows shortening the (electrical) distance between the TDI and TDO pins by providing a single-bit register as a shortcut during scan operations involving multiple IC components that are connected in series. EXTEST is the "workhorse" instruction that allows driving the signal values on the component's output pads from the boundary register while capturing the input values into their respective cells in the boundary register. This is followed by shifting the captured values out (using the TDO output) while simultaneously shifting in the new driving values (using the TDI input).

An alternative to using boundary scan is to use a "traditional" in-circuit tester that uses a special "bed-of-nails" fixture. In this approach [Parker, 1987], every net on a PCB would be probed using a tester pin which comes in physical contact with that net such that the current signal value of the net can be observed by the tester. The tester can also be used to control the signal values of the individual nets by injecting appropriate currents through the tester pins. However, since each net is already connected to an output pin of a component on the PCB, this approach amounts to *backdriving* the output drivers of IC components and therefore poses a potential risk of damage to the IC components. This approach is becoming more difficult and/or costly to implement as the number of nets goes up and IC pin spacing is reduced. Furthermore, due to fixturing difficulties, double-sided PCBs cannot be tested in this manner. The IEEE 1149.1 boundary-scan standard [IEEE, 1990] helps solve these problems by providing convenient direct access to the I/O pins of an IC component without requiring the traditional bed-of-nails fixture.

The "CrossCheck" Technique. The CrossCheck approach [Gheewala, 1989] uses cells with built-in test points to observe critical signal values. The test points are connected to an underlining grid structure using very small FETs called *cross-point switches.* An on-chip test control circuit generates the necessary signals to address the individual probe lines and capture the results in a *multi-input signature register* (MISR). Test patterns can be

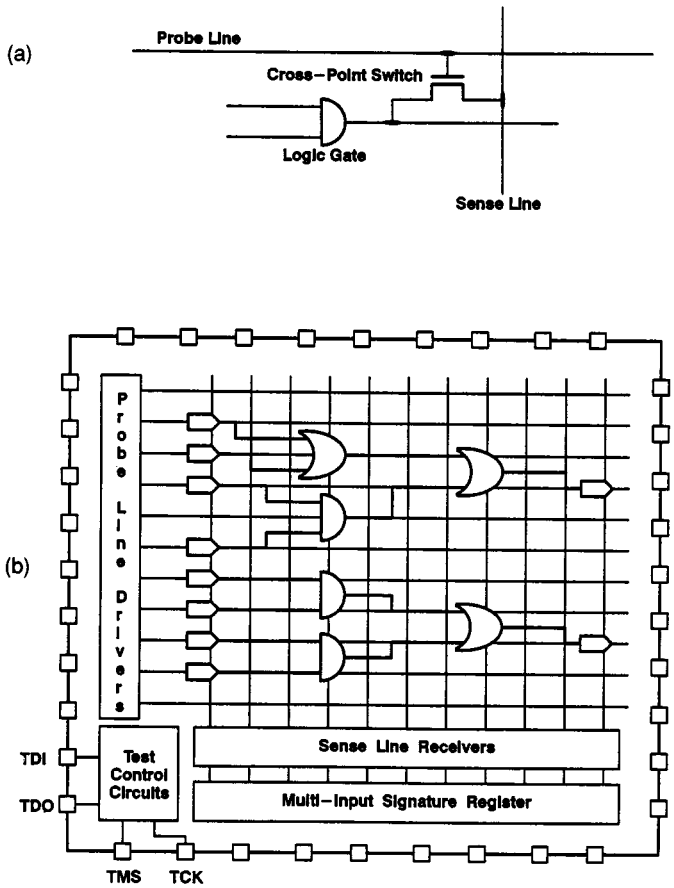


FIGURE 85.19 (a) Cross-point switch implementation. (b) Overview of the CrossCheck technique.

generated externally or by using an on-chip pattern generator, and the final test signature (i.e., contents of MISR) can be accessed using dedicated test pins, such as by providing an IEEE 1149.1 TAP (see previous subsection). Figure 85.19 shows how the CrossCheck technique is implemented on an ASIC.

CrossCheck methodology provides a high degree of observability of the ASIC. Since it is not possible to provide observability of all signals of a design, careful analysis must be performed to determine the most effective points for inserting the cross-point switches. Similarly, the size of the grid structure for the probe lines might be chosen to be design-dependent. However, in many instances it may be better to implement the probe lines as part of the IC master slice in order to reduce the amount of customization to a minimum.

The benefit offered by the CrossCheck technique is due to the potential for the reduced number of test patterns necessary to test an ASIC. This is due to the fact that as observability of internal nodes is increased it becomes easier to generate efficient test patterns which can detect many faults simultaneously. Furthermore, increased observability of internal nodes also improves diagnosability and may help determine the root cause of a failure sooner. On the negative side, the CrossCheck technique does not help improve controllability of internal nodes as achieved using scan-path techniques. Also, a primary disadvantage of the CrossCheck methodology is area penalty due to routing channels that must be set aside for the grid structure. Furthermore, added capacitance of the cross-point switches may affect performance, especially in high-speed applications. In addition, since the technique offers very good observability but no controllability of the internal nodes, it lacks the advantage offered by scan-based approaches for system debug and internal **path-delay testing** [Dervisoglu and Stong, 1991]. However, recent advances have been made that improve the controllability of internal nodes using the CrossCheck technique in gate-array ICs.

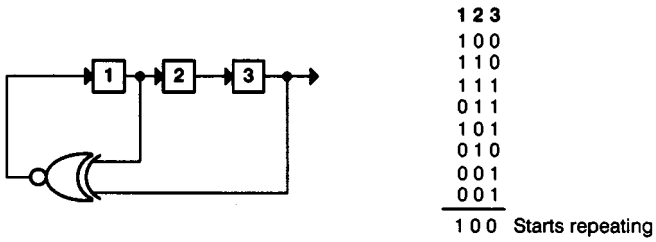


FIGURE 85.20 Three-bit maximal-length LFSR.

Built-in Self-Test (BIST) Techniques. The term **built-in self-test** (or BIST) is a generic name given to any test technique in which an external test resource (e.g., component tester) is not needed to apply test patterns and check a circuit's response to those patterns. This implies that the test patterns must be preloaded into the target device or be generated by the target device itself, in real time. For example, dedicating a section of an IC component for implementing a ROM-based sequencer to apply prestored patterns to test another section of that IC would be classified as a BIST technique. It is often more cost effective to generate the test patterns in real time (i.e., during testing), but in general it is not possible to develop real-time test pattern generation techniques that generate arbitrarily selected test patterns without additionally generating unnecessary ones. Note that whereas storing the test patterns in a ROM might be acceptable in some cases, the size of ROM necessary to store the test patterns prevents this technique being used for implementing BIST in large/complex digital circuits.

One approach to test vector generation is to ignore the specifics of the target circuit and enumerate all possible permutations of inputs. Thus, using *exhaustive* testing, an *n*-input combinational logic cone would be tested by checking its response to all 2**n permutations of input values. In this case, a binary counter can be used as the test pattern generator (TPG). Other, more efficient counter forms (such as a *maximal-length linear feedback shift register,* LFSR) may also be used as the TPG. An LFSR is a special kind of circular-shift register where the serial data input is determined by an EXCLUSIVE-OR function of some of the bit positions. Bit positions which are included in the feedback EXCLUSIVE-OR function are referred to as the tap positions. For any given *degree* (i.e., number of bits) n of LFSR there is at least one set of tap positions that result in the LFSR going through all nonzero n-bit permutations when it is started in any nonzero state. An LFSR that can go through all 2**n states is called a maximal-length LFSR. Figure 85.20 shows a 3-bit maximal-length LFSR and the state sequence that it produces. Exhaustive testing guarantees that all *detectable* faults which do not transform a combinational circuit into a sequential circuit will be detected. Depending upon the clock frequency, this approach becomes impractical to apply when the number of input variables goes up (usually above 22) [McCluskey, 1984].

In cases where the number of test patterns necessary to achieve exhaustive testing is too large to be applicable, a related technique, called **pseudo-random testing**, may be used. Pseudo-random testing achieves many of the benefits of exhaustive testing but requires much fewer test patterns. This is achieved by generating the test patterns in random fashion from among the 2**n possible patterns. However, the random generation of test patterns is done using a deterministic algorithm that produces test patterns in repeatable sequence. Before pseudo-random testing is chosen, it is necessary to examine the pseudo-random test resistance of the circuit. For example, if 500,000 pseudo-random test patterns are applied to a 20-input AND gate, there is only a 0.00004% probability that an essential test pattern (which sets all 20 inputs to logic 1) will be included among them.

Yet another related technique is to use *pseudo-exhaustive* testing that aims at breaking a circuit into separate partitions and testing each partition exhaustively [Barzilai et al., 1985; Dervisoglu, 1985; Bardell and McAnney, 1984]. Pseudo-exhaustive testing uses the same techniques used in exhaustive testing for testing the individual partitions without generating test patterns that cover the entire circuit. Mathematical considerations for pseudo-random/pseudo-exhaustive testing are too complex to describe here. The following example is presented for illustration purposes only. Figure 85.21 depicts the combinational portion of a digital circuit consisting of a number of overlapping logic cones that each produce a single output signal. All inputs are assumed to be

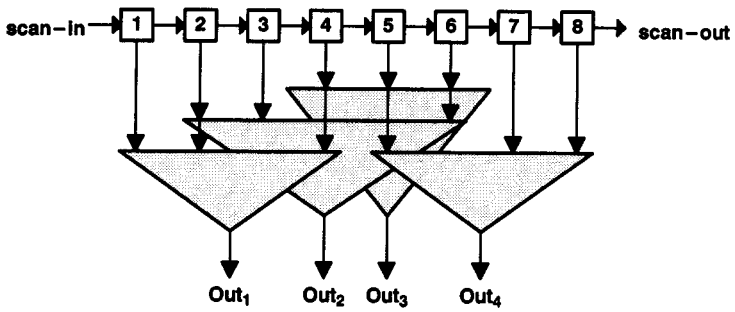


FIGURE 85.21 Overlapping logic cones connected to a common scan path.

connected to scannable flip-flops (i.e., pseudo-inputs) or to primary input pins of the component such that all inputs are 100% controllable either by controlling the values in the flip-flops or the primary input pins. All flip-flops are assumed to be scannable and are arranged into a single *scan path* such that the logic cones have n or fewer inputs all of which lie within k consecutive bits along the scan path. Outputs from the individual logic cones connect (not shown here) to the inputs of flip-flops and/or primary output pins. Thus, all logic cone outputs are also 100% observable. Now, assume that the serial output from the LFSR shown in Fig. 85.20 is connected as the "scan-in" input to the scan-path register shown in Fig. 85.21. In this case any *consecutive* 3-bit partition of the scan-path register will go through the same state sequence as the LFSR itself, delayed from it by the number of flip-flops between that partition and the output bit of the LFSR. For example, the third logic cone that has inputs from flip-flops 4, 5, and 6 will see all input permutations except the all-zeros case which can be applied separately as a special case. On the other hand, the first logic cone, with inputs from flip-flops 1, 2, and 4, will not receive all possible nonzero permutations of three input variables. This is because the first logic cone receives its three inputs from three *nonconsecutive* positions of the scan-path register. In this case only input permutations that have even parity across positions 1, 2, and 4 will be received by the first logic cone. Furthermore, the fourth logic cone that also receives inputs from three nonconsecutive bit positions which are 4 bits apart will receive all 3-bit nonzero input permutations. Analysis of which set of input permutations may be generated across nonconsecutive n bits of a scan-path register which receives the outputs from an mth degree ($m \geq n$) LFSR is based on *linear dependence* and is outside the scope of this section. However, the problem may also be approached statistically by choosing the degree of the LFSR to be higher than n but smaller than k which is the largest span of inputs to any logic cone. For example, in Fig. 85.21 the degree of the LFSR may be chosen as 4. In this case, the probability that a logic cone which has 4 or fewer inputs separated by k bits (here, $k = 5$) may be calculated [Lempel and Cohn, 1985]. It should be noted that a logic cone may be tested in full even when it has not received all 2**n input permutations.

BIST also requires ability to capture the test results without the need for an external tester. This is often achieved by using a *multi-input signature register* (MISR) to capture individual test results and compress these into an overall value called the test *signature*. Figure 85.22 shows a sample signature register that can compress test results captured from four separate outputs into a single 4-bit signature. Provided that the test circuit has deterministic behavior, a signature register can be started in a given starting state, and its final value may be compared to a known good signature to determine pass/fail status. However, compressing test results into a single overall signature may prevent proper fault detection if multiple erroneous outputs (which may result

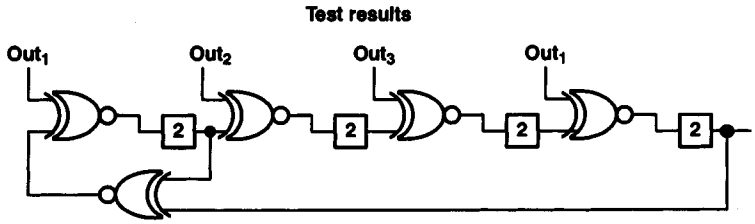


FIGURE 85.22 A four-bit parallel-input signature register.

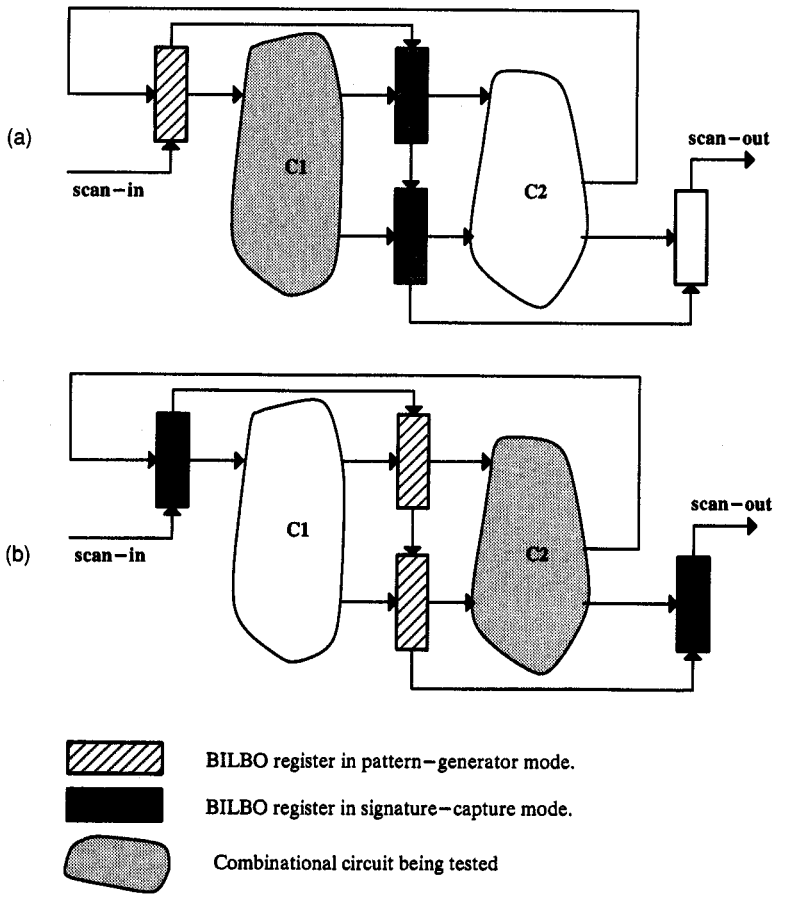


FIGURE 85.23 Using BILBO technique to partition and test a large circuit. (a) Testing combinatorial circuit C_1. (b) Testing combinatorial circuit C_2.

from the same fault being detected on multiple test vectors) causes the final test signature to be correct even though interim signatures were wrong. The probability that a faulty circuit signature will be the same as the good circuit signature is known as aliasing probability. It can be shown that if the test length is sufficiently long, aliasing probability diminishes toward 2^{-t}, where t is the number of bits of the signature register [Dervisoglu, 1985].

The two constructs of LFSR and the MISR can be merged into a single multipurpose register in a *built-in logic block observation* (BILBO) approach [Konemann et al., 1979] where each register can have multiple modes of operation including the LFSR mode, MISR mode, SCAN mode, and NORMAL mode. In this case an on-chip test-control circuit may be used to control the modes of operation of the BILBO registers so that, in turn, each register is used as a test pattern generator or signature register to test a digital component. Figure 85.23 illustrates how to use the BILBO scheme in a stepwise fashion to test a large digital circuit.

Path-Delay Testing

Path-delay testing is aimed at testing whether a given component/system operates at a specified performance level that is often measured as the maximum system clock frequency. For example, the lower bound for the maximum clock frequency which a microprocessor IC is specified that it can reach needs to be verified. However, due to the very large number of different operations that a microprocessor can perform it is not practical to verify correct behavior of such a component operating at maximum clock frequency for every possible single operation or sequence of operations that it is designed to perform. On the other hand, it may be possible to examine the structure of the design to discover its *logic paths* and verify that signals can be propagated along

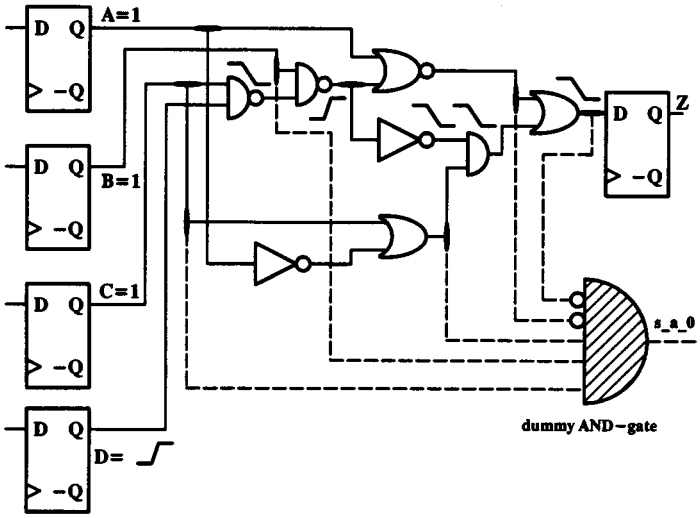


FIGURE 85.24 Circuit example to illustrate path-delay test pattern generation (all flip-flops are clocked using a common clock signal that has not been shown).

these paths within a specified propagational delay time between the initiation of a signal transition at the beginning of the path and the arrival of the final values at the end of that path. This is called *path-delay* testing. A modern IC component with typical complexity would contain many hundreds of thousands of logic paths, so that it becomes impractical to test all of them for at-speed operation. All *synchronous* digital circuits are designed so that there is a fixed clock period resulting from the use-constant frequency clock signals to time their operation. Obviously, the clock period constitutes an upper bound for the propagational delay through any logic path, since otherwise clock pulses may arrive at the flip-flops while their data input signals may still be transitioning. On the other hand, propagational delay through some logic paths may be very close to this upper bound (i.e., clock period) value whereas others may have more slack in them. It is therefore important to identify the *critical* paths and perform path-delay testing on these. Hence path-delay testing can be broken into the two phases of critical-path selection and path-delay test pattern generation.

Several different approaches can be used in identifying the critical paths, including:

1. Select sufficiently large number of paths selected at random from a list of all logic paths.
2. Calculate worst-case timing for all logic paths and select a certain percentage of the slowest paths.
3. First identify certain key nodes and then select paths that pass through those nodes using either of the two approaches listed in (1) and (2) above.

The more challenging problem is to generate the test patterns to verify that none of the signal propagations along a given logic path require longer than the clock-period time to complete. A path-delay test pattern is a pair of patterns that generates the desired signal transition(s) and provides the sensitization of the signal paths whereby the generated transition(s) is (are) sensitized through the combinational circuit to the input of a flip-flop where it will be captured when the system clock is applied. For example, Fig. 85.24 shows a combinational circuit and identifies a specific signal path for which the path delay is to be measured. To determine the appropriate path-delay test patterns, a dummy AND gate is first added to the circuit as shown. An input to the AND gate is derived from the output of the combinational circuit through which the input signal transition is to be propagated. This signal is used in its true or complemented form depending upon whether the final value of the signal transition is a logic 1 or logic 0, respectively. Other inputs to the dummy AND gate come from all remaining inputs of gates through which the desired signal transitions must flow. If the desired signal transition is flowing through an AND or NAND gate, the remaining inputs of these gates are also fed to the inputs of the dummy AND gate, whereas if the desired signal transitions flow through OR or NOR gates, their remaining inputs are inverted and then connected to the inputs of the dummy AND gate. The dummy AND gate is not actually implemented as part of the combinational logic but rather acts as a convenient place to

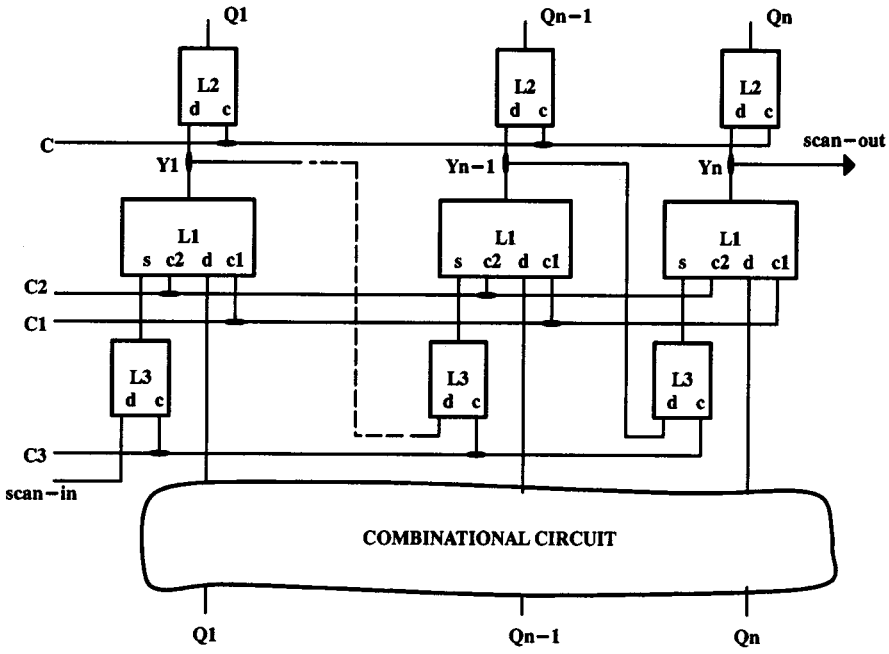


FIGURE 85.25 Using a three-latch flip-flop design to enable path-delay testing.

collect all the necessary conditions for sensitizing the transitions. For example, in the example given above the first pattern requires input flip-flops *A, B,* and *C* all to be set to the logic 1 value in order to sensitize a l*ow-to-high* transition at the *D* input, whereas the second test pattern requires *A, B,* and *C* all to remain at logic 1 while *D* is changed from logic 0 to the logic 1 value. This way the transitions created on input *D* will travel through the identified signal path to reach the destination flip-flop *Z*.

Path-delay test patterns become much easier to generate and also apply to a circuit if the circuit is designed using scannable flip-flops that are additionally capable of storing two arbitrarily selected values in them. This can be done in such a fashion that the initial value available at the flip-flop output will be replaced by the second value when a first clock pulse is applied, and the flip-flop will revert to its normal mode of operation before the second clock pulse is applied. This way the pair of test patterns that form a path-delay test are first loaded into the flip-flops (using scan) and then two clock pulses are applied at speed. The final result captured by the second clock pulse is then scanned out and examined to determine pass/fail status. It is also possible to get an actual measurement of the path delays by repeating the same test over and over again while systematically reducing the time distance between the two clock pulses to determine the minimum separation of the two clock pulses required for proper operation.

Figure 85.25 shows a modified LSSD latch design [Malaiya and Narayanaswamy, 1983] that can be used to enable path-delay testing as described above. Using this design, it is possible to load any two arbitrary test vectors to the combinational circuit in rapid succession. First, test vector Q_1, Q_2,…,Q_n would be scanned into the L1 latches outputs by using clocks C_3 and C_2. Next, the test vector would be moved into the L2 latches by applying a single *C* clock. This way the flip-flop outputs would be set to their initial values defined by Q_1, Q_2,…, Q_n. Following this, the second test vector Y_1, Y_2,…,Y_n would be scanned into the L1 latches using clock signals C_3 and C_2. Now applying the *C* clock causes the first test vector (Q_i) to be replaced by the second test vector (Y_i), and if the C_1 clock is applied next, the response of the combinational circuit will be captured in the L1 latches. This way, the minimum delay between the clock signals *C* and C_1 that is necessary to allow the signals to propagate through the combinational circuit can be determined. Other flip-flop designs with built-in features to support *double-strobe* testing are also possible [Dervisoglu and Stong, 1991].

A different and more difficult-to-use approach for generating test patterns for path-delay measurement is to perform scan-in to load the internal flip-flops with a special pattern that prior circuit analysis will have determined will be transformed into the actually intended test pattern when the first functional clock pulse is

applied. The circuit analysis required to use this approach amounts to performing simulation in reverse time flow to determine what state the device under test should be placed in (using scan) so that its next state corresponds to the desired test pattern.

Future for Design for Test

Present-day trends for striving to achieve shorter time to market while at the same time meeting competitive cost demands are going to continue into the foreseeable future. Design for testability is one of several areas that manufacturers from IC components to complete systems are paying increased emphasis to in order to meet their product goals. Twenty years ago some product managers considered testing as being necessary to weed out the bad from the good but did not consider DFT to be adding value to a product. However, since testing is essential, the value of DFT is seen in reducing the cost of an essential item. Hence DFT adds value to a product at least by an amount equal to the savings in test costs that it brings about. Furthermore, DFT improves time to market by making it possible to identify initial production problems at an earlier point in time. For example, initial productions of high-performance ASIC components may contain flaws that prevent their at-speed operation under certain circumstances. If these flaws are not discovered in a timely manner, they may turn into "showstopper" issues causing serious delays in revenue shipments of products. Whereas no "guaranteed" solutions exist to prevent and/or find a solution for all types of problems, design for testability is a rapidly maturing field of digital design.

Defining Terms

Boundary scan: A technique for applying scan design concepts to control/observe values of signal pins of IC components by providing a dedicated boundary-scan register cell for each signal I/O pin.

Built-in self-test (BIST): Any technique for applying prestored or real-time-generated test cases to a subcircuit, IC component, or system and computing an overall pass/fail signature without requiring external test equipment.

Path-delay testing: Any one of several possible techniques to verify that signal transitions created by one clock event will travel through a particular logic/path in a subcircuit, IC component, or system and will reach their final steady-state values before a subsequent clock event.

Pseudo-random testing: A technique that uses a linear feedback shift register (LFSR) or similar structure to generate binary test patterns with statistical distribution of values (0 and 1) across the bits; these patterns are generated without considering the implementation structure of the circuit to which they will be applied.

Scan design: A technique whereby storage elements (i.e., flip-flops) in an IC are connected in series to form a shift-register structure that can be entered into a test mode to load/unload data values to/from the individual flip-flops.

Related Topic

23.2 Testing

References

M. Abramovici, M. A. Breuer, and A. D. Friedman, *Digital Systems Testing and Testable Design,* Rockville, Md.: Computer Science Press, 1990.

Advanced Micro Devices Inc. [AMDI], "Am29C818 CMOS Pipeline Register with SSR Diagnostics," product specification, Bus Interface Products Data Book, 1987, pp. 47–55.

H. Ando, "Testing VLSI with random access scan," in digest of papers, COMPCON, February 1980, pp. 50–52.

P. H. Bardell and W. H. McAnney, "Parallel pseudorandom test sequences for built-in test," in Proc. International Test Conference, October 1984, pp. 302–308.

P. H. Bardell, W. H. McAnney, and J. Savir, *Built-In Test for VLSI. Pseudorandom Techniques,* New York: Wiley, 1978.

Z. Barzilai, D. Coppersmith, and A. L. Rosenberg, "Exhaustive generation of bit patterns with applications to VLSI self-testing," *IEEE Trans. on Computers,* vol. C-32, no. 2, pp. 190–194, February 1985.
M. A. Breuer and A. D. Friedman, *Diagnosis and Reliable Design of Digital Systems,* Rockville, Md.: Computer Science Press, 1976, pp. 139–146, 156–160.
B. I. Dervisoglu, "VLSI self-testing using exhaustive bit patterns," in Proc. IEEE International Conference on Computer Design, October 1985, pp. 558–561.
B. I. Dervisoglu and G. E. Stong, "Design for testability: Using scanpath techniques for path-delay test and measurement," in Proc. International Test Conference, October 1991, pp. 364–374.
E. B. Eichelberger and T. W. Williams, "A logic design structure for LSI testability," *Journal of Design Automation and Fault-Tolerant Computing,* vol. 2, no. 2, pp. 165–178, 1978.
S. Funatsu, N. Wakatsuki, and T. Arima, "Test generation systems in Japan," in Proc. 12th Design Automation Symposium, June 1975, pp. 114–122.
T. Gheewala, "CrossCheck: A cell based VLSI testability solution," in Proc. 26th Design Automation Conference, 1989, pp. 706–709.
"IEEE Standard Test Access Port and Boundary-Scan Architecture," IEEE Std. 1149.1-1990, May 1990.
B. Konemann, J. Mucha, and G. Zwiehoff, "Built-in logic block observation technique," in digest of papers, International Test Conference, October 1979, pp. 37–41.
A. Lempel and M. Cohn, "Design of universal test sequences for VLSI," *IEEE Trans. on Information Theory,* vol. IT-31, no. 1, pp. 10–17, 1985.
Y. K. Malaiya and R. Narayanaswamy, "Testing for timing faults in synchronous sequential integrated circuits," in Proc. International Test Conference, 1983, pp. 560–571.
E. J. McCluskey, "Verification testing. A pseudoexhaustive test technique," *IEEE Trans. on Computers,* vol. C-33, no. 6, pp. 541–546, June 1984.
K. P. Parker, *Integrating Design and Test,* New York: IEEE Computer Society Press, 1987.
J. H. Stewart, "Future testing of large LSI circuit cards," in Proc. Semiconductor Test Symposium, Cherry Hill, N.J., October 1977, pp. 6–15.

Further Information

An excellent treatment of design for testability topics is found in Abramovici et al. [1990]. Also, Breuer and Friedman [1976] provide a very good treatment of pseudo-random test topics.

C. M. Maunder and R. E. Tulloss (*The Test Access Port and Boundary-Scan Architecture,* IEEE Computer Society Press Tutorial, 1990) provide a user's guide for boundary-scan and the IEEE 1149.1 Standard.

B. I. Dervisoglu ("Using Scan Technology for Debug and Diagnostics in a Workstation Environment," in Proc. International Test Conference, 1988, pp. 976–986) provides a very good example of applying DFT techniques all the way from the IC component level to the system level. Also, B. I. Dervisoglu ("Scan-Path Architecture for Pseudorandom Testing," *IEEE Design & Test of Computers,* vol. 6, no. 4, pp. 32–48, August 1989) describes using pseudo-random testing at the system level. Similarly, P. H. Bardell and M. J. Lapointe ("Production Experience with Built-in Self-Test in the IBM ES/9000 System," in Proc. International Test Conference, October 1991, pp. 28–36) describe application of BIST for testing a commercial product at the system level.

Oldfield, J.V., Oklobdzija, V.G. "Section IX – Computer Engineering"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

The ViewSonic® VP140 ViewPanel is about as flat as a board. This new active-matrix LCD flat-panel has a 14-in. viewing area. Measuring at only 2.5 in. deep and weighing just 12.1 lb, the VP140 weighs only a fraction of standard displays and uses 90% less desktop space. The display unit supports a maximum noninterlaced resolution of 1024×768 pixels at a 75-Hz refresh rate. Additionally, the VP140 monitor can be configured for both desktop or slim line wall displays and it supports up to 16.7 million color images for both PC and Macintosh® environments. This revolutionary view panel represents the future of monitors. (Photo courtesy of ViewSonic Corporation.)

IX

Computer Engineering

86 Organization *R. F. Tinder, V. G. Oklobdzija, V. C. Hamacher, Z. G. Vranesic, S. G. Zaky, J. Raymond*
Number Systems • Computer Arithmetic • Architecture • Microprogramming

87 Programming *J. M. Feldman, E. W. Czeck, T. G. Lewis, J. J. Martin*
Assembly Language • High-Level Languages • Data Types and Data Structures

88 Memory Systems *D. Berger, J. R. Goodman, G. S. Sohi*
Memory Hierarchies • Cache Memories • Parallel and Interleaved Memories • Virtual Memory • Research Issues

89 Input and Output *S. Sherr, R. C. Durbeck, W. Suryn, M. Veillette*
Input Devices • Computer Output Printer Technologies • Smart Cards

90 Software Engineering *C. A. Argila, C. Jones, J. J. Martin*
Tools and Techniques • Testing, Debugging, and Verification • Programming Methodology

91 Computer Graphics *E. P. Rozanski*
Graphics Hardware • Graphics Software

92 Computer Networks *T. G. Robertazzi*
Local Area Networks • Metropolitan Area Networks • Wide Area Networks • The Future

93 Fault Tolerance *B. W. Johnson*
Hardware Redundancy • Information Redundancy • Time Redundancy • Software Redundancy • Dependability Evaluation

94 Knowledge Engineering *M. Abdelguerfi, R. Eskicioglu, J. Liebowitz*
Databases • Rule-Based Expert Systems

95 Parallel Processors *T. Feng*
Classifications • Types of Parallel Processors • System Utilization

96 Operating Systems *J. Boykin*
Types of Operating Systems • Distributed Computing Systems • Fault-Tolerant Systems • Parallel Processing • Real-Time Systems • Operating System Structure • Industry Standards

97 Computer Security and Cryptography *J. A. Cooper, O. Goldreich*
Computer and Communications Security • Fundamentals of Cryptography

98 Computer Reliability *C. G. Guy*
Definitions of Failure, Fault, and Error • Failure Rate and Reliability • Relationship Between Reliability and Failure Rate • Mean Time to Failure • Mean Time to Repair • Mean Time Between Failures • Availability • Calculation of Computer System Reliability • Markov Modeling • Software Reliability • Reliability Calculations for Real Systems

99 The Internet and its Role in the Future *G. L. Hawke*
History • The Internet Today • The Future

John V. Oldfield
Syracuse University

Vojin G. Oklobdzija
University of California

COMPUTER ENGINEERING is a discipline that deals with the engineering knowledge required to build digital computers and special systems that communicate and/or process or transmit data. As such, computer engineering is a multi-disciplinary field because it involves many different aspects of engineering that are necessary in designing such complex systems. To illustrate this point one can think of all the various parts of engineering that are involved in a design of a digital computer system. One can start with the knowledge of the material science that is necessary to process the materials of which the integrated circuits are made. One also has to deal with the devices and device physics to make the most efficient transistors of which computing sytems are built. The knowledge of electrical engineering and electronic circuits in particular is necessary in order to design fast and efficient integrated circuits. One level further in the hierarchy of the required knowledge is a logic design which is an implementation of the digital functions. Digital design involves not only an intimate knowledge of electrical engineering but also the use of computer aided design tools and algorithms for efficient implementation of computational structures. Building a complex computer system is similar to building a house — at the very beginning one cannot be bothered with all the details involved in the process, such as plumbing and electrical wiring. Similarly a process of designing an electronic computer starts with an architecture that specifies the functionality and major blocks. Much like building a house, those blocks are later designed by teams of engineers using the architectural specifications of the computer. Computer architecture is on a cross-road between electrical engineering and computer science. On one hand, one does not need to specify all the details of implementation while defining an architecture. However, if one does not know the important aspects of the design which require the knowledge of electrical engineering, the architecture may not be good. Given that the implementation of the architecture has to serve as a platform for various applications, the knowledge of software, compilers, and high-level languages is also necessary.

Computer engineering is not only a very diverse discipline, but as such it is a subject of very rapid changes reflecting high rate of progress in a variety of disciplines encompassed by computer engineering. The performance of digital computers has been doubling steadily every two years while the capacity of the semiconductor memory has been quadrupling every three years. The price-performance figure has dropped for two orders of magnitude in the last ten years. This trend has radically changed the way the computer is perceived today. From exclusive and expensive machines, affordable to only a few, it has become a commodity. For example, an average automobile today contains in the order of 20 processors controlling various aspects of the machine function, brake system, navigation, etc.

Some of the technology-specific aspects of computer engineering were covered in Section VIII. This section, however, is concerned with higher-level aspects which are substantially independent of circuit technology. Chapter 86 reviews organizational matters which particularly affect computer processor design, such as the arithmetic and logical functions required. The next chapter considers the major topic of programming, which may be different in each "layer," using the previous analogy. Programming too has long been dominated by a particular paradigm, the so-called *imperative* model, in which the programmer expresses an algorithm, i.e., a process for solving a problem, as a sequence of instructions—either simple or complex, depending on the type of programming required. Recently others have emerged, such as rule-based programming, which has a *declarative* model, i.e., the user specifies the facts and rules of a situation and poses a question, leaving the computer ("knowledge-engine") to make its own inferences en route to finding a solution or set of solutions.

Computer memory systems are considered in Chapter 88. Early purists preferred the term *storage systems,* since the organization of a computer memory bears little resemblance to what we know of the organization of the human brain. For economic reasons, computer memories have been organized as a hierarchy of different technologies, with decreasing cost per bit as well as increased access times as one moves away from the central processor. The introduction of virtual memory in the Manchester Atlas project (c. 1964) was a major breakthrough in removing memory management from the tasks of the programmer, but recently the availability of vast quantities of semiconductor memory at ultralow prices has reduced the need for this technique.

The topic of Chapter 89 is the input and output of information. Early computers were confined almost exclusively to character information, but "input/output" now refers to any form of *transducer,* to choose an engineering term, which allows any form of information to be sensed whether in analog or digital form, entered

into a computer system, and be output by it in a correspondingly useful form. Information may vary in time such as a temperature indication, in two dimensions such as the user's action in moving a mouse, or even in three dimensions, and output may be as simple as closing a contact or drawing a picture containing a vast range of colors.

Software engineering as discussed in Chapter 90 refers to the serious problem of managing the complexity of the layers of software. This problem has few parallels in other walks of life and is exacerbated by the rate of change in computing. It is dominated by the overall question "Is this computer system reliable?" which will be referred to in Chapter 98. Some parallels can be drawn with other complex human organizations, and, fortunately, the computer itself can be applied to the task.

Graphical input and output is the topic of Chapter 91. Early promise in the mid-1960s led to the pessimistic observation a decade later that this was "a solution looking for a problem," but as computer display technology improved in quality, speed, and, most importantly, cost, attention was focused on visualization algorithms, e.g., the task of producing a two-dimensional representation of a three-dimensional object. This is coupled with the need to provide a natural interface between the user and the computer and has led to the development of interactive graphical techniques for drawing, pointing, etc., as well as consideration of the human factors involved.

As computers have extended their scope, it has become necessary for a computer to communicate with other computers, whether nearby, such as a file server, or across a continent or ocean, such as in electronic mail. Chapter 92 reviews the major concepts of both local and wide area computer networks.

Many engineers were skeptical as to whether early computers would operate sufficiently long before a breakdown would prevent the production of useful results. Little recognition has been given to the pioneers of component and circuit reliability that have made digital systems virtually, but still not totally, fault-free. Critical systems, whether in medicine or national defense, must operate even if components and subsystems fail. The next chapter reviews the techniques employed to make computer systems fault-tolerant.

The idea of a rule-based system, referred to earlier, is covered in Chapter 94. Application software naturally reflects the nature of the application, and the term *knowledge engineering* has been coined to include languages and techniques for particularly demanding tasks, which cannot readily be expressed in a conventional scientific or business programming language.

Parallel systems are emerging as the power of computer systems is extended by using multiple units. The term *unit* may correspond to anything from a rudimentary processor, such as a "smart word" in a massively parallel "fine grain" architecture, to a full-scale computer, in a coarse-grain parallel system with a few tens of parallel units. Chapter 95 discusses the hardware and software approaches to a wide variety of parallel systems.

Operating systems, which are described in the next chapter, turn a "raw" computer into an instrument capable of performing useful, low-level tasks, such as creating a file or starting a process corresponding to an algorithm, or transferring its output to a device such as a printer, which may be busy with other tasks.

As society has become more dependent upon the computer and computer technology, it has become increasingly concerned with protecting the privacy of individuals and maintaining the integrity of computer systems against infiltration—by individuals, groups, and even on occasion by governments. Techniques for protecting the security of a system and ensuring individual privacy are discussed in Chapter 97.

Chapter 98 discusses the overall reliability of computer systems, based on the inevitable limitations of both hardware and software mentioned earlier. Given the inevitability of failure, human or component, what can be said about the probability of a whole computer system failing? This may not be an academic issue for a passenger reading this section while flying in a modern jet airliner, which may spend over 95% of a flight under the control of an automatic pilot. He or she may be reassured to know, however, that the practitioners of reliability engineering have reduced the risk of system failure to truly negligible proportions.

Nomenclature

Symbol	Quantity	Unit
A	area	m2
A_m	main amplifier gain	
A_p	preamplifier gain	
A_v	availability	
BW	bandwidth	Mbyte/s
C	capacitance	F
d	distance	m
E_L	illuminance	
f	proportionality factor	
h	Planck's constant	$6.625 \times 10–34$ J·s
L	latencyns	
λ	failure rate	
μ_f	flip-flop sensitivity	
μ_p	photodetector sensitivity	
μ_s	Schmitt trigger sensitivity	
ν	hardware utilization	
ω	angular velocity	rad/s
P	parallelism	
P_c	character pitch	
q	drop charge	
R	wheel radius	
R_1	shaft radius	
S	sensitivity	fL
S	speed-up ratio	
t_L	optical loss	
V_b	band velocity	m/s
ξ	rotation angle	rad
$z(t)$	hazard rate	

Tinder, R.F., Oklobdzija, V.G., Hamacher, V.C., Vranesic, Z.G., Zaky, S.G., Raymond, J. “Organization”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

86

Organization

Richard F. Tinder
Washington State University

Vojin G. Oklobdzija
University of California, Davis

V. Carl Hamacher
Queen's University, Canada

Zvonko G. Vranesic
University of Toronto

Safwat G. Zaky
University of Toronto

Jacques Raymond
University of Ottawa

86.1 Number Systems
Positional and Polynomial Representations • Unsigned Binary Number System • Unsigned Binary-Coded Decimal, Hexadecimal, and Octal Systems • Conversion between Number Systems • Signed Binary Numbers • Floating-Point Number Systems

86.2 Computer Arithmetic
Number Representation • Algorithms for Basic Arithmetic Operations • Implementatino of Addition • Implementation of the Multiplication Algorithm • Floating-Point Representation

86.3 Architecture
Functional Units • Basic Operational Concepts • Performance • Multiprocessors

86.4 Microprogramming
Levels of Programming • Microinstruction Structure • Microprogram Development • High-Level Languages for Microprogramming • Emulation • Applications of Microprogramming

86.1 Number Systems

Richard F. Tinder

Number systems provide the basis for conveying and quantifying information. Weather data, stocks, pagination of books, weights and measures—these are just a few examples of the use of numbers that affect our daily lives. For this purpose we find the decimal (or arabic) number system to be reliable and easy to use. This system evolved presumably because early humans were equipped with a crude type of calculator, their ten fingers. A number system that is appropriate for humans, however, may be intractable for use by a machine such as a computer. Likewise, a number system appropriate for a machine may not be suitable for human use.

Before concentrating on those number systems that are useful in computers, it will be helpful to review the characteristics that are desirable in any number system. There are *four* important characteristics in all:

- Distinguishability of symbols
- Arithmetic operations capability
- Error control capability
- Tractability and speed

To one degree or another the decimal system of numbers satisfies these characteristics for hard-copy transfer of information between humans. Roman numerals and **binary** are examples of number systems that do not satisfy all four characteristics for human use. On the other hand, the binary number system is preferable for use in digital computers. The reason is simply put: current digital electronic machines recognize only two identifiable states physically represented by a high voltage level and a low voltage level. These two physical states are logically interpreted as the binary symbols 1 and 0.

A fifth desirable characteristic of a number system to be used in a computer should be that it have a minimum number of easily identifiable states. The binary number system satisfies this condition. However, the digital computer must still interface with humankind. This is done by converting the binary data to a decimal and character-based form that can be readily understood by humans. A minimum number of identifiable characters (say 1 and 0, or true and false) is not practical or desirable for direct human use. If this is difficult to understand, imagine trying to complete a tax form in binary or in any number system other than decimal. On the other hand, use of a computer for this purpose would not only be practical but, in many cases, highly desirable.

Positional and Polynomial Representations

The *positional form* of a number is a set of side-by-side (juxtaposed) digits given generally in *fixed-point* form as

$$N_r = (\underbrace{\overset{\text{MSD}\downarrow}{a_{n-1}} \ldots a_3 a_2 a_1 a_0}_{\text{Integer}} \overset{\text{Radix Point}\downarrow}{.} \underbrace{a_{-1} a_{-2} a_{-3} \ldots \overset{\text{LSD}\downarrow}{a_{-m}}}_{\text{Fraction}})_r \tag{86.1}$$

where the **radix** (or base) r is the total number of digits in the number system and a is a digit in the set defined for radix r. Here, the radix point separates n integer digits on the left from m fraction digits on the right. Notice that a_{n-1} is the most significant (highest-order) digit, called MSD, and that a_{-m} is the least significant (lowest-order) digit, denoted by LSD.

The *value* of the number in Eq. (86.1) is given in *polynomial form* by

$$\begin{aligned} N_r &= \sum_{i=-m}^{n-1} a_i r^i \\ &= \begin{pmatrix} a_{n-1} r^{n-1} + \cdots + a_2 r^2 + a_1 r^1 + a_0 r^0 \\ + a_{-1} r^{-1} + a_{-2} r^{-2} + \cdots + a_{-m} r^{-m} \end{pmatrix}_r \end{aligned} \tag{86.2}$$

where a_i is the digit in the ith position with a *weight* r^i.

Application of Eqs. (86.1) and (86.2) follows directly. For the decimal system $r = 10$, indicating that there are 10 distinguishable characters recognized as decimal numerals $0, 1, 2, \ldots, r - 1 (= 9)$. Examples of the positional and polynomial representations for the decimal system are

$$\begin{aligned} N_{10} &= (d_3 d_2 d_1 d_0 . d_{-1} d_{-2} d_{-3})_{10} \\ &= 3017.528 \end{aligned}$$

and

$$\begin{aligned} N_{10} &= \sum_{i=-3}^{n-1} d_i 10^i \\ &= 3 \times 10^3 + 0 \times 10^2 + 1 \times 10^1 + 7 \times 10^0 + 5 \times 10^{-1} + 2 \times 10^{-2} + 8 \times 10^{-3} \\ &= 3000 + 10 + 7 + 0.5 + 0.02 + 0.008 \end{aligned}$$

where d_i is the decimal digit in the ith position. Exclusive of possible leading and trailing zeros, the MSD and LSD for this number are 3 and 8, respectively. This number could have been written in a form such as $N_{10} =$ 03017.52800 without altering its value but implying greater accuracy of the fraction portion.

Unsigned Binary Number System

Applying Eqs. (86.1) and (86.2) to the binary system requires that $r = 2$, indicating that there are two distinguishable characters, typically 0 and $(r - 1) = 1$, that are used. In positional representation these characters (numbers) are called *binary digits* or *bits*. Examples of the positional and polynomial notations for a binary number are

$$N_2 = (b_{n-1} \ldots b_3 b_2 b_1 b_0 \,.\, b_{-1} b_{-2} b_{-3} \ldots b_{-m})_2$$
$$= 101101.101_2$$

MSB — LSB

and

$$N = \sum_{i=-m}^{n-1} b_i 2^i$$
$$= 1\times 2^5 + 0\times 2^4 + 1\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}$$
$$= 32 + 8 + 4 + 1 + 0.5 + 0.125$$
$$= 45.625_{10}$$

where b_i is the bit in the ith position. Thus, the bit positions are weighted …, 16, 8, 4, 2, 1, ½, ¼, ⅛, … for any number consisting of integer and fraction portions. Binary numbers so represented are sometimes referred to as *natural* binary. In positional representation the bits on the extreme left and extreme right are called the MSB (most significant bit) and LSB (least significant bit), respectively. Notice that by obtaining the value of a binary number a conversion from binary to decimal has been performed. The subject of radix (base) conversion will be dealt with more extensively later.

For reference purposes Table 86.1 provides the binary-to-decimal conversion for two-, three-, four-, five-, and six-bit binary. The six-bit binary column is only halfway completed for brevity.

TABLE 86.1 Binary-to-Decimal Conversion

Two-Bit Binary	Decimal Value	Three-Bit Binary	Decimal Value	Four-Bit Binary	Decimal Value	Five-Bit Binary	Decimal Value	Six-Bit Binary	Decimal Value
00	0	000	0	0000	0	10000	16	100000	32
01	1	001	1	0001	1	10001	17	100001	33
10	2	010	2	0010	2	10010	18	100010	34
11	3	011	3	0011	3	10011	19	100011	35
		100	4	0100	4	10100	20	100100	36
		101	5	0101	5	10101	21	100101	37
		110	6	0110	6	10110	22	100110	38
		111	7	0111	7	10111	23	100111	39
				1000	8	11000	24	101000	40
				1001	9	11001	25	101001	41
				1010	10	11010	26	101010	42
				1011	11	11011	27	101011	43
				1100	12	11100	28	101100	44
				1101	13	11101	29	101101	45
				1110	14	11110	30	101110	46
				1111	15	11111	31	101111	47
								.	.
								.	.
								.	.

TABLE 86.2 NBCD Bit Patterns and Decimal Equivalent

NBCD Bit Pattern	Decimal	NBCD Bit Pattern	Decimal
0000	0	1000	8
0001	1	1001	9
0010	2	1010	NA
0011	3	1011	NA
0100	4	1100	NA
0101	5	1101	NA
0110	6	1110	NA
0111	7	1111	NA

NA = not allowed.

In the natural binary system the number of bits in a unit of data is commonly assigned a name. Examples are:

- 4-data-bit unit: nibble (or half-byte)
- 8-data-bit unit: byte
- 16-data-bit unit: two bytes (or half-word)
- 32-data-bit unit: word (or four bytes)
- 64-data-bit unit: double-word

  etc.

The word size for a computer is determined by the number of bits that can be manipulated and stored in registers. The foregoing list of names would be applicable to a 32-bit computer.

Unsigned Binary-Coded Decimal, Hexadecimal, and Octal Systems

While the binary system of numbers is most appropriate for use in computers, it has several disadvantages when used by humans who have become accustomed to the decimal system. For example, binary machine code is long, difficult to assimilate, and tedious to convert to decimal. However there exist simpler ways to represent binary numbers for conversion to decimal representation. Three examples, commonly used, are natural binary-coded decimal (NBCD), binary-coded **hexadecimal** (BCH), and binary-coded **octal** (BCO). These number systems are useful in applications where a digital device, such as a computer, must interface with humans. The NBCD code representation is also useful in carrying out computer arithmetic.

The NBCD Representation

The BCD system as used here is actually an 8, 4, 2, 1 weighted code called *natural* BCD or NBCD. This system uses patterns of four bits to represent each decimal position of a number and is one of several such weighted BCD code systems. The NBCD code is converted to its decimal equivalent by polynomials of the form

$$N_{10} = b_3 \times 2^3 + b_2 \times 2^2 + b_1 \times 2^1 + b_0 \times 2^0$$
$$= b_3 \times 8 + b_2 \times 4 + b_1 \times 2 + b_0 \times 1$$

for any $b_3b_2b_1b_0$ code integer. Thus, decimal 6 is represented as $(0\times8) + (1\times4) + (1\times2) + (0\times1)$, or 0110 in NBCD code. Like natural binary, NBCD code is also called "natural" because its bit positional weights are derived from integer powers of 2^n. Table 86.2 shows the NBCD bit patterns for decimal integers 0 through 9.

The NBCD code is currently the most widely used of the BCD codes. There are many excellent sources of information on BCD codes. One, in particular, provides a fairly extensive coverage of both weighted and unweighted BCD codes [Tinder, 1991].

Decimal numbers greater than 9 or less than 1 can be represented by the NBCD code if each digit is given in that code and if the results are combined. For example, the number 63.98 is represented by (or converted to) NBCD code as

$$\begin{aligned} & \quad 6 \quad 3 \quad . \quad 9 \quad 8 \\ 63.98_{10} &= 0110 \;\; 0011 \;\; . \;\; 1001 \;\; 1000)_{\text{NBCD}} \\ &= 1100011.10011_{\text{NBCD}} \end{aligned}$$

Here, the code weights are 80, 40, 20, 10; 8, 4, 2, 1; 0.8, 0.4, 0.2, 0.1; and 0.08, 0.04, 0.02, 0.01 for the tens, units, tenths, and hundredths digits, respectively, representing four decades. Conversion between binary and NBCD requires conversion to decimal as an intermediate step. For example, to convert from NBCD to binary requires that groups of four bits be selected in both directions from the radix point to form the decimal number. If necessary, zeros are added to the leftmost or rightmost ends to complete the groups of four bits as in the above example. Negative NBCD numbers can be represented either in sign-magnitude notation or 1's or 2's **complement** notation as discussed later.

Another BCD code that is used for number representation and manipulation is called excess 3 BCD (or XS3 NBCD, or simply XS3). XS3 is an example of a *biased-weighted* code (a bias of 3). This code is formed by adding 0011_2 ($= 3_{10}$) to the NBCD bit patterns in Table 86.2. Thus, to convert XS3 to NBCD code, 0011 must be subtracted from XS3 code. In four-bit quantities the XS3 code has the useful feature that when adding two numbers together in XS3 notation a carry will result and yield the correct value any time a carry results in decimal (i.e., when 9 is exceeded). This feature is not shared by either natural binary or NBCD addition.

The Hexadecimal and Octal Systems

The hexadecimal number system requires that $r = 16$ in Eqs. (86.1) and (86.2), indicating that there are 16 distinguishable characters in the system. By convention, the permissible hexadecimal digits are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, and F for decimals 0 through 15, respectively. Examples of the positional and polynomial representations for a hexadecimal number are

$$\begin{aligned} N_{16} &= (h_{n-1} \ldots h_3 h_2 h_1 h_0 \; . \; h_{-1} h_{-2} h_{-3} \ldots h_{-m})_{16} \\ &= (\text{AF3.C8})_{16} \end{aligned}$$

with a decimal value of

$$\begin{aligned} N &= \sum_{i=-m}^{n-1} h_i 16^i \\ &= 10 \times 16^2 + 15 \times 16^1 + 3 \times 16^0 + 12 \times 16^{-1} + 8 \times 16^{-2} \\ &= 2803.78125_{10} \end{aligned}$$

Here, it is seen that a hexadecimal number has been converted to decimal by using Eq. (86.2).

The octal number system requires that $r = 8$ in Eqs. (86.1) and (86.2), indicating that there are eight distinguishable characters in this system. The permissible octal digits are 0, 1, 2, 3, 4, 5, 6, and 7, as one might expect. Examples of the application of Eqs. (86.1) and (86.2) are

$$\begin{aligned} N_8 &= (o_{n-1} \ldots o_3 o_2 o_1 o_0 \; . \; o_{-1} o_{-2} o_{-3} \ldots o_{-m})_8 \\ &= 501.74_8 \end{aligned}$$

with a decimal value of

TABLE 86.3 The BCH and BCO Number Systems

Binary	BCH	BCO	Decimal	Binary	BCH	BCO	Decimal
0000	0	0	0	1010	A	12	10
0001	1	1	1	1011	B	13	11
0010	2	2	2	1100	C	14	12
0011	3	3	3	1101	D	15	13
0100	4	4	4	1110	E	16	14
0101	5	5	5	1111	F	17	15
0110	6	6	6	10000	10	20	16
0111	7	7	7	11011	1B	33	27
1000	8	10	8	110001	31	61	49
1001	9	11	9	1001110	4E	116	78

$$
\begin{aligned}
N &= \sum_{i=-m}^{n-1} o_i 8^i \\
&= 5 \times 8^2 + 0 \times 8^1 + 1 \times 8^0 + 7 \times 8^{-1} + 4 \times 8^{-2} \\
&= 321.9375_{10}
\end{aligned}
$$

When the hexadecimal and octal number systems are used to represent bit patterns in binary, they are called binary-coded hexadecimal (BCH) and binary-coded octal (BCO), respectively. These two number systems are examples of *binary-derived radices*. Table 86.3 lists several selected examples showing the relationships between BCH, BCO, binary, and decimal.

What emerges on close inspection of Table 86.3 is that each hexadecimal digit corresponds to four binary digits and that each octal digit corresponds to three binary digits. The following example illustrates the relationships between these number systems:

$$
\begin{aligned}
10110111111.11011_2 &= \begin{array}{cccccccc} 5 & B & F & . & D & 8 \\ 0101 & 1011 & 1111 & . & 1101 & 1000 \end{array} \\
&= 5BF.D8_{16} \\
&= \begin{array}{cccccccc} 2 & 6 & 7 & 7 & . & 6 & 6 \\ 010 & 110 & 111 & 111 & . & 110 & 110 \end{array} \\
&= 2677.66_8 \\
&= 1471.84375_{10}
\end{aligned}
$$

To separate the binary digits into groups of four (for BCH) or groups of three (for BCO), counting must begin from the radix point and continue outward in both directions. Then, where needed, zeros are added to the leading and trailing ends of the binary representation to complete the MSDs and LSDs for the BCH and BCO forms.

Conversion between Number Systems

It is not the intent of this section to cover all methods for radix (base) conversion. Rather, the plan is to provide general approaches, separately applicable to the integer and fraction portions, followed by specific examples.

Conversion of Integers

Since the polynomial form of Eq. (86.2) is a geometrical progression, the integer portion can be represented in *nested radix* form. In source radix *s*, the nested representation is

$$
\begin{aligned}
N_s &= \left(a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_1 s^1 + a_0 s^0\right)_s \\
&= a_0 + s(a_1 + s(a_2 + \cdots + a_{n-1})))))_s \\
&= a_0 + s\left(\sum_{i=1}^{n-1} a_i s^{i-1}\right)
\end{aligned}
\tag{86.3}
$$

for digits a_i having integer values from 0 to $s-1$. The nested radix form not only suggests a conversion process but also forms the basis for computerized conversion.

Consider that the number in Eq. (86.3) is to be represented in nested radix r form

$$
\begin{aligned}
N_r &= b_0 + r(b_1 + r(b_2 + \cdots + b_{m-1})))))_r \\
&= b_0 + r\left(\sum_{i=1}^{m-1} b_i r^{i-1}\right)
\end{aligned}
\tag{86.4}
$$

where, in general, $m \neq n$. Then, if N_s is divided by r, the results are of the form

$$
\frac{N_s}{r} = Q + \frac{R}{r} \tag{86.5}
$$

where Q is the integer quotient rearranged as $Q_0 = b_1 + r(b_2 + \cdots + b_{m-1})))_r$ and R is the remainder $R_0 = b_0$. A second division by r yields $Q_0/r = Q_1 + R_1/r$, where Q_1 is arranged as $Q_1 = b_2 + r(b_3 + \cdots + b_{m-1}))_r$ and $R_1 = b_1$. Thus, by repeated division of the integer result Q_i by r, the remainders yield $(b_0, b_1, b_2, \ldots, b_{m-1})_r$ in that order.

The conversion method just described, called the *radix divide method,* can be used to convert between any two integers of different radices. However, the requirement is that *the arithmetic required by N_s/r must be carried out in source radix, s.* Except for source radices 10 and 2, this poses a severe problem for humans. Table 86.4 provides the recommended procedures for integer conversion. The radix divide method is suitable for computer conversion providing, of course, that the computer is programmed to carry out the arithmetic in different radices.

TABLE 86.4 Summary of Recommended Methods for Integer Conversion by Noncomputer Means

Integer Conversion	Conversion Method
$N_{10} \rightarrow N_r$	Radix division by radix r using Eq. (86.5)
$N_s \rightarrow N_{10}$	Eq. (86.2) or Eq. (86.3)
$N_s)_{s\neq 10} \text{—>} N_r)_{r\neq 10}$	$N_s \rightarrow N_{10}$ by Eq. (86.2) or (86.3) $N_{10} \rightarrow N_r$ radix division by r using Eq. (86.5)
Special Cases for Binary Forms	
$N_2 \rightarrow N_{10}$	Positional weighting
$N_2 \rightarrow N_{BCH}$	Partition N_2 into groups of four bits starting from radix point, then apply Table 86.3
$N_2 \rightarrow N_{BCO}$	Partition N_2 into groups of three bits starting from radix point, then apply Table 86.3
$N_{BCH} \rightarrow N_2$	Reverse of $N_2 \rightarrow N_{BCH}$
$N_{BCO} \rightarrow N_2$	Reverse of $N_2 \rightarrow N_{BCO}$
$N_{BCH} \rightarrow N_{BCO}$	$N_{BCH} \rightarrow N_2 \rightarrow N_{BCO}$
$N_{BCO} \rightarrow N_{BCH}$	$N_{BCO} \rightarrow N_2 \rightarrow N_{BCH}$
$N_{NBCD} \rightarrow N_{XS3}$	Add 0011_2 $(= 3_{10})$ to N_{NBCD}
$N_{XS3} \rightarrow N_{NBCD}$	Subtract 0011_2 $(= 3_{10})$ from N_{NBCD}

The integer conversion methods of Table 86.4 can be illustrated by the following simple examples:

Example 1. $139_{10} \rightarrow N_2$

N/r	Q	R	
$139/2 =$	69	1	
$69/2 =$	34	1	
$34/2 =$	17	0	
$17/2 =$	8	1	
$8/2 =$	4	0	
$4/2 =$	2	0	
$2/2 =$	1	0	
$1/2 =$	0	1	$139_{10} = 10001011_2$

Example 2. $10001011_2 \rightarrow N_{10}$. By positional weights,

$$N_{10} = 128 + 8 + 2 + 1 = 139_{10}$$

Example 3. $139_{10} \rightarrow N_8$

N/r	Q	R	
$139/8 =$	17	3	
$17/8 =$	2	1	
$2/8 =$	0	2	$139_{10} = 213_8$

Example 4. $10001011_2 \rightarrow N_{\text{BCO}}$

2	1	3	
010	001	011	$= 213_{\text{BCO}}$

Example 5. $213_{\text{BCO}} \rightarrow N_{\text{BCH}}$

2	1	3			8	B	
$213_{\text{BCO}} =$ 010	001	011	$= 10001011_2 =$	1000	1011	$= 8B_{16}$	

Example 6. $213_8 \rightarrow N_5$

$$213_8 = 2 \times 8^2 + 1 \times 8^1 + 3 \times 8^0 = 139_{10}$$

N/r	Q	R	
$139/5 =$	27	4	
$27/5 =$	5	2	
$5/5 =$	1	0	
$1/5 =$	0	1	$213_8 = 1024_5$

Check: $1 \times 5^3 + 2 \times 5^1 + 4 \times 5^0 = 125 + 10 + 4 = 139_{10}$

Conversion of Fractions

By extracting the fraction portion from Eq. (86.2) one can write

$$\begin{aligned} .N_s &= (a_{-1}s^{-1} + a_{-2}s^{-2} + \cdots + a_{-m}s^{-m})_s \\ &= s^{-1}(a_{-1} + s^{-1}(a_{-2} + \cdots + a_{-m})))))_s \\ &= s^{-1}(a_{-1} + \sum_{i=2}^{m} a_{-i}s^{-i+1})_s \end{aligned} \tag{86.6}$$

in radix s. This is called the *nested inverse radix* form that provides the basis for computerized conversion.

TABLE 86.5 Summary of Recommended Methods for Fraction Conversion by Noncomputer Means

Fraction Conversion	Conversion Method
$.N_{10} \rightarrow .N_r$	Radix multiplication by using Eq. (86.8)
$.N_s \rightarrow .N_{10}$	Equation (86.2) or Eq. (86.6)
$.N_r)_{s\neq 10} \rightarrow .N_r)_{r\neq 10}$	$N_s \rightarrow N_{s10}$ by Eq. (86.2) or Eq. (86.6) $N_{10} \rightarrow N_r$ radix multiply by Eq. (86.8)
Special Cases for Binary Forms	
$.N_2 \rightarrow .N_{BCH}$	Partition $.N_2$ into groups of four bits from radix point, then apply Table 86.3
$.N_2 \rightarrow .N_{BCO}$	Partition $.N_2$ into groups of three bits from radix point, then apply Table 86.3
$.N_{BCH} \rightarrow .N_2$	Reverse of $.N_2 \rightarrow .N_{BCH}$
$.N_{BCO} \rightarrow .N_2$	Reverse of $.N_2 \rightarrow .N_{BCO}$
$.N_{BCH} \rightarrow .N_{BCO}$	$.N_{BCH} \rightarrow .N_2 \rightarrow .N_{BCO}$
$.N_{BCO} \rightarrow .N_{BCH}$	$.N_{BCO} \rightarrow .N_2 \rightarrow .N_{BCH}$

If the fraction in Eq. (86.6) is represented in nested inverse radix r form, then

$$\begin{aligned} .N_r &= r^{-1}(b^{-1} + r^{-1}(b^{-2} + \cdots + b^{-p})))))_r \\ &= r^{-1}(b_{-1} + \sum_{i=2}^{p} b_{-i} r^{-i+1})_r \end{aligned} \tag{86.7}$$

for any fraction represented in radix r. Now, if N_s is multiplied by r, the result is of the form

$$.N_s \times r = I + F \tag{86.8}$$

where I is the product integer, $I_1 = b_{-1}$, and F_0 is the product fraction arranged as $F_1 = r^{-1}(b_{-2} + r^{-1}(b_{-3} + \ldots + b_{-p})))_r$. By repeated multiplication by r of the remaining fractions F_i, the resulting integers yield $(b_{-1}, b_{-2}, b_{-3}, \ldots b_{-m})_r$, in that order.

The conversion just described is called the *radix multiply method* and is perfectly general for converting between fractions of different radices. However, as in the case of integer conversion, the requirement is that *the arithmetic required by $.N_s \times r$ must be carried out in source radix, s.* For noncomputer use by humans, this procedure is usually limited to fraction conversions $N_{10} \rightarrow N_r$, where the source radix is 10 (decimal). The recommended methods for converting between fractions of different radices are given in Table 86.5. The radix multiply method is well suited to computer use.

For any integer of radix s, there exists an exact representation in radix r. This is not the case for a fraction whose conversion is a geometrical progression that never converges. Terminating a fraction conversion at n digits (to the right of the radix point) results in an error or uncertainty. In decimal, this error is given by

$$\begin{aligned} \epsilon_{10} &= a_{-n} r^{-n} + a_{-(n+1)} r^{-(n+1)} + a_{-(n+2)} r^{-(n+2)} + \cdots \\ &= r^{-n} \left[a_{-n} + \sum_{i=1}^{\infty} a_{-(n+i)} r^{-(n+i)} \right]_r \end{aligned}$$

where the quantity in brackets approaches the value of $a_{-n} + 1$. Therefore, terminating a fraction conversion at n digits from the radix point results in an error with bounds

$$0 < \epsilon_{10} \leq r^{-n}(a_{-n} + 1) \tag{86.9}$$

in decimal. Equation (86.9) is useful in deciding when to terminate a fraction conversion.

Often, it is desirable to terminate a fraction conversion at $(n + 1)$ digits and then round off to n from the radix point. A suitable method for rounding to n digits in radix r is: Perform the fraction conversion to $(n + 1)$ digits from the radix point, then drop the $(n + 1)$ digit if $a_{-(n+1)} < r/2$, or add $r^{-(n-1)}$ to the result if $a_{-(n+1)} \geq r/2$.

After rounding off to n digits, the maximum error becomes the difference between the rounded result and the smallest value possible. By using Eq. (86.9), this difference is

$$\begin{aligned}\epsilon_{\max} &= r^{-n}(a_{-n} + 1) - r^{-n}(a_{-n} + a_{-(n+1)}/r) \\ &= r^{-n}(1 - a_{-(n+1)}/r)\end{aligned}$$

Then, by rounding to n digits, there results an error with bounds

$$0 < \epsilon_{10} \leq r^{-n}(1 - a_{-(n+1)}/r) \tag{86.10}$$

in decimal. If $a_{-(n+1)} < r/2$ and the $(n + 1)$ digit is dropped, the maximum error is r^{-n}. Note that for $N_s \rightarrow N_{10} \rightarrow N_r$ type conversions, the bounds of errors aggregate.

The following examples illustrate the fraction conversion methods of Table 86.5.

Example 7. $0.654_{10} \rightarrow N_2$ rounded to eight bits

$.N_s \times r$	F	I	
0.654×2	0.308	1	
0.308×2	0.616	0	
0.616×2	0.232	1	
0.232×2	0.464	0	
0.464×2	0.928	0	
0.928×2	0.856	1	$0.654_{10} = 0.10100111_2$
0.856×2	0.712	1	
0.712×2	0.424	1	
0.424×2	0.848	0	$\epsilon_{\max} = 2^{-8}$

Example 8. $0.654_{10} \rightarrow N_8$ terminated at four digits

$.N_s \times r$	F	I	
0.654×8	0.232	5	
0.232×8	0.856	1	$0.654_{10} = 5166_8$
0.856×8	0.848	6	with error bounds
0.848×8	0.784	6	$0 < \epsilon_{10} \leq 7 \times 8^{-4} = 1.71 \times 10^{-3}$

Example 9. $0.5166_8 \rightarrow N_2$ rounded to eight bits and let $0.5166_8 \rightarrow N_{10}$ be rounded to four decimal places.

$$\begin{aligned}0.5166_8 &= 5 \times 8^{-1} + 1 \times 8^{-2} + 6 \times 8^{-3} + 6 \times 8^{-4} \\ &= 0.625000 + 0.015625 + 0.011718 + 0.001465 \\ &= 0.6538 \text{ rounded to four decimal places; } \epsilon_{10} \leq 10^{-4}\end{aligned}$$

$.N_s \times r$	F	I
0.6538×2	0.3076	1
0.3076×2	0.6152	0
0.6152×2	0.2304	1
0.2304×2	0.4608	0
0.4608×2	0.9216	0

$0.9216 \times 2 \quad 0.8432 \quad 1$
$0.8432 \times 2 \quad 0.6864 \quad 1 \qquad 0.5166_8 = 0.10100111_2$(compare with Example 7)
$0.6864 \times 2 \quad 0.3728 \quad 1$
$0.3728 \times 2 \quad 0.7457 \quad 0 \qquad \epsilon_{10} \leq 10^{-4} + 2^{-8} = 0.0040$

Example 10. $0.10100111_2 \rightarrow N_{BCH}$

$$0.10100111_2 = \underset{\text{A}}{0.1010} \cdot \underset{7}{0111} = 0.\text{A7}_{BCH}$$

Signed Binary Numbers

To this point only unsigned numbers (assumed to be positive) have been considered. However, both positive and negative numbers must be used in computers. Several schemes have been devised for dealing with negative numbers in computers, but only four are commonly used:

- Signed-magnitude representation
- Radix complement representation
- Diminished radix complement representation
- Excess (offset) code representation

Of these, the radix 2 complement representation, called 2's complement, is the most widely used system in computers.

Signed-Magnitude Representation

A signed-magnitude number consists of a magnitude together with a symbol indicating its sign (positive or negative). Such a number lies in the decimal range of $-(r^{n-1} - 1)$ through $+(r^{n-1} - 1)$ for n integer digits in radix r. A fraction portion, if present, would consist of m digits to the right of the radix point.

The most common examples of signed-magnitude numbers are those in the decimal and binary systems. The sign symbols for decimal (+ or –) are well known. In binary it is established practice to use 0 = plus and 1 = minus for the sign symbols and to place one of them in the MSB position for each number. Examples in eight-bit binary are

$$+45.5_{10} = \underset{\text{Sign bit}}{0} \quad \overbrace{101101.1_2} \quad +0_{10} = 0 \;\; 0000000_2$$

$$-123_{10} = \underset{\text{Sign bit}}{1} \quad \overbrace{1111011_2} \quad -0_{10} = 1 \;\; 0000000_2$$

Although the sign-magnitude system is used in computers, it has two drawbacks. There is no unique zero, as indicated by the examples, and addition and subtraction calculations require time-consuming decisions regarding operation and sign as, for example, (–7) minus (–4). Even so, the sign-magnitude representation is commonly used in **floating-point** number systems.

Radix Complement Representation

The *radix complement* of an n-digit number N_r is obtained by subtracting it from r^n, that is $r^n - N_r$. The operation $r^n - N_r$ is equivalent to complementing the number and adding 1 to the LSD. Thus, the radix complement is $\overline{N}_r + 1_{LSD}$ where $\overline{N}_r = r^n - 1 - N_r$ is the complement of a number in radix r. Therefore, one may write

$$\text{Radix complement of } N_r = r^n - N_r$$

$$= \overline{N}_r + 1 \tag{86.11}$$

The complements $\overline{N}_r$ for digits in three commonly used number systems are given in Table 86.6. Notice that the complement of a binary number is formed simply by replacing the 1's with 0's and 0's with 1's as required by $2^n - 1 - N_2$.

With reference to Table 86.6 and Eq. (86.11), the following examples of radix complement representation are offered.

Example 11. The 10's complement of 47.83 is

$$\overline{N}_{10} + 1_{\text{LSD}} = 52.17$$

Example 12. The 2's complement of 0101101.101 is

$$\overline{N}_2 + 1_{\text{LSB}} = 1010010.011$$

Example 13. The 16's complement of A3D is

$$\overline{N}_{16} + 1_{\text{LSD}} = 5\text{C}2 + 1 = 5\text{C}3$$

The decimal value of Eq. (86.11) can be found from the polynomial expression

$$N_{\text{radix compl.}})_{10} = -(a_{n-1}r^{n-1}) + \sum_{i=-m}^{n-2} a_i r^i \tag{86.12}$$

for any n-digit number of radix r. In Eqs. (86.11) and (86.12) the MSD is taken to be the position of the sign symbol.

2's Complement Representation. The radix complement for binary is the 2's complement representation. In 2's complement the MSB is the sign bit, 1 indicating a negative number or 0 if positive. The decimal range of representation for n-integer bits in 2's complement is from $-(2^{n-1})$ through $+(2^{n-1})$. From Eq. (86.11), the 2's complement is formed by

$$N_2)_{\text{2's compl.}} = 2^n - N_2 = \overline{N}_2 + 1 \tag{86.13}$$

A few examples in eight-bit binary are shown in Table 86.7. Notice that application of Eq. (86.13) changes the sign of the decimal value of a binary number (+ to –, and vice versa) and that only one zero representation exists.

Application of Eq. (86.12) gives the decimal value of any 2's complement number, including those containing a radix point. For example, the pattern $N_{\text{2's compl.}} = 11010010.011$ has a decimal value

$$\begin{aligned} N_{\text{2's compl.}})_{10} &= -1 \times 2^7 + 1 \times 2^6 + 1 \times 2^4 + 1 \times 2^1 + 1 \times 2^{-2} + 1 \times 2^{-3} \\ &= -128 + 64 + 16 + 2 + 0.25 + 0.125 \\ &= -45.625_{10} \end{aligned}$$

The same result could have easily been obtained by first applying Eq. (86.13) to $N_{\text{2's compl.}}$ followed by the use of positional weighting to obtain the decimal value. Thus,

TABLE 86.6 Complements for Three Commonly Used Number Systems

	Complement ($-N_r$)		
Digit	Binary	Decimal	Hexadecimal
0	1	9	F
1	0	8	E
2		7	D
3		6	C
4		5	B
5		4	A
6		3	9
7		2	8
8		1	7
9		0	6
A			5
B			4
C			3
D			2
E			1
F			0

TABLE 86.7 Examples of Eight-Bit 2's and 1's Complement Representations (MSB = Sign Bit)

Decimal Value	2's Complement	1's Complement
–128	10000000	
–127	10000001	10000000
–31	11100001	11100000
–16	11110000	11101111
–15	11110001	11110000
–3	11111101	11111100
–0	00000000	11111111
+0	00000000	00000000
+3	00000011	00000011
+15	00001111	00001111
+16	00010000	00010000
+31	00011111	00011111
+127	01111111	01111111
+128		

$$\begin{aligned} N_{\text{2's compl.}} &= 00101101.101 \\ &= 32 + 8 + 5 + 0.5 + 0.125 \\ &= 45.625_{10} \end{aligned}$$

which is known to be a negative number, -45.625_{10}.

Negative NBCD numbers can be represented in 2's complement. The foregoing discussion on 2's complement applies to NBCD with consideration of how NBCD is formed from binary. As an example, -59.24_{10} is represented by

$$0101\ \ 1001.0010\ \ 0100)_{\text{NBCD}} = 10100110.11011100)_{\text{2's compl. NBCD}}$$

In a similar fashion, negative NBCD numbers can also be represented in 1's complement following the procedure given in the next paragraph. Sign-magnitude representation of a negative NBCD number simply requires the addition of a sign bit to the NBCD magnitude.

Diminished Radix Complement Representation

The diminished radix complement of a number is obtained by

$$\begin{aligned} N_r)_{\text{dim. rad. compl.}} &= r^n - N_r - 1 \qquad (86.14) \\ &= \overline{N}_{\text{r}} \end{aligned}$$

Thus, the complement of a number is its diminished radix complement. It also follows that the radix complement of a number is the diminished radix complement with 1 added to the LSD as in Eq. (86.13). The range of representable numbers is $-(r^{n-1} - 1)$ through $+(r^{n-1} - 1)$ for radix r.

In the binary and decimal number systems, the diminished radix complement representations are the 1's complement and 9's complement, respectively. Examples of 1's complement are shown in Table 86.7 for comparison with those of 2's complement. Notice that in 1's complement there are two representations for zero, one for +0 and the other for –0. This fact limits the usefulness of the 1's complement representation for computer arithmetic.

Excess (Offset) Representations

Other systems for representing negative numbers use *excess* or *offset* codes. Here, a bias B is added to the true value N_r of the number to produce an excess number N_{xs} given by

$$N_{xs} = N_r + B \tag{86.15}$$

When $B = r^{n-1}$ exceeds the usable bounds of negative numbers, N_{xs} remains positive. Perhaps the most common use of the excess representation is in floating-point number systems—the subject of the next section.

Two examples are given below in eight-bit excess 128 code.

Example 14.

-43_{10}	11010101	$N_{2\text{'s compl.}}$
$+128_{10}$	10000000	B
85_{10}	01010101	$N_{xs} = -43_{10}$ in excess 128 code

Example 15.

27_{10}	00011011	$N_{2\text{'s compl.}}$
$+128_{10}$	10000000	B
155_{10}	10011011	$N_{xs} = 27_{10}$ in excess 128 code

The representable decimal range for an excess 2^{n-1} number system is -2^{n-1} through $+(2^{n-1} - 1)$ for an n-bit binary number. However, if $N_2 + B > 2^{n-1} - 1$, *overflow* occurs and 2^{n-1} must be subtracted from $(N_2 + B)$ to give the correct result in excess 2^{n-1} code.

Floating-Point Number Systems

In fixed-point representation [Eq. (86.1)], the radix point is assumed to lie immediately to the right of the integer field and at the left end of the fraction field. The fixed-point system is the most commonly used system for representing bounded orders of magnitude. For example, with 32 bits a binary number could represent decimal numbers with upper and lower bounds of the order of $\pm 10^{10}$ and $\pm 10^{-10}$. However, for greatly expanded bounds of representation, as in scientific notation, the *floating-point* representation is needed.

A floating-point number (FPN) in radix r has the general form

$$\text{FPN})_r = F \times r^E \tag{86.16}$$

where F is the *fraction* (or **mantissa**) and E is the *exponent*. Only fraction digits are used for the mantissa! Take, for example, Planck's constant $h = 6.625 \times 10^{-34}$ J·s. This number can be represented many different ways in floating point notation:

$$\begin{aligned} \text{Planck's constant } h &= 0.625 \times 10^{-33} \\ &= 0.0625 \times 10^{-32} \\ &= 0.00625 \times 10^{-31} \end{aligned}$$

All three adhere to the form of Eq. (86.16) and are, therefore, legitimate floating-point numbers in radix 10. Thus, as the radix point *floats* to the left, the exponent is *scaled* accordingly. The first form for h is said to be *normalized* because the MSD of F is nonzero, a means of standardizing the radix point position. Notice that the sign for F is positive while that for E is negative.

In computers the FPN is represented in binary where the normalized representation requires that the MSB for F always be 1. Thus, the range in F in decimal is

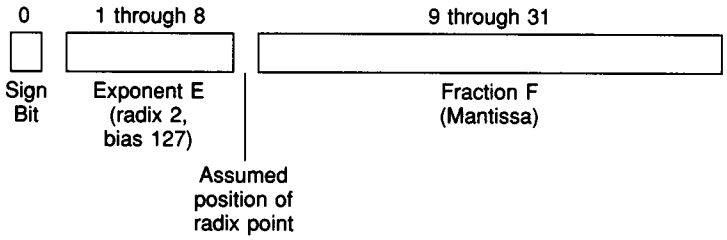


FIGURE 86.1 IEEE standard bit format for normalized floating-point representation.

$$0.5 \le F < 1$$

Also, the mantissa F is represented in sign-magnitude from. The normalized format for a 32-bit floating-point number in binary, which agrees with the IEEE standard [IEEE, 1985], is shown in Fig. 86.1. Here, the sign bit (1 if negative or 0 if positive) is placed at bit position 0 to indicate the sign of the fraction. Notice that the radix point is assumed to lie between bit positions 8 and 9 to separate the E bit-field from the F bit-field.

Before two FPNs can be added or subtracted in a computer, the E fields must be compared and equalized and the F fields adjusted. The decision-making process can be simplified if all exponents are converted to positive numbers by using the excess representation given by Eq. (86.15). For a q-digit number in radix r, the exponent in Eq. (86.16) becomes

$$E_{xs} = E_r + r^{q-1} \tag{86.17}$$

where E is the actual exponent augmented by a bias of $B = r^{q-1}$. The range in the actual exponent E_r is usually taken to be

$$-(r^{q-1} - 1) \le E_r \le +(r^{q-1} - 1)$$

In the binary system, required for computer calculations, Eq. (86.17) becomes

$$E_{xs} = E_2 + 2^{q-1} \tag{86.18}$$

with a range in actual exponent of $-(2^{q-1}-1) \le E_2 \le +(2^{q-1} - 1)$. In 32-bit normalized floating-point form, the exponent is stored in excess 128 code, while the number is stored in sign-magnitude form.

There still remains the question of how the number 0 is to be represented. If the F field is zero, then the exponent can be anything and the number will be zero. However, in computers the normalized FPN_2 limits F to $0.5 \le F < 1$ since the MSB for F is always 1. The solution to this problem is to assume that the number is zero if the exponent bits are all zero regardless of the value of the mantissa. This leads, however, to a discontinuity in normalized FPN_2 representation at the low end.

The IEEE standard for normalized FPN_2 representation attempts to remove the problem just described. The IEEE system stores the exponent in excess $2^{q-1} - 1$ code and limits the decimal range of the actual exponent to

$$-(2^{q-1} - 2) \le E_2 \le +(2^{q-1} - 1)$$

For 32-bit FPN representation, the exponent is stored in excess 127 code as indicated in Fig. 86.1. Thus, the allowable range of representable exponents is from

$$-126_{10} = 00000001_2 \quad \text{through} \quad +127_{10} = 11111110_2$$

This system reserves the use of all 0's or all 1's in the exponent for special conditions [IEEE, 1985; Pollard, 1990]. So that the F field magnitude can diminish linearly to zero when $E = -126$, the MSB = 1 for F is not specifically represented in the IEEE system but is implied.

The following example attempts to illustrate the somewhat confusing aspects of the IEEE normalized representation:

The number 101101.11001_2 is to be represented in IEEE normalized FPN_2 notation.

$$101101.11001_2 = .10110111001 \times 2^6$$

$$\text{Sign bit} = 0 \text{ (positive)}$$

$$E_{xs} = 6 + 127 = 133_{10} = 10000101_2$$

$$F = 0110111001\ldots00 \text{ (the MSB} = 1 \text{ is not shown)}$$

Therefore, the IEEE normalized FPN is

$$FPN_2 = 0 \quad 10000101 \quad 0110111001\ldots0$$

Still other forms of FPNs are in use. In addition to the IEEE system, there are the IBM, Cray, and DEC systems of representation, each with its own single- and double-precision forms.

Defining Terms

Binary: Representation of quantities in base 2.
Complement: Opposite form of a number system.
Floating point: Similar to "scientific notation" except used to represent binary operations in a computer.
Hexadecimal: Base 16 number system.
Mantissa: Fraction portion of a floating-point number.
Octal: Base 8 number system.
Radix: Base to which numbers are represented.

Related Topic

86.2 Computer Arithmetic

References

H.L. Garner, "Number systems and arithmetic," in *Advances in Computers*, vol. 6, F.L. Alt et al., Eds., New York: Academic, 1965, pp. 131–194.

IEEE, *IEEE Standard for Binary Floating-Point Arithmetic*, ANSI/IEEE Std. 754–1985.

D.E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms*, vol. 2, Reading, Mass: Addison-Wesley, 1969.

C. Tung, "Arithmetic," in *Computer Science*, A.F. Cardenas et al., Eds., New York: Wiley-Interscience, 1972, chap. 3.

Further Information

K. Hwang, *Computer Arithmetic*, New York: Wiley, 1978.

L.H. Pollard, *Computer Design and Architecture*, Englewood Cliffs, N.J.: Prentice-Hall, 1990.

R.F. Tinder, *Digital Engineering Design: A Modern Approach*, Englewood Cliffs, N.J.: Prentice-Hall, 1991.

86.2 Computer Arithmetic

Vojin G. Oklobdzija

As the ability to perform computation increased from the early days of computers up to the present, so has the knowledge of how to utilize the hardware and software to perform computation. Digital computer arithmetic emerged from that period in two ways: as an aspect of logic design and as a development of efficient **algorithms** to utilize the available hardware.

Given that numbers in a digital computer are represented as a string of zeros and ones and that hardware can perform only a relatively simple and primitive set of Boolean operations, all the arithmetic operations performed are based on a hierarchy of operations that are built upon the very simple ones.

What distinguishes computer arithmetic is its intrinsic relation to technology and the ways things are designed and implemented in a digital computer. This comes from the fact that the value of a particular way to compute, or a particular algorithm, is directly evaluated from the actual speed with which this computation is performed. Therefore, there is a very direct and strong relationship between the technology in which digital logic is implemented to compute and the way the computation is structured. This relationship is one of the guiding principles in the development of computer arithmetic.

The subject of computer arithmetic can be, for simpler treatment, divided into number representation; basic, arithmetic operations (such as addition, multiplication, and division), and evaluation of functions.

Number Representation

The only way to represent information in a digital computer is via a string of bits, i.e., zeros and ones. The number of bits being used depends on the length of the *computer word,* which is a quantity of bits on which hardware is capable of operating (sometimes also a quantity that is brought to the CPU from memory in a single access). The first question is what relationship to use in establishing correspondence between those bits and a number. Second, we need to make sure that certain properties that exist in the corresponding **number representation system** are satisfied and that they directly correspond to the operations being performed in hardware over the taken string of bits.

This relationship is defined by the rule that associates one numerical value designated as X (in the text we will use capital X for the numerical value) with the corresponding bit string designated as x.

$$x = \{x_{n-1}, x_{n-2}, \ldots, x_0\}$$

where

$$x_i \in 0, 1$$

In this case the associated word (the string of bits) is n bits long.

When for every value X there exists one and only one corresponding bit string x, we define the number system as **nonredundant.** If however, we could have more than one bit string x that represents the same value X, the number system is **redundant.**

Most commonly we are using numbers represented in a *weighted* number system, where a numerical value is associated with the bit string x according to the equation

$$x = \sum_{i=0}^{n-1} x_i w_i$$

where

$$w_0 = 1 \qquad \text{and} \qquad w_i = (w_i - 1)(r_i - 1)$$

TABLE 86.8 The Relationship between the Implicit Value and the Explicit Value

Implied Attributes: Radix Point, Negative Number Representation, Others	Expression for Implicit Value X_i as a Function of Explicit Value x_e	Numerical Implicit Value X_i (in Decimal)
Integer magnitude	$X_i = x_e$	27
Integer, two's complement	$X_i = -2^5 + x_e$	–5
Integer, one's complement	$X_i = -(2^5 - 1) + x_e$	–4
Fraction, magnitude	$X_i = -2^{-5}x_e$	27/32
Fraction, two's complement	$X_i = -2^{-4}(2^{-5} + x_e)$	–5/16
Fraction, one's complement	$X_i = -2^{-4}(2^{-5} + 1 + x_e)$	–4/16

Source: A. Avizienis, "Digital computer arithmetic: A unified algorithmic specification," in *Symp. Computers and Automata,* Polytechnic Institute of Brooklyn, April 13–15, 1971.

The value r_i is an integer designated as the *radix*, and in a nonredundant number system it is an integer equal to the number of allowed values for x_i. In general x_i could consist of more than one bit. The numerical value associated with x is designated as the *explicit value* of x_e. In conventional number systems the radix r_i is the same positive integer for all the digit positions x_i and with the canonical set of digit values

$$\sum i = \{0, 1, 2, 3, \ldots, r_i - 1\} \qquad \text{for } 0 \le i \le n - 1$$

An example of a weighted number system with a mixed radix would be the representation of time in weeks, days, hours, minutes, and seconds with a range for representing 100 weeks:

$$r = 10, 10, 7, 24, 60, 60$$

In digital computers the radices encountered are 2, 4, 10, and 16, with 2 being the most commonly used one.

The digit set x_i can be *redundant* and *nonredundant.* If the number of different values x_i can assume is $n_x \le r$, then we have a *nonredundant* digit set. Otherwise, if $n_x > r$, we have a *redundant* digit set. Use of the *redundant* digit set has its advantages in efficient implementation of algorithms (multiplication and division in particular).

Other number representations of interest are *nonweighted* number systems, where the relative position of the digit does not affect the weight so that the appropriate interchange of any two digits will not change the value x. The best example of such a number system is the residue number system (RNS).

We also define **explicit value x_e** and **implicit value X_i** of a number represented by a bit string x. The *implicit value* is the only value of interest to the user, while the *explicit value* provides the most direct interpretation of the bit string x. Mapping of the *explicit value* to the *implicit value* is obtained by an arithmetic function that defines the number representation used. It is a task of the arithmetic designer to devise algorithms that result in the correct implicit value of the result for the operations on the operand digits representing the explicit values. In other words, the arithmetic algorithm needs to satisfy the *closure* property.

The relationship between the *implicit value* and the *explicit value* is best illustrated by Table 86.8.

Representation of Signed Integers

The two most common representations of signed integers are sign and magnitude (SM) representation and true and complement (TC) representation. While SM representation might be easier to understand and convert to and from, it has its own problems. Therefore, we will find TC representation to be more commonly used.

Sign and Magnitude Representation (SM). In SM representation signed integer X_i is represented by sign bit x_s and magnitude x_m (x_s, x_m). Usually 0 represents the positive sign (+) and 1 represents the negative sign (–). The magnitude of the number x_m can be represented in any way chosen for the representation of positive integers. A disadvantage of SM representation is that two representations of zero exist, positive and negative zero: $x_s = 0$, $x_m = 0$ and $x_s = 1$, $x_m = 0$.

True and Complement Representation (TC). In TC representation there is no separate bit used to represent the sign. Mapping between the explicit and implicit value is defined as

TABLE 86.9 True and Complement Mapping

x_e	X_i
0	0
1	1
2	2
M	M
$C/2 - 1$	$C/2 - 1$
$C/2 + 1$	$-(C/2 + 1)$
M	M
$C - 2$	-2
$C - 1$	-1
C	0

TABLE 86.10 Mapping of the Explicit Value x_e into RC and DRC Number Representations

x_e	X_i(RC)	X_i(DRC)
0	0	0
1	1	1
2	2	2
M	M	M
$1/2\,r^n - 1$	$1/2\,r^n - 1$	$1/2\,r^n - 1$
$1/2\,r^n$	$-1/2\,r^n$	$-(1/2\,r^n - 1)$
M	M	M
$r^n - 2$	-2	-1
$r^n - 1$	-1	0

$$X_i = \begin{cases} x_e & x_e < C/2 \\ x_e - C & x_e > C/2 \end{cases}$$

The illustration of TC mapping is given in Table 86.9. In this representation positive integers are represented in the *true form*, while negative integers are represented in the *complement form*.

With respect to how the complementation constant C is chosen, we can further distinguish two representations within the TC system. If the complementation constant is chosen to be equal to the range of possible values taken by x_e, $C = r^n$ in a conventional number system where $0 \le x_e \le r^n - 1$, then we have defined the *range complement* (RC) system. If, on the other hand, the complementation constant is chosen to be $C = r^n - 1$, we have defined the *diminished radix complement* (DRC) (also known as the *digit complement* [DC]) number system. Representations of the RC and DRC number representation systems are shown in Table 86.10.

As can be seen from Table 86.10, the RC system provides for one unique representation of zero because the complementation constant $C = r^n$ falls outside the range. There are two representations of zero in the DRC system, $x_e = 0$ and $r^n - 1$. The RC representation is not symmetrical, and it is not a closed system under the change of sign operation. The range for RC is $[-1/2\,r^n, 1/2\,r^n - 1]$. The DC is symmetrical and has the range of $[-(1/2\,r^n - 1), 1/2\,r^n - 1]$.

For the radix $r = 2$, RC and DRC number representations are commonly known as *two's complement* and *one's complement* number representation systems, respectively. Those two representations are illustrated by an example in Table 86.11 for the range of values $-(4 \le X_i \le 3)$.

Algorithms for Basic Arithmetic Operations

The algorithms for the arithmetic operation are dependent on the number representation system used. Therefore, their implementation should be examined for each number representation system separately, given that the complexity of the algorithm, as well as its hardware implementation, is dependent on it.

Addition and Subtraction in Sign and Magnitude System

In the *SM number system* addition/subtraction is performed on pairs (u_s, u_m) and (w_s, w_m) resulting in a sum (s_s, s_m), where u_s and w_s are sign bits and u_m and w_m are magnitudes. The algorithm is relatively complex because it requires comparisons of the signs and magnitudes as well. Extending the addition algorithm in order to perform subtraction is relatively easy because it only involves change of the sign of the operand being subtracted. Therefore, we will consider only the addition algorithm.

The algorithm can be described as

if $u_s = w_s$ (signs are equal) **then**
$s_s = u_s$ and $s_m = u_m + w_m$ (operation includes checking for the overflow)

TABLE 86.11 Two's Complement and One's Complement Representation

	Two's Complement, $C = 8$		One's Complement, $C = 7$	
X_i	x_e	X_i (2's complement)	x_e	X_i (1's complement)
3	3	011	3	011
2	2	010	2	010
1	1	001	1	001
0	0	000	0	000
–0	0	000	7	111
–1	7	111	6	110
–2	6	110	5	101
–3	5	101	4	100
–4	4	100	3	—

if $u_s \neq w_s$ **then**
  if $u_m > w_m$: $s_m = u_m - w_m$ $s_s = u_s$
  else: $s_m = w_m - u_m$ $s_s = w_s$

Addition and Subtraction in True and Complement System

Addition in the *TC system* is relatively simple. It is sufficient to perform modulo addition of the explicit values; therefore,

$$s_e = (u_e + w_e) \bmod C$$

Proof will be omitted.

In the *RC number system* this is equivalent to passing the operands through an adder and discarding the carry-out of the most significant position of the adder which is equivalent to performing the modulo addition (given that $C = r^n$).

In the *DRC number system* the complementation constant is $C = r^n - 1$. Modulo addition in this case is performed by subtracting r^n and adding 1. It turns out that this operation can be performed by simply passing the operands through an adder and feeding the carry-out from the most significant digit position into the carry-in at the least significant digit position. This is also called addition with *end-around carry.*

Subtracting two numbers is performed by simply changing the sign of the operand to be subtracted preceding the addition operation.

Change of Sign Operation. The change of sign operation involves the following operation:

$$W_i = -Z_i$$
$$w_e = (-z_e) = (-z_e) \bmod C = C - Z_i \bmod C = C - z_e$$

which means that the change of sign operation consists of subtracting the operand z_e from the complementation constant C.

In the *DRC system* complementation is performed by simply complementing each digit of the operand Z_i with respect to $r - 1$. In the case of $r = 2$, this results in a simple inversion of bits.

In the *RC system* the complementation is performed by complementing each digit of the operand Z_i with respect to $r - 1$ and adding 1 to the resulting z_e.

Implementation of Addition

Carry Look-Ahead Adder (CLA)

The first significant speed improvement in the implementation of a parallel adder was a carry-look-ahead adder (CLA) developed by Weinberger and Smith [1958] in 1958. The CLA is one of the fastest schemes used for the addition of two numbers even today given that the delay incurred to add two numbers is logarithmically

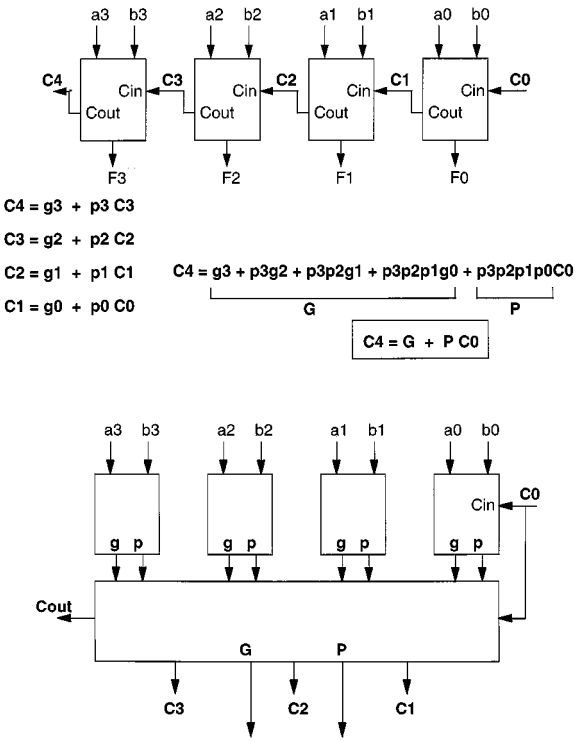


FIGURE 86.2 Carry Look-Ahead adder structure. (a) Generation of carry generate and propagate signals and (b) generation of group signals G, P and intermediate carries.

dependent on the size of the operands (delay = log[N]). The concept of CLA is illustrated in Fig. 86.2a and b. For each bit position of the adder, a pair of signals (p_i, g_i) is generated in parallel. It is possible to generate local carries using (p_i, g_i) as seen in the equations. Those signals are designated as: p_i = carry-propagate and g_i = carry-generate because they take part in propagation and generation of carry signal C_{i-1}. However, each bit position requires an incoming carry signal C_{i-1} in order to generate the outgoing carry C_i. This makes the addition slow because the carry signal has to ripple from stage to stage as shown in Fig. 86.2a. The adder can be divided into the groups and the carry-generate and carry-propagate signals can be calculated for the entire group (G,P). This will take an additional time equivalent to and AND-OR delay of the logic. However, now we can calculate each group's carry signals in an additional AND-OR delay. for the generation of the carry signal from the adder only the incoming carry signal into the group is now required. Therefore, the rippling of the carry is limited only to the groups. In the next step we may calculate generate and propagate signals for the group of groups (G*, P*) and continue in that fashion until we have only one group left generating the C_{out} signal from the adder. This process will terminate in log number of steps given that we are generating a tree structure for generation of carries. The computation of carries within the groups is done individually as illustrated in Fig. 86.2a and this process requires only the incoming carry into the group.

The logarithmic dependence on the delay (delay = log[N]) is only valid under the assumption that the gate delay is constant without depending on the fan-out and fan-in of the gate. In practice this is not true and even when the bipolar technology (which does not exhibit strong dependence on the fan-out) is used to implement CLA structure, the further expansion of the carry-block is not possible given the practical limitations on the fan-in of the gate.

In CMOS technology this situation is much different given that CMOS gate has strong dependency not only on fan-in but on fan-out as well. This limitation takes away much of the advantages gained by using the CLA scheme. However, by clever optimization of the critical path and appropriate use of dynamic logic the CLA scheme can still be advantageous, especially for the adders of a larger size.

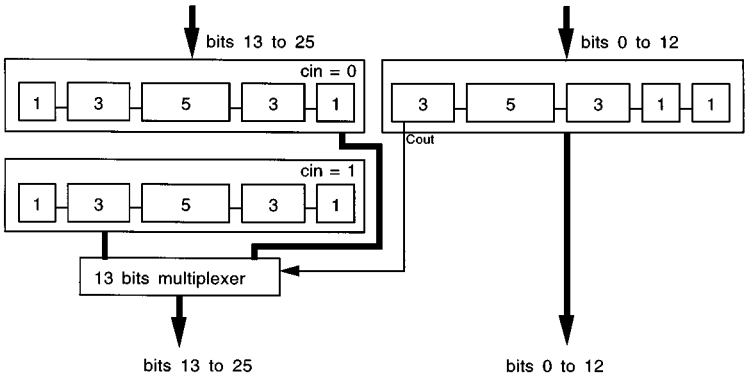


FIGURE 86.3 26-bit carry-select adder.

Conditional-Sum Addition

Another one of the fast schemes for addition of two numbers that predates CLA is conditional-sum addition (CSA) proposed by Sklansky [1960] in 1960. The essence of the CSA scheme is the realization that we can add two numbers without waiting for the carry signal to be available. Simply, the number were added in two instances: one assuming $C_{in} = 0$ and the other assuming $C_{in} = 1$. The results: Sum_0, Sum_1 and $Carry_0$, $Carry_1$ are presented at the input of a multiplexer. The final values are being selected at the time C_{in} arrives at the "select" input of a multiplexer. As in CLA the input bits are divided into groups which are added "conditionally".

It is apparent that starting from the least significant bit (LSB) position the hardware complexity starts to grow rapidly. Therefore, in practice, the full-blown implementation of the CSA is not often seen.

However, the idea of adding the most significant bit (MSB) portion conditionally and selecting the results once the carry-in signal is computed in the LSB portion is attractive. Such a scheme (which is a subset of CSA) is known as "carry-select adder". A 26-b carry-select adder consisting of two 13-bit portions is shown in Fig. 86.3.

Multiplication Algorithm

The multiplication operation is performed in a variety of forms in hardware and software. In the beginning of computer development any complex operation was usually programmed in software or coded in the microcode of the machine. Some limited hardware assistance was provided. Today it is more likely to find full hardware implementation of the multiplication for reasons of speed and reduced cost of hardware. However, in all of them, multiplication shares the basic algorithm with some adaptations and modifications to the particular implementation and number system used. For simplicity we will describe a basic multiplication algorithm that operates on positive n-bit-long integers X and Y resulting in the product P, which is $2n$ bits long:

$$P = XY = X \times \sum_{i=0}^{n-1} y_i r^i = \sum_{i=0}^{n-1} X \times y_i r^i$$

This expression indicates that the multiplication process is performed by summing n terms of a *partial product*: $X \times y_i r^i$. This product indicates that the ith term is obtained by a simple arithmetic left shift of X for the i positions and multiplication by the single digit y_i. For the binary radix $r = 2$, y_i is 0 or 1 and multiplication by the digit y_i is very simple to perform. The addition of n terms can be performed at once, by passing the partial products through a network of adders (which is the case of full hardware multiplier), or sequentially, by passing the *partial product* through an adder n times. The algorithm to perform multiplication of X and Y can be described as [Ercegovac, 1985]

$$p^{(0)} = 0$$
$$p^{j+1} = 1/r(p^j + r^n X y_j) \qquad \text{for } j = 0, \ldots, n-1$$

It can be easily proved that this recurrence results in $p^{(n)} = XY$.

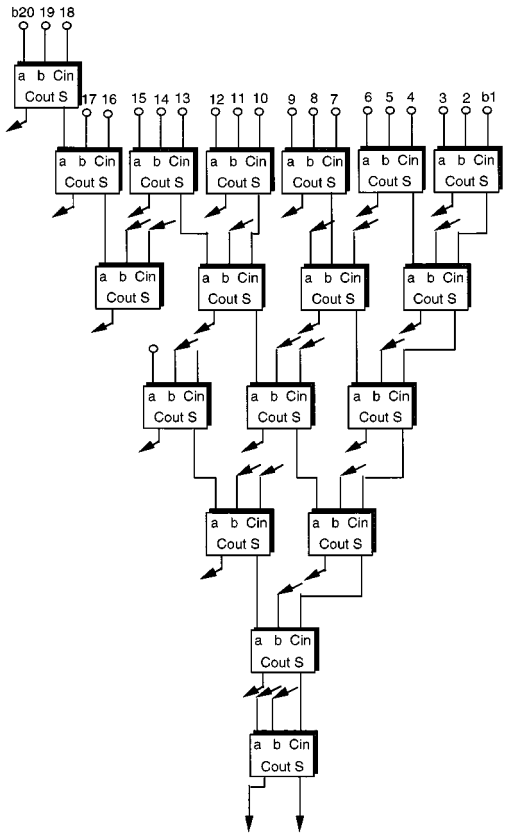


FIGURE 86.4 Wallace Tree.

Various modifications of the multiplication algorithm exist; one of the most famous is the *modified Booth recoding algorithm* described by Booth in 1951. This algorithm allows for the reduction of the number of partial products, thus speeding up the multiplication process. Generally speaking, the Booth algorithm is a case of using the redundant number system with the radix higher than 2.

Implementation of the Multiplication Algorithm

Speed of multiply operation is of utmost importance in digital signal processors (DSP) today as well as in the general purpose processors. Therefore, research in building a fast parallel multiplier has been going on since such a paper was published by Wallace [1964] in 1964. In his historic paper, Wallace introduced a way of summing the partial product bits in parallel using a tree of carry-save adders which became generally known as the "Wallace Tree" (Fig. 86.4).

A suggestion for speed improvement of such a process of adding partial product bits in parallel followed in the paper published by Dadda [1965]. In his 1965 paper, Dadda introduced a notion of a counter structure that will take a number of bits p in the same bit position (of the same "weight") and output a number q that represents the count of ones in the input. Dadda has introduced a number of ways to compress the partial product bits using such a counter, which later became known as Dadda's counter.

The quest for making the parallel multiplier even faster continued for almost 30 years. The search for producing a fastest "counter" did not result in a general structure that yielded a faster partial product summation than that which used a full-adder (FA) cell, or 3:2 counter. Therefore, the use of the Wallace Tree was almost prevalent in the implementation of the parallel multipliers. In 1981, Weinberger disclosed a structure he called "4-2 carry-save module". This structure contained a combination of FA cells in an intricate interconnection structure that was yielding faster partial product compression than the use of 3:2 counters.

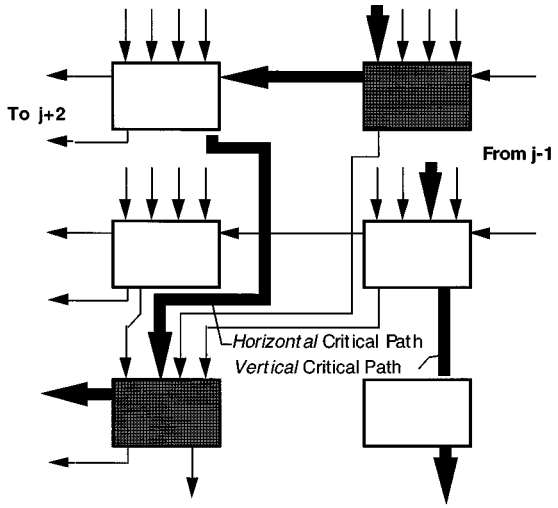


FIGURE 86.5 4:2 compressor.

The 4:2 compressor (Fig. 86.5) actually compresses five partial product bits into three; however, it is connected in such a way that four of the inputs are coming from the same bit position of the weight *j* while one bit is fed from the neighboring position *j*-1 (known as carry-in). The output of such a 4-2 module consists of one bit in the position *j* and two bits in the position *j*+1. This structure does not represent a counter (though it became erroneously known as a "4-2 counter") but a "compressor" which would compress four partial product bits into two (while using one bit laterally connected between adjacent 4-2 compressors). The efficiency of such a structure is higher (it reduces the number of partial product bits by one half). The speed of such a 4-2 compressor has been determined by the speed of 3 XOR gates in series (in the redesigned version of 4-2 compressor) making such a scheme more efficient that the one using 3:2 counters in a regular Wallace Tree. The other equally important feature of the use of a 4-2 compressor is that the interconnections between such cells follow more regular patterns than in the case of the Wallace Tree.

Booth Encoding

Booth's algorithm [Booth, 1951] is widely used in the implementations of hardware or software multipliers because its application makes it possible to reduce the number of partial products. It can be used for both sign-magnitude numbers as well as 2's complement numbers with no need for a correction term or a correction step.

A modification of the Booth algorithm was proposed by MacSorley [1961] in which a triplet of bits is scanned instead of two bits. This technique has the advantage of reducing the number of partial products by one half regardless of the inputs. This is summarized in Table 86.12.

TABLE 86.12 Modified Booth Recoding

$x_{i+2}x_{i+1}x_i$	Add to Partial Product
000	+0Y
001	+1Y
010	+1Y
011	+2Y
100	–2Y
101	–1Y
110	–1Y
111	–0Y

The recoding is performed within two steps: encoding and selection. The purpose of the encoding is to scan the triplet of bits of the multiplier and define the operation to be performed on the multiplicand, as shown in Table 86.8. This method is actually an application of a sign-digit representation in radix 4. The Booth-MacSorley algorithm, usually called the Modified Booth algorithm or simply the Booth algorithm, can be generalized to any radix.

Booth recoding necessitates the internal use of 2's complement representation in order to efficiently perform subtraction of the partial products as well as additions. However, floating point standard specifies sign magnitude representation which is followed by most of the non-standard floating point numbers in use today. The advantage of Booth recoding is that it generates only a half of the partial products as compared to the multiplier implementation which does not use Booth recoding. However, the benefit achieved comes at the expense of

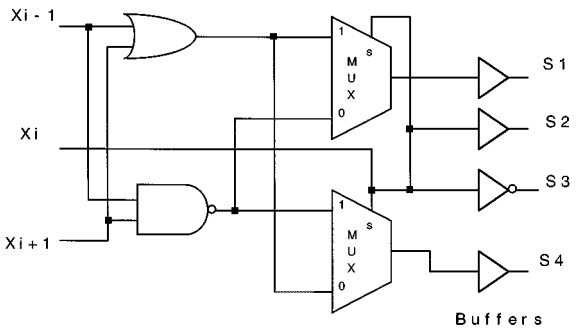


FIGURE 86.6 Booth encoder.

increased hardware complexity. Indeed, this implementation requires hardware for the encoding and for the selection of the partial products (0, ±Y, ±2Y). An optimized encoding is shown in Fig. 86.6.

Division Algorithm

Division is a more complex process to implement because, unlike multiplication, it involves *guessing* the digits of the quotient. Here, we will consider an algorithm for division of two positive integers designated as *dividend* Y and *divisor* X resulting in a *quotient* Q and an integer *remainder* Z according to the relation given by

$$Y = XQ + Z$$

In this case the dividend contains $2n$ integers and the divisor has n digits in order to produce a quotient with n digits.

The algorithm for division is given with the following recurrence relationship [Ercegovac, 1985]:

$$z^{(0)} = Y$$
$$z^{(j+1)} = rz^{(j)} - Xr^{n}Q_{n-1-j} \qquad \text{for } j = 0, \ldots, n-1$$

this recurrence relation yields

$$z^{(n)} = r^{n}(Y - XQ)$$
$$Y = XQ + z^{(n)}r^{-n}$$

which defines the division process with remainder $Z = z^{(n)}r^{-n}$.

The selection of the quotient digit is done by satisfying that $0 \le Z < X$ at each step in the division process. This selection is a crucial part of the algorithm, and the best known are *restoring* and *nonrestoring* division algorithms. In the former the value of the *tentative partial remainder* $z^{(j)}$ is restored after the wrong guess is made of the quotient digit q_j. In the latter this correction is not done in a separate step, but rather in the step following. The best-known division algorithm is the so-called SRT algorithm independently developed by Sweeney, Robertson, and Tocher. Algorithms for a higher radix were further developed by Robertson and his students, most notably Ercegovac.

Defining Terms

4:2 Compressor: A structure used in the partial product reduction tree of a parallel multiplier for achieving faster and more efficient reduction of the partial product bits.

Algorithm: Decomposition of the computation into subcomputations with an associated precedence relation that determines the order in which these subcomputations are performed [Ercegovac, 1985].

Booth-MacSorley algorithm: An algorithm used for recoding of the multiplier such that the number of partial products is roughly reduced by a factor of two. It is a special case of the application of the redundant number system to represent the multiplier.

Carry look-ahead adder: An implementation technique of addition that accelerates the propagation of the carry signal, thus increasing the speed of addition operation.

Dadda's counter: A generalized structure used to produce a number (count) representing the number of bits that are "one". It is used for efficient reduction of the partial products in a parallel multiplier.

Explicit value x_e: A value associated with the bit string according to the rule defined by the number representation system being used.

Implicit value X_i: The value obtained by applying the arithmetic function defined for the interpretation of the explicit value x_e.

Nonredundant number system: The system where for each bit string there is one and only one corresponding numerical value x_e.

Number representation system: A defined rule that associates one numerical value x_e with every *valid* bit string x.

Redundant number system: The system in which the numerical value x_e could be represented by more than one bit string.

SRT algorithm: An algorithm for division of binary numbers which uses redundant number representation.

Wallace tree: A technique for summing the partial product bits of a parallel multiplier in a carry-save fashion using full-adder cells.

Related Topic

86.1 Number Systems

References

A. Avizienis, "Digital computer arithmetic: A unified algorithmic specification," in *Symposium on Computers and Automata,* Polytechnic Institute of Brooklyn, April 13–15, 1971.

A. D. Booth, "A signed binary multiplication technique", *Quarterly J. Mechan. Appl. Math.,* vol. IV, 1951.

L. Dadda, "Some schemes for parallel multipliers," *Alta Frequenza,* 34, 349–356, 1965.

M. Ercegovac, *Digital Systems and Hardware/Firmware Algorithms*, New York: Wiley, 1985, chap. 12.

O. L. MacSorley, "High speed arithmetic in binary computers", *Proc. IRE,* 49(1), 1961.

V. G. Oklobdzija and E. R. Barnes, "Some optimal schemes for ALU implementation in VLSI technology", Proceedings of 7th Symposium on Computer Arithmetic, Urbana, Ill.: University of Illinois, June 4–6, 1985.

Sklanski, "Conditional-sum addition logic", *IRE Trans. Electron. Computers,* EC-9, 226–231, 1960.

C. S. Wallace, "A suggestion for a fast multiplier," *IEEE Trans. Electron. Computers,* EC-13, 14–17, 1964.

S. Waser and M. Flynn, *Introduction to Arithmetic for Digital Systems Designers*, New York: Holt, 1982.

Weinberger and J. L. Smith, "A logic for high-speed addition", *National Bureau of Standards, Circulation 591,* p. 3–12, 1958.

Further Information

For more information about specific arithmetic algorithms and their implementation see K. Hwang, *Computer Arithmetic: Principles, Architecture and Design*, New York: Wiley, 1979 and also E. Swartzlander, *Computer Arithmetic*, vols. I and II, Los Alamitos, Calif.: IEEE Computer Society Press, 1980.

Publications in *IEEE Transactions on Electronic Computers* and *Proceedings of the Computer Arithmetic Symposia* by various authors are very good sources for detailed information on a particular algorithm or implementation.

86.3 Architecture[1]

V. Carl Hamacher, Zvonko G. Vranesic, and Safwat G. Zaky

Computer architecture can be defined here to mean the functional operation of the individual hardware units in a computer system and the flow of information and control among them. This is a somewhat more general definition than is sometimes used. For example, some articles and books refer to instruction set architecture or the system bus architecture.

The main functional units of a single-processor system, a basic way to interconnect them, and features that are used to increase the speed with which the computer executes programs will be described. Following this, a brief introduction to systems that have more than one processor will be provided.

Functional Units

A digital computer, or simply a computer, accepts digitized input information, processes it according to a list of internally stored *machine instructions,* and produces the resultant output information. The list of instructions is called a *program*, and internal storage is called *computer memory.*

A computer has five functionally independent main parts: input, memory, arithmetic and logic, output, and control. The input unit accepts digitally encoded information from human operators, through electromechanical devices such as a keyboard, or from other computers over digital communication lines. The information received is usually stored in the memory and then operated on by the arithmetic and logic unit circuitry under the control of a program. Results are sent back to the outside world through the output unit. All these actions are coordinated by the control unit. The arithmetic and logic unit, in conjunction with the main control unit, are referred to as the **processor.**

Input and output equipment is usually combined under the term **input-output unit** (*I/O unit*). This is reasonable because some standard equipment provides both input and output functions. The simplest example of this is the video terminal consisting of a keyboard for input and a cathode-ray tube for output. The control circuits of the computer recognize two distinct devices, even though the human operator may associate them as being part of the same physical unit.

The **memory unit** stores programs and data. There are two main classes of memory devices called *primary* and *secondary* memory. Primary storage, or main memory, is an electronic storage device, constructed from integrated circuits that consist of millions of semiconductor storage cells, each capable of storing one bit of information. These cells are accessed in groups of fixed size called *words.* The main memory is organized so that the contents of one word can be stored or retrieved in one basic operation called a *memory cycle.*

To provide direct access to any word in the main memory in a short and fixed amount of time, a distinct address number is associated with each word location. A given word is accessed by specifying its address and issuing a control command that starts the storage or retrieval process. The number of bits in a word is called the *word length* of the computer. Word lengths vary from 16 to 64 bits. Small machines such as personal computers or workstations, may have only a few million words in the main memory, while larger machines have tens of millions of words. The time required to access a word for reading or writing is less than 100 ns.

Although primary memory is essential, it tends to be expensive and volatile. Thus cheaper, more permanent, magnetic media secondary storage is used for files of information that contain programs and data. A wide selection of suitable devices is available, including magnetic disks, drums, diskettes, and tapes.

Execution of most operations within a computer takes place in the **arithmetic and logic unit** (ALU) of a processor. Consider a typical example. Suppose that two numbers located in the main memory are to be added, and the sum is to be stored back into the memory. Using a few instructions, each consisting of a few basic steps, determined by the **control unit**, the operands are first fetched from the memory into the processor. They are then added in the ALU, and the result is stored back in memory. Processors contain a number of high-speed storage elements called *registers,* which are used for temporary storage of operands. Each register contains one

[1]Adapted from V.C. Hamacher, Z.G. Vranesic, and S.G. Zaky, *Computer Organization,* 4th ed., New York: McGraw-Hill, 1996. With permission.

The First Digital Computers

Of all the new technologies to emerge from World War II, none was to have such profound and pervasive impacts as the digital computer. As early as the 1830s, the Englishman Charles Babbage conceived of an "Analytical Engine" that would perform mathematical operations using punched cards, hundreds of gears, and steam power. Babbage's idea was beyond the capabilities of 19th-century technology, but his vision represented a goal that many were to pursue in the next century and a half.

In the mid-1920s, MIT electrical engineer Vannevar Bush devised the "product integraph", a semi-automatic machine for solving problems in determining the characteristics of complex electrical systems. This was followed a few years later by the "differential analyzer", the first general equation-solving machine. These machines were mechanical, analog devices, but at the same time that they were being built and copied, the principles of electrical, digital machines were being laid out.

In 1937, Claude Shannon published in the *Transactions* of the AIEE the circuit principles for an "electric adder to the base of two", and George Stibitz of Bell Labs built such an adding device on his own kitchen table. In that same year, Howard Aiken, then a student at Harvard, proposed a gigantic calculating machine that could be used for everything from vacuum tube design to problems in relativistic physics. With support from Thomas J. Watson, president of IBM, Aiken was able to build his machine, the "Automatic Sequence Controlled Calculator", or "Mark I". When it was finished in 1944, the Mark I was quickly pressed into war service, calculating ballistics problems for the Navy.

In 1943, the government contracted with John W. Mauchly and J. Presper Eckert of the University of Pennsylvania to build the "Electronic Numerical Integrator and Computer"—the first true electronic digital computer. When the ENIAC was finally dedicated in February 1946, it was both a marvel and a monster—weighing 30 tons, consuming 150 kW of power, and using 18,000 vacuum tubes. With all of this, it could perform 5,000 additions or 400 multiplications per second, which was about one thousand

word of data and its access time is about 10 times faster than main memory access time. Large-scale micro-electronic fabrication techniques allow whole processors to be implemented on a single semiconductor chip containing a few million transistors.

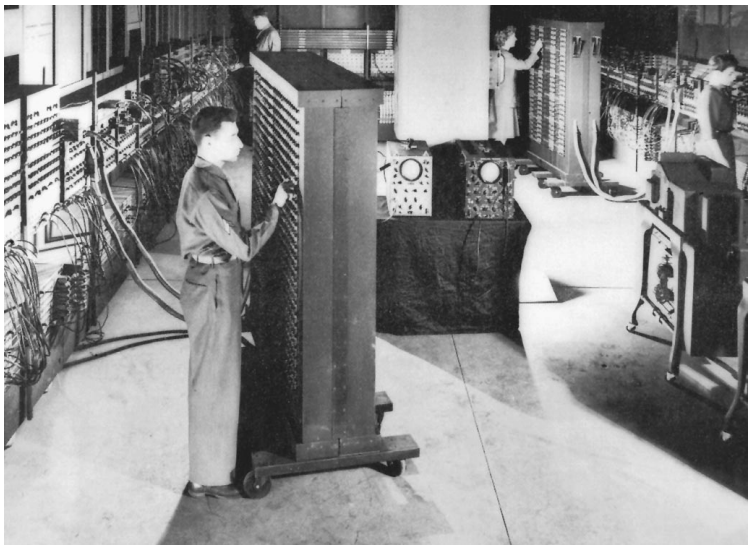
Basic Operational Concepts

To perform a given computational task, an appropriate program consisting of a set of machine instructions is stored in the main memory. Individual instructions are brought from the memory into the processor for execution. Data used as operands are also stored in the memory. A typical instruction may be

MOVE MEMLOC, Ri

This instruction loads a copy of the operand at memory location MEMLOC into the processor register Ri. The instruction requires a few basic steps to be performed. First, the instruction must be transferred from the memory into the processor, where it is decoded. Then the operand at location MEMLOC must be fetched into the processor. Finally, the operand is placed into register R1. After operands are loaded into the processor registers in this way, instructions such as ADD Ri, Rj, Rk can be used to add the contents of registers Ri and Rj, and then place the result into register Rk.

Instruction set design has been intensively studied in recent years to determine the effectiveness of the various alternatives. See Patterson and Hennessey [1994] for a thorough discussion.



The ENIAC, pictured above, was the first true electronic digital computer. Early programmers set up problems by plugging in cables and setting switches. ENIAC could perform calculations about one thousand times faster than any other machine of its day. (Photo courtesy of the IEEE Center for the History of Electrical Engineering.)

times faster than any other machine of the day. The ENIAC showed the immense possibilities of digital electronic computers.

These possibilities occupied engineers and mathematicians for the coming decades. For electrical engineers, the computer represented a challenge and responsibility for the most powerful new machine of the twentieth century. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

The connection between the main memory and the processor that allows for the transfer of instructions and operands is called the **bus**, as shown in Fig. 86.7. A bus consists of a set of address, data, and control lines. The bus also allows program and data files to be transferred from their long-term location on magnetic disk storage to the main memory. Long distance digital communication with other computers is also enabled by transfers over the bus to the Communication Line Interface, as shown in the figure. The bus interconnects a number of devices, but only two devices (a sender and a receiver) can use it at any one time. Therefore, some control circuitry is needed to manage the orderly use of the bus when a number of devices wish to use it.

Normal execution of programs may sometimes be preempted if some I/O device requires urgent control action or servicing. For example, a monitoring device in a computer-controlled industrial process may detect a dangerous condition that requires the execution of a special service program dedicated to the device. To cause this service program to be executed, the device sends an *interrupt* signal to the processor. The processor temporarily suspends the program that is being executed and executes the special *interrupt service routine.* After providing the required service, the processor switches back to the interrupted program. To appreciate the complexity of the computer system software programs needed to control such switching from one program task to another and to manage the general movement of programs and data between primary and secondary storage, consult Tanenbaum [1990].

The need often arises during program loading and execution to transfer blocks of data between the main memory and a disk or other secondary storage I/O devices. Special control circuits are provided to manage these transfers without detailed control actions from the main processor. Such transfers are referred to as *direct memory access* (DMA). Assuming that accesses to the main memory from both I/O devices (such as disks) and

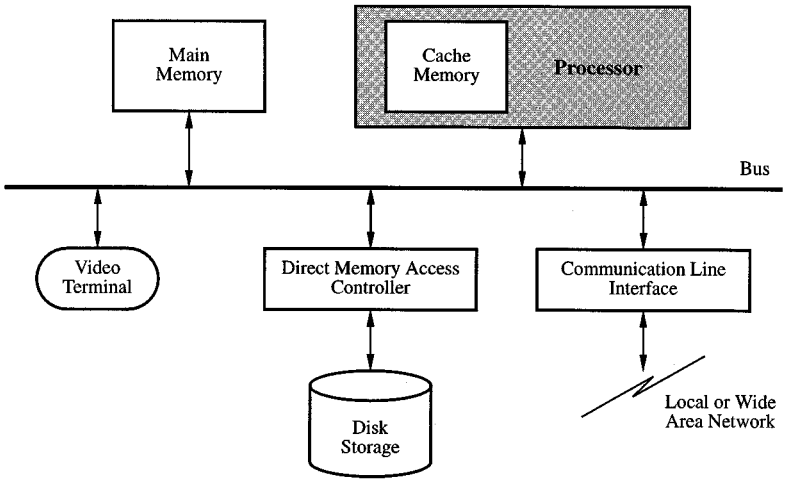


FIGURE 86.7 Interconnection of major components in a computer system.

the main processor can be appropriately interwoven over the bus, I/O-memory transfers and computation in the main processor can proceed in parallel, and performance of the overall system is improved.

Performance

A major performance measure for computer systems is the time, T, that it takes to execute a complete program for some task. Suppose N machine instructions need to be executed to perform the task. A program is typically written in some high-level language, translated by a compiler program into machine language, and stored on a disk. An operating system software routine then loads the machine language program into the main memory, ready for execution. Assume that, on average, each machine language instruction requires S basic steps for its execution. If basic steps are executed at the rate of R steps per second, then the time to execute the program is

$$T = (N \times S)/R$$

The main goal in computer architecture is to develop features that minimize T.

We will now give an outline of main memory and processor design features that help to achieve this goal. The first concept is that of a **memory hierarchy**. We have already noted that access to operands in processor registers is significantly faster than access to the main memory. Suppose that when instructions and data are first loaded into the processor, they are stored in a small, fast, **cache memory** on the processor chip itself. If instructions and data in the cache are accessed repeatedly within a short period of time, as happens often with program loops, then program execution will be speeded up. The cache can only hold small parts of the executing program. When the cache is full, its contents are replaced by new instructions and data as they are fetched from the main memory. A variety of *cache replacement algorithms* are in use. The objective of these algorithms is to maximize the probability that the instructions and data needed for program execution are found in the cache. This probability is known as the cache *hit ratio*. A higher hit ratio means that a larger percentage of the instructions and data are being found in the cache, and do not require access to the slower main memory. This leads to a reduction in the memory access basic step time components of S, and hence to a smaller value of T.

The basic idea of a cache can be applied at different points in a computer system, resulting in a hierarchy of storage units. A typical memory hierarchy is shown in Fig. 86.8. Some systems have two levels of cache to take the best advantage of size/speed/cost tradeoffs. The main memory is usually not large enough to contain all of the programs and their data. Therefore, the highest level in the memory hierarchy is usually magnetic disk storage. As the figure indicates, it has the largest capacity, but the slowest access time. Segments of a program, often called *pages*, are transferred from the disk to the main memory for execution. As other pages are needed, they may replace the pages already in the main memory if the main memory is full. The orderly,

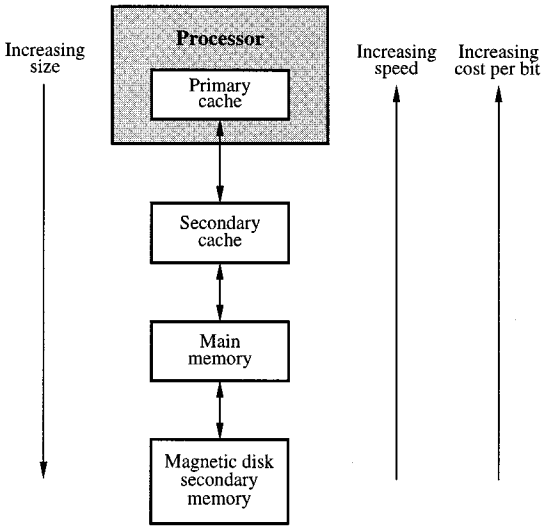


FIGURE 86.8 Memory hierarchy.

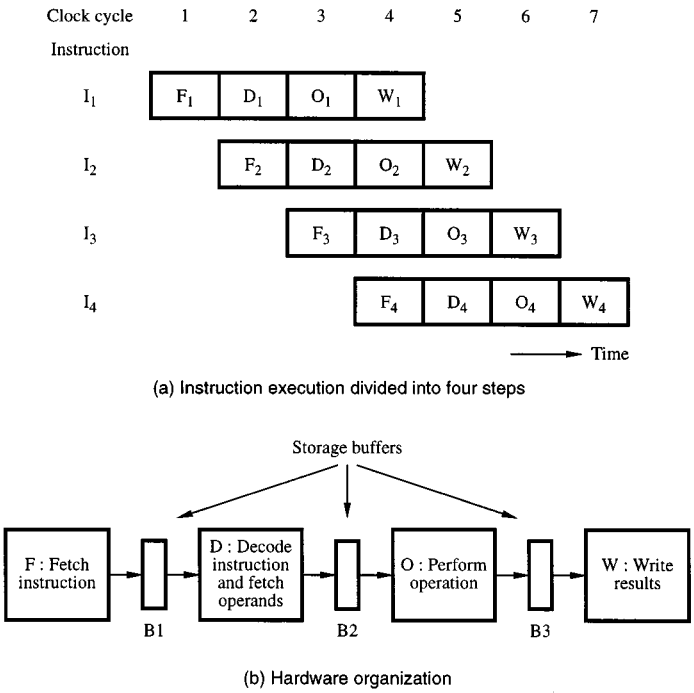


FIGURE 86.9 Pipelining of instruction execution.

automatic movement of large program and data segments between the main memory and the disk, as programs execute, is managed by a combination of operating system software and control hardware. This is referred to as **memory management**.

We have implicitly assumed that instructions are executed one after another. Most modern processors are designed to allow the execution of successive instructions to overlap, using a technique known as **pipelining**. In the example in Fig. 86.9, each instruction is broken down into 4 basic steps—fetch, decode, operate, and write—and a separate hardware unit is provided to perform each of these steps. As a result, the execution of

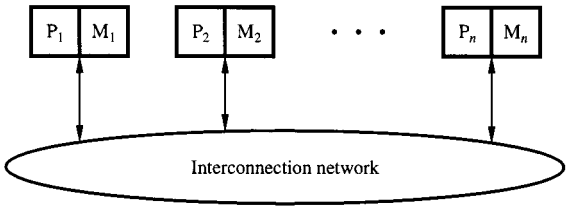


FIGURE 86.10 A multiprocessor system.

successive instructions can be overlapped as shown, resulting in an instruction completion rate of one per basic time step. If the execution overlap pattern shown in the figure can be maintained for long periods of time, the effective value of S tends toward 1.

When the execution of some instruction I depends on the results of a previous instruction, J, which is not yet completed, instruction I must be delayed. The pipeline is said to be *stalled,* waiting for the execution of instruction J to be completed. While it is not possible to eliminate such situations altogether, it is important to minimize the probability of their occurrence. This is a key consideration in the design of the instruction set of modern processors and the design of the compilers that translate high-level language programs into machine language.

Now, imagine that multiple functional units are provided in the processor so that more than one instruction can be in the operate stage. This *parallel* execution capability, when added to pipelining of the individual instructions, means that execution rates of more than one instruction completion per basic step time can be achieved. This mode of enhanced processor performance is called **superscalar processing**.

The rate, R, of performing basic steps in the processor is usually referred to as the processor clock rate; and it is of the order of 100 to 200 million steps per second in current high-performance VLSI processors. This rate is determined by the technology used in fabricating the processors, and is strongly related to the size or area occupied by individual transistors. This size feature, which is currently in the submicron range, has been steadily decreasing as fabrication techniques improve, allowing increases in R to be achieved.

Multiprocessors

Physical limits on electronic speeds prevent single processors from being speeded up indefinitely. A major design trend has seen the development of systems that consist of a large number of processors. Such multiprocessors can be used to speed up the execution of large programs by executing subtasks in parallel. The main difficulty in achieving this type of speedup is in being able to decompose a given task into its parallel subtasks and assign these subtasks to the individual processors in such a way that communication among the subtasks can be done efficiently. Fig. 86.10 shows a block diagram of a multiprocessor system, with the interconnection network needed for data sharing among the processors Pi. Parallel paths are needed in this network in order for parallel activity to proceed in the processors as they access the global memory space represented by the multiple memory units Mi.

Defining Terms

Arithmetic and logic unit: The logic gates and register storage elements used to perform the basic operations of addition, subtraction, multiplication, and division of numeric operands, and the comparison, shifting, and alignment operations on more general forms of numeric and nonnumeric data.

Bus: The collection of data, address, and control lines that enables exchange of information, usually in word-size quantities, among the various computer system units. In practice, a large number of units can be connected to a single bus. These units contend in an orderly way for the use of the bus for individual transfers.

Cache memory: A high-speed memory for temporary storage of copies of the sections of program and data from the main memory that are currently active during program execution.

Computer architecture: The functional operation of the individual hardware units in a computer system and the flow of information and control among them.

Control unit: The circuits required for sequencing the basic steps needed to execute machine instructions.

Input-output unit (I/O): The equipment and controls necessary for a computer to interact with a human operator, to access mass storage devices such as disks and tapes, or to communicate with other computer systems over communication networks.

Memory hierarchy: The collection of cache, primary, and secondary memory units that comprise the total storage capability in the computer system.

Memory management: The combination of operating system software and hardware controls that is needed to access and move program and data segments up and down the memory hierarchy during program execution.

Memory unit: The unit responsible for storing programs and data. There are two main types of units: primary memory, consisting of millions of bit storage cells fabricated from electronic semiconductor integrated circuits, used to hold programs and data during program execution; and secondary memory, based on magnetic disks, diskettes, and tapes, used to store permanent copies of programs and data.

Multiprocessor: a computer system comprising multiple processors and main memory unit modules, connected by a network that allows parallel activity to proceed efficiently among these units in executing program tasks that have been sectioned into subtasks and assigned to the processors.

Pipelining: The overlapped execution of the multiple steps of successive instructions of a machine language program, leading to a higher rate of instruction completion than can be attained by executing instructions strictly one after another.

Processor: The arithmetic and logic unit combined with the control unit needed to sequence the execution of instructions. Some cache memory is also included in the processor.

Superscalar processing: The ability to execute instructions at a completion rate that is faster than the normal pipelined rate, by providing multiple functional units in the pipeline to allow a small number of instructions to proceed through the pipeline process in parallel.

Related Topics

86.2 Computer Arithmetic • 86.4 Microprogramming

References

V.C. Hamacher, Z.G. Vranesic, and S.G. Zaky, *Computer Organization*, 4th ed., New York: McGraw-Hill, 1996.

D. A. Patterson and J. L. Hennessey, *Computer Organization and Design—The Hardware/Software Interface,* San Mateo, Calif.: Morgan Kaufman, 1994.

A.S. Tanenbaum, *Structured Computer Organization*, 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1990.

Further Information

The IEEE magazines *Computer*, *Micro*, and *Software* all have interesting articles on subjects related to computer architecture, including software aspects. Also, articles on computer architecture occasionally appear in *Scientific American.*

86.4 Microprogramming

Jacques Raymond

Since the 1950s when Wilkes et. al. [1958] defined the term and the concept, microprogramming has been used as a clean and systematic way to define the instruction set of a computer. It has also been used to define a virtual architecture out of a real hardwired one.

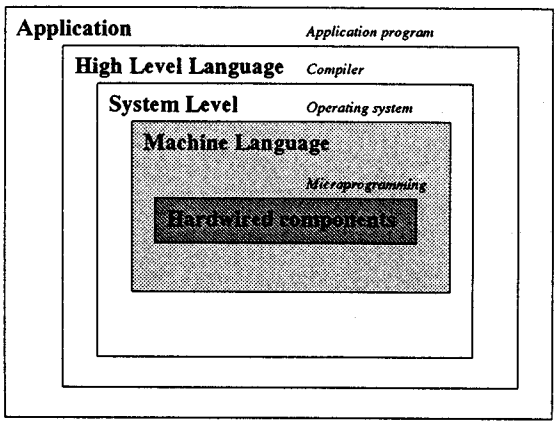


FIGURE 86.11 Levels of programming in a computer system.

Levels of Programming

In Fig. 86.12, we see that a computer application is usually realized by programming a given algorithm in a high-level language. A system offering a high-level language capability is implemented at the system level via a compiler. The operating system is (usually) implemented in a lower-level language. The machine instruction set can be hardwired (in a hardware implementation) or implemented via microprogramming (Fig. 86.11).

Therefore, microprogramming is simply an extra level in the general structure. Since it is used to define the machine instruction set, it can be considered at the *hardware* level. Since this definition is done via a program at a low level, but still eventually modifiable, it can also be considered to be at the *software* level. For these reasons, the term **firmware** has been coined to name sets of microprograms. In short, **microinstructions** that specify hardware functions (*microoperations* such as Open a path, Select operation) are used to form a more complex instruction (Convert to binary, Add decimal). The machine instruction set is defined via a set of microprogram routines and a microprogrammed instruction decoder.

In a microprogrammed machine, the hardware is designed in terms of its capabilities (ALU, data paths, I/O, processing units) with little concern for how these capabilities will have to be accessed by the programmers. The microoperations are decoded from microinstructions. The way programmers *view* the machine is defined at the microprogramming level.

This approach offers some differences over the hardwired approach. The advantages are that it is more systematic in implementation, modifiable, economical on most designs, and easier to debug. The disadvantages are that it is uneconomical on simple machines, slower, and needs support software. Like all programs, microprograms reside in memory. The term "control memory" is commonly used for microprograms.

Microinstruction Structure

On a given hardware, many processing functions are available. In general a subset *O* of these functions can be performed in parallel, for example, carrying on an addition between two registers while copying a register on an I/O bus. These functions are called **microcommands.**

Horizontal Microinstructions

Each of the fields **f** of a microinstruction specifies a microcommand. If the format of the microinstruction is such that all possible microcommands can be specified, the instruction is called *horizontal*. Most of the time, it is wasteful in memory as, in a microprogram, not every possible microcommand is specified in each microinstruction. However, it permits the microprogrammer to fully take advantage of all possible parallelisms and to build faster machines.

For example, the horizontal specification of an ALU operation,

ALUOperation	SourcePathA	SourcePathB	ResultPath	CvtDecimal

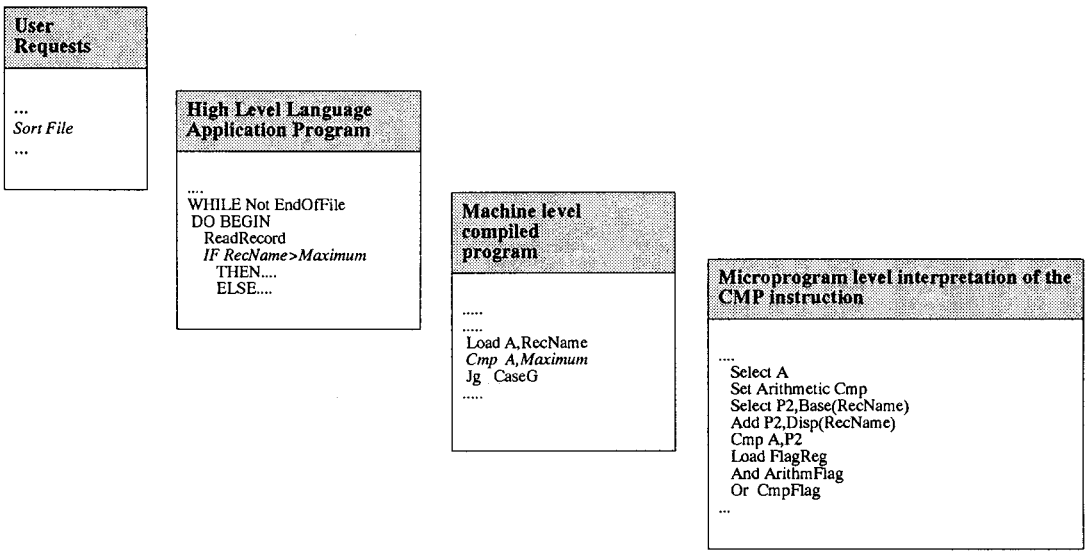


FIGURE 86.12 A view of computer system levels.

specifies both operands, which register will contain the result, whether or not the result is to be converted to decimal, and the operation to be performed. If this instruction is, for example, part of a microprogram defining a 32-bit addition instruction, assuming a 16-bit path, it is wasteful to specify twice the source and result operands.

In some cases, it is possible to design a microinstruction that specifies more microcommands than can be executed in parallel. In that case, the execution of the microcommand is carried out in more than one clock cycle. For this reason they are called *polyphase* microinstructions (as opposed to *monophase*).

Vertical Microinstructions

At the other extreme, if the microinstruction allows only the specification of a single microcommand at a time, the instruction is then called *vertical*. In that case, only the necessary commands for a particular program are specified, resulting in smaller **control memory** requirements. However, it is not possible to take advantage of possible parallelism offered by the hardware, since only one microcommand is executed at a time. For example, the vertical specification of an ALU operation is as follows:

SourceA	Reg#	1st Operand
SourceB	Reg#	2nd Operand
Result	Reg#	Result
ALU	Op	Operation

Diagonal Microinstructions

Most cases fit in between these two extremes (see Fig. 86.13). Some parallelism is possible; however, microcommands pertaining to a given processing unit are regrouped. This results in shorter microprograms than in the vertical case and may still allow some optimization. For example, a diagonal specification of an ALU operation is as follows:

SelectSources	RegA#	RegB#	Select Operands
SelectResult	Reg#	Dec/Bin	Result place and format
Select ALU Operation			Perform the operation

Optimization

Time and space optimization studies can be performed before designing the microinstruction format. The reader is referred to Das et al. [1973] and Agerwala [1976] for details and more references.

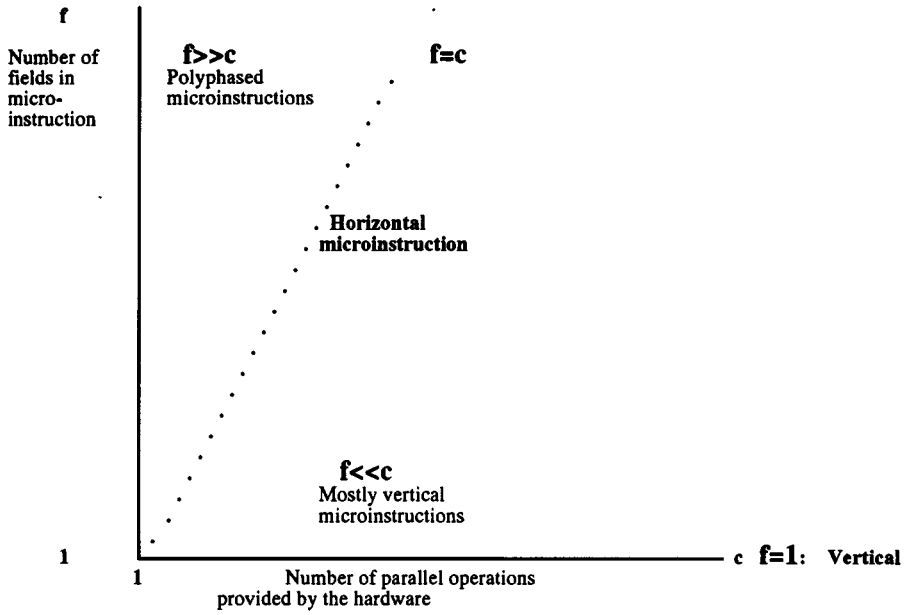


FIGURE 86.13 Microinstruction fields versus microcommands.

Microprogram Development

Microassemblers

The first level of specification of microinstructions is, just like its counterpart at the machine level, the assembler. Although the process and philosophy is exactly the same, it is traditionally called a microassembler. A microassembler is a software program (it is not relevant to know in which language it is written) whose function is to translate a source program into the binary code equivalent. Obviously, to write a source program, a language has to be designed. At assembly level, languages are usually very close to the hardware structure, and the objects defined are microregisters, gate level controls, and paths. Operations are the microoperations (sometimes slightly more sophisticated with a microassembler with macrofacilities).

This level provides an easily readable microprogram and does much to help avoid syntax errors. In binary, only the programmer can catch a faulty 1 or 0; the microassembler can catch syntax errors or some faulty register specifications. No microassembler exists that can catch all logic errors in the implementation of a given instruction algorithm. It is still very easy to make mistakes. It should be noted that this level is a good compromise between convenience and cost.

The following is a typical example of a microprogram in the microassembler (it implements a 16-bit add on an 8-bit path and ALU):

```
CLC   Clear Carry
Lod   A       Get first part of first operand
Add   B       Add to first part of second operand
Sto   C       Give low byte of final result
Lod   a       Get second part of first operand
Adc   b       Add to second part of second operand and to carry bit
Sto   c       Give high byte of final result
JCS   Error   Jump to error routine if Result >65536
Jmp   FetchNext
```

High-Level Languages for Microprogramming

Many higher-level languages have been designed and implemented; see a discussion of some design philosophies in Malik and Lewis [1978]. In principle, a high-level language program is oriented toward the application it supports and is farther away from the hardware-detailed implementation of the machine it runs on. The applications supported are mostly other machine definitions (**emulators**) and implementations of some algorithms.

The objects defined and manipulated by high-level languages for microprogramming are therefore the virtual components of virtual machines. They are usually much the same as their real counterparts: registers, paths, ALUs, microoperations, etc. Furthermore, writing a microprogram is usually defining a machine architecture. It involves a lot of intricate and complicated details, but the algorithms implemented are mostly quite simple. The advantages offered by high-level languages to write better algorithms without getting lost in implementation details are therefore not exploited.

Firmware Implementation

Microprogramming usually requires the regular phases of design, coding, test, conformance acceptance, documentation, etc. It differs from other programming activities when it comes to the deliverable. The usual final product takes the form of hardware, namely, a control memory, PROM, ROM, or other media, containing the bit patterns equivalent to the firmware. These implementation steps require special hardware and software support. They include a linker, loader, PAL programmer, or ROM burner; a test program in a control memory test bench is also necessary.

Supporting Software

It is advisable to test the microprogram before its actual hard implantation, if the implantation process is irreversible or too costly to repeat. Software simulators have been implemented to allow thorough testing of the newly developed microprogram. Needless to say these tools are very specialized to a given environment and therefore costly to develop, as their development cost cannot be distributed over many applications.

Emulation

Concept

In a microprogrammed environment, a computer architecture is softly (or firmly) defined by the microprogram designed to implement its operations, its data paths, and its machine-level instructions. It is easy to see that if one changes the microprogram for another one, then a new computer is defined. In this environment, the desired operation is simulated by the execution of the firmware, instead of being the result of action on real hardwired components.

Since the word *simulation* was already in use for simulation of some system by software, the word *emulation* was chosen to mean simulation of an instruction set by firmware. Of course, "simulation" by hardware is not a simulation but the *real thing*.

The general structure of an emulator consists of the following pseudocode algorithm:

```
BEGIN
Initialize Machine Components
Repeat
Fetch Instruction
Emulate Operation of the current instruction
Process interrupts
Update instruction counter
Until MachineIsOff
Perform shutdown procedure
END
```

Many variations exist, in particular to process interrupts within the emulation of a lengthy operation or to optimize throughput, but the general principle and structure are fairly constant.

Emulation of CPU Operation

One of the advantages of microprogramming is that the designer can implement his or her *dream instructions* simply by emulating its operation. We have seen already the code for a typical 16-bit adder, but it is not difficult to code a parity code generator, a cyclic redundancy check calculator, or an instruction that returns the eigenvalues of an $n \times n$ matrix. This part is straight programming. One consideration is to make sure that the machine is still listening to the outside world (interruptions) or actively monitoring it (I/O flags) in order not to loose asynchronous data while looking for a particular pattern in a 1 megabyte string. Another consideration is to optimize memory usage by combining common processes for different operations. For example, emulating a 32-bit add instruction and emulating a 16-bit add instruction have common parts. This is, however, a programming concern not specific to emulation.

I/O System and Interrupts

Programming support for I/Os and interrupts is more complicated than for straight machine instructions. This is due to the considerable speed differences between I/O devices and a CPU, the need for synchronization, the need for not losing any external event, and the concerns for optimizing processing time. Microprogramming offers considerable design flexibility, as these problems are more easily handled by programming than with hardware components.

Applications of Microprogramming

The main application of microprogramming is the emulation of a virtual machine architecture on a different host hardware machine. It is, however, easy to see that the concept of emulation can be broadened to other functions than the traditional hardware operation.

It is mainly a matter of point of view. Emulation and simulation are essentially the same process but viewed from different levels. Realizing a 64-bit addition and implementing a communication controller are qualitatively the same type of task. Once this is considered, there are theoretically no limits to the uses of microprogramming.

From the programmer's point of view, programming is the activity of producing, in some language, an implementation of some algorithm. If the language is at the very lowest level, as is the case with microprogramming, and at the same time the algorithm is filled with intricate data structures and complex decisions, the task might be enormous, but nothing says it cannot be done (except, maybe, experience). With this perspective of the field, we now look at some existing applications of microprogramming.

Operating System Support

One of the first applications, besides emulation, was to support some operating system functions. Since microprograms are closer to the hardware and programming directly in microcode removes the overhead of decoding machine-level instructions, it was thought that directly coding operating system (OS) functions would improve their performance. Success was achieved in some areas, such as virtual memory. In general, people write most OS functions in assembly language, probably because the cost is not offset by the benefits, especially with rapidly changing OS versions. The problems raised by the human side of programming have changed the question "Should it be in microcode or in assembler?" to the question "Should it be in assembler or in C?" This parallels the CISC/RISC debate.

High-Level Languages Support

Early research was done also in the area of support for high-level languages. Support can be in the form of microprogrammed implementations of some language primitive (for example, the trigonometric functions) or support for the definition and processing of data structures (for example, trees and lists primitives). Many interesting research projects have led to esoteric laboratory machines. More common examples include the translate instructions, string searches and compares, or indexing multidimensional arrays.

Paging, Virtual Memory

An early and typical application of microprogramming is the implementation of the paging algorithm for a virtual memory system. It is a typical application since it is a low-level function that must be time optimized and is highly hardware dependent. Furthermore, the various maintenance functions which are required by the

paging algorithms and the disk I/Os can be done during the idle time of the processing of other functions or during part of that processing in order to avoid I/O delays.

Diagnostics

Diagnostic functions have also been an early application of microprogramming. A firmware implementation is ideally suited to test the various components of a computer system, since the gates, paths, and units can be exercised in an isolated manner, therefore allowing one to precisely pinpoint the trouble area.

Controllers

Real-time controllers benefit from a microprogrammed implementation, due to the speed gained by programming only the required functions, therefore avoiding the overhead of general-purpose instructions. Since the microprogrammer can better make use of the available parallelism in the machine, long processes can still support the asynchronous arrival of data by incorporating the interrupt polling at intervals in these processes.

High-Level Machines

Machines that directly implement the constructs of high-level languages can be easily implemented via microprogramming. For example, Prolog machines and Lisp machines have been tried. It is also possible to conceive an application directly microcoded. Although this could provide a high performance hardware, human errors and software engineering practice seem to make such a machine more of a curiosity than a maintainable system.

Defining Terms

Control memory: A memory containing a set of microinstructions (a microprogram) that defines the instruction set and operations of a CPU.

Emulator: The firmware that simulates a given machine architecture.

Firmware: Meant as an intermediate between software, which can be modified very easily, and hardware, which is practically unchangeable (once built); the word *firmware* was coined to represent the microprogram in control memory, i.e., the modifiable representation of the CPU instruction set.

High-level language for microprogramming: A high-level language more or less oriented toward the description of a machine. Emulators can more easily be written in a high-level language; the source code is compiled into the microinstructions for actual implementation.

Horizontal microinstruction: Theoretically, a completely horizontal microinstruction is made up of all the possible microcommands available in a given CPU. In practice, some encoding is provided to reduce the length of the instruction.

Microcommand: A small bit field indicating if a gate is open or closed, if a function is enabled or not, if a control path is active or not, etc. A microcommand is therefore the specification of some action within the control structure of a CPU.

Microinstruction: The set of microcommands to be executed or not, enabled or not. Each field of a microinstruction is a microcommand. The instruction specifies the new state of the CPU.

Vertical microinstruction: A completely vertical microinstruction would contain one field and therefore would specify one microcommand. An Op code is used to specify which microcommand is specified. In practice, microinstructions that typically contain three or four fields are called vertical.

Related Topic

86.3 Architecture

References

T. Agerwala, "Microprogram optimization: a survey," *IEEE Trans. Comput.*, vol. C25, no. 10, pp. 862–873, 1976.
J.D. Bagley, "Microprogrammable virtual machines," *Computer*, pp. 38–42, 1976.
D.K. Banerji and J. Raymond, *Elements of Microprogramming*, Englewood Cliffs, N.J.: Prentice-Hall, 1982.
G.F. Casaglia, "Nanoprogramming vs. microprogramming," *Computer*, pp. 54–58, 1976.

S.R. Das, D. K. Banerji, and A. Chattopadhyay, "On control memory minimization in microprogrammed digital computers," *IEEE Trans. Comput.,* vol. C22, no. 9, pp. 845–848, 1973.

L.H. Jones, "An annotated bibliography on microprogramming," *SIGMICRO Newsletter,* vol. 6, no. 2, pp. 8–31, 1975.

L.H. Jones, "Instruction sequencing in microprogrammed computers," *AFIPS Conf. Proc.,* vol. 44, pp. 91–98, 1975.

K. Malik and T.J. Lewis, "Design objectives for high level microprogramming languages," in *Proceedings of the 11th Annual Microprogramming Workshop,* Englewood Cliffs, N.J.: Prentice-Hall, 1978, pp. 154–160.

J. Raymond and D.K. Banerji, "Using a microprocessor in an intelligent graphics terminal," *Computer,* pp. 18–25, 1976.

M.V. Wilkes, W. Renwick, and D.J. Wheeler, "The design of the control unit of an electronic digital computer," *Proc. IEE,* pp. 121–128, 1958.

Further Information

J. Carter, *Microprocessor Architecture and Microprogramming, a State Machine Approach,* Englewood Cliffs, NJ: Prentice-Hall, 1996.

S. Habib, Ed., *Microprogramming and Firmware Engineering Methods,* New York: Van Nostrand Reinhold, 1988.

H. Okuno, N. Osato, and I. Takeuchi, "Firmware approach to fast lisp interpreter. Twentieth annual workshop on microprogramming", (MICRO-20), *ACM,* 1987.

A. J. Van der Hoeven, P. Van Prooijen, E. F. Deprettere, and P. M. Dewilde, "A hardware design system based on object-oriented principles", *IEEE,* pp. 459–463, 1993.

Feldman, J.M., Czeck, E.W., Lewis, T.G., Martin, J.J. "Programming"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

87

Programming

James M. Feldman
Northeastern University

Edward W. Czeck
Northeastern University

Ted G. Lewis
Navel Postgraduate School

Johannes J. Martin
University of New Orleans

87.1 Assembly Language
NumberCount( ) • Comparisons Down on the Factory Floor • Compiler Optimization and Assembly Language

87.2 High-Level Languages
What Is a HLL? • How High Is a HLL? • HLLs and Paradigms

87.3 Data Types and Data Structures
Abstract Data Types • Fundamental Data Types • Type Constructors • Dynamic Types • More Dynamic Data Types • Object-Oriented Programming

87.1 Assembly Language

James M. Feldman and Edward W. Czeck

The true language of computers is a stream of 1s and 0s—bits. Everything in the computer, be it numbers or text or program, spreadsheet or database or 3-D rendering, is nothing but an array of bits. The meaning of the bits is in the "eye of the beholder"; it is determined entirely by context. Bits are not a useful medium for human consumption. Instead, we insist that what *we* read be formatted spatially and presented in a modest range of visually distinguishable characters. 0 and 1 arranged in a dense, page-filling array do not fulfill these requirements *in any way*. The several languages that are presented in this handbook are all intended to make something readable to two quite different readers. On the one hand, they serve the human reader with his/her requirements on symbols and layout; on the other, they provide a grammatically regular language for interpretation by a **compiler**. A compiler, of course, is normally a program running on a computer, but human beings can and sometimes do play both sides of this game. They want to play with both the input and output. Such accessibility requires that not only the input but the output of the compilation process be comfortably readable by humans. The language of the input is called a **high-level language (HLL)**. Examples are C, Pascal, Ada and Modula II. They are designed to express both regularly and concisely the kinds of operations and the kinds of constructs that programmers manipulate. The output end of the compiler generates **object code**—a generally unreadable, binary representation of machine language, lacking only the services of a **linker** to turn it into true machine language. The language that has been constructed to represent object code for human consumption is *assembly language*. That is the subject of this section.

Some might object to our statement of purpose for assembly language. While few will contest the concept of assembly language as the readable form of object code, some see writing assembly code as the way to "get their hands on the inner workings of the machine." They see it as a "control" issue. Since most HLLs today give the user reasonably direct ways to access hardware, where does the "control" issue arise? What assembly proponents see as the essential reason for having an assembly language is the option to optimize the "important" sections of a program by doing a better job of machine code generation than the compiler does. This perspective was valid enough when compilers were mediocre optimizers. It was not unlike the old days when a car came with a complete set of tools because you needed them. The same thing that has happened to cars has happened to compilers. They are engineered to be "fuel efficient" and perform their assigned functions with remarkable

ability. When the cars or compilers get good enough and complex enough, the tinkerer may do more harm than good. IBM's superscalar RISC computer—the RS6000—comes with superb compilers *and no* ***assembler*** *at all.* The Pentagon took a long look at their costs of programming their immense array of computers. Contrary to popular legend, they decided to save money. The first amendment not withstanding, their conclusion was: "Thou shalt not assemble."

The four principal reasons for *not* writing assembly language are

- Any sizable programming job gets done at least four times faster in a HLL.
- Most modern compilers are good optimizers of code; some are superb.
- Almost all important code goes through revisions—maintenance. Reworking old assembly code is similar to breaking good encryption; it takes forever.
- Most important of all is portability. To move any program to a different computer, you must generate machine code for that new platform. With a program in a HLL, a new platform is almost free; all it requires is another pass through the compiler for the target platform. With assembly code, you are back to square one. Assembly code is unique to the platform.

Given all of that, the question naturally arises: Why have an article on assembly language? We respond with two reasons, both of which we employ in our work as teachers and programmers:

- An essential ingredient in understanding computer hardware and in designing new computer systems and compilers is a detailed appreciation of the operations of central processing units (CPUs). These are best expressed in assembly language. Our undergraduate Computer Engineering courses include a healthy dose of assembly language programming for this specific reason.
- If you are concerned about either CPU design or compiler effectiveness, you have to be able to look in great detail at the interface between them—*machine language.* As we have said, the easiest way to read machine language is by translating it to assembly language. This is one way to get assembly language, not by writing in it as a source of code but by running the object code itself through a backward translator called a **disassembler**. While many compilers will oblige you by providing an assembly listing if asked, often that listing does not include optimizations that occur only when the several modules are linked together, providing opportunities for truly global optimization. Some compilers "help" the reader by using **macros** (names for predefined blocks of code) in place of the real machine instructions and register assignments. The absence of the optimizations and the inclusion of unexpected macros can make the assembly listing almost useless for obtaining insight into the program's fine detail. The compilers that we have used on the DECstations and SPARC machines do macro inclusion. To see what is really going on in these machines, you must disassemble the machine code. That is precisely what the Think C® compiler on the Macintosh does when you ask for machine code. It disassembles what it just did in compiling and linking the whole program. What you see is what is really there. The code we present for the 68000 was obtained in that way.

These are important applications. Even if most or all other programming needs can be better met in HLLs, these applications are sufficient reason for many engineers to want to know something about assembly language.

There are other applications of assembly language, but they tend to be specific to rather specialized and infrequent tasks. For example, the back end of most HLL compilers is a machine code generator. To write one of those, you certainly must know something about assembly language. On rare occasions, you may find some necessary machine-specific transaction which is not supported by the HLL of choice or which requires some special micro optimization. A "patch" of assembly code is a way to fit this inexpressible thought into the program's vocabulary. These are rare events. The reason why we recommend to you this section on assembly code is that it improves your understanding of HLLs and of computer architecture.

We will take a single subroutine which we express in C and look at the machine code that is generated on two representative machines. The machines include two widely used *complex instruction set computers* (**CISCs**) and one *reduced instruction set computer* (**RISC**). These are the 68000®, the VAX®, and a SPARC®. We will have two objectives:

- To see how a variety of paradigms in HLLs are translated (or, in other words, to see what is really going on when you ask for a particular HLL operation)
- To compare the several architectures to see how they are the same and how they differ

The routine attempts to get a count of the number of numbers which occur in a block of text. Since we are seeking numbers and not *digits*, the task is more complex than you might first assume. This is why we say "attempts." The function that we present below handles all of the normal text forms:

- Integers, such as 123 or –17
- Numbers written in a fixed-point format, such as 12.3 or 0.1738
- Numbers written in a floating-point format, such as –12.7e+19 or 6.781E2

If our program were to scan the indented block of code above, it would report finding six numbers. The symbols that the program recognizes as potentially part of a number include the digits 0 to 9 and the symbols 'e', 'E', ' . ', '–' and '+'. Now it is certainly possible to include other symbols in legitimate numbers, such as HEX numbers or the like, but this little routine will not properly deal with them. Our purpose was not to handle all comers but to provide a routine with some variety of expression and possible application. Let us begin.

NumberCount( )

We enter the program at the top with one pointer passed from the calling routine and a set of local variables comprising two integers and eight Boolean variables. Most of the Boolean variables will be used in pairs. The first element of a pair, for instance, *ees* of *ees* and *latche*, indicates that the current character is one of a particular class of non-numeric characters which might be found inside a number. If you consider that the number begins at the first digit, then these characters can occur legally only once within a given number. *ees* will be set true if the current character is the *first instance* of either 'e' or 'E'. The paired variable, *latche*, is set true if there has ever been one of those characters in the current number. The other pairs are *period* and *latchp* and *sign* and *latchs*.

There is also a pair of Booleans which indicate if the current character is a *digit* and if the scanner is currently *inside* a number. Were you to limit your numbers to integers, these two are the only Booleans which would be needed. At the top of the program, all Booleans are reset (made FALSE). Then we step through the block looking for numbers. The search stops when we encounter the first null [char(0)] marking the end of the block. Try running through the routine with text containing the three forms of number. You will quickly convince yourself that the routine works with all normal numbers. If someone writes "3..14" or "3.14ee6", the program will count 2 numbers. That is probably right in the first two cases. Who knows in the third?

Let us look at this short routine in C.

```
# define blk_length 20001
int NumberCount(char block[])
{  int count=0,inside=0,digit;
   int ees=0, latche=0, latchp=0, period=0, latchs=0, sign=0;
   char *source;

   source = block;
   do {
     digit = (*source >= '0') && (*source <= '9');
     period = (*source=='.') && inside && !latchp; && !latche;
     latchp = (latchp || period);
     ees = ((*source=='E') || (*source=='e')) && inside && !latche;
     latche = (latche || ees);
     sign = ((*source=='+') || (*source=='-')) && inside && latche && !latchs;
     latchs = (latchs || sign);
     if (inside) {
       if (!(digit || ees || period || sign)) inside=latchp=latche=latchs=0;
     }
     else if (digit) {
```

```
    count++;
    inside = 1;
   }
 source++;
 }
 while ((*source != '\0') && ((source-block)<blk_length+1));
 return count;
 }
```

To access values within the character array, the normal C paradigm is to step a pointer along the array. *Source* points at the current character in the array; **source* is the character ("what *source* points at"). *source* is initialized at the top of the program before the loop (source = block;) and incremented (source++;) at the bottom of the loop. Note the many repetitions of **source*. Each one means the same current character. If you read that expression as *the character which source is pointing to*, it looks like an invitation to fetch the same character from memory eight times. A compiler that optimizes by removing *common subexpressions* should eliminate all but the first such fetch. This optimization is one of the things that we want to look for.

For those less familiar with C, the meanings of the less familiar symbols are:

`==`	equal (in the logical sense)
`!`	not
`!=`	not equal
`&&`	and
`||`	or
`count++`	increment *count* by 1 unit (after using it)

C uses 0 as FALSE and anything else as TRUE.

Comparisons Down on the Factory Floor

Now let us see what we can learn by running this program through compilers on several quite different hosts. The items that we wish to examine include:

I. Subroutine operations comprising:
   A. Building the call block
   B. The call itself
   C. Obtaining memory space for local variables
   D. Accessing the call block
   E. Returning the function value
   F. Returning to the calling routine

II. Data operations
   A. Load and store
   B. Arithmetic
   C. Logical
   D. Text

III. Program control
   A. Looping
   B. if and the issue of multiple tests

Our objectives are to build three quite different pictures:

- An appreciation for the operations underlying the HLL statements
- An overview of the architectures of several important examples of CISC and RISC processors
- An appreciation for what a HLL optimizer should be doing for you

We will attempt to do all three all of the time.

Let us begin with the calling operations. Our first machine will be the MC68000, one of the classical and widely available CISC processors. It or one of its progeny is found in many machines and forms the heart of the Macintosh (not the PowerMac) and the early Sun workstations. Programmatically, the 68000 family shares a great deal with the very popular VAX family of processors. Both of these CISC designs derive in rather linear fashion from DEC's PDP-11 machines that were so widely used in the 1970s. Comparisons to that style of machine will be done with the SPARC, a RISC processor found in Sun, Solbourne, and other workstations.

Memory and Registers

All computers will have data stored in memory and some space in the CPU for manipulating data. Memory can be considered to be a long list of bytes (8-bit data blocks) with *addresses* (locations in the list) spanning some large range of numbers from 0 to typically 4 billion (4 GB). The memory is constructed physically by grouping chips so that they appear to form enormously deep columns of bytes, as shown in Fig. 87.1. Since each column can deliver one byte on each request, the number of adjacent columns determines the number of bytes which may be obtained from a single request. Machines today have 1, 2, 4, or 8 such columns. (Some machines, the 68000 being our current example, have only 2 columns but arrange to have the CPU ask for two successive transfers to get a total of 4 bytes.) In general, the CPU may manipulate in a single step a datum as wide as the memory. For all of the machines which we will consider, that maximum datum size is 32 bits or 4 bytes. While convention would have us call this biggest datum a *word*, historical reason has led both the VAX and MC68000 to call it a *longword*. Then, 2 bytes is either a *halfword* or a *word*. We will use the VAX/68000 notation (longword, word, and byte) wherever possible to simplify the reading. To load data from memory, the CPU sends the address and the datum size to the memory and gets the datum as the reply. To store data, the address is sent and then the datum and datum size.

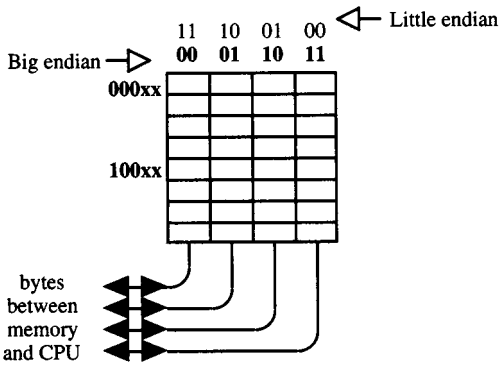


FIGURE 87.1 Memory arranged as 4 columns of bytes. The binary addresses are shown in the two formats widely used in computers. The illustration shows only 32 bytes in a 4 3 8 array, but a more realistic span would be 4 3 1,000,000 or 4 3 4,000,000 (4 MB to 16 MB).

Some machines require that the datum be properly aligned with the stacking order of the memory columns in Fig. 87.1. Thus, on the SPARC, a longword must have an address ending in 00 (xxx00 in Fig. 87.1), and a word address must end in 0. The programmer who arranges to violate this rule will be greeted with an **address error**. Since the MC68000 has only two columns, it complains only if you ask for words or longwords with odd addresses. Successor models of that chip (68020, 30, and 40), like the VAX, accept any address and have the CPU read two longwords and do the proper repacking.

The instruction explicitly specifies the size and indicates how the CPU should calculate the address. An instruction to load a byte, for example, is `LB`, `MOVE.B`, or `MOVB` on the SPARC, MC68000, and VAX, respectively. These are followed immediately by an expression which specifies an address. We will discuss how to specify an address later. First, we must introduce the concept of a register.

The space for holding data and working on it in the CPU is the *register set*. Registers are a very important resource. Bringing data in from memory is quite separate from any operations on that data. Data in memory must first be fetched, then acted upon. Data in registers can be acted on immediately. Thus, the availability of registers to store very active variables and intermediate results makes a processor inherently faster. In some machines, most or all of the registers are tied to specific uses. The most prevalent example would be Intel's 80x86 processors, which power the ubiquitous PC. Such architectures, however, are considered quite old-fashioned. All of the machines that we are considering are of a type called *general register machines* in that they have a large group of registers which may be used for any purpose. The machines that we include have either 16 or 32 registers, with only a few tied to specific machine operations.

Table 87.1 shows the general register resources in the three machines. The SPARC is a little strange. The machine provides eight *global* registers and then a *window blind* of 128 registers which sits behind a frame

TABLE 87.1 General Registers in the Three Machines

	Reg	Special	Names	Comments
MC68000	16	1	D0..D7 A0..A7	A(ddress) register operations are 32 bits wide. Address generation uses A registers as bases. D (data) registers allow byte, word, and longword operations. A7 is SP.
VAX	16	4	r0..r11 AP, FP, SP, PC	AP,FP,SP and PC hold the addresses of the argument block, the frame, the stack and the current place in the program, respectively. All data instructions can use any register.
SPARC	32 (136)	4	zero, g1..g7, i0..i5, FP, RA, l0..l7, o0..o5, SP, o7	The 4 groups of eight registers comprise: global (g), incoming parameters (i), local (l) and outgoing parameters (o). g0 is a hardwired 0 as a data source and a wastebasket as a destination. The registers are arranged as a window blind (see text) with the g's always visible and the others moveable in multiple overlapping frames of 24.

The special registers are within the set of general registers. Where a PC is not listed, it exists as a special register and can be used as an address when the program uses *program-relative* addressing.

which exposes 24 of the 128. A program can ask the machine to raise or lower the blind by 16 registers. That leaves an overlap of eight between successive yanks or rewinds. This arrangement is called a *multiple overlapping register set* (MORS). If you think of starting with register r8 at the bottom and r31 at the top, a yank of 16 on the blind will now have r49 at the top and r24 at the bottom. r24 to r31 are shared between the old set and the new. To avoid having to keep track of which registers are showing, the set of 24 are divided into what came *in* from the last set, those that are only *local*, and those that will go *out* to the next set. These names apply to going toward increasing numbers. In going the other direction, the *ins* of the current set will become the *outs* of the next set. Almost all other machines keep their registers screwed down to the local masonry, but you will see in a moment how useful a MORS can be. (Like other useful but expensive accessories, the debate is always on whether it is worth it [Patterson and Hennessy, 1989].)

Stack. Most subroutines define a number of local variables. NumberCount in C, for example, defines 10 local variables. While these local variables will often be created and kept in register, there is always some need for a bit of memory for each *invocation of* (call to) a subroutine. In the "good old days," this local storage was often tied to the block of code comprising the subroutine. However, such a fixed block means that a subroutine could never call itself or be called by something that it called. To avoid that problem (and for other purposes) a memory structure called a *stack* was invented which got its name because it behaved like the spring-loaded plate stack in a restaurant. Basically, it is a *last-in-first-out* (LIFO) structure whose top is defined by a pointer (address) which resides in a register commonly called the *stack pointer* or SP.

Heap. When a subroutine needs space to store local variables, it acquires that space on the stack. When the subroutine finishes, it returns that stack space for use by other routines. Thus, local variable allocations live and die with their subroutines. It is often necessary to create a data structure which is passed to other routines whose lives are independent of the creating routine. This kind of storage must be independent of the creator. To meet this need, the *heap* was invented. This is an expandable storage area managed by the system. You get an allocation by asking for it [*malloc* (*structure_size*) in C]. You get back a pointer to the allocation and the routine can pass that pointer to any other routine and then go away. When it comes time to dispose of the allocation—that is, return the space for other uses—the program must do that actively by a deallocation call [*free*(*pointer*) in C]. Thus, one function can create a structure, several may use it, and another one can return the memory for other uses, all by passing the pointer to the structure from one to another.

Both heap and stack provide a mechanism to obtain large (or small) amounts of storage dynamically. Thus, large structures which are created only at run time need not have static space stored for them in programs that are stored on disk nor need they occupy great chunks of memory when the program does not need them. Dynamic allocation is very useful and all modern HLLs provide for it.

Since there are two types of dynamic storage, there must be some way to lay out memory so that unpredictable needs in either stack or heap can be met at all times. The mechanism is simplicity itself. The program is stuffed into low addresses in memory along with any static storage (e.g., globals) which are declared in the program. The entire remaining space is then devoted to dynamic storage. The heap starts right after the program and

grows toward higher addresses; the stack goes in at the top of memory and grows down. The system is responsible to see that they never collide (a *stack crash*). When it all goes together, it looks like Fig. 87.2 [Aho et al., 1986].

There is one last tidbit that an assembly programmer must be aware of in looking at memory. Just as some human alphabets are written left to right and some right to left (not to mention top to bottom), computer manufacturers have chosen to disagree on how to arrange words in memory. The two schemes are called *big-endian* and *little-endian* (after which end of a number goes in the lowest-numbered byte and also after a marvelous episode in *Gulliver's Travels*). The easiest way to perceive how it is done in the two systems is to think of all numbers as being written in conventional order (left to right), but for big-endian you start counting on the upper left of the page and on little-endian you start counting on the upper right (see Fig. 87.1). Since each character in a text block is *a number* of length 1 byte, this easy description makes big-endian text read in normal order (left to right) but little-endian text reads from right to left. Figure 87.3 shows the sentence "This is a sentence" followed by the two hexadecimal (HEX) numbers 01020304 and 0A0B0C0D written to consecutive bytes in the two systems. Why must we bring this up? Because anyone working in assembly language must know how the bytes are arranged. Furthermore, two of the systems we are considering are big-endian and one (the VAX) is little-endian. Which is the better system? Either one. It is having both of them that is a nuisance.

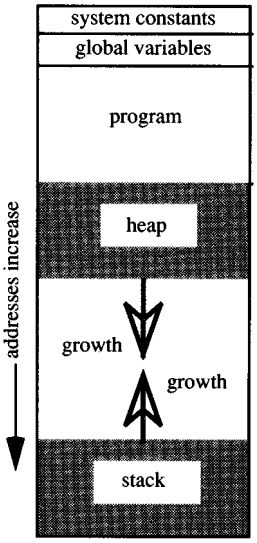


FIGURE 87.2 Layout of a program, static storage, and dynamic storage in memory.

As you look at Fig. 87.3, undoubtedly you will prefer big-endian, but that is only because it appeals to your prejudices. In truth, either works well. What is important is that you be able to direct your program to go fetch the item of choice. In both systems, you use the lowest-numbered byte to indicate the item of choice. Thus, for the number 01020304, the address will be 13. For the big-endian system, 13 will point to the byte containing 04 and for the little-endian system, it will point at the byte containing 01.

Figure 87.3 contains a problem for some computers which we alluded to in the discussion of Fig. 87.1. We have arranged the bytes to be four in a row as in Fig. 87.1. That is the way that the memory is arranged in two of our three machines. (In the 68000, there are only two columns.) A good way to look at the fetch operation is that the memory always delivers a whole row and then the processor must acquire the parts that it wants and then properly arrange them. (This is the effect if not always the method.) Some processors—the VAX being a conspicuous example—are willing to undertake getting a longword by fetching two longwords and then piecing together the parts that it wants. Others (in our case, the 68000 and the SPARC) are not so accommodating. Those machines opt for simplicity and speed and require that the program keep its data aligned. To use one of those machines, you (or the compiler or assembler) must rearrange Fig. 87.3 by inserting a null byte into Fig. 87.2. This modification is shown in Fig. 87.4. With this modification, all three machines could fetch the two numbers in one operation without rearrangement.

Look closely at the numbers 01020304 and 0A0B0C0D in Fig. 87.4. Notice that for both configurations, the numbers read from left to right and that (visually) they appear to be in the same place. Furthermore, as pointed out in the discussion of Fig. 87.3, the "beginning" or address of each of the numbers is identical. However, the byte that is pointed at by the address is not the same and the internal bytes do not have the same addresses. Getting big-endian and little-endian machines in a conversation is not easy. It proves to be even more muddled than these figures suggest. A delightful and cogent discussion of the whole issue is found in Cohen [1981].

The principal objective in this whole section has been accomplished if looking at Fig. 87.4 and given the command to load a byte from location 0000 0019, you get the number 0B in the big-endian machine and 0C in the little-endian machine.

If you are not already familiar with storing structures in memory, look at the string (sentence) and ask how those letters get in memory. To begin with, every typeable symbol and all of the unprintable actions such as tabbing and carriage returns have been assigned a numerical value from the *ASCII code*. Each assignment is a byte-long number. What "This" really looks like (HEX, left to right) is 54 68 69 73. The spaces are HEX 20; the period 2E. With the alignment null byte at the end, this list of characters forms a proper C string. It is a structure of 20 bytes. A structure of any number of bytes can be stored, but from the assembly point of view,

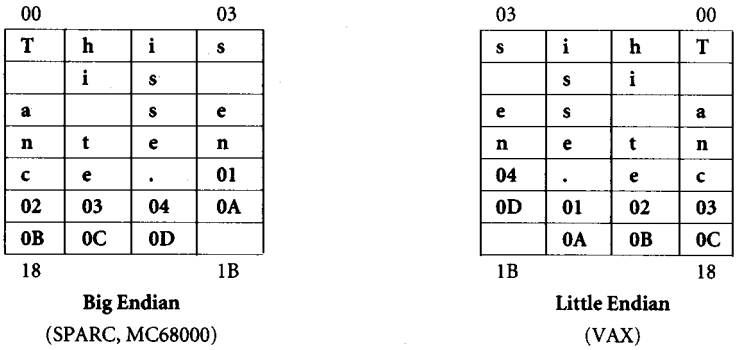


FIGURE 87.3 Byte numbering and number placement for big- and little-endian systems. Hexadecimal numbers are used for the memory addresses.

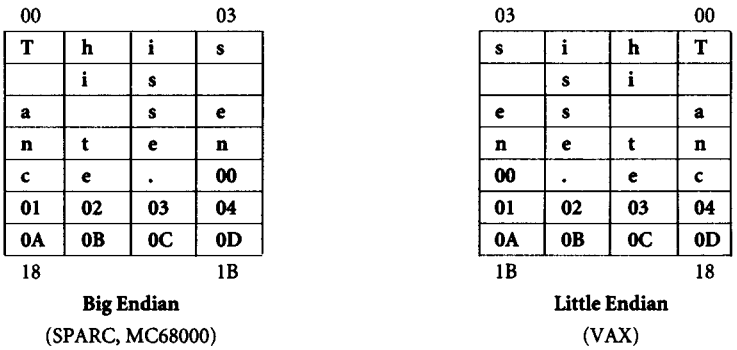


FIGURE 87.4 The same items as in Fig. 87.3, but with justification of the long integers to begin on a longword boundary.

it is all just a list of bytes. You may access them two at a time, four at a time, or one at a time. Any interpretation of those bytes is entirely up to the program. Unlike the HLL which requires that you tell it what each named variable is, assembly language knows only bytes and groups of bytes. In assembly language, the "T" can be thought of as a letter or the number 54 (HEX). Your choice. Or, more importantly, your program's choice.

Addressing

Now that we have both memory and addresses, we should next consider how these processors require that programmers specify the data that is to be acted upon by the instructions.

All of these machines have multiple modes of address. The VAX has the biggest vocabulary; the SPARC the smallest. Yet all can accomplish the same tasks. Four general types of address specification are quite common among assembly languages. These are shown in Table 87.2. They are spelled out in words in the table, but their usage is really developed in the examples which follow in this and the succeeding sections.

In Table 87.2, formats 1.4 and 1.5 and the entries in 4 require some expansion. The others will be clear in the examples we will present. Base-index addressing is the mechanism for dealing with subscripts. The base points at the starting point of a data structure, such as a string or a vector; the index measures the offset from the start of the structure to the element in question. For most machines, the index is simply a separate register which counts the bytes from the base to the item in question. If the items in the list are 4 bytes long, then to increment the index, you add 4. While that is not hard to remember, the VAX does its multiplication by the item length for you. Furthermore, it allows you to index any form of address that you can write. To show you what that means, consider expanding numbers stored in words into numbers stored in longwords. The extension is to preserve sign. The VAX provides specific instructions for conversions. If we were moving these words in one array to longwords in another array, we would write:

TABLE 87.2 Addressing Modes

1.	**Explicit addresses**	**Example**	
	1.1. Absolute addressing	765	The actual address written into the instruction.
	1.2. Register indirect	(r3)	Meaning "the address is in register 3."
	1.3. Base-displacement	–12(r3)	Meaning "12 bytes before the address in register 3."
	1.4. Base-index	(r3,r4)	Meaning make an address by adding the contents of r3 and r4. This mode has many variations which are discussed below.
	1.5. Double indirect	@5(r4)	Very uncommon! Means calculate an address as in 1.3, then fetch the longword there, and then use it as the address of what you really want.
2.	**Direct data specification**		
	2.1. Immediate/literal	#6 or 6	Meaning "use 6 as the datum." In machines which use #6, 6 without # means address 6. This is called "absolute addressing."
3.	**Program-relative**		
	3.1. Labels	loop:	The label (typically an alphanumeric ending in a colon) is a marker in the program which the assembler and linker keep track of. The common uses are to jump to a labeled spot or to load labeled constants stored with the program.
4.	**Address-modifying forms** (CISC only)		
	4.1. Postincrement	(sp)+	Same as 1.2 except that, after the address is used, it is incremented by the size of the datum in bytes and returned to the register from which it came.
	4.2. Predecrement	–(sp)	The value in SP is decremented by the size of the datum in bytes, used as the address and returned to the register from which it came.

```
CVTWL (r4)[r5],(r6)[r5]    ;convert the words starting at (r4) to longwords starting at (r6)
```

Note that the same index, [r5], is used for both arrays. On the left, the contents of r5 are multiplied by 2 and added to r4 to get the address; on the right, the address is r5*4+r6. You would be saying: "Convert the 4th word to the 4th longword." This is undoubtedly compact and sometimes convenient. It is also unique to the VAX.

For the 68000, the designers folded both base-displacement and base-index into one mode and made room for word or longword indices. It looks like:

```
add.1 64(A3,D2.w),D3    ;address = (A3+64) +sign-extended(D2)
```

The 68000 limits the displacement to a signed byte, but other than that, it is indeed a rather general indexing format. If you do not want the displacement, set it to 0.

For the powerful but simple SPARC, the simple base-index form shown in 1.4 is all that you have (or need).

The double-indirect format, 1.5, is so rarely used that it has been left out of almost all designs but the VAX. What makes it occasionally useful is that subroutines get pointers to "pass by pointer" variables. Thus, if you want to get the variable, first you must load the address and then the variable. The VAX allows you to do this in one instruction. While that sounds compact, it is expensive in memory cycles. If you want to use that pointer again, it pays to have it in register.

The two items under heading 4 are strange at first. Their principal function is adding items to and removing them from a dynamic stack, or for C, to execute the operation *X++ or *(– –X). The action may be viewed with the code below and the illustration of memory in Fig. 87.2:

```
movl r4, –(sp)          ;make room on the stack (subtract 4 from SP) and put
                         the contents of r4 in that spot

movl (sp)+, r4          ;take a longword off the stack, shorten the stack by 4
                         bytes, and put the longword in r4
```

RISCs abhor instructions which do two unrelated things. Instead of using a dynamic stack, they use a quasi-static stack. If a subroutine needs 12 bytes of stack space, it explicitly subtracts 12 from SP. Then it works from there with the base-displacement format (1.3) to reference any place in the block of bytes just defined. If you want to use a pointer and then increment the pointer, RISCs will do that as two independent instructions.

Let us consider one short section of MC68000 code from our sample program in C to see how these modes work and to sample some of the flavor of the language:

```
;ees = ((*source=='E') || (*source=='e')) && inside && !latche;

            CMPI.B      #$45,(A4)   ; 'E'   "compare immediate"    literal hex 45,
                                                                    what A4 points at
            BEQ         first       ;       "branch if equal"     to label first
            CMPI.B      #$65,(A4)   ; 'e'   "compare immediate"    literal hex 65,
                                                                    what A4 points at
            BNE         second      ;       "branch if not equal" to label second
first:
            TST.W       D6          ;       "test word" (subtract 0) D6 ('inside')
            BEQ         second      ;       "branch if equal"     to label second
            TST.W       D3          ;       "test word" (subtract 0) D3 ('latche')
            BEQ         third       ;       "branch if equal"     to label third
second:
            MOVEQ       #00,D0      ;       "move quick"          literal 0 to D0
            BRA         fourth      ;       "branch always"       to label fourth
third:
            MOVEQ       #$01,D0     ;       "move quick"          literal 1 to D0
fourth:
            MOVE.W      D0,-6(A6)   ;       "move word"           from D0 to -6(FP)
```

There are all sorts of little details in this short example. For example, a common way to indicate a comment is to start with a ";". The assembler will ignore your comments. The "#" indicates a literal, and the "$" that the literal is written in hexadecimal notation. The VAX would use #^x to express the same idea. "Compare" means "subtract but save only the **condition codes** of the result" (*v* or *overflow*, *n* or *negative*, *z* or *zero*, and *c* or *carry*). Thus, the first two lines do a subtraction of whatever A4 is pointing at (*source) from the ASCII value for 'E' and then, if the two were equal (the result, zero), the program jumps to line 5. If *source is not 'E', then it simply goes to the next line, line 3. The instruction, TST.W D6, is quite equivalent to CMPI.W D6, #0, but the TST instruction is inherently shorter and faster. On a SPARC, where it would be neither shorter nor faster, TST does not exist.

Exactly what the assembler or linker does to replace the label references with proper addresses, while interesting, is not particularly germane to our current topic. Note that the range of the branch is somewhat limited. In the 68000, the maximum branch is ±32K and in the VAX a mere +127 to –128. If you need to go further, you must combine a branch with a jump. For example, if you were doing BEQ farlabel, you would instead do:

```
    BNE     nearlabel
    jmp     farlabel    ;   this instruction can go any distance
nearlabel:
```

Follow through the example above until the short steps of logic and the addressing modes are clear. Then progress to the next section where we use the addressing modes to introduce the general topic of subroutine calling conventions.

Calling Conventions

Whenever you invoke a subroutine in a HLL, the calling routine (*caller*) must pass to the called routine (*callee*) the parameters that the subroutine requires. These parameters are defined at compile time to be either *pass-by-value* or *pass-by-pointer* (or *pass-by-reference*), and they are listed in some particular order. The convention for passing the parameters varies from architecture to architecture and HLL to HLL, but basically it always consists of building a *call block* which contains all of the parameters and which will be found where the recipient expects to find it.

Along with the passing of parameters, for each system, a convention is defined for register and stack use which establishes:

- Which registers must be returned from callee to caller with the same contents that the callee received (such registers are said to be *preserved across a call*)
- Which registers may be used without worrying about their contents (such registers are called *scratch registers*)
- Where the return address is to be found
- Where the value returned by a function will be found

The convention may be supported by hardware or simply a gentlemanly rule of the road. However the rules come into being, they define the steps which must be accomplished coming into and out of a subroutine. The whole collection of such rules forms the *calling convention* for that machine. In this section, we look at our three different machines to see how all accomplish the same tasks but by rather different mechanisms.

The two CISCs do almost all of their passing and saving on the stack. The call block will be built on the stack; the return address will be put on the stack; saved registers will be put on the stack. Only a few stack references are passed forward in register; the value returned by the function will be passed back in register.

How different is the SPARC! The parameters to be passed are placed in the *out* registers (six are available for this purpose). Only the overflow, if any, would go on the stack. In general, registers are saved by window-blinding rather than moving them to the stack. On return, data is returned in the *in* registers and the registers restored by reverse window-blinding.

MC68000 Call and Return. Let us look at the details for two of the machines. We start with the 68000, because that is the most open and "conventional." We continue with the function NumberCount. Only a single parameter must be passed—the pointer to the text block. The HLL callee sees NumberCount(block) as an integer (i.e., what will be returned), but the assembly program must do a call and then use the returned integer as instructed. A typical assembly routine would be:

```
MOVE.L    A2,-(SP)       ; move pointer to block onto the stack
JSR       NumberCount    ; save return address on the stack and start
                         ; executing NumberCount
                         ; do something with value returned in D0
```

The first instruction puts the pointer to the block, which is in A2, on the stack. It first must make room, so the "–" in –(A7) first subtracts 4 from A7 (making room for the longword) and then moves the longword into the space pointed to by the now-modified A7. The one instruction does two things: the decrementing of SP and the storing of the longword in memory.

```
MOVE.L    A2,–(A7)     A7 ⇐ A7–4      ;A7 = SP
                       M(A7) ⇐ A2     ;M(x) = memory(address x)
```

The next instruction, *jump subroutine* (JSR), does three things. It decrements SP (i.e., A7) by 4, stores the return address on the top of the stack, and puts the address of NumberCount in the *program counter*. We have just introduced two items which need specific definition:

Return address (RA): This will always be the address of the instruction which the callee should return to. In the 68000 and the VAX (and all other CISCs), the RA points to the first instruction after the JSR. In the SPARC and almost any RISC, RA will point to the second instruction after JSR. That curious difference will be discussed later.

Program counter (PC): This register (which is a *general register* on the VAX but a special register on the other machines) points to the place (memory location) in the machine language instruction stream where the program is currently operating. As each instruction is fetched, the PC is automatically incremented. The action of the JSR is to save the last version of the PC—the one for the next fetch—and replace it with the starting address of the routine to be jumped to.

Summing up these transactions in algebraic form:

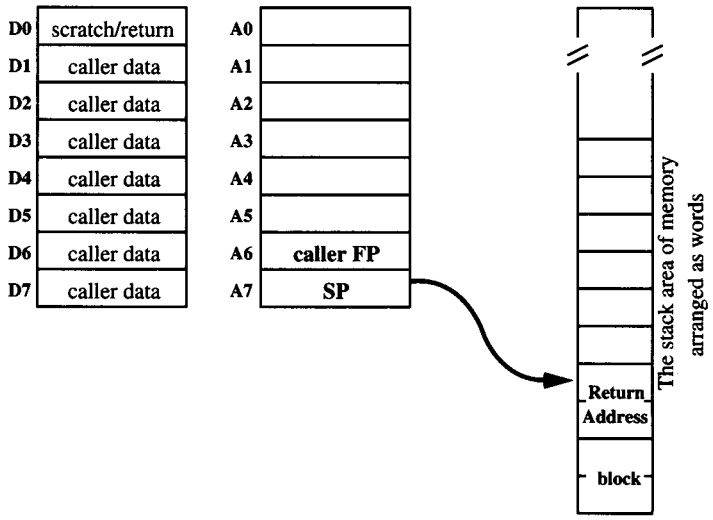


FIGURE 87.5 The stack area of the 68000's memory and the register assignments that the called subroutine sees as it is entered at the top. The registers all hold longwords, the size of an address. In typical PC/Macintosh compilers, integers are defined as 16-bit words. Accordingly, the stack area of memory is shown as words, or half the width of a register.

```
JSR NumberCount     SP ⇐ SP-4      ;A7 = SP
                    M(SP) ⇐ PC     ;M(x) = memory(address x)
                    PC ⇐ address of NumberCount
```

Should you wonder how the address of NumberCount gets in there, the *linker*, which assigns each section of code to its proper place in memory and therefore knows where all the labels are, will insert the proper address in place of the name.

This completes the call as far as building the call block, doing the call itself, and picking up the result. Had there been more parameters to pass, that first instruction would have been replicated enough times to *push* all of the parameters, one at a time, onto the stack. Now let us look at the conventions from the point of view of the callee. The callee has more work.

When the callee picks up the action, the stack and registers are as shown in Figure 87.5. With the exception of D0 and A7, the callee has no registers . . . yet. The callee must make room for local variables in either register or memory. If it wants to use registers, it must save the user's data from the registers. The subroutine can get whatever space it needs on the stack. Only after the setup will it get down to work. The entire section of stack used for local variables and saving registers is called the callee's frame. It is useful to have a pointer (FP) to the bottom of the frame to provide a static reference to the return address, the passed parameters, and the subroutine's local variable area on the stack. In the 68000, the convention is to use A6 as FP. When our routine, NumberCount, begins, the address in A6 points to the start of the caller's frame. The first thing the callee must do is to establish a local frame. It does that with the instruction LINK.

Typical of a CISC, each instruction does a large piece of the action. The whole entry operation for the 68000 is contained in two instructions:

```
LINK      A6,#$FFF8
MOVEM.L   D3-D7/A4,-(SP)
```

The first instruction does the frame making; the second does the saving of registers. There are multiple steps in each. Each double step of decrementing SP and moving a value onto the stack will be called a *push*. The steps are as follows:

```
LINK      A6,#$FFF8        ;push A6 (A7 ‹ A7-4, M(A7) ‹ A6)
                           ;move A7 to A6 (SP to FP)
                           ;add FFF8 (-8) to SP (4 words for local variables)
```

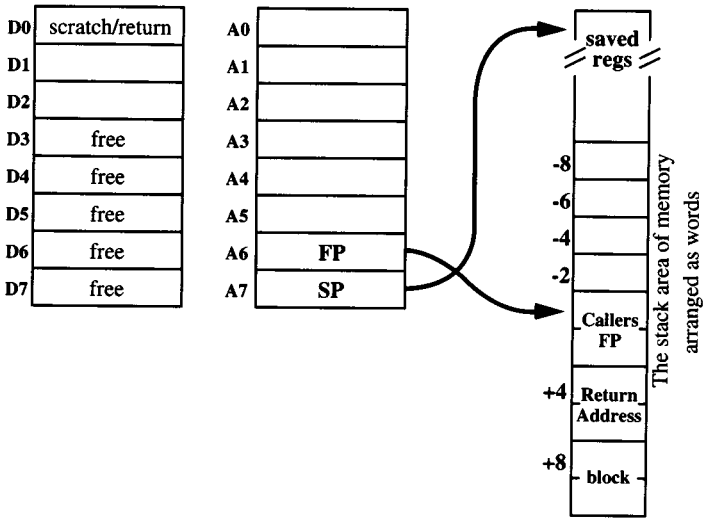


FIGURE 87.6 The stack area of the 68000's memory and the register situation just after MOVEM has been executed. The memory area between the two arrows is the subroutine's *frame*.

```
MOVEM.L   D3-D7/A4,-(A7)   ;push 5 data registers (3..7) and 1 address
                           ;register (A4)
```

At this point, the stack looks like Fig. 87.6.

The subroutine is prepared to proceed. How it uses those free registers and the working space set aside on the stack is the subject of the section on optimization in this chapter. For the moment, however, we simply assume that it will do its thing in exemplary fashion, get the count of the numbers, and return. We continue in this section by considering the rather simple transaction of getting back.

The callee is obliged to put the answer back where the caller expects to find it. Two paradigms for return are common. The one that our compiler uses is to put the answer in D0. The other common paradigm is to put the answer back on the stack. The user will have left enough room for that answer at FP+8, whether or not that space was also used for transferring parameters in. Using our paradigm, the return becomes:

```
MOVE.W $FFFC(A6),D0        ;answer from callee's stack frame [-4(FP)] to D0
MOVEM.L (A7)+,D3-D7/A4     ;registers restored to former values
UNLK A6                    ;SP ‹ FP, FP ‹ M(SP), SP ‹ SP+4
RTS                        ;PC ‹ M(SP), SP ‹ SP+4
```

When all of this has transpired, the machine is back to the caller with SP pointing at *block*. The registers look like Fig. 87.5 except for two important changes. SP is back where the caller left it and D0 contains the answer that the caller asked for.

Transactional Paradigms

The final topic in this section is the description of some of the translations of the simple and ordinary phrases of the HLLs into assembly language. We will show some in each of our three machines to show both the similarities and slightly different flavors that each machine architecture gives to the translation.

The paradigms that we will discuss comprise:

- Arithmetic
- Replacement
- Testing and branching, particularly multiple Boolean expressions
- Stepping through a structure

Many studies have shown that most computer arithmetic is concerned with addressing, testing, and indexing. In NumberCount there are several examples of each. For example, near the bottom of the program, there are statements such as:

```
count++;
```

For all three machines, the basic translation is the same: Add an *immediate* (a constant stored right in the instruction) to a number in register. However, for the CISCs, one may also ask that the number be brought in and even put back in memory. The three translations of this pair comprise:

MC68000	**VAX**	**SPARC**
ADDQ.W #$1,$FFFE(A6)	INCL R0	add %o2,1,%o2

Typical of the VAX, it makes a special case out of adding 1. There is no essential difference in asking it to add 1 by saying "1," but if one has a special instruction, it saves a byte of program length. With today's inexpensive memories, a byte is no longer a big deal, but when the VAX first emerged (1978), they were delivered with less memory than a PC or Mac would have today. The VAX, of course, can say ADDL #1, r0 just like the 68000, and for any number other than 1 or 0, it would. Note also that the VAX compiler chose to keep *count* in register, while in Think C® decided to put it on the stack (–2(SP)). A RISC has no choice. If you want arithmetic, your numbers must be in register. However, once again, we are really talking about the length of the code, not the speed of the transaction. *All* transactions take place from registers. The only issues are whether the programmer can see the registers and whether a single instruction can include both moving the operands and doing the operand arithmetic. The RISC separates the address arithmetic (e.g., –2(SP)) from the operand arithmetic, putting each in its own instruction. Both get the job done.

The next items we listed were *replacement* and *testing and branching*. We have both within the statement:

```
digit = (*source >= '0') && (*source <= '9');
```

The translation requires several statements:

MC68000	**VAX**	**SPARC**
MOVE.B (A4),D3	clrb r1	add %g0,0,%o1
CMPI.B #$30,D3	cmpb @4(ap),#48	ldsb [%o2],%o0
BLT ZERO	blss ZERO	subcc %o0, 47,%g0
CMPI.B #$39,D3	cmpb @4(ap),#57	ble ZERO
BLE ONE	bgtr ZERO	nop
	incb r1	subcc %o0,57,%g0
		bg ZERO
		nop
		add %g0, 1,%o1
		add %o1,0,%l3
ZERO:	**ZERO:**	**ZERO:**
MOVEQ #$00,D0		
BRA DONE		
ONE:		
MOVEQ #$01,D0		
DONE:		
MOVE.W D0,$FFF6(A6)		

To begin with, all three do roughly the same thing. The only noticeable difference in concept is that the SPARC compiler chose to compare the incoming character (*source) to 47 (the character before '0') and then branch if the result showed the letter to be "less than or equal," while the other two compared it to '0' as asked and then branched if the result was "less than." No big deal. But let us walk down the several columns to see the specific details. Prior to beginning, note that all three must bring in the character, run one or two tests, and then set an integer to either *zero* (false) or *not zero* (true). Also, let it be said that each snatch of code is

purportedly optimized, but at least with the small sample that we have, it looks as if each could be better. We begin with three parallel walkdowns. Notes as needed are provided below.

MC68000	**VAX**	**SPARC**
character from M → D#	Set (byte) DIGIT to 0	Set (byte) DIGIT to 0
Is (D3-'0') <=0?	Is (*source-'0') <=0?	character from Mfi out1
If <, branch to label ZERO	If <, branch to ZERO	Is (*source-'/') <=0?
Is (D3-'9')<=0?	Is (*source-'9') <=0?	If <=, branch to ZERO
If <=. branch to label ONE	If neither, branch to ZERO	Is (*source-'9') <=0?
	Add (byte) 1 to DIGIT	If neither, branch to
ZERO:	**ZERO:**	Add 1 to DIGIT
Put a longword 0 in D0		**ZERO:**
Branch to label DONE		
ONE:		
Put a longword 1 in D0		
DONE:		
Put value in D0 into DIGIT		

Notes:

1. Moving the character into register to compare it with '0' and '9':
   a. The first 68000 line moves the next character *as a byte* into register D3. The other 3 bytes will be ignored in the byte operations. Remember that the program had already moved the pointer to the string into A4.
   b. The SPARC does the same sort of thing with a pointer in %o2, except with the difference that it sign-extends the byte to a longword. Sign extension simply copies the *sign-bit* into the higher-order bits, effectively making 3E into 0000 003E or C2 into FFFF FFC2. That is what the mnemonic means: "LoaD Signed Byte."
   c. The VAX compiler takes a totally different approach—a rather poor one, actually. It leaves not only the byte in memory but even the pointer to the byte. Thus, every time it wants to refer to the byte—and it does so numerous times—it must first fetch the address from memory and then fetch the byte itself. This double memory reference is what @4(ap) means: "At the address you will find at address 4(ap)." The only thing that makes all this apparent coming and going even remotely acceptable is that the VAX will cache (place in a fast local memory) both the address and the character the first time that it gets them. Then, it can refer to them rapidly again. Cache references, however, are not as fast as register references.
2. Testing the character:
   The next line (3rd for the SPARC) does a comparison between 48 (or 47) and the character. *Compare* is an instruction which subtracts one operand from the other, but instead of putting the results somewhere, it stores only the facts on whether the operation delivered a negative number or zero or resulted in either an overflow or a carry. These are stored in **flags**, single bits associated with the arithmetic unit. The bits can contain only one result at a time. The 68000 and VAX must test immediately after the comparison or they will lose the bits. The SPARC changes the bits only when the instruction says so (the CC on the instruction — "change condition codes"). Thus, the subtraction can be remote from the test-and-branch.

   The SPARC is explicit about where to store the subtraction—in %g0. %g0 is a pseudo-register. It is always a 0 as a source and is a garbage can as a destination. The availability of the 0 and the garbage can serves all the same functions that the special instructions for zeros and comparisons do on the CISCs.
3. The differences in the algorithm to do the tests:
   There are two different paradigms expressed in these three examples. One says: "Figure out which thing you want to do and then do that one thing." The other says: "First set the result false and then figure out if you should set it true." While the second version would seem to do a lot of unnecessary settings to zero, the other algorithm will execute one less branch. That would make it roughly equivalent.

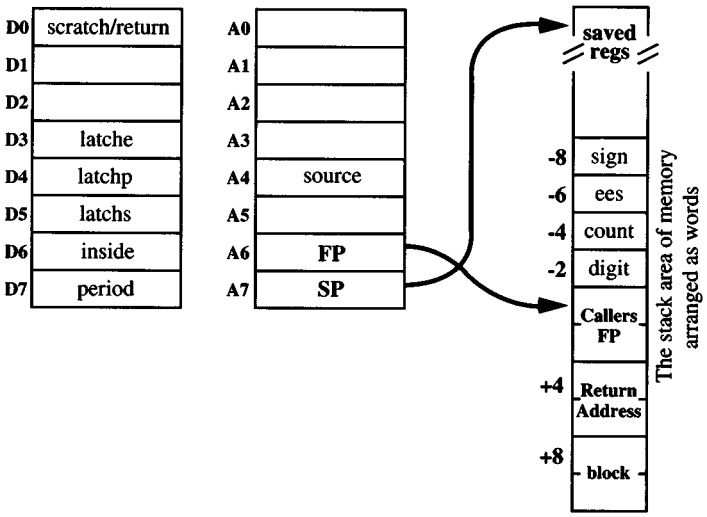


FIGURE 87.7 The stack area of the 68000's memory and the register assignments that the Think C® compiler made with global optimization *turned off*. The stack is shown just after the MOVEM instruction. The items in bold are as they would be after that instruction. While the registers all hold longwords, in typical PC/Macintosh compilers, integers are defined as words. This figure is the programmatic successor to Fig. 87.6.

> However, the 68000 algorithm is definitely longer—uses more memory for code. That is not really much of an issue, but why put the result first into a temporary register and then where you really want it?

Compiler Optimization and Assembly Language

Compiler Operations

To understand the optimizing features of compilers and their relation to assembly language, it is best to understand some of the chores for the compiler. This section examines variable allocation and how it can be optimized, and the optimization task of constant expression removal. Examples of how compilers perform these operations are taken from the various systems used in the article.

Variable Allocation. Variables in high-level languages are an abstraction of memory cells or locations. One of the compiler's tasks is to assign the abstract variables into physical locations—either registers within the processor or locations within memory. Assignment strategies vary, but an easy and often-used strategy is to place *all* variables in memory. Easy, indeed, but wasteful of execution time in that it requires memory fetches for all HLL variables. Another assignment strategy is to assign as many variables to the registers as possible and then assign any remaining variables to memory; this method is typically sufficient, except when there is a limited number of registers, such as in the 68000. In these cases, the best assignment strategy is to assign registers to the variables which have the greatest use and then assign any remaining variables to memory. In examining the compilers and architecture used in this article, we find examples of all these methods.

In the unoptimized mode, VAX and Sparc compilers are among the many which take the easy approach and assign variables only to memory locations. In Figs. 87.6 and 87.7, the variable assignments are presented for the unoptimized and optimized options. Note that only one or two registers are used, both as scratch pads, in the unoptimized option, whereas the optimization assigns registers to all variables. The expected execution time savings is approximately 42 of the 50 memory references per loop iteration. That does not include additional savings caused by compact code. Detailed comparisons are not presented since the interpretation of architectural comparisons is highly subjective.

Unlike the VAX and Sparc compilers, the 68000 compiler assigns variables to registers in both the unoptimized and unoptimized options; these assignments are depicted in Figs. 87.7 and 87.8. Since there are only eight general-purpose data registers in the 68000 and two are assigned as scratch pads, only six of the program's ten

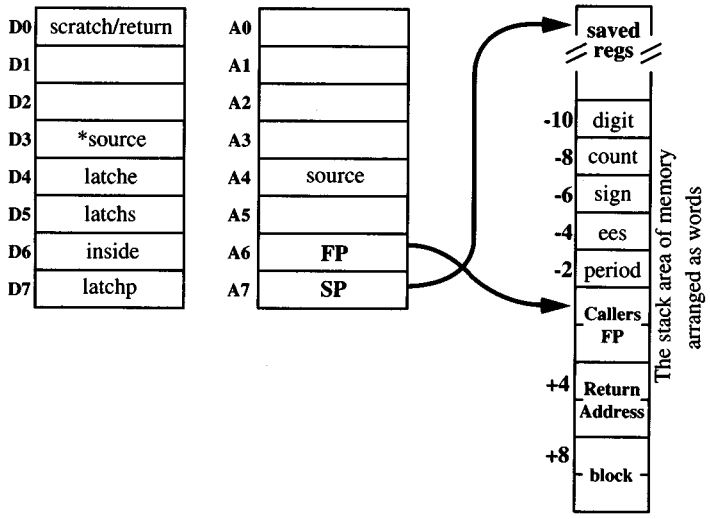


FIGURE 87.8 The stack area of the 68000's memory and the register assignments that the Think C® compiler made with global optimization *turned on*. The stack is shown just after the MOVEM instruction. The items in bold are as they would be after that instruction. This figure should be compared with Fig. 87.7.

variables can be assigned to registers. The question is how the 68000 compiler chose which variables to assign to registers and which to leave in memory. As might be expected, the compiler assigned registers based on their usage for the unoptimized option as well as the optimized. The exact counting strategy is unknown. However, a fair estimate, which yields the same assignment as the compiler, is to count only the variable's usage in the loop—the likely place for the program to spend most of its execution time. There are other ways to estimate the variable usage such as assigning weights to a reference based on its placement within a loop, its placement within a conditional (if-then-else) statement, etc. These estimates and how they are applied within a compiler can change the variable allocation as well as the efficiency of the code.

In the optimized case, a slightly different register assignment is used. This is because the optimizer created another character variable—*source—which it assigned to a register. The motivation for its creation and its assignment to a register is shown in the next section on constant expression removal.

Even though the assignment of variables to registers gives an improvement in performance, it is not always possible to assign a variable to a register. In C, one operation is to assign the address of a variable to a pointer-type variable (e.g., ip = & i). If i were assigned to a register, the operation would become invalid, because a register does not have a memory address. Although this example appears limited to the C language, the use of a variable's address is widespread when subroutine parameters are passed by reference. (Variables sent to a subroutine are either passed by reference, where the address of the variable is passed to the subroutine, allowing modifications to the original variable, or they are passed by value, where a copy of the variable is passed to the subroutine.) When a parameter is passed by reference its address must be obtained and passed to the subroutine, an action commonly found in most languages. This action compounds the task of selecting candidate variables for register assignment.

Constant Expression Removal. The task of allocating program variables to physical locations is accomplished by all compilers; we have shown that there are many ways to achieve this goal with varying ease or run-time performance. This section explores a compiler task which is done strictly for optimization—the removal of constant expressions. In exploring this task, we show strategies for the recognition of this optimization and also some caveats in their application.

Constant expressions are subexpressions within a statement which are unchanged during its execution. An obvious example is the expression *vector[x]* in the following conditional statement.

```
if (vector[x] < 100) && (vector[x] > 0) then ...
```

SPARC GCC no optimization

```
L2:   add 0, %g0, %o0          ; put 0 in o0 — i.e., assume a false result
      ld [%fp-92], %o1         ; get source from memory
      ldsb [%o1], %o1          ; get *source
      add 47, %g0, %o2         ; move 47 into o2
      subcc %o1, %o2, %g0      ; compare *source to '0'
      ble L5                   ; branch if less then, digit is false
      nop
      ld [%fp-92], %o1         ; get source again
      ldsb [%o1], %o1          ; get *source again
      add 57, %g0, %o2         ;
      subcc %o1, %o2, %g0      ; compare to '9'
      bg L5
      nop
      add 1, %g0, %o0          ; results is a true, change temporary
L5:                            ; jump target if result is false
      st %o0, [%fp-36]         ; move the result to variable.
```

SPARC GCC optimization

```
L2:   add 0, %g0, %o1          ; assume a false result from statement.
      ldsb [%o2], %o0          ; get *source save in register o0
      addcc -47, %o0, %g0      ; compare to '0'
      ble L5
      nop
      addcc -57, %o0, %g0      ; reuse *source and compare to '9'
      bg L5
      nop
      add 1, %g0, %o1          ; results is a true, so change temporary
L5:                            ; jump target if result is false
      add %o1, %g0, %l3        ; move the result to variable.
```

i7 return address	r31
i6 frame pointer	r30
i5 param #5	r29
i4 param #4	r28
i3 param #3	r27
i2 param #2	r26
i1 param #1	r25
i0 param #0	r24
l7 local	r23
l6 local	r22
l5 local	r21
l4 local	r20
l3 local	r19
l2 local	r18
l1 local	r17
l0 local	r16
o7 scratch	r15
o6 stack pointer	r14
o5 param out #5	r13
o4 param out #4	r12
o3 param out #3	r11
o2 param out #2	r10
o1 param out #1	r9
o1 param out #1	r8
g7 global	r7
g6 global	r6
g5 global	r5
g4 global	r4
g3 global	r3
g2 global	r2
g1 global	r1
g0 src=0,dst=WB	r0

FIGURE 87.9 SPARC register assignments.

An astute coder who does not trust compilers would not allow two memory references for vector[x] in the same conditional statement and would rewrite the code by assigning vector[x] to a temporary variable and using that variable in the conditional. An astute compiler would also recognize the constant expression and assign vector[x] to a scratch pad register and use this register for both comparisons. This type of optimization, where small sections of code (typically one source line) are optimized, is called *peep-hole* optimization.

Within the example program, the assignment statement which checks if the character is a digit within the range from '0' to '9' is a statement which can benefit from this type of optimization. The C code lines, with the unoptimized and optimized SPARC assembly code, are listed below. Note that in addition to the constant expression removal the optimized code also assigns variables to registers.

```
digit = (*source >= '0') && (*source <= '9') ;
```

In translating this line on the SPARC, Fig. 87.9 shows the 32 registers visible at any moment in the window-blinding SPARC. The top 24 shift by 16 in a normal call. The eight globals remain the same. The shift of the registers is accompanied by copying SP to o6 and the call instruction puts the return address into o7. Accordingly, a call wipes out the caller's o7. Register g0 serves as a 0 (as a source) and as a *wastebasket* (as a destination). *ld* loads a longword, and *ldsb* sign-extends a byte into a longword. The instruction after a branch is executed whether the branch is taken or not (delayed branching). An instruction such as add 47, %g0, %o2 adds a constant to 0 and puts it in the register. This is equivalent to move.l #47, d4 on the 68000. An *add* or *sub* with *cc* appended changes the condition codes. To do a *compare*, one uses *addcc* or *subcc* and puts the result in g0 (the wastebasket).

The same type of constant expression can be found and removed with a global perspective, typically from within loops. A simple example is the best way to describe how they can be removed and to offer some caveats when the compiler cannot see the global picture. The following example code updates each element in the vector by adding the first element scaled by a constant y. An observation shows that the subexpression, vector[0] * y, is constant throughout all executions of the loop. An obvious improvement in code is to calculate the product of vector[0] * y outside of the loop and store its value in a temporary variable. This is done in the second example.

Constant expression present	**Constant expression removed**
for(i= 0 ; i < size; i++) { vector[i] = vector[i] + (vector[0] * y); }	temp = vector[0] * y ; for(i= 0 ; i < size; i++) { vector[i] = vector[i] + temp ; }

Ideally, the compiler should find and remove these constant expressions from loops, but this is not as obvious as it may seem. Consider the above example if the following line were inserted in the loop:

```
y = vector[i-1] + vector[i+1]
```

If each source line is taken in isolation, y appears constant, but y is dependent on the loop index i. Hence before removing constant expressions, the compiler must map the dependencies of each variable on the other variables and the loop index. Additionally, other not-so-obvious dependencies—such as when two pointers modify the same structure—are difficult to map and can result in erroneous object code. This is one of the difficulties in optimizing compiler operation and why its extent is limited.

A subtle example for constant expression removal is found in our sample program in the reference to *source. In these statements, the character referenced (addressed) by source is obtained from memory. The pointer (address) *source* is changed only at the bottom of the loop and the memory contents addressed by *source* are static. A global optimization should obtain the character once at the top of each pass of the loop and save on subsequent memory references throughout. The 68000 C compiler with the optimization option determined *source to be constant throughout the loop and assigned register D3 to hold its contents. This saved seven of the eight memory accesses to *source in each loop pass. The unoptimized 68000 option, the SPARC, and the VAX compilers did not use global constant expression removal and must fetch the operand from memory before its use.

The Problems. With optimization yielding more efficient code resulting in improved system performance, why would you not use it? Our favorite, among the several reasons, is the following quote from compiler documentation: "Compiling with optimization may produce incorrect object code." Problems are caused by assumptions used by the compiler which are not held by the programmer. For example, an optimization which assumes that memory contents do not change with time is erroneous for multi-tasking systems which share memory structures and also for memory- mapped I/O devices, where memory contents are changed by external events. For these cases, the data in register may not match the newer data in memory.

Additionally, HLL debuggers do not always work well with the optimization option since the one-to-one correspondence between HLL code and the object code may have been removed by optimization. Consider the reassignment of *source to a data register which is performed by the 68000 C compiler. If a debugger were to modify the contents of *source, then it would have to know about the two locations where it is stored: the memory and the register. Other types of optimizations which may cause problems are when unneeded variables are removed or when code is resequenced. If a HLL debugger tries to single-step through HLL code, there may not be corresponding assembly code, and its execution will appear erroneous.

High-Level Language and Assembly Language Relations

In comparing the various assembly languages from the compiler-generated code, we have not presented a full vocabulary of assembly languages or the minutiae of the underlying machines. Exploring only the code generated by the compilers may lead one to believe that all assembly languages and processor architectures are pretty much the same. This is not really the case. What we have shown is that compilers typically use the same assembly language instructions regardless of the underlying machines. The compiler writer's motivation for this apparent similarity is not because all architectures are the same, but because it is difficult—arguably even nonproductive—for the compiler to take advantage of the complex features which some CPU architectures offer. An argument may be made that compilers generate the best code by developing code rather independently of the

underlying architecture. Only in the final stages of code generation is the underlying platform's hardware architecture specifically considered [see Aho et al., 1986]. Differences in the architectures and assembly languages are plentiful; compilers typically do not and probably should not take advantage of such features.

The VAX is one of the best examples of an architecture having an almost extraordinary vocabulary, which is why it is often considered the prototypical *CISC* machine. What were the motivations for having this rich vocabulary if compilers simply ignore it? Early computer programming was accomplished through slightly alphabetized machine language—mnemonics for opcodes and sometimes for labels. Assembly language represented a vast improvement in readability, but even though FORTRAN, COBOL and Algol were extant at the same time as assembly language, their crude or absent optimization abilities led to the popular belief that *really good* programming was always done in assembly. This was accepted lore until early studies of optimization began to have an impact on commercial compilers. It is fair to say that this impact did not occur until the early 1980s. The VAX and the 68000 were products of the middle and late 1970s. It is no great surprise then to find that CISC computer architectures were designed to enable and assist the assembly language programmer. Such an objective promotes the inclusion of many complex features which the programmer might want to utilize. However, two facts emerged in the late 1970s which suggested that this rich vocabulary was provided at too high a cost for its benefit (for a more complete discussion of the emergence of the RISC concept, which really began with Seymour Cray in the 1960s, see [Feldman and Retter, 1994]):

- It was widely observed that the generation, testing, and maintenance of large programs in assembly code was extremely expensive when compared to doing the same task in a HLL. Furthermore, the compilers were improving to the point where they competed quite well with typical assembly code.
- Although initially not widely accepted, it was observed that the rich vocabulary made it very difficult to set up an efficient processing pipeline for the instruction stream. In essence, the assembly line was forced to handle too many special cases and slowed down under the burden. When the analysis of compiled programs showed that only a limited span of instructions was being used, these prescient designers decided to include only the heavily used instructions and to restrict even these instructions so that they would flow in unblemished, uniform streams through the production line. Because this focus resulted in noticeably fewer instructions— though that was not the essential objective—the machines were called *RISC*, a sobriquet that came out of a VLSI design project at Berkeley [Patterson and Hennessy, 1989].

Even though RISC hardware designs have increased performance in essence by reducing the complexity and richness of the assembly language, back at the ranch the unrepentant assembly language programmer still desired complex features. Some of these features were included in the assembly languages not as native machine instructions but essentially as a *high-level* extension to assembly language. A universal extension is the inclusion of *macros*. In some sense, a macro looks like a subroutine, but instead of a call to a distant block of code, the macro results in an inline insertion of a predefined block of code. Formally, a macro is a name which identifies a particular sequence of assembly instructions. Then, wherever the name of the macro appears in the text, it is replaced by the lines of code in the macro definition. In some sense, macros make assembly language a little more like a HLL. It makes code more readable, makes code maintenance a little faster and more reliable (fix the macro definition and you fix all of the invocations of the macro), and it speeds up the programmer's work.

Another extension to some assembly languages is extended mnemonics. Here the coder places a mnemonic in place of specific assembly language instructions; during code assembly the mnemonic is automatically translated to an optimal and correct instruction or instruction sequence. These free the coder from the management of low-level details, leaving the task to a program where it is better suited. Examples of extended mnemonics include *get* and *put*, which generate memory transfers by selecting the addressing mode as well as the specific instructions based on the operand locations. An increasingly common feature of assembly languages is the inclusion of structured control statements which emulate high-level language control-flow constructs such as: `if .. then .. else`, `for` loops, `while .. do` loops, `repeat .. until` loops, `break`, and `next`. These features remove the tedium from the programmer's task, allow for a more readable code, and reduce the cost of code development. An amusing set of examples are found in the assemblers that we have used on the SPARC. Architecture not withstanding, the assembly programmers wanted VAX assembly code! In spite of the absence of such constructs in the SPARC architecture, you find expressions such as CMP (compare) and MOV. Since these are easily translated to single lines of real SPARC code, their only raison d'etre

is to keep the old assembly language programmers happy. Presumably, those who knew not the VAX are not writing SPARC assembly code.

Summary

After all this fuss over compilers and how they generate assembly code, the obvious question is "Why bother to write any assembly code at all?" Three reasons why "some assembly may be required" follow.

- A human writing directly in assembly language is probably better than an optimizing compiler in extracting the last measure of resources (e.g., performance, hardware usage) from the machine. That inner loop—the code where the processor spends most of its execution cycles—may need to be hand-optimized to achieve acceptable performance. Real-time systems, where the expedient delivery of the data is as critical as its correctness, are another area where the required optimization may be greater than that achievable by compilers. The disparity in performance between human optimizers and their machine competitors comes from two special capabilities of the human programmer. These are the ability to know what the program will be doing—forward vision based on the program's intent rather than its structure—and the ability to take advantage of special quirks or tricks that have no general applicability. If you really need to extract this last full measure of performance, assembly language is the route. The cost of doing such hand-optimization is much greater than the hours spent in doing it and getting it debugged. Special quirks and tricks expressible only in assembly language will not translate to another machine and may disappear even in an "upgrade" of the intended processor.
- There is overhead in using HLL conventions, some of which can be eliminated by directly coding in assembly language. A typical embedded processor does not need the full span of HLL conventions and support, such as parameter passing or memory and stack allocation. One can *get away with* such dangerous things as global variables which do not have to be passed at all. By eliminating these conventions, increased performance is obtained. It should be pointed out that code written without such standard conventions is likely to be *very* peculiar, bug-prone, and hard to maintain.
- HLLs provide only limited access to certain hardware features of the underlying machine. Assembly language may be required to access these features. Again, this makes the code unportable and hard to maintain, but small stubs of assembly code may be required to invoke hardware actions which have no representation in a HLL. For example, setting or clearing certain bits in a special register may not be expressible in C. While any address can be explicitly included in C code, how do you reference a register which has no address? An example of such usage is writing or reading into or out of the status register. Some machines map these transactions into special addresses so that C could be used to access them, but for the majority of machines which do not provide this route to the hardware, the only way to accomplish these actions is with assembly code. To this end, some C compilers provide an inline assembler. You can insert a few lines of assembly language right in the C code, get your datum into or out of the special register, and move right back to HLL. Those compilers which provide this nonstandard extension also provide a rational paradigm for using HLL variable names in the assembly statements. Where necessary, the name gets expanded to allow the variable to be fetched and then used.

These reasons are special; they are not valid for most applications. Using assembly language loses development speed, loses portability, and increases the maintenance costs. While this caveat is well taken and widely accepted, at least for the present, few would deny the existence of situations where assembly language programming provides the best or only solution.

Defining Terms

Address error: An exception (error interrupt) caused by a program's attempt to access unaligned words or longwords on a processor which does not accommodate such requests. The address error is detected within the CPU. This contrasts with problems which arise in accessing the memory itself, where a logic circuit external to the CPU itself must detect and signal the error to cause the CPU to process the exception. Such external problems are called *bus errors.*

Assembler: A computer program (*application*) for translating an assembly-code text file to an *object* file suitable for linking to become an executable image (*application*) in machine language. Some HLL compilers include an inline assembler, allowing the programmer to drop into and out of assembly language in the midst of a HLL program.

CISC: *Complex instruction set computer,* a name to mean "not a RISC," but generally one that offers a very rich vocabulary of computer operations at a cost of making the processor which must handle this variety of operations more complex, expensive, and often slower than a RISC designed for the same task. One of the benefits of a CISC is that the code tends to be very compact. When memory was an expensive commodity, this was a substantial benefit. Today, speed of execution rather than compactness of code is the dominant force.

Compiler: A computer program (*application*) for translating a HLL text file to an *object* file suitable for linking to become an executable image (*application*) in machine language. Some compilers do both compilation and linking, so their output is an application.

Condition codes: Many computers provide a mechanism for saving the characteristics of results of a particular calculation. Such characteristics as *sign, zero result, carry* or *borrow,* and *overflow* are typical of integer operations. The program may reference these flags to determine whether to branch or not.

Disassembler: A computer program which can take an executable image and convert it back into assembly code. Such a reconstruction will be true to the machine language but normally loses much of the convenience factors, such as *macros* and name equivalencies, that an original assembly language program may contain.

Executable image: A program in pure machine code and including all of the necessary header information that allows an operating system to load it and start running it. Since it can be run directly, it is *executable.* Since it represents the original HLL or assembly program it is an *image.*

Flags: See *Condition codes.*

High-level language (HLL): A computer programming language generally designed to be efficient and succinct in expressing human programming concepts and paradigms. To be contrasted with low-level programming languages such as *assembly language.*

Linker: A computer program which takes one or more object files, assembles them into blocks which are to fit in particular blocks in memory, and resolves all external (and possibly internal) references to other segments of a program and to libraries of precompiled subroutines. The output of the linker is a single file called an *executable image* which has all addresses and references resolved and which the operating system can load and run on request.

Macro: A single line of code-like text, defined by the programmer, which the assembler will then recognize and which will result in an inline insertion of a predefined block of code. In most cases, the assembler allows both hidden and visible local variables and local labels to be used within a macro. Macros also appear in some HLLs, such as C (the *define* paradigm).

Object code: A file comprising an intermediate description of a segment of a program. The object file contains binary data, machine language for program, tables of offsets with respect to the beginning of the segment for each label in the segment, and data that would be of use to debugger programs.

RISC: *Reduced instruction set computer,* a name coined by Patterson et al. at the University of California at Berkeley to describe a computer with an instruction set designed for maximum execution speed on a particular class of computer programs. Such designs are characterized by requiring separate instructions for load/store operations and arithmetic operations on data in registers. The earliest computers explicitly designed by these rules were designs by Seymour Cray at CDC in the 1960s. The earliest development of the RISC philosophy of design was given by John Cocke in the late 1970s at IBM. See *CISC* above for the contrasting approach.

Related Topic

87.2 High-Level Languages

References

A.V. Aho, R. Sethi, and J.D. Ullman, *Compiler Principles, Techniques and Tools,* Reading, Mass.: Addison-Wesley, 1986. A detailed text on the principles of compiler operations and tools to help you write a compiler. This text is good for those wishing to explore the intricacies of compiler operations.

D. Cohen, "On holy wars and a plea for peace," *Computer,* pp. 11–17, Sept. 1981. A delightful article on the comparisons and motivations of byte ordering in memory.

J. Feldman and C. Retter, *Computer Architecture: A Designer's Text Based on a Generic RISC,* New York: McGraw-Hill, 1994.

D. Patterson and J. Hennessy, *Computer Architecture, A Quantitative Approach,* San Mateo, Calif.: Morgan Kaufman, 1989. An excellent though rather sophisticated treatment of the subject. The appendices present a good summary of several seminal RISC designs.

87.2 High-Level Languages

Ted G. Lewis

High-level languages (**HLLs**), also known as higher-order languages (**HOLs**), have a rich history in the annals of computing. From their inception in the 1950s until advances in the 1970s, HLLs were thought of as simple mechanical levers for producing machine-level instructions (see Table 87.3). Removing the details of the underlying machine, and automatically converting from a HLL statement to an equivalent machine-level statement, releases the programmer from the drudgery of the computer, allowing one to concentrate on the solution to the problem at hand.

TABLE 87.3 Each Statement of a HLL Translates into More than One Statement in a Machine-Level Language Such as Assembler

Language	Typical Number of Machine-Level Statements
FORTRAN	4–8
COBOL	3–6
Pascal	5–8
APL	12–15
C	3–5

Over the years, HLLs evolved into a field of study of their own, finding useful applications in all areas of computing. Some HLLs are designed strictly for solving numerical problems, and some for symbolic problems. Other HLLs are designed to control the operation of the computer itself, and yet even more novel languages have been devised to describe the construction of computer hardware. The number of human-crafted languages has multiplied into the hundreds, leading to highly special-purpose HLLs.

This evolution is best characterized as a shift away from the mechanical lever view of a HLL toward HLLs as notations for encoding **abstractions**. An abstraction is a model of the real world whose purpose is to de-emphasize mundane details and highlight the important parts of a problem, system, or idea. Modern HLLs are best suited to expressing such abstractions with little concern for the underlying computer hardware.

Abstraction releases the HLL designer from the bounds of a physical machine. A HLL can adopt a metaphor or arbitrary model of the world. Such unfettered languages provide a new interface between human and computer, allowing the human to use the machine in novel and powerful ways. Abstractions rooted in logic, symbolic manipulation, database processing, or operating systems, instead of the instruction set of a central processing unit (CPU), open the engineering world to new horizons. Thus, the power of computers depends on the expressiveness of HLLs.

To illustrate the paradigm shifts brought on by HLLs over the past 30 years, consider PROLOG, LISP, SQL, C++, and various operating system command languages. PROLOG is based on first-order logic. Instead of computing a numerical answer, PROLOG programs derive a conclusion. LISP is based on symbolic processing instead of numerical processing and is often used to symbolically solve problems in calculus, robotics, and artificial reasoning. SQL is a database language for manipulating large quantities of data without regard for whether it is numeric or symbolic. C++ is based on the **object-oriented paradigm**, a model of the world that is particularly powerful for engineering, scientific, and business problem solving.

None of these modern languages bear much resemblance to the machines they run on. The idea of a mechanical lever has been pushed aside by the more powerful idea of language as world builder. The kinds of worlds that can be constructed, manipulated, and studied are limited only by the HLL designer's formulation of the world as a paradigm.

In this section we answer some fundamental questions about HLLs: What are they? What do we mean by "high level"? What constitutes a paradigm? What are the advantages and disadvantages of HLLs? Who uses HLLs? What problems can be solved with these languages?

What Is a HLL?

At a rudimentary level, all languages, high and low, must obey a finite set of rules that specify both their syntax and semantics. **Syntax** specifies legal combinations of symbols that make up statements in the language. **Semantics** specifies the meanings attached to a syntactically correct statement in the language. To illustrate the difference between these two fundamental traits of all languages consider the statement, "The corners of the round table were sharp." This is syntactically correct according to the rules of English grammar, but what does it mean? Round tables do not have sharp corners, so this is a meaningless statement. We say it is semantically incorrect.

Statements of a language can be both syntactically and semantically correct and still be unsuitable for computer languages. For example, the phrase ". . . time flies . . ." has two meanings: one as an expression of clock speed, and the other as a reference to a species of insects. Therefore, we add one other requirement for computer languages: there must be only one meaning attached to each syntactically correct statement of the language. That is, the language must be *unambiguous*.

This definition of a computer language does not separate a HLL from all other computer languages. To understand the features of HLLs that make them different from other computer languages, we must understand the concepts of mechanical translation and abstraction. Furthermore, to understand the differences among HLLs, we must know how abstractions are used to change the computing paradigm. But first, what is a HLL in terms of translation and abstraction?

Defining the syntax of a HLL is easy. We simply write rules that define all legal combinations of the symbols used by the language. Thus, in FORTRAN, we know that arithmetic statements obey the rules of algebra, with some concessions to accommodate keyboards. A **metalanguage** is sometimes used as a kind of shorthand for defining the syntax of other languages, thus reducing the number of cases to be listed.

Defining the semantics of a language is more difficult because there is no universally accepted metalanguage for expressing semantics. Instead, semantics is usually defined by another program that translates from the HLL into some machine-level language. In a way, the semantics of a certain HLL is defined by writing a program that unambiguously maps each statement of the HLL into an equivalent sequence of machine-level statements. For example, the FORTRAN statement below is converted into an equivalent machine-level sequence of statements as shown to the right:

```
X = (B**2 – 4*A*C)      PUSH B
                        PUSH #2
                        POWER       //B**2
                        PUSH #4
                        PUSH A
                        PUSH C
                        MULT        //A*C
                        MULT        //4*(A*C)
                        SUB         //(B**2)–(4*(A*C))
                        POP   X     //X=
```

In this example, we assume the presence of a **pushdown stack** (see Fig. 87.10). The PUSH and POP operations are machine-level instructions for loading/storing the top element of the stack. The POWER, MULT, and SUB instructions take their arguments from the top of the stack and return the results of exponentiation, multiplication, and subtraction to the top of the stack. The symbolic expression of calculation in fortran becomes a sequence of low-level machine instructions which often bear little resemblance to the HLL program.

The foregoing example illustrates the mechanical advantage provided by FORTRAN because one FORTRAN statement is implemented by many

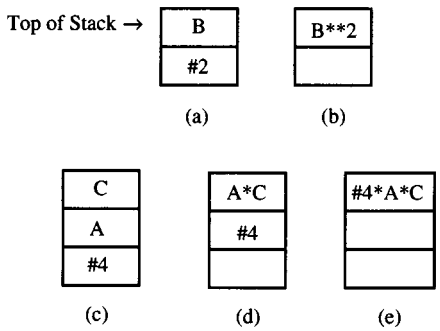


FIGURE 87.10 (a) The stack after PUSH #4 and PUSH B; (b) the stack after POWER; (c) the stack after PUSH #4, PUSH A, and PUSH C; (d) the stack after MULT; and (e) the stack after MULT a second time.

machine-level statements. Furthermore, it is much easier for a human programmer to read and write FORTRAN than to read and write machine-level instructions. One major advantage of a HLL is the obvious improvement in program creation and, later on, its maintenance. As the size of the program increases, this advantage becomes larger as we consider the total cost to design, code, test, and enhance an application program.

The FORTRAN program containing the example statement is treated like input data by the translating program which produces a machine-level sequence as output. In general, the input data is called the **source program**, and the resulting translated output is called the **object program**. There are two ways to obtain an object program from a source program: compiling and interpreting.

In most cases, FORTRAN is translated by a compiler program. The idea behind a **compiler** is that the translator converts the source program in its entirety before any part of the resulting object program is actually run on the computer. That is, compiling is a two-step process. In some HLLs, however, it is impossible to entirely convert a source program into an object program until the program executes.

Suppose the storage for A, B, and C in the previous example is not known at the time the program is compiled. We might want to allocate storage on-the-fly while the program is running, because we do not know in advance that the storage is needed. This is an example of **delayed binding** of a variable to its storage location in memory.

Powerful languages such as Pascal and C permit a limited amount of delayed binding, as illustrated in the following example written in Pascal. This example also illustrates a limited amount of abstraction introduced by the HLL.

```
type    rnumber = real;                 {template}
        rptr = ^rnumber;                {pointer}
var     Aptr, Bptr, Cptr : rptr;        {instance}
...
{later in the program...}
new(Aptr); read( Aptr^);                {binding}
new(Bptr); read( Bptr^);
new(Cptr); read( Cptr^);
X := (Bptr^) * (Bptr^) – 4 * (Aptr^) * (Cptr^);
```

The **type** statement is an abstraction that defines a template and access mechanism for the variables A, B, and C that are to be created on-the-fly. The **var** statement is similar to the DIMENSION statement in that it tells the translator to allocate space for three pointers: Aptr, Bptr, and Cptr. Each of these allocations will point to the actual values of A, B, and C according to the previous **type** statement.

The actual allocation of storage is not known until the program executes the sequence of new() functions in the body of the program. Each new() function allocates space according to the **type** statement and returns a pointer to that space. To access the actual values in these storage spaces, the up arrow, ^, is written following the variable name. Thus, the read() function gets a number from the keyboard and puts it in the space pointed to by the pointer variable. Similarly, the value of X is computed by indirect reference to each value stored at the newly allocated memory location.

The purpose of this example is to illustrate the use of delayed binding in a HLL. Languages such as LISP and C++ require even greater degrees of delayed binding because of the abstractions they support. When the amount of delayed binding becomes so great that very little of the program can be compiled, we say that the HLL is an *interpreted language*, and the translator becomes an **interpreter** rather than a compiler. This crossover is often obscure, so some HLLs are translated by both a compiler and an interpreter. BASIC is a classic example of a HLL that is both interpreted and compiled.

The purpose of delayed binding is to increase the level of a HLL by introducing abstraction. Abstraction is the major differentiating feature between HLLs and other computer languages. Without abstraction and delayed binding, most HLLs would be no more powerful than a macro **assembler** language. However, with abstraction, HLLs permit a programmer to express ideas that transcend the boundaries of the physical machine.

We can now define HLL based on the concept of abstraction. *A HLL is a set of symbols which obey unambiguous syntactic and semantic rules: the syntactic rules specify legal combinations of symbols, and the semantic rules specify legal meanings of syntactically correct statements relative to a collection of abstractions.*

The notion of abstraction is very important to understanding what a HLL is. The example above illustrates a simple abstraction, e.g., that of data structure abstraction, but other HLLs employ much more powerful abstraction mechanisms. In fact, the *level of abstraction* of a HLL defines how high a HLL is. But, how do we measure the level of a HLL? What constitutes a HLL's height?

How High Is a HLL?

There have been many attempts to quantify the level of a programming language. The major obstacle has been to find a suitable measure of level. This is further complicated by the fact that nearly all computer languages contain some use of abstraction, and therefore nearly all languages have a "level." Perhaps the most interesting approach comes from information theory.

Suppose a certain HLL program uses P operators and Q operands to express a solution to some problem. For example, a four-function pocket calculator uses $P = 4$ operators for addition, subtraction, multiplication, and division. The same calculator might permit $Q = 2$ operands by saving one number in a temporary memory and the other in the display register. In a HLL the number of operators and operands might number in the hundreds or thousands.

We can think of the set of P operators as a grab bag of symbols that a working programmer selects one at a time and places in a program. Suppose each symbol is selected with probability $1/P$, so the information content of the entire set is

$$-\sum_{1}^{P} \frac{1}{P} \log\left(\frac{1}{P}\right) = \log(P)$$

Assuming the set is not depleted, the programmer repeats this process P times, until all of the operators have been selected and placed in the program. The information content contributed by the operators is $P \log(P)$, and if we repeat the process for selecting and placing all Q operands, we get $Q \log(Q)$ steps again. The sum of these two processes yields $P \log(P) + Q \log(Q)$ symbols. This is known as Halstead's **metric** for *program length* [Halstead, 1977].

Similar arguments can be made to derive the volume of a program, V, level of program abstraction, L, and level of the HLL, l, as follows.

P = Number of distinct operators appearing in the program
p = Total number of operators appearing in the program
Q = Number of distinct operands appearing in the program
q = Total number of operands appearing in the program
N = Number of operators and operands appearing in the program
V = Volume = $N \log_2(P + Q)$
L = Level of abstraction used to write the program $\approx (2/P)^*(Q/q)$
l = Level of the HLL used to write the program = L^2V
E = Mental effort to create the program = V/L

Halstead's theory has been applied to English (*Moby Dick*) and a number of programs written in both HLL and machine-level languages. A few results based on the values reported in Halstead [1977] are given in Table 87.4. This theory quantifies the level of a programming language: *PL/I* is higher level than Algol-58, but lower level than English.

In terms of the mental effort required to write the same program in different languages, Table 87.4 suggests that a HLL is about twice as high level as assembler language. That is, the level of abstraction of *PL/I* is more than double that of assembler. This abstraction is used to reduce mental effort and solve the problem faster.

TABLE 87.4 Comparison of Languages in Terms of Level, l, and Programming Effort, E

Language	Level, l	Effort, E
English	2.16	1.00
PL/I	1.53	2.00
Algol-58	1.21	3.19
FORTRAN	1.14	3.59
Assembler	0.88	6.02

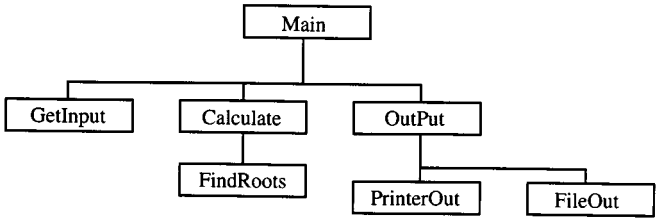


FIGURE 87.11 Hierarchical decomposition of procedural program.

HLLs and Paradigms

A programming **paradigm** is a way of viewing the world, e.g., an idealized model. HLLs depend on paradigms to guide their design and use. In fact, one might call HLL designers *paradigm engineers* because a good HLL starts with a strong model. Without such a model, the abstraction of a HLL is meaningless. In this section we examine the variety of paradigms embodied in a number of HLLs.

The **procedural paradigm** was the earliest programming paradigm. It is the basis of COBOL, FORTRAN, Pascal, C, BASIC, and most early languages. In this paradigm the world is modeled by an algorithm. Thus, an electrical circuit's behavior is modeled as a system of equations. The equations are solved for voltage, current, and so forth by writing an algorithmic procedure to numerically compute these quantities.

In the procedural paradigm a large system is composed of modules which encapsulate procedures which in turn implement algorithms. Hierarchical decomposition of a large problem into a collection of subordinate problems results in a hierarchical program structure. Hence, a large FORTRAN or C program is typically composed of a collection of procedures (subroutines in FORTRAN and functions in C) organized in layers, forming a tree structure, much like the organization chart of a large corporation (see Fig. 87.11).

Hierarchy is used in the procedural paradigm to encapsulate low-level algorithms, thus abstracting them away. That is, algorithm abstraction is the major contributor to leveling in a procedural HLL. Figure 87.11 illustrates this layering as a tree where each box is a procedure and subordinate boxes represent procedures called by parent boxes, the top-most box is the most abstract, and the lowest boxes in the tree are the most concrete.

Intellectual leverage is limited to control flow encapsulation in most procedural languages. Only the execution paths through the program are hidden in lower levels. While this is an improvement over machine-level languages, it does not permit much flexibility. For example, algorithmic abstraction is not powerful enough to easily express non-numerical ideas. Thus, a C program is not able to easily model an electronic circuit as a diagram or object that can be reasoned about, symbolically.

One of the reasons procedural HLLs fail to fully hide all details of an abstraction is that they typically have weak models of data. Data is allowed to flow across many boundaries, which leads to problems with encapsulation. In FORTRAN, BASIC, Pascal, and C, for example, access to any data is given freely through globals, parameter passing, and files. This is called **coupling** and can have disastrous implications if not carefully controlled.

One way to reduce coupling in a procedural language is to eliminate side-effects caused by unruly access to data. Indeed, if procedures were prohibited from directly passing and accessing data altogether, many of the problems of procedural languages would go away. An alternative to the procedural paradigm is the **functional paradigm**. In this model of the world, everything is a function that returns a value. Data is totally abstracted away so that algorithms are totally encapsulated as a hierarchical collection of functions. LISP is the most popular example of a functional HLL [Winston and Horn, 1989].

A LISP statement that limits data access usually consists of a series of function calls; each function returns a single value which is used as an argument by another function and so on until the calculation is finished. For example, the FORTRAN statement X = (B**2 – 4*A*C) given earlier is written in functional form as follows:

```
ASSIGN( X, MINUS(SQUARE(B), TIMES( 4, TIMES(A,C))))
```

This statement means to multiply A times C, then multiply the result returned by TIMES by 4, then subtract this from the result returned by SQUARE, and so forth. The final result is assigned to X.

One of the most difficult concepts to adjust to when using the procedural paradigm is the idea that all things are functions. The most significant implication of this kind of thinking is the replacement of loops with **recursion** and branches with guards. Recall that everything is a function that must return a value—even control structures. To illustrate, consider the functional (non-LISP) equivalent of the summation loop in FORTRAN, below.

```
S=0
DO 20 I=1,10          SUM( XList, N):
S=S+X(I)              N>0 |
20 CONTINUE                   N is N–1,
                              SUM is Head( XList ) + SUM (TAIL(Xlist), N)
```

The functional form will seem strange to a procedural programmer because it is higher level, e.g., more abstract. It hides more details and uses functional operators HEAD (for returning the first element of XList), TAIL (for returning the N–1 tail elements of XList), and **is** for binding a value to a name. Also, notice the disappearance of the loop. Recursion on SUM is used to run through the entire list, one element at a time. Finally, the guard N>0 prevents further recursion when N reaches zero.

In the functional program, N is decremented each time SUM is recursively called. Suppose N = 10, initially; then SUM is called 10 times. When N > 0 is false, the SUM routine does nothing, thus terminating the recursion. Interestingly, the additions are not performed until the final attempt to recurse fails. That is, when N = 0, the following sums are collected as the nested calls unwind:

```
SUM     : XList(10)
        : SUM+Xlist(9)
 …      …
        : SUM+XList(1)
```

Functional HLLs are higher level than procedural languages because they reduce the number of symbols needed to encode a solution as a program. The problem with functional programs, however, is their high execution overhead caused by the delayed binding of their interpreters. This makes LISP and PROLOG, for example, excellent **prototyping** languages but expensive production languages. LISP has been confined to predominantly research use; few commercial products based on LISP have been successfully delivered without first rewriting them in a lower-level language such as C. Other functional languages such as PROLOG and STRAND88 have had only limited success as commercial languages.

Another alternative is the **declarative paradigm.** Declarative languages such as Prolog and STRAND88 are both functional and declarative. In the declarative paradigm, solutions are obtained as a byproduct of meeting limitations imposed by constraints. Think of the solution to a problem as the only (first) solution that satisfies all constraints declared by the program.

An example of the declarative paradigm is given by the simplified PROLOG program below for declaring an electrical circuit as a list of constraints. All of the constraints must be true for the circuit() constraint to be true. Thus, this program eliminates all possible R, L, C, V circuits from consideration, except the one displayed in Fig. 87.12. The declarations literally assert that Circuit(R, L, C, V) is a thing with "R connected to L, L connected to C, L connected to R, C connected to V, and V connected to R." This eliminates "V connected to L," for example, and leaves only the solution shown in Fig. 87.12.

```
Circuit(R, L, C, V) :
        Connected(R, L)
        Connected(L, C)
        Connected(L, R)
        Connected(C, V)
        Connected(V, R)
```

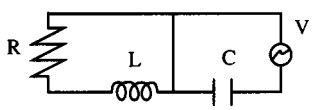


FIGURE 87.12 Solution to declaration for Circuit(R, L, C, V).

One interesting feature of declarative languages is their ability to represent infinite calculations. A declaration might constrain a solution to be in an infinite series of numbers, but the series may not need to be fully computed to arrive at an answer.

Another feature of such languages is their ability to compute an answer when in fact there may be many answers that meet all of the constraints. In many engineering problems, the first answer is as good as any other answer.

The declarative paradigm is a very useful abstraction for unbounded problems. Adding abstraction to the functional paradigm elevates declarative languages even higher. Solutions in these languages are arrived at in the most abstract manner, leading to comparatively short, powerful programs.

Perhaps the most common use of declarative languages is for construction of expert systems [Smith, 1988]. These kinds of applications are typically diagnostic. That is, they derive a conclusion based on assertions of fact. An electrical circuit board might be diagnosed with an expert system that takes symptoms of the ailing board as its input and derives a conclusion based on rules of electronic circuits—human rules of thumb given it by an experienced technician—and declarative reasoning. In this example, the rules are constraints expressed as declarations. The expert system program may derive more than one solution to the problem because many solutions may fit the constraints.

Declarative languages have the same inefficiencies as functional languages. For this reason, expert system applications are usually developed in a specialized declarative system called an *expert system shell*. A shell extracts the declarative or constraint-based capability from functional languages such LISP and PROLOG to improve efficiency. Often it is possible to simplify the shell so that early binding is achieved, thus leading to compiling translators rather than interpreters. Very large and efficient expert systems have been developed for commercial use using this approach.

Yet another paradigm used as the basis of modern languages such as C++ and Object Pascal is the *object-oriented programming (OOP) paradigm* [Budd, 1991]. OOP merges data and procedural abstractions into a single concept. In OOP, an object has both storage capacity and algorithmic functionality. These two abstractions are encapsulated in a construct called a **class**. One or more objects can be created by cloning the class. Thus, an **object** is defined as an instance of a class [Lewis, 1991].

OOP actually represents a culmination of ideas of procedural programming that have evolved over the past three to four decades. It is a gross oversimplification to say that OOP is procedural programming, because it is not, but consider the following evolution.

Procedure = Algorithm + Data Structures
Abstract Data Structure = Implementation Part + Interface Part
Class = Abstract Data Structure + Functions
Object = Class + Inheritance

The first "equation" states that a procedure treats algorithms and data separately, but the programmer must understand both the data structure and the algorithms for manipulating the data structures of an application. This separation between algorithms and data is a key feature of the procedural paradigm. During the 1970s structured programming was introduced to control the complexity of the procedural paradigm. While only partially successful, structured programming limited procedures to less powerful control structures by eliminating the GOTO and programs with labels. However, structured programming did not go far enough.

The next improvement in procedural programming came in the form of increased abstraction, called **ADT** (abstract data structures). An ADT separates the interface part of a procedure from its implementation part. Modula II and Ada™ were designed to support ADTs in the form of modules and packages. The interface part is an abstraction that hides the details of the algorithm. Programming in this form of the procedural paradigm reduces complexity by elevating a programmer's thoughts to a higher level of abstraction, but it still does not remove the problem of how procedures are related to one another.

Classes group data together into clusters that contain all of the functions that are allowed to access and manipulate the data. The class concept is a powerful structuring concept because it isolates the data portion of a program, thus reducing coupling and change propagation.

The class construct invented by the designers of Simula67 enforced the separation of interface and implementation parts of a module, and in addition introduced a new concept. Inheritance is the ability to do what another module can do. Thus, inheritance relates modules by passing on the algorithmic portion of a module

TABLE 87.5 Procedural versus Object-Oriented Thinking

Procedural	Object-Oriented
Instructions and data are separated.	Objects consist of both data and instructions.
Software design is linear, e.g., it progresses from design through coding and testing. This means change is difficult to accommodate.	Software design is interactive with coding and testing. This means change is easier to accommodate.
Programs are top-down decompositions of procedures, e.g., trees.	Programs are networks of objects that send messages to one another without concern for tree structure.
Program components are the real world, thus making programming more of a magic art.	Program components have abstractions of corres-pondence with the real world, thus making programming more of a discipline.
New programs are mostly custom built with little reuse from earlier programs. This leads to high construction costs and errors.	New programs are mostly specializations of earlier programs through reuse of their components. This leads to low construction costs and higher quality.

to other modules. Inheritance in a programming language like SmallTalk, Object Pascal, and C++ means even greater abstraction because code can be reused without being understood.

An object is an instance of a class. New objects inherit all of the functions defined on all of the classes used to define the parent of the object. This simple idea, copied from genetics, has a profound impact on both design and programming. It changes the way software is designed and constructed, i.e., it is a new paradigm.

Object-oriented thinking greatly differs from procedural thinking (see Table 87.5). In OOP a problem is decomposed into objects which mimic the real world. For example, the objects in Fig. 87.12 are resistor, inductor, capacitor, and voltage source. These objects have real-world features (state) such as resistance, inductance, capacitance, and voltage. They also have behaviors defined by sinusoidal curves or phase shifts. In short, the real-world objects have both state and function. The state is represented in a program by storage and the function is represented by an algorithm. A resistor is a program module containing a variable to store the resistance and a function to model the behavior of the resistor when it is subjected to an input signal.

The objects in an object-oriented world use inheritance to relate to one another. That is, objects of the same class all inherit the same functions. These functions are called **methods** in SmallTalk and **member functions** in C++[Ellis and Stroustrup, 1990]. However, the state or storage attributes of objects cloned from the same class differ. The storage components of an object are called **instance variables** in SmallTalk and **member fields** in C++.

The wholism of combining data with instructions is known as *ADTs*; the concept of sending messages instead of calling procedures is the *message-passing paradigm*; the concept of interactive, nonlinear, and iterative development of a program is a consequence of an object's **interface specification** being separated from its **implementation part**; the notion of modeling the real world as a network of interacting objects is called *OOD* (object-oriented design); the concept of specialization and **reuse** is known as *inheritance;* and OOP is the act of writing a program in an object-oriented language while adhering to an object-oriented view of the world.

Perhaps an analogy will add a touch of concreteness to these vague concepts. Suppose automobiles were constructed using both technologies. Table 87.6 repeats the comparison of Table 87.5 using an automobile design and manufacturing analogy.

We illustrate these ideas with a simple C++ example. The following code declares a class and two subclasses which inherit some properties of the class. The code also shows how interface and implementation parts are separated and how to override unwanted methods. Figure 87.13 depicts the inheritance and class hierarchy for this example.

```
class Node{
   public:                          //The interface part...
        Node() {}                   //Constructor function
        virtual ~Node() {}          //Destructor function
        virtual int eval() { error(); return 0;}    //Override this function
}
```

TABLE 87.6 Analogy with an Automobile Manufacturer

Procedural	Object-Oriented
Vendors and Assemblers work from their own plans. There is little coordination between the two.	Vendors and Assemblers follow the same blueprints; thus the resulting parts are guaranteed to fit together.
Manufacturing and design are sequential processes. A change in the design causes everyone to wait while the change propagates from designers to workers on the production line.	Design interacts with production. Prototypes are made and discarded. Production workers are asked to give suggestions for improvement or on how to reduce costs.
Changes on the manufacturing floor are not easily reflected as improvements to the design on the drafting board.	Changes in implementation rarely affect the design as interfaces are separated from implementation. Thus, the materials may change, but not the need for the parts themselves.
New cars are designed and constructed from the ground up, much like the first car ever produced.	New cars are evolutionary improvement to existing base technology, plus specializations that improve over last year's model.

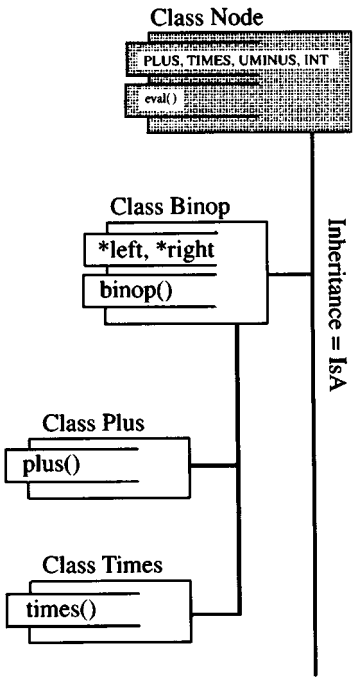


FIGURE 87.13 Partial class hierarchy for a C++ program that simulates a pocket calculator. The shaded Node class is an abstract class that all other classes use as a template.

The Node class consists of public functions that are to be overridden by descendants of the class. We know this because the functions are virtual, which in C++ means we expect to replace them later. Therefore we call this an **abstract class**. Also, Node() is the name of both the constructor and destructor member functions. A constructor is executed when a new object is created from the class, and the destructor is executed when the object is destroyed. These two functions take care of initialization and garbage collection which must be performed before and after dynamic binding of objects to memory space. Figure 87.13 shows this Node as an abstract class from which all other subclasses of this application are derived.

Now, we create a subclass that inherits the properties (interface) of Node() and adds a new property, e.g., Binop. Binop is an abstraction of the binary operators of a pocket calculator, which is the real-world object being simulated by this example program. The expression to be calculated is stored in a binary tree, and Binop sprouts a new left and right subtree when it is created and deletes this subtree when it is disposed.

```
class Binop : public Node {          //Derive Binop from Node
  public:
      Node *left, *right;            //Pointers to left and right subtrees
      ~Binop() { delete left; delete right;} //Collect garbage
      Binop( Node *lptr, Node *rptr) {left = lptr; right=rptr;}
}
```

Next, we define further specializations of Binop: one for addition of two numbers, Plus(), and the other for multiplication, Times(). The reader can see how to extend this to other operators.

```
class Plus: public Binop {
  public:                            //Add member functions to Binop
      Plus( Node *lptr, Node *rptr) : Binop( lptr, rptr) {} //Use Binop
      int eval() { return left->eval()+right->eval();}        //Do Addition
};
class Times: public Binop {
  public:                            //Add member functions to Binop
      Times( Node *lptr, Node *rptr) : Binop( lptr, rptr) {}  //Use Binop
      int eval() { return left->eval()*right->eval();}            //Do Multiply
};
```

In each case, the special-purpose operations defined in Plus() and Times() reuse Binop's code to perform the pointer operations. Then they add a member function eval() to carry out the operation. This illustrates reuse and the value of inheritance.

At this point, we have a collection of abstractions in the form of C++ classes. An object, however, is an instance of a class. Where are the objects in this example? We must dynamically create the required objects using the new function of C++.

```
Node *ptr = new Plus ( lptr, rptr );      //Create an object and point to it
int result = ptr->eval();                 //Add
delete ptr;
```

The foregoing code instantiates an object that ptr points to, sends a message to the object telling it to perform the eval() function, and then disposes of the object. This example assumes that lptr and rptr have already been defined elsewhere.

Clearly, the level of abstraction is greatly raised by OOP. Once a class hierarchy is established, the actual data processing is hidden or abstracted away. This is the power of the OOP paradigm.

The pure object-oriented languages such as SmallTalk80 and CLOS have achieved only modest success due to their unique syntax and heavy demands on computing power. Hybrid HLLs such as Object Pascal and C++ have become widely accepted because they retain the familiar syntax of procedural languages, and they place fewer demands on hardware.

Although OOP is an old technology (circa 1970), it began to gain widespread acceptance in the 1990s because of the growing power of workstations, the increased use of graphical user interfaces, and the invention of hybrid object-oriented languages such as C++. Typically, C++ adds 15–20% overhead to an application program due to delayed binding of objects to their methods. Given that the power of the hardware increases more than 20% per annum, this is an acceptable performance penalty. In addition, OOP is much more suitable for the design of graphical user-interface-intensive applications because the display objects correspond with programming objects, thus simplifying design and coding. Finally, if you know C, it is a small step to learn C++.

Summary and Conclusions

HLLs: What are they? What do we mean by "high level"? What constitutes a paradigm? What are the advantages and disadvantages of HLLs? Who uses HLLs? What problems can be solved with these languages?

HLLs are human inventions that allow humans to control and communicate with machines. They obey rules of syntax and unambiguous semantics which are combined to express abstract ideas. HLLs are called "high level" because they express abstractions.

We have chosen to define the level of a HLL in terms of the information content of its syntax and semantics. The Halstead measure of language level essentially says that the higher a HLL is, the fewer symbols are needed to express an idea. Thus, if language A is higher than language B, a certain program can be expressed more succinctly in A than B. This is clearly the case when comparing HLLs with various machine-level languages, where a single statement in the HLL requires many statements in the machine-level language.

HLLs differ from one another in the abstractions they support. Abstract views of the world are called paradigms, and the guiding principle of any HLL is its programming paradigm.

We have compared the following programming paradigms: procedural, functional, declarative, and object-oriented. Procedural programming has the longest history because the first HLLs were based on low-level abstractions that are procedural. FORTRAN, COBOL, C, and Pascal are classical examples of the procedural languages.

Functional and declarative languages employ higher levels of abstraction by restricting the world view to simple mechanisms: mathematical functions and constraints. It may seem odd that such restrictions increase the level of abstraction, but languages like LISP and PROLOG hide much of the detail found to be necessary in the procedural paradigm. This increases the measure of level defined in this section.

Object-oriented programming embraces a novel abstraction that seems to fit the world of computing: objects. In this paradigm, the world is modeled as a collection of objects that communicate by sending messages to one another. The objects are related to each other through an inheritance mechanism that passes on the algorithmic behavior from one class of objects to another class. Inheritance permits reuse and thus raises the programming abstraction to a level above previous paradigms.

The future of HLLs is uncertain and unpredictable. It is unlikely that anyone in 1970 would have predicted the acceptance of functional, declarative, or object-oriented paradigms in the 1990s. Therefore, it is unlikely that the following predictions bear much relationship to computing in the year 2000. However, it is instructive to project a few scenarios and explain their power.

Functional and declarative programming result in software that can be mathematically analyzed, thus leading to greater assurances that the software actually works. Currently these paradigms consume too much memory and machine cycles. However, in the year 2000, very high-speed machines will be commonplace. What will we use these powerful machines for? One answer is that we will no longer be concerned with the execution efficiency of a HLL. The drawbacks of functional and declarative languages will fade, to be replaced by concern for the correctness and expressiveness of the HLL. If this occurs, functional and declarative languages will be the preferred HLLs because of the elevated abstractions supported by the functional and declarative paradigms. Applications constructed from these HLLs will exhibit more sophisticated logic, communicate in non-numeric languages such as speech and graphics, and solve problems that are beyond the reach of current HLLs.

Object-orientation is an appealing idea whose time has come. OOP will be to the 1990s what structured programming was to the 1970s. Computer hardware is becoming increasingly distributed and remote. Networks of workstations routinely solve problems in concert rather than as stand-alone systems. This places greater demands on flexibility and functionality of applications. Consider the next step beyond objects—servers:

$$\text{Server} = \text{Object} + \text{Process}$$

A server is an object that is instantiated as an operating system process or task. The server sends messages to other servers to get work done. The servers "live" on any processor located anywhere on the network. Software is distributed and so is the work. OOP offers the greatest hope for distributing applications in this fashion without loss of control. Should this scenario come true, the OOP paradigm will not only be appropriate, but contribute to greater HLL leverage through reusable objects, distributed servers, and delayed binding of methods to these servers.

Object-oriented languages, databases, and operating systems are on the immediate horizon. Graphical user-interface servers such as X-Windows already exist, lending credibility to this scenario. At least for the near future, HLLs are most likely to become identical with the object-oriented paradigm.

Defining Terms

Abstract class: A class consisting of only an interface specification. The implementation part is unspecified, because the purpose of an abstract class is to establish an interface.

Abstraction: Abstraction in computer languages is a measure of the amount of separation between the hardware and an expression of a programming idea. The level of abstraction of a high-level language defines the level of that language.

ADT: An abstract data type (ADT) is a software module that encapsulates data and functions allowed to be performed on that data. ADTs also separate the interface specification of a module from the implementation part to minimize coupling among modules.

Assembler: A computer program for translating symbolic machine instructions into numerical machine instructions. Assemblers are considered low-level languages for programming a computer.

Class: A specification for one or more objects that defines state (data) and functions (algorithms) that all objects may inherit when created from the class. A class is a template for implementing objects.

Compiler: A computer program that translates the source program statements of a high-level language into lower-leveled object program statements. Compilers differ from interpreters in that they do not immediately perform the operations specified in the source program. Instead, a compiler produces an object program that in turn performs the intended operations when it is run.

Coupling: A measure of the amount of interaction between modules in a computer program. High coupling means that a change in one module is likely to affect another module. Low coupling means there is little impact on other modules whenever a change is made in one module.

Declarative paradigm: A programming paradigm in which the world is modeled as a collection of rules and constraints.

Delayed binding: The process of postponing the meaning of a programming object until the object is manipulated by the computer. Delayed binding is used by interpreters and compilers, but more often it is employed by interpreters.

Functional paradigm: A programming paradigm in which the world is modeled as a collection of mathematical functions.

HLL (also **HOL**)**:** A HLL is a set of symbols which obey unambiguous syntactic and semantic rules: the syntactic rules specify legal combinations of symbols, and the semantic rules specify legal meanings of syntactically correct statements relative to a collection of abstractions.

Implementation part: The definition or algorithm for a programming module which gives the details of how the module works.

Instance variables: Data encapsulated by an object.

Interface specification: The definition of a programming module without any indication of how the module works.

Interpreter: A computer program that translates and performs the intended operations of the source statements of a high-level language program. Interpreters differ from compilers in that they immediately perform the intended operations specified in the source program, and they do not produce an object program.

Member fields: Instance variables of a C++ object.

Member functions: Methods defined on a C++ object.

Metalanguage: A formal language for defining other languages. A metalanguage is typically used to define the syntax of a high-level language.

Methods: Functions allowed to be performed on the data of an object.

Metric: A measure of a computer program's complexity, clarity, length, difficulty, etc.

Object: An instance of a class. Objects have state (data) and function (algorithms) that are allowed to manipulate the data.

Object-oriented paradigm: A programming paradigm in which the world is modeled as a collection of self-contained objects that interact by sending messages. Objects are modules that contain data and all functions that are allowed to be performed on the encapsulated data. In addition, objects are related to one another through an inheritance hierarchy.

Object program: Machine form of a computer program, which is the output from a translator.

Paradigm: An idealized model, typically used as a conceptual basis for software design. Programming paradigms dictate the approach taken by a programmer to organize, and then write, a computer program.

Procedural paradigm: A programming paradigm in which the world is modeled as a collection of procedures which in turn encapsulate algorithms.

Prototyping: A simplified version of a software system is a prototype. Prototyping is the process of designing a computer program through a series of versions; each version becomes a closer approximation to the final one.

Pushdown stack: A data structure containing a list of elements which are restricted to insertions and deletions at one end of the list, only. Insertion is called a push operation and deletion is called a pull operation.

Recursion: A procedure is called recursive if it calls itself.

Reuse: Programming modules are reused when they are copied from one application program and used in another. Reusability is a property of module design that permits reuse.

Semantics: The part of a formal definition of a language that specifies the meanings attached to a syntactically correct statement in the language.

Source program: Symbolic form of a computer program, which is the input to a translator.

Syntax : The part of a formal definition of a language that specifies legal combinations of symbols that make up statements in the language.

Related Topic

87.1 Assembly Language

References

T. Budd, *Object-Oriented Programming,* Reading, Mass.: Addison-Wesley, 1991.

M.A. Ellis and B. Stroustrup, *The Annotated C++ Reference Manual,* Reading, Mass.: Addison-Wesley, 1990.

M.H. Halstead, *Elements of Software Science*, New York: Elsevier North-Holland, 1977.

T.G. Lewis, *CASE: Computer-Aided Software Engineering*, New York: Van Nostrand Reinhold, 1991.

P. Smith, *Expert System Development in Prolog and Turbo-Prolog,* New York: Halsted Press, 1988.

P.H. Winston and B.K.P. Horn, *LISP,* Reading, Mass.: Addison-Wesley, 1989.

87.3 Data Types and Data Structures

Johannes J. Martin

The study of *data types* and *data structures* is a part of the discipline of computer programming. The terms refer to the two aspects of data objects: their usage and their implementation, respectively. The study of data *types* deals with the *identification* of (abstract) data objects in the context of a programming project and with methods of their more or less formal *specification*; the study of data *structures*, on the other hand, is concerned with the *implementation* of such objects using already existing data objects as raw material.

Concretely, the area addresses a basic problem of programming: the reduction of complex objects, such as vectors, tensors, text, graphic images, sound, functions, directories, maps, corporate organizations, models of ecosystems or machinery, or anything else that a program may have to deal with, to the only native objects of digital computers: arrays of binary digits (bits). The fundamental problem of this reduction is managing program complexity. Two organizational tools are essential to its solution: abstraction and hierarchical structuring. Abstraction refers to the separation of *what* computational objects are used for from *how* they are reduced to (i.e., implemented by means of) simpler ones. Hierarchical structuring refers to breaking this reduction up into small manageable steps. Through several steps of abstraction more and more complex objects are constructed, each one reduced to the previous, simpler generation of objects. This process ends when the desired objects have been composed.

Abstract Data Types

An **abstract data type** is one or more *sets* of computational objects together with some basic operations on those objects. One of these sets is defined by the type either by enumeration or by generating operations and

is called the *carrier set* of the type. Customarily it is given the same name as the type. All other sets are called auxiliary sets of the type. In exceptional cases a type may have more than one carrier set.

The heart of the specification of an abstract data type is the definition of its functions, their syntax and semantics. Their syntax is specified by their **functionalities** and their semantics by algebraic axioms. [For more details see, e.g., Martin, 1986].

With sets A and B, the expression $A \rightarrow B$ denotes the set of all functions that have the domain A and the codomain B. Functions $f \in A \rightarrow B$ (traditionally denoted by $f: A \rightarrow B$) are said to have the functionality $A \rightarrow B$.

The collection of basic operations does not need to be minimal. It should, however, be rich enough so that all other operations that one might wish to perform on the objects of the carrier set can be expressed exclusively by these basic operations. The type *Boolean*, for example, consists of the set of Boolean values, Boolean = {true, false}, with, e.g., the operations *not*, *and*, and *or*.

In general, things are not quite this simple. To be useful for programming purposes, even the type Boolean requires at least one additional function. This function, called a *conditional expression*, provides a choice of one of two given values depending on a given Boolean value. It has the form:

f: Boolean 3 SomeType 3 SomeType → SomeType

and, with a, b ∈ SomeType, is defined by:

f (true, a, b) = a and
f (false, a, b) = b

The syntactical form of conditional expressions varies for different programming languages that provide this construct. For example, in the language C it has the form:

```
Boolean ? SomeType : SomeType.        /* with the result type of SomeType */
```

The set SomeType is an auxiliary set of the type Boolean.

Fundamental Data Types

The fundamental types listed next are supported by almost all modern high-level programming languages (reference books on Pascal, Modula II, C, and Ada are listed among the references at the end of this section):

Integer, Real (sometimes called Float), Character, and Boolean

Since their carrier sets are ordered (one of the operations of these types is ≤), these types are also called scalar types. All provide operations for comparing values; in addition, Integer and Real come equipped with the usual arithmetic operations (+, –, *, /) and Boolean with the basic logical operations (not, and, or). Most computers support these operations by hardware instructions. Thus, while bit arrays are the original native objects of a digital computer, the fundamental scalar types may be viewed as the given elementary building blocks for the construction of all other types.

Type Constructors

Enumerated Types

Beginning with Pascal, modern languages provide a rather useful constructor for scalar types, called enumerated types. Enumerated types have finite (small) carrier sets that the programmer defines by enumerating the constants of the type (specified as identifiers). For example, if the type Boolean were not available in Pascal, its carrier set could simply be defined by:

type Boolean = (false, true)

In Pascal, enumerated types are automatically equipped with operations for comparison as well as with the functions succ and pred, i.e., successor and predecessor. In the above example, succ(false) = true, pred(true) = false; succ(true) and pred(false) are not allowed.

Records

Values of scalar types can be arranged into tuples by a construct called a *record* (also called a *structure* in some languages). Into programming languages, records introduce the equivalent of the Cartesian product. Tuples can be viewed as abstract data types. Consider pairs as an example:

The type *Pairs:*

Carrier set:	*Pairs*,
Auxiliary sets:	*A, B*;
Operations:	pair $\in$ *A* 3 *B* $\rightarrow$ *Pairs*;
	first $\in$ *Pairs* $\rightarrow$ *A*;
	scnd $\in$ *Pairs* $\rightarrow$ *B*;

where $\forall$ a $\in$ *A*, b $\in$ *B*

first (pair (a, b)) = a; scnd (pair(a, b)) = b;

Using Pascal notation this type is defined by:

type Pairs = **record** first: A; scnd: B **end**

By providing the so-called field names, first and scnd, the programmer implicitly chooses the names for the selector functions. Pascal does not provide the function "pair." Instead, it permits the declaration of variables of type Pairs whose component values may be set (by assignment) and retrieved:

p: Pairs;	{declaration of p as a variable of type Pairs}
p.first := a; p.scnd := b;	{p now has the value of "pair(a,b)" above}
if p.first = x then . . . else . . .	{p.first is Pascals notation for "first(p)" }

The sets A and B can be of any type including records. Furthermore, since there is no restriction on the number of fields records may have, they can represent arbitrary tuples.

Arrays

Arrays permit the construction of simple function spaces, I $\rightarrow$ A. The domain I, a scalar type—in some languages restricted to a subset {0, 1, ..., *n*} of the integers—is called the index set of the array; A is an arbitrary type. In Pascal, the mathematical notation f $\in$ I $\rightarrow$ A assumes the form:

f**: array**[I] **of** A

I has to be a finite scalar type (e.g., 0 .. 40). The function f can now be defined by associating values of type A with the values of the domain I using the assignment operation:

f[i] := a; where i $\in$ I and a $\in$ A

Application of f to a value j $\in$ I is expressed by f[j]. This expression has a value of type A and can be used as such.

As with records, Pascal allows the naming of the function space I $\rightarrow$ A by the definition:

type myFunctions = **array**[I] **of** A

and the subsequent declaration of a specific function (array) by:

f : myFunctions

Functions of several arguments are represented by so-called multidimensional arrays:

f $\in$ I_1 3 . . . 3 I_n $\rightarrow$ A

is defined by

f : **array**[I_1, . . ., I_n] **of** A

Variant Records

Variant records model *disjoint* (also called *tagged*) *unions*. In contrast to an ordinary union C = A ∪ B, a disjoint union D = A + B is formed by tagging the elements of A and B before forming D such that elements of D can be recognized as elements of A or of B. In programming, this amounts to creating variables that can be used to house values of both type A and type B. A tag field, (usually) part of the variable, is used to keep track of the type of the value currently stored in the variable. In Pascal, D = A + B is expressed by:

```
type    tagType = (inA, inB) {an enumerated type};
        D       = record
                  case kind: tagType of
                    inA: (aValue: A);
                    inB: (bValue:B);
                  end.
```

Variables of type D are now used as follows:

```
            mix:    D;              {mix is declared to be of type D}
mix.kind            :=      inA;
mix.aValue          :=      a;
. . .
if mix.kind = inA
   then {do something with mix.aValue, which is of type A}
   else {do something with mix.bValue, which is of type B}
```

Conceptually, only one of the two fields, mix.aValue or mix.bValue, exists at any one time. The proper use of the tag is policed in some languages (e.g., Ada) and left to the programmer in others (e.g., Pascal).

An Example of a User-Defined Abstract Data Type

Most carrier sets are assumed to contain a distinguished value: *error*. *Error* is not a proper computational object: a function is considered to compute *error* if it does not return to the point of its invocation due to some error condition. Functions are called **strict** if they compute the value *error* whenever one or more of their arguments have the value *error*.

The following example models a cafeteria tray stack with the following operations:

1. create a new stack with n trays;
2. remove a tray from a stack;
3. add a tray to the stack;
4. check if the stack is empty;

Specification:

```
Cts (cafeteria tray stacks)          is the carrier set of the type;
Boolean and Integer                  are auxiliary sets;
newStack, remove, add, isEmpty       are the operations of the type
where
        newStack        ∈ Integer    → Cts; {create a stack of n trays}
        remove, add     ∈ Cts        → Cts;
        isEmpty         ∈ Cts        → Boolean;
```

Axioms (logical expressions that describe the semantics of the operations): All functions are strict and, for all non-negative values of n,

1. remove (newStack(n)) = if n = 0 then *error* else newStack(n − 1);
2. add(newStack(n)) = newStack(n + 1);
3. isEmpty (newStack(n)) = (n = 0).

These axioms suffice to describe the desired behavior of Cts exactly, i.e., using these axioms, arbitrary expressions built with the above operations can be reduced to *error* or newStack(m) for some m.

Implementation:
For the representation of Cts (i.e. its data structure) we will choose the type Integer.

```
type      Cts = integer;

function newStack(n: Integer):Integer;
    begin if n < 0     then error ('n must be >= 0')     else newStack := n end;

function remove (s: Cts):Cts;
    begin if s = 0     then error ('stack empty')        else remove := s - 1 end;

function add (s: Cts):Cts;                    begin add := s + 1 end;

function isEmpty (s: Cts):Boolean;            begin isEmpty := (s = 0) end;
```

Above, "error" is a function that prints an error message and stops the program.

Dynamic Types

The carrier sets of dynamic types contain objects of vastly different size. For these types, variables (memory space) must be allocated dynamically, i.e., at run time, when the actual sizes of objects are known. Examples for dynamic types are character strings, lists, tree structures, sets, and graphs. A classical example of a dynamic type is a special type of a list: a queue. As the name suggests, a queue object is a sequence of other objects with the particular restrictive property that objects are inspected at and removed from its front and added to its rear.

Specification of Queues

Carrier set: *Queues*
Auxiliary sets: *Boolean*, *A* (*A* contains the items to be queued)
Operations: newQueue, isEmpty, queue, pop, front

1. newQueue	$\in$ *Queues*;		{a new, empty queue}
2. isEmpty	$\in$ *Queues*	$\rightarrow$ *Boolean*;	{check if a queue is empty}
3. queue	$\in A \times$ *Queues*	$\rightarrow$ *Queues*;	{add an object to the rear of a queue}
4. front	$\in$ *Queues*	$\rightarrow A$;	{return front element for inspection}
5. pop	$\in$ *Queues*	$\rightarrow$ *Queues*;	{remove front element from a queue}

Axioms: All functions are strict and, for a $\in$ A and s $\in$ Queues,

```
      isEmpty (newQueue);       (i.e. isEmpty (newQueue)    is true)
not   isEmpty (queue(a, s));    (i.e. isEmpty (queue(a, s))  is false)
      pop (newQueue)        =   error;
      pop (queue(a, s))     =   if s = newQueue
                                      then newQueue else queue(a, pop(s));
      front(newQueue)       =   error;
      front(queue(a, s))    =   if s = newQueue
                                      then a else front(s).
```

Implementation of Queues

The following implementation represents queues of the form queue(a, s) by the ordered pair (a, s) and a new, empty queue by the null pair *nil*. For the moment we assume that a data type, *Pairs*, which provides pairs on demand at run time, already exists and is defined as follows:

Specification of Pairs

Carrier Set: *Pairs*
Auxiliary sets: *Boolean*, *A*

Operations: nil, isNil, pair, first, scnd

1. nil	∈ Pairs,		{a distinguished pair, the null pair}
2. isNil	∈ Pairs	→ Boolean,	{test for the null pair}
3. pair	∈ A 3 Pairs	→ Pairs,	{combine an item and a pair to a new pair}
4. first	∈ Pairs	→ A,	{the first component, i.e., the item}
5. scnd	∈ Pairs	→ Pairs,	{the second component, i.e., the pair}

Axioms: All functions are strict and, for a ∈ A and p ∈ Pairs,

	isNil(nil);	(i.e. isNil(nil)	is true)
not	isNil(pair(a,p));	(i.e. isNil(pair(a,p))	is false)
	first (nil) = error;		
	first(pair(a,p)) = a;		
	scnd (nil) = error;		
	scnd(pair(a,p)) = p;		

With pairs, queues may now be implemented as follows:

```
type   Queues = Pairs;
function newQueue    : Queues;                  begin  newQueue  := nil          end;
function isEmpty (s : Queues)    : Boolean;     begin  isEmpty   := (s = nil)    end;
function queue (x : A; s : Queues)   : Queues;  begin  queue    := pair(x, s)   end;
function pop (s : Queues) : Queues;
begin
  if isNil(s)
       then error ('cannot pop empty queue')
       else if scnd(s) = nil
            then pop := nil
            else pop := pair (first(s), pop (scnd(s)))
  end;
function front (s : Queues) : A;
begin
  if isNil(s)
       then error ('an empty queue does not have a front')
       else if scnd(s) = nil
            then front := first(s)
            else front := front(scnd(s))
  end;
```

The logic of these programs echoes the axioms. Such implementations are sometimes not very efficient but useful for prototype programs, since the probability of their correctness is very high. The queues behave as *values*, i.e., the functions queue(a,s) and pop(s) do not modify the queues s but compute new queues; after the execution of, e.g., s1 := pop(s) there are two independent queue values, s and s1. This is exactly the behavior postulated by the axioms. However, practical applications frequently deal with *mutable objects*, objects that can be modified. With mutable objects memory may often be used more efficiently, since it is easier to decide when a memory cell, used, e.g., for storing an ordered pair, is no longer needed and thus may be recycled. If queues are viewed as mutable objects, the operations *queue* and *pop* are implemented as procedures that modify a queue. In order to apply the style of axioms introduced above for the description of mutable objects, these objects are best viewed as containers of values. The *procedure* queue(a, qobj), for example, takes the queue value, e.g., s, out of the container qobj, applies the *function* queue(a, s) (described by the axioms), and puts the result back into qobj.

If more than one place in a program needs to maintain a reference to such an object, the object must be implemented using a *head cell:* a storage location that represents the object and that is not released. The following

implementation uses a head cell with two fields, one pointing to the front and one to the rear of the queue. We assume that the type *Pairs* has two additional functions:

```
6. pairH ∈ Pairs 3 Pairs   → Pairs;   {create a pair with 2 pair fields}
7. firstH ∈ Pairs          → Pairs;   {retrieve the first field of such a 2-pair cell}
```

and three procedures:

```
8. setfirstH ( p: Pairs; q: Pairs);     {change the firstH field of q to p}
9. setscnd ( p: Pairs; q: Pairs);       {change the scnd field of q to p}
10. delete (s: Pairs)                   {free the storage space occupied by s}
```

```
type    Queues = Pairs;
procedure newQueue(var q : Queues);      begin q          := pairH(nil, nil)    end;
function isEmpty (s : Queues) : Boolean; begin isEmpty    := (firstH(s) = nil)  end;
procedure queue (x : A; s : Queues);
   var temp: Pairs;
   begin temp := pair(x, nil);
      if isNil(firstH(s)) then setfirstH(temp, s) else setscnd(temp, scnd(s));
      setscnd (temp, s);
      end;
function pop (s : Queues) : Queues;
   var temp : Pairs;
   begin
      if isNil(firstH(s))
         then error ('cannot pop empty queue')
         else begin temp := firstH(s); setfirstH(scnd(temp), s); delete(temp) end
      end;
function front (s : Queues) : A;
   begin
      if isNil(firstH(s))
         then error ('an empty queue does not have a front')
         else front := first(firstH(s))
      end;
```

Compared to the value implementation given earlier, this implementation improves the performance of *front* and *pop* from O(n) to O(1).

An algorithm has O(f(n)) (pronounced: order f(n) or proportional to f(n)) time performance if there exists a constant c, such that, for arbitrary values of the input size n, the time that the algorithm needs for its computation is $t \leq c \cdot f(n)$.

Most modern programming languages support the implementation of the type *pairs* (*n-tuples*) whose instances can be created dynamically. It requires two operations, called *new* and *dispose* in Pascal, that dynamically allocate and deallocate variables, respectively. These operations depend on the concept of a *reference* (or *pointer*), which serves as a name for a variable. References always occupy the same storage space independently of the type of variable they refer to. The following implementation of *Pairs* explains the concept.

```
type CellKind  = (headCell, bodyCell);
     Pairs     = ^PairCell; {Pairs are references to PairCells}
     PairCell  = record     tail: Pairs;
                   case  kind: CellKind of
                         headCell: (frnt: Pairs);
                         bodyCell: (val: A)
                   end;
function pair(item: A; next:Pairs):Pairs;
  var p: Pairs;
```

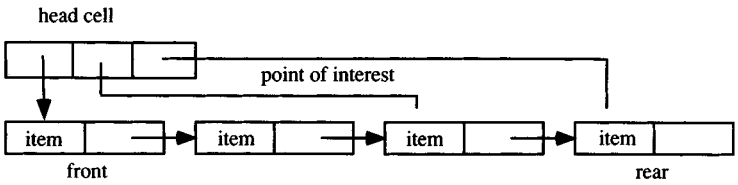


FIGURE 87.14 A list implementation with a point of interest and access to front and rear.

```
    begin
      new(p, bodyCell); {a new "bodyCell" is created and accessible through p}
      p^.kind := bodyCell; p^.val := item; p^.tail := next; pair := p
      end;
  function first(p: Pairs):A;
    begin
      if p = nil then error('...')
          else if p^.kind = bodyCell then first := p^.val else error('...')
      end;
  procedure setfirstH(p, q:Pairs);
    begin
      if q = nil then error('...')
          else if q^.kind = headCell then q^.frnt := p else error('...')
      end;
```

(Note: The Pascal constant **nil** denotes the null pointer, a reference to nothing.)

```
  function isNil(p: Pairs): Boolean; begin isNil := (p = nil) end;
```

The reader should have no difficulty filling in the rest of the implementation of Pairs. Most of the algorithms on dynamic data structures have been developed in the 1960s; still, an excellent reference is Knuth [1973].

More Dynamic Data Types

Stacks and Lists with a Point of Interest

A queue is the type of (linear) list used to realize first-come-first-served behavior. In contrast, another linear type, the *stack*, realizes last-come-first-served behavior. Sometimes it is necessary to scan a list object without dismantling it. This is accomplished by giving the list a point of interest (see Fig. 87.14). This requires four additional operations:

```
restart(l: List);                 {moves point of interest to beginning of list l}
current(l:List):A;                {returns object at point of interest}
advance(l:List);                  {advances point of interest by one toward end of list l}
endOfList(l : List): Boolean;     {true, if end of list has been reached}
```

The type can be extended further by allowing insertions and deletions at the point of interest.

N-ary, Binary, and General Trees

An **n-ary tree** is the smallest set containing the *empty tree* and all ordered n+1-tuples $t = (a, t_1, ..., t_n)$ where a is member of some auxiliary set and the t_i are n-ary trees. The element a is called the *root element* or simply the root of t and the t_i are called *subtrees* of t.

Note that in this sense, a list is a unary tree. Binary trees used as searchtrees access finite ordered sets.

A **binary searchtree** is a tree that accesses a set. A tree t *accesses* a set s if the root of t is some element a of s, and s_1, called the left subtree of t, accesses the subset $\{x | x \in s \text{ and } x < a\}$ and s_2, *called the right subtree of t, accesses the subset* $\{x | x \in s \text{ and } x > a\}$.

If the left and right subtrees of the above definition are of similar size, then the time for finding an element in the set is proportional to log(n) where n is the cardinality of the set.

Quaternary trees (usually called quad trees) and octonary trees (called oct trees) are used to access two-dimensionally and three-dimensionally organized data, respectively. As with lists, the implementation of n-ary trees is based on n+1-tuples. A minimal set of operations for binary trees includes:

nilTree	∈ Trees;		{the empty tree, represented by nil}
isNil	∈ Trees	→ Boolean;	{test if tree is empty}
tree	∈ A 3 Trees 3 Trees	→ Trees;	{build tree from an item and subtrees}
root	∈ Trees	→ A;	{retrieve root item of tree}
left	∈ Trees	→ Trees;	{retrieve left subtree}
right	∈ Trees	→ Trees;	{retrieve right subtree}

A **general tree** is the smallest set containing all order pairs $t = (a,s)$ where a is a member of some auxiliary set and s is a possibly empty list of general trees. The element a is called the root element or simply the root of t and the trees in s are called subtrees of t.

Note that there is no empty general tree; the simplest tree has a root and an empty list of subtrees. General trees are useful for the representation of hierarchical organizations such as the table of contents of a book or the organization of a corporation.

Functions, Sets, Relations, Graphs

Functions with reasonably small domains can be represented by arrays, as described earlier. Similarly, sets formed from a reasonably small universal set, relations on small domains, and graphs with not too many vertices can be represented by their characteristic functions implemented as bit arrays. In fact, Pascal provides a type constructor for sets that are derived from small universal sets.

Frequently domains are far too large for this approach. For example, the symbol table that a compiler of a programming language maintains is a function from the set of valid identifiers to some set of attributes. The set of valid identifiers is infinite, or, if some length limitation is imposed, finite but exceedingly large (there are nearly 300 billion identifiers of eight characters or less). For most of its domain a symbol table returns the default value *new* or *not found*. It is therefore economical to store the mapping only for those domain values that map to a value different from the default value. A function of this sort is usually specified as follows:

Specification of Functions:

Carrier Set:	*Functions*	
Auxiliary sets:	*Dom*, *Cod*	(domain and codomain)
Operations:	newFun, apply, update	

1. newFun	∈ Functions,	(returns default everywhere)	
2. apply	∈ Functions 3 Dom		→ Cod,
3. update	∈ Functions 3 Dom 3 Cod		→ Functions;

Axioms: All functions are strict and, for x,z ∈ Dom, y ∈ Cod and f ∈ Functions,

apply(newFun, x)	= default;
apply (update(f,x,y), z)	= if x = z then y else apply (f, z);

An implementation based on these axioms amounts to representing a function as a list of those of its individual mappings that differ from *default* and leads to an O(1) performance for *update* and an **O(n) performance** of *apply*. Better is an implementation by binary searchtrees with a performance of O(log(n)) for both *apply* and *update*.

Hashing

The fastest method for the implementation of functions is *hash coding* or *hashing*. By means of a *hash function,* h ∈ Dom → 0 .. k–1, the domain of the function is partitioned into k sections and each section is associated with an index. For each partition a simple list implementation is used and the lists are stored in an array A:

array[1 .. k–1]. In order to compute apply(f,x) or update(f,x,y), the list at A[hash(x)] is searched or updated. If the hash function has been properly chosen and if k and the number of function values different from default are of similar size, then the individual lists can be expected to be very short and independent of the number of nondefault entries of the function; thus performance for apply and update is O(1).

The above discussion applies also to sets, relations, and graphs, since these objects can be represented by their characteristic functions.

Object-Oriented Programming

In languages that support object-oriented programming, *classes* (i.e., types) of objects are defined by specifying (1) the variables that each object will own as *instance variables* and (2) operations, called *methods,* applicable to the objects of the class. As a difference in style, these methods are not invoked like functions or procedures, but are *sent* to an object as a *message.* The expression [*window moveTo* : *x* : *y*] is an example of a message in the programming language *Objective C*, a dialect of C. Here the object *window*, which may represent a window on the screen, is instructed to apply to itself the method *moveTo* using the parameters *x* and *y.*

New objects of a class are created—usually dynamically—by *factory methods* addressed to the class itself. These methods allocate the equivalent of a record whose fields are the instance variables of the object and return a reference to this record, which represents the new object. After its creation an object can receive messages from other objects.

To data abstraction, object-oriented programming adds the concept of inheritance: from an existing class new (sub)classes can be derived by adding additional instance variables and/or methods. Each subclass inherits the instance variables and methods of its superclass. This encourages the use of existing code for new purposes. As an example, consider a class of a *list* objects. Each object has two instance variables pointing to the front and the rear of the list. In Objective C, the specification of the interface for this list class, i.e., the declaration of the instance variables and headers (functionalities) of the methods, has the following form:

```
@interface MyLists : Object     /* Object is the universal (system) class from which
                                   all classes are derived directly or indirectly */
{                               /* declaration of the instance variables;
   listRef front;                  listRef is the type of a pointer to a list assumed
   listRef rear;                   to be defined elsewhere */
}
– initList;                     /* initializes instance variables with null pointers */
– (BOOL) isEmpty;               /* test for empty list; note: parameter list is implied */
– add : (item) theThing;        /* item is the type of things on the list */
– pop;
– (item) front;
@end
```

As a companion of the *interface* file there is also an *implementation* file that contains the executable code for the methods of the class. A list with a point of interest can be defined as a subclass of MyList as follows:

```
@interface   ScanList : MyList   /* ScanList is made a subclass of MyList */

{   listRef pointOfInterest; }

– restart;
– (BOOL)endOfList;
– advance;
– (Item)current;
@end
```

If we also need a list that can add and delete at the point of interest, we define:

```
@interface      InsertionList : ScanList
{ }                             /* there are no new instance variables */
```

```
– insert : (item) theThing;
– shrink;                    /* removes item at the point of interest */
@end
```

If a subclass defines a method already defined in the superclass, the new definition overrides the old one. Suppose we need a list where items are kept in ascending order:

```
@interface  SortedList : MyList
{ }
– add : (item) theThing;     /* this version of add inserts theThing at the proper
                                place to keep the list sorted */
@end
```

Defining Terms

Abstract data type: One or more sets of computational objects together with some basic operations on those objects. One of these sets is defined by the type either by enumeration or by generating operations and is called the *carrier set* of the type. Customarily it is given the same name as the type. All other sets are called auxiliary sets of the type. In exceptional cases a type may have more than one carrier set.

Binary searchtree: A tree that accesses a set. A tree t accesses a set s if the root of t is some element a of s, and s_1, called the left subtree of t, accesses the subset $\{x \mid x \in s \text{ and } x < a\}$, *and* s_2, *called the right subtree of t, accesses the subset* $\{x \mid x \in s \text{ and } x > a\}$.

Functionality: With sets A and B, the expression $A \rightarrow B$ denotes the set of all functions that have the domain A and the codomain B. Functions $f \in A \rightarrow B$ (traditionally denoted by $f: A \rightarrow B$) are said to have the functionality $A \rightarrow B$.

General tree: The smallest set containing all ordered pairs $t = (a, s)$ where a is member of some auxiliary set and s is a possibly empty list of general trees. The element a is called the root element or simply the root of t and the trees in s are called subtrees of t.

n-ary tree: The smallest set containing the *empty tree* and all ordered n+1-tuples $t = (a, t_1, \ldots, t_n)$ where a is member of some auxiliary set and the t_i are n-ary trees. The element a is called the root element or simply the root of t and the t_i are called subtrees of t.

O(f(n)) performance: An algorithm has O(f(n)) (pronounced: order f(n) or proportional to f(n)) time performance if there exists a constant c, such that, for arbitrary values of the input size n, the time that the algorithm needs for its computation is $t \leq c \cdot f(n)$.

Strictness: Most carrier sets are assumed to contain a distinguished value: *error*. *Error* is not a proper computational object: a function is considered to compute *error* if it does not return to the point of its invocation due to some error condition. Functions are called *strict* if they compute the value *error* whenever one or more of their arguments have the value *error*.

Related Topic

90.3 Programming Methodology

References

A.V. Aho, *Data Structures and Algorithms,* Reading, Mass.: Addison-Wesley, 1988.

T.H. Cormen, *Introduction to Algorithms,* New York: McGraw-Hill, 1995.

K. Jensen and N. Wirth, *Pascal: User Manual and Report,* Berlin, Springer-Verlag, 1974.

B.W. Kernighan and D.M. Ritchie, *The C Programming Language,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.

D.E. Knuth, *The Art of Computer Programming,* vol. 1, Reading, Mass.: Addison-Wesley, 1973, chap. 2.

J.J. Martin, *Data Types and Data Structures*, C.A.R. Hoare, Series Ed., Englewood Cliffs, N.J.: Prentice-Hall International, 1986.

NeXT Step Concepts, Chapter 3, *Objective C*, NeXT Developers' Library, NeXT Inc.

S. Sahni, *Fundamentals of Data Structures in C++,* Computer Science Press, 1995.
R. Sedgewick, *Algorithms in C++,* Reading, Mass.: Addison-Wesley, 1993.
B. Stroustrup, *The C++ Programming Language*, Reading, Mass.: Addison-Wesley, 1991.
United States Department of Defense, *Reference Manual for the Ada® Programming Language,* Washington D.C.: U.S. Government Printing Office, 1983.
N. Wirth, *Programming in Modula-2,* Berlin: Springer-Verlag, 1983.

Further Information

There is a wealth of textbooks on data structures. Papers on special aspects of data types and their relation to programming languages are found regularly in periodicals such as *ACM Transactions on Programming Languages and Systems, the Journal of the Association for Computing Machinery, IEEE Transactions on Computers, IEEE Transactions on Software Engineering,* or *Acta Informatica.*

Berger, D., Goodman, J.R., Sohi, G.S. "Memory Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

88

Memory Systems

Doug Burger
University of Wisconsin-Madison

James R. Goodman
University of Wisconsin-Madison

Gurindar S. Sohi
University of Wisconsin-Madison

88.1 Introduction
88.2 Memory Hierarchies
88.3 Cache Memories
88.4 Parallel and Interleaved Memories
88.5 Virtual Memory
88.6 Research Issues

88.1 Introduction

A *memory system* serves as a repository of information (data) in a computer system. The processor [also called the central processing unit (CPU)] accesses (reads or loads) data from the memory system, performs computations on them, and stores (writes) them back to memory. The memory system is a collection of storage locations. Each storage location, or *memory word*, has a numerical *address*. A collection of storage locations forms an *address space*. Figure 88.1 shows the essentials of how a processor is connected to a memory system via address, data, and control lines.

When a processor attempts to load the contents of a memory location, the request is very urgent. In virtually all computers, the work soon comes to a halt (in other words, the processor *stalls*) if the memory request does not return quickly. Modern computers are generally able to continue briefly by overlapping memory requests, but even the most sophisticated computers will frequently exhaust their ability to process data and stall momentarily in the face of long memory delays. Thus, a key performance parameter in the design of any computer, fast or slow, is the effective speed of its memory.

Ideally, the memory system must be both infinitely large so that it can contain an arbitrarily large amount of information and infinitely fast so that it does not limit the processing unit. Practically, however, this is not possible. There are three properties of memory that are inherently in conflict: speed, capacity, and cost. In general, technology tradeoffs can be employed to optimize any two of the three factors at the expense of the third. Thus it is possible to have memories that are (1) large and cheap, but not fast; (2) cheap and fast, but small; or (3) large and fast, but expensive. The last of the three is further limited by physical constraints. A large-capacity memory that is very fast is also physically large, and speed-of-light delays place a limit on the speed of such a memory system.

The **latency** (L) of the memory is the delay from when the processor first requests a word from memory until that word arrives and is available for use by the processor. The latency of a memory system is one attribute of performance. The other is **bandwidth** (BW), which is the rate at which information can be transferred from the memory system. The bandwidth and the latency are related. If R is the number of requests that the memory can service simultaneously, then:

$$BW = \frac{R}{L} \tag{88.1}$$

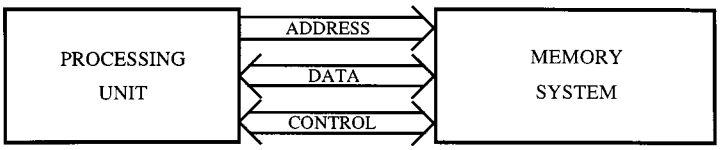


FIGURE 88.1 The memory interface.

From Eq. (88.1) we see that a decrease in the latency will result in an increase in bandwidth, and vice versa, if R is unchanged. We can also see that the bandwidth can be increased by increasing R, if L does not increase proportionately. For example, we can build a memory system that takes 20 ns to service the access of a single 32-bit word. Its latency is 20 ns per 32-bit word, and its bandwidth is

$$\frac{32}{20 \times 10^{-9}} \frac{\text{bits}}{\text{sec}}$$

or 200 Mbytes/s. If the memory system is modified to accept a new (still 20 ns) request for a 32-bit word every 5 ns by overlapping results, then its bandwidth is

$$\frac{32}{5 \times 10^{-9}} \frac{\text{bits}}{\text{sec}}$$

or 800 Mbytes/s. This memory system must be able to handle four requests at a given time.

Building an ideal memory system (infinite capacity, zero latency and infinite bandwidth, with affordable cost) is not feasible. The challenge is, given a set of cost and technology constraints, to engineer a memory system whose abilities match the abilities that the processor demands of it. That is, engineering a memory system that performs as close to an ideal memory system (for the given processing unit) as is possible. For a processor that stalls when it makes a memory request (some current microprocessors are in this category), it is important to engineer a memory system with the lowest possible latency. For those processors that can handle multiple outstanding memory requests (vector processors and high-end CPUs), it is important not only to reduce latency, but also to increase bandwidth (over what is possible by latency reduction alone) by designing a memory system that is capable of servicing multiple requests simultaneously.

Memory hierarchies provide decreased average latency and reduced bandwidth requirements, whereas parallel or **interleaved** memories provide higher bandwidth.

88.2 Memory Hierarchies

Technology does not permit memories that are cheap, large, and fast. By recognizing the nonrandom nature of memory requests, and emphasizing the *average* rather than worst case latency, it is possible to implement a hierarchical memory system that performs well. A small amount of very fast memory, placed in front of a large, slow memory, can be designed to satisfy most requests at the speed of the small memory. This, in fact, is the primary motivation for the use of registers in the CPU: in this case, the programmer or compiler makes sure that the most commonly accessed variables are allocated to registers.

A variety of techniques, employing either hardware, software, or a combination of the two, can be employed to assure that most memory references are satisfied by the faster memory. The foremost of these techniques is the exploitation of the *locality of reference* principle. This principle captures the fact that some memory locations are referenced much more frequently than others. *Spatial locality* is the property that an access to a given memory location greatly increases the probability that neighboring locations will soon be accessed. This is largely, but not exclusively, a result of the tendency to access memory locations sequentially. *Temporal locality* is the property that an access to a given memory location greatly increases the probability that the same location

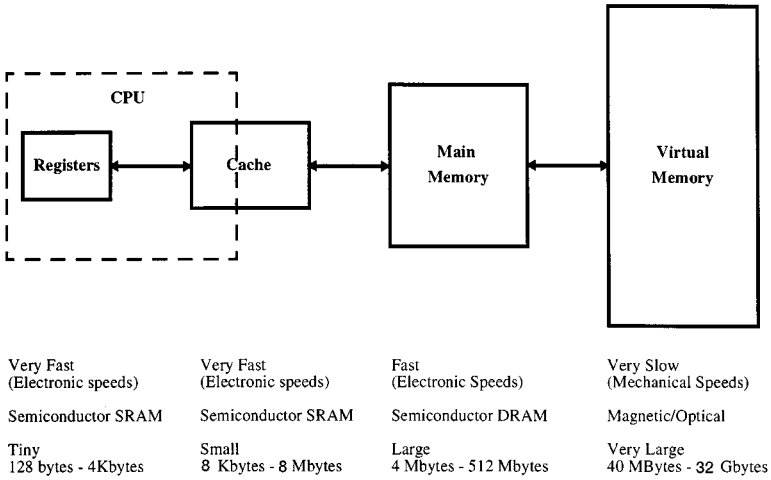


FIGURE 88.2 A memory hierarchy.

will be accessed again soon. This is largely, but not exclusively, a result of the frequency of looping behavior of programs. Particularly for temporal locality, a good predictor of the future is the past: the longer a variable has gone unreferenced, the less likely it is to be accessed soon.

Figure 88.2 depicts a common construction of a memory hierarchy. At the top of the hierarchy are the CPU registers, which are small and extremely fast. The next level down in the hierarchy is a special, high-speed semiconductor memory known as a **cache memory**. The cache can actually be divided into multiple distinct levels; most current systems have between one and three levels of cache. Some of the levels of cache may be on the CPU chip itself, they may be on the same module as the CPU, or they all may be entirely distinct. Below the cache is the conventional memory, referred to as *main memory,* or *backing storage.* Like a cache, main memory is semiconductor memory, but it is slower, cheaper, and denser than a cache. Below the main memory is the virtual memory, which is generally stored on magnetic or optical disk. Accessing the virtual memory can be tens of thousands times slower than accessing the main memory because it involves moving mechanical parts.

As requests go deeper into the memory hierarchy, they encounter levels that are larger (in terms of capacity) and slower than the higher levels (moving left to right in Fig. 88.2). In addition to size and speed, the bandwidth in-between adjacent levels in the memory hierarchy is smaller for the lower levels. The bandwidth in-between the registers and top cache level, for example, is higher than that between cache and main memory or main memory and virtual memory. Since each level presumably intercepts a fraction of the requests, the bandwidth to the level below need not be as great as that to the intercepting level.

A useful performance parameter is the *effective latency.* If the needed word is found in a level of the hierarchy, it is a *hit;* if a request must be sent to the next lower level, the request is said to *miss.* If the latency L_{HIT} is known in the case of a hit and the latency in the case of a miss is L_{MISS}, the effective latency for that level in the hierarchy can be determined from the *hit ratio* (H), the fraction of memory accesses that are hits:

$$L_{\text{average}} = L_{\text{HIT}} \cdot H + L_{\text{MISS}} \cdot (1 - H) \tag{88.2}$$

The portion of memory accesses that miss is called the *miss ratio* ($M = 1 - H$). The hit ratio is strongly influenced by the program being executed, but is largely independent of the ratio of cache size to memory size. It is not uncommon for a cache with a capacity a few thousand bytes to exhibit a hit ratio greater than 90%.

88.3 Cache Memories

The basic unit of construction of a semiconductor memory system is a *module* or *bank.* A memory bank, constructed from several memory chips, can service a single request at a time. The time that a bank is busy servicing a request is called the *bank busy time.* The bank busy time limits the bandwidth of a memory bank.

Both caches and main memories are constructed in this fashion, although caches have significantly shorter bank busy times than do main memory banks.

The hardware can dynamically allocate parts of the cache memory for addresses deemed most likely to be accessed soon. The cache contains only redundant copies of the address space. The cache memory is *associative,* or *content-addressable.* In an associative memory, the address of a memory location is stored, along with its content. Rather than reading data directly from a memory location, the cache is given an address and responds by providing data which may or may not be the data requested. When a cache miss occurs, the memory access is then performed with respect to the backing storage, and the cache is updated to include the new data.

The cache is intended to hold the most active portions of the memory, and the hardware dynamically selects portions of main memory to store in the cache. When the cache is full, bringing in new data must be matched by deleting old data. Thus a strategy for cache management is necessary. Cache management strategies exploit the principle of locality. Spatial locality is exploited by the choice of what is brought into the cache. Temporal locality is exploited in the choice of which block is removed. When a cache miss occurs, hardware copies a large, contiguous block of memory into the cache, which includes the word requested. This fixed-size region of memory, known as a cache *line* or *block*, may be as small as a single word, or up to several hundred bytes. A block is a set of contiguous memory locations, the number of which is usually a power of two. A block is said to be *aligned* if the lowest address in the block is exactly divisible by the block size. That is to say, for a block of size B beginning at location A, the block is aligned if

$$A \ modulo \ B = 0 \tag{88.3}$$

Conventional caches require that all blocks be aligned.

When a block is brought into the cache, it is likely that another block must be evicted. The selection of the evicted block is based on some attempt to capture temporal locality. Since prescience is so difficult to achieve, other methods are generally used to predict future memory accesses. A least-recently-used (LRU) policy is often the basis for the choice. Other replacement policies are sometimes used, particularly because true LRU replacement requires extensive logic and bookkeeping.

The cache often comprises two conventional memories: the data memory and the tag memory, shown in Fig. 88.3. The address of each cache line contained in the data memory is stored in the tag memory, as well as other information (*state* information), particularly the fact that a valid cache line is present. The state also keeps track of which cache lines the processor has modified. Each line contained in the data memory is allocated a corresponding entry in the tag memory to indicate the full address of the cache line.

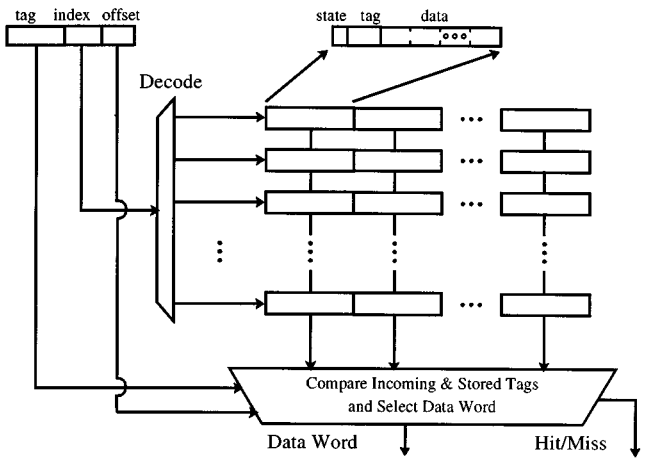


FIGURE 88.3 Components of a cache memory. (*Source:* Adapted from M. D. Hill, "A case for direct-mapped caches," *IEEE Computer,* 21(12), 27, 1988.)

The requirement that the cache memory be associative (content-addressable) complicates the design. Addressing data by content is inherently more complicated than by its address. All the tags must be compared concurrently, of course, because the whole point of the cache is to achieve low latency. The cache can be made simpler, however, by introducing a mapping of memory locations to cache cells. This mapping limits the number of possible cells in which a particular line may reside. The extreme case is known as *direct mapping*, in which each memory location is mapped to a single location in the cache. Direct mapping makes many aspects of the design simpler, since there is no choice of where the line might reside, and no choice as to which line must be replaced. However, direct mapping can result in poor utilization of the cache when two memory locations are alternately accessed and must share a single cache cell.

A hashing algorithm is used to determine the cache address from the memory address. The conventional mapping algorithm consists of a function of the form

$$A_{\text{cache}} = \frac{A_{\text{memory}} \bmod cache\_size}{cache\_line\_size} \tag{88.4}$$

where A_{cache} is the address within the cache for main memory location A_{memory}, *cache_size* is the capacity of the cache in addressable units (usually bytes), and *cache_line_size* is the size of the cache line in addressable units. Since the hashing function is simple bit selection, the tag memory need only contain the part of the address not implied by the hashing function. That is,

$$A_{\text{tag}} = A_{\text{memory}} \text{ div } size\_of\_cache \tag{88.5}$$

where A_{tag} is stored in the tag memory and *div* is the integer divide operation. In testing for a match, the complete address of a line stored in the cache can be inferred from the tag and its storage location within the cache.

A *two-way set-associative* cache maps each memory location into either of two locations in the cache and can be constructed essentially as two identical direct-mapped caches. However, both caches must be searched at each memory access and the appropriate data selected and multiplexed on a tag match (hit). On a miss, a choice must be made between the two possible cache lines as to which is to be replaced. A single LRU bit can be saved for each such pair of lines to remember which line has been accessed more recently. This bit must be toggled to the current state each time either of the cache lines is accessed.

In the same way, an *M-way associative* cache maps each memory location into any of *M* memory locations in the cache and can be constructed from *M* identical direct-mapped caches. The problem of maintaining the LRU ordering of *M* cache lines quickly becomes hard, however, since there are *M*! possible orderings, and therefore it takes at least

$$\lceil \log_2 (M!) \rceil \tag{88.6}$$

bits to store the ordering. In practice, this requirement limits true LRU replacement to 3- or 4-way set associativity.

Figure 88.4 shows how a cache is organized into sets, blocks, and words. The cache shown is a 2-Kbyte, 4-way set-associative cache, with 16 sets. Each set consists of four blocks. The cache block size in this example is 32 bytes, so each block contains eight 4-byte words. Also depicted at the bottom of Fig. 88.4 is a 4-way interleaved main memory system (see the next section for details). Each successive word in the cache block maps into a different main memory bank. Because of the cache's mapping restrictions, each cache block obtained from main memory will be loaded into its corresponding set, but may appear anywhere within that set.

Write operations require special handling in the cache. If the main memory copy is updated with each write operation—a technique known as *write-through* or *store-through*—the writes may force operations to stall while the write operations are completing. This can happen after a series of write operations even if the processor is allowed to proceed before the write to the memory has completed. If the main memory copy is not updated

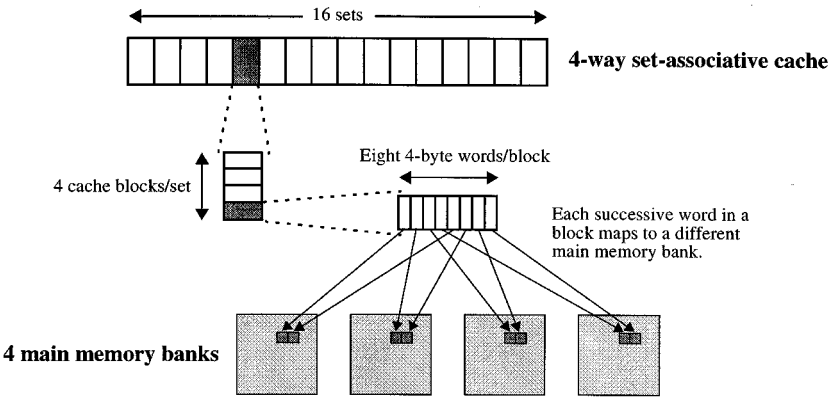


FIGURE 88.4 Organization of a cache.

with each write operation—a technique known as *write-back* or *copy-back* or *deferred writes*—the main memory locations become stale, that is, memory no longer contains the correct values and must not be relied upon to provide data. This is generally permissible, but care must be exercised to make sure that it is always updated before the line is purged from the cache and that the cache is never bypassed. Such a bypass could occur with DMA (*direct memory access*), in which the I/O system writes directly into main memory without the involvement of the processor.

Even for a system that implements write-through, care must be exercised if memory requests bypass the cache. While the main memory is never stale, a write that bypasses the cache, such as from I/O, could have the effect of making the cached copy stale. A later access by the CPU could then provide an incorrect value. This can only be avoided by making sure that cached entries are invalidated even if the cache is bypassed. The problem is relatively easy to solve for a single processor with I/O, but becomes very difficult to solve for multiple processors, particularly so if multiple caches are involved as well. This is known in general as the cache *coherence* or *consistency* problem.

The cache exploits spatial locality by loading an entire cache line after a miss. This tends to result in bursty traffic to the main memory, since most accesses are filtered out by the cache. After a miss, however, the memory system must provide an entire line at once. Cache memory nicely complements an interleaved, high-bandwidth main memory (described in the next section), since a cache line can be interleaved across many banks in a regular manner, thus avoiding memory conflicts, and thus can be loaded rapidly into the cache. The example main memory shown in Fig. 88.3 can provide an entire cache line with two parallel memory accesses.

Conventional caches traditionally could not accept requests while they were servicing a miss request. In other words, they *locked up* or *blocked* when servicing a miss. The growing penalty for cache misses has made it necessary for high-end commodity memory systems to continue to accept (and service) requests from the processor while a miss is being serviced. Some systems are able to service multiple miss requests simultaneously. To allow this mode of operation, the cache design is *lockup-free* or *non-blocking* [Kroft, 1981]. Lockup-free caches have one structure for each simultaneous outstanding miss that they can service. This structure holds the information necessary to correctly return the loaded data to the processor, even if the misses come back in a different order than that in which they were sent.

Two factors drive the existence of multiple levels of cache memory in the memory hierarchy: access times and a limited number of transistors on the CPU chip. Larger banks with greater capacity are slower than smaller banks. If the time needed to access the cache limits the clock frequency of the CPU, then the first-level cache size may need to be constrained. Much of the benefit of a large cache may be obtained by placing a small first-level cache above a larger second-level cache; the first is accessed quickly and the second holds more data close to the processor. Since many modern CPUs have caches on the CPU chip itself, the size of the cache is limited by the CPU silicon real-estate. Some CPU designers have assumed that system designers will add large off-chip caches to the one or two levels of caches on the processor chip. The complexity of this part of the memory hierarchy may continue to grow as main memory access penalties continue to increase.

Caches that appear on the CPU chip are manufactured by the CPU vendor. Off-chip caches, however, are a commodity part sold in large volume. An incomplete list of major cache manufacturers is Hitachi, IBM Micro, Micron, Motorola, NEC, Samsung, SGS-Thomson, Sony, and Toshiba. Although most personal computers and all major workstations now contain caches, very high-end machines (such as multi-million dollar supercomputers) do not usually have caches. These ultra-expensive computers can afford to implement their main memory in a comparatively fast semiconductor technology such as static RAM (SRAM), and can afford so many banks that cacheless bandwidth out of the main memory system is sufficient. Massively parallel processors (MPPs), however, are often constructed out of workstation-like nodes to reduce cost. MPPs therefore contain cache hierarchies similar to those found in the workstations on which the nodes of the MPPs are based.

Cache sizes have been steadily increasing on personal computers and workstations. Intel Pentium-based personal computers come with 8 Kbyte each of instruction and data caches. Two of the Pentium chip sets, manufactured by Intel and OPTi, allow level-two caches ranging from 256 to 512 Kbyte and 64 Kbyte to 2 Mbyte, respectively. The newer Pentium Pro systems also have 8 Kbyte, first-level instruction and data caches, but they also have either a 256 Kbyte or a 512 Kbyte second-level cache on the same module as the processor chip. Higher-end workstations—such as DEC Alpha 21164-based systems—are configured with substantially more cache. The 21164 also has 8 Kbyte, first-level instruction and data caches. Its second-level cache is entirely on-chip, and is 96 Kbyte. The third-level cache is off-chip, and can have a size ranging from 1 to 64 Mbyte.

For all desktop machines, cache sizes are likely to continue to grow—although the rate of growth compared to processor speed increases and main memory size increases is unclear.

88.4 Parallel and Interleaved Memories

Main memories are comprised of a series of semiconductor memory chips. A number of these chips, like caches, form a *bank*. Multiple memory banks can be connected together to form an **interleaved** (or parallel) memory system. Since each bank can service a request, an interleaved memory system with K banks can service K requests simultaneously, increasing the peak bandwidth of the memory system to K times the bandwidth of a single bank. In most interleaved memory systems, the number of banks is a power of two, that is, $K = 2^k$. An n-bit memory word address is broken into two parts: a k-bit bank number and an m-bit address of a word within a bank. Though the k bits used to select a bank number could be any k bits of the n-bit word address, typical interleaved memory systems use the low-order k address bits to select the bank number; the higher order $m = n - k$ bits of the word address are used to access a word in the selected bank. The reason for using the low-order k bits will be discussed shortly. An interleaved memory system which uses the low-order k bits to select the bank is referred to as a *low-order* or a *standard* interleaved memory.

There are two ways of connecting multiple memory banks: *simple interleaving* and *complex interleaving*. Sometimes simple interleaving is also referred to as *interleaving*, and complex interleaving as *banking*.

Figure 88.5 shows the structure of a simple interleaved memory system. m address bits are simultaneously supplied to every memory bank. All banks are also connected to the same read/write control line (not shown in Fig. 88.5). For a read operation, the banks start the read operation and deposit the data in their latches. Data can then be read from the latches, one by one, by setting the switch appropriately. Meanwhile, the banks could be accessed again, to carry out another read or write operation. For a write operation, the latches are loaded, one by one. When all the latches have been written, their contents can be written into the memory banks by supplying m bits of address (they will be written into the same word in each of the different banks). In a simple interleaved memory, all banks are cycled at the same time; each bank starts and completes its individual operations at the same time as every other bank; a new memory cycle can start (for all banks) once the previous cycle is complete. Timing details of the accesses can be found in *The Architecture of Pipelined Computers*, [Kogge, 1981].

One use of a simple interleaved memory system is to back up a cache memory. To do so, the memory must be able to read blocks of contiguous words (a cache block) and supply them to the cache. If the low-order k bits of the address are used to select the bank number, then consecutive words of the block reside in different banks, and they can all be read in parallel, and supplied to the cache one by one. If some other address bits are used for bank selection, then multiple words from the block might fall in the same memory bank, requiring multiple accesses to the same bank to fetch the block.

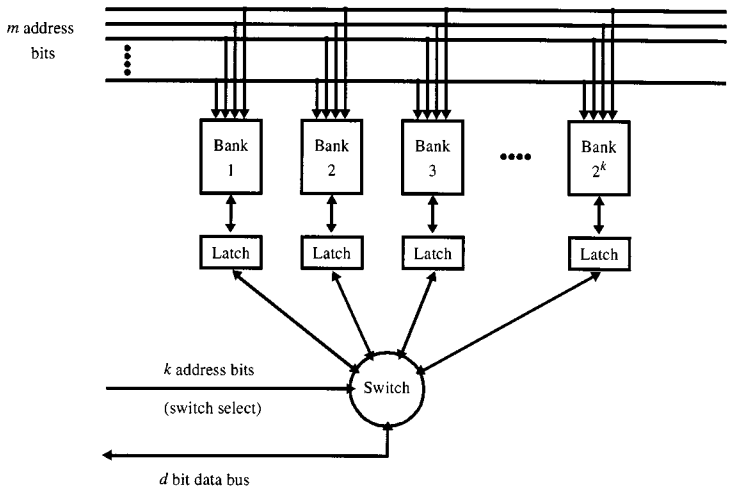


FIGURE 88.5 A simple interleaved memory system. (*Source:* Adapted from P. M. Kogge, *The Architecture of Pipelined Computers,* 1st ed., New York: McGraw-Hill, 1981, p. 41.)

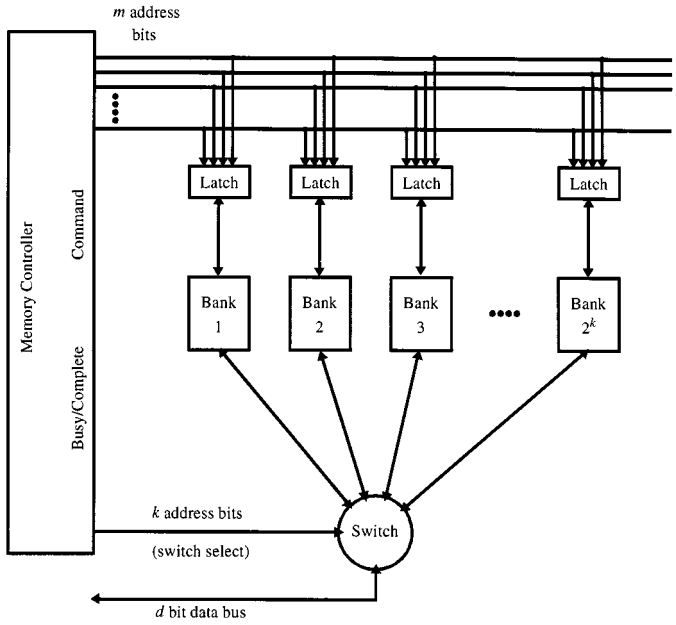


FIGURE 88.6 A complex interleaved memory system. (*Source:* Adapted from P. M. Kogge, *The Architecture of Pipelined Computers,* 1st ed., New York: McGraw-Hill, 1981, p. 42.)

Figure 88.6 shows the structure of a complex interleaved memory system. In such a system, each bank is set up to operate on its own, independent of the operation of the other banks. For example, bank 1 could carry out a read operation on a particular memory address, and bank 2 carries out a write operation on a completely unrelated memory address. (Contrast this with the operation in a simple interleaved memory where all banks are carrying out the same operation, read or write, and the locations accessed within each bank represent a contiguous block of memory.) Complex interleaving is accomplished by providing an address latch and a read/write command line for each bank. The *memory controller* handles the overall operation of the interleaved memory. The processing unit submits the memory request to the memory controller, which determines which bank needs to be accessed. The controller then determines if the bank is busy (by monitoring a busy line for

each bank). The controller holds the request if the bank is busy, submitting it later when the bank is available to accept the request. When the bank responds to a read request, the switch is set by the controller to accept the request from the bank and forward it to the processing unit. Details of the timing of accesses can be found in *The Architecture of Pipelined Computers* [Kogge, 1981].

A typical use of a complex interleaved memory system is in a *vector processor*. In a vector processor, the processing units operate on a vector, for example a portion of a row or a column of a matrix. If consecutive elements of a vector are present in different memory banks, then the memory system can sustain a bandwidth of one element per clock cycle. By arranging the data suitably in memory and using standard interleaving (for example, storing the matrix in row-major order will place consecutive elements in consecutive memory banks), the vector can be accessed at the rate of one element per clock cycle as long as the number of banks is greater than the bank busy time.

Memory systems that are built for current machines vary widely, the price and purpose of the machine being the main determinant of the memory system design. The actual memory chips, which are the components of the memory systems, are generally commodity parts built by a number of manufacturers. The major commodity DRAM manufacturers include (but certainly are not limited to) Hitachi, Fujitsu, LG Semicon, NEC, Oki, Samsung, Texas Instruments, and Toshiba.

The low-end of the price/performance spectrum is the personal computer, presently typified by Intel Pentium systems. Three of the manufacturers of Pentium-compatible chip sets (which include the memory controllers) are Intel, OPTi, and VLSI Technologies. Their controllers provide for memory systems that are simply interleaved, all with minimum bank depths of 256 Kbyte, and maximum system sizes of 192 Mbyte, 128 Mbyte, and 1 Gbyte, respectively.

Both higher-end personal computers and workstations tend to have more main memory than the lower-end systems, although they usually have similar upper limits. Two examples of such systems are workstations built with the DEC Alpha 21164, and servers built with the Intel Pentium Pro. The Alpha systems, using the 21171 chip set, are limited to 128 Mbyte of main memory using 16 Mbit DRAMs, although they will be expandable to 512 Mbyte when 64-Mbit DRAMs are available. Their memory systems are 8-way simply interleaved, providing 128 bits per DRAM access. The Pentium Pro systems support slightly different features. The 82450KX and 82450GX chip sets include memory controllers that allow reads to bypass writes (performing writes when the memory banks are idle). These controllers can also buffer eight outstanding requests simultaneously. The 82450KX controller permits 1- or 2-way interleaving, and up to 256 Mbyte of memory when 16-Mbit DRAMs are used. The 82450GX chip set is more aggressive, allowing up to four separate (complex-interleaved) memory controllers, each of which can be up to 4-way interleaved and have up to 1 Gbyte of memory (again with 16 Mbit DRAMs).

Interleaved memory systems found in high-end *vector supercomputers* are slight variants on the basic complex interleaved memory system of Fig. 88.6. Such memory systems may have hundreds of banks, with multiple memory controllers that allow multiple independent memory requests to be made every clock cycle. Two examples of modern vector supercomputers are the Cray T-90 series and the NEC SX series. The Cray T-90 models come with varying numbers of processors—up to 32 in the largest configuration. Each of these processors is coupled with 256 Mbyte of memory, split into 16 banks of 16 Mbyte each. The T-90 has complex interleaving among banks. the largest configuration (the T-932) has 32 processors, for a total of 512 banks and 8 Gbyte of main memory. The T-932 can provide a peak of 800 Gbyte/s bandwidth out of its memory system. NEC's SX-4 product line, their most recent vector supercomputer series, has numerous models. Their largest single-node model (with one processor per node) contains 32 processors, with a maximum of 8 Gbyte of memory, and a peak bandwidth of 512 Gbyte/s out of main memory. Although the sizes of the memory systems are vastly different between workstations and vector machines, the techniques that both use to increase total bandwidth and minimize bank conflicts are similar.

88.5 Virtual Memory

Cache memory contains portions of the main memory in dynamically allocated cache lines. Since the data portion of the cache memory is itself a conventional memory, each line present in the cache has two addresses associated with it: its main memory address and its cache address. Thus, the main memory address of a word

can be divorced from a particular storage location and abstractly thought of as an element in the address space. The use of a two-level hierarchy—consisting of main memory and a slower, larger disk storage device—evolved by making a clear distinction between the address space and the locations in memory. An address generated during the execution of a program is known as a *virtual address*, which must be translated to a *physical address* before it can be accessed in main memory. The total address space is simply an abstraction.

A **virtual memory** address is mapped to a physical address, which indicates the location in main memory where the data actually reside [Denning, 1970]. The mapping is maintained through a structure called the *page table*, which is maintained in software by the operating system. Like the tag memory of a cache memory, the page table is accessed through a virtual address to determine the physical (main memory) address of the entry. Unlike the tag memory, however, the table is usually sorted by virtual addresses, making the translation process a simple matter of an extra memory access to determine the real physical address of the desired item. A system maintaining the page table in the way analogous to a cache tag memory is said to have *inverted page tables*. In addition to the real address mapped to a virtual page, and an indication of whether the page is present at all, a page table entry often contains other information. For example, the page table may contain the location on the disk where each block of data is stored when not present in main memory.

The virtual memory can be thought of as a collection of blocks. These blocks are often aligned and of fixed size, in which case they are known as *pages*. Pages are the unit of transfer between the disk and main memory, and are generally larger than a cache line—usually thousands of bytes. A typical page size for machines in 1997 is 4 Kbyte. A page's virtual address can be broken into two parts, a virtual page number and an offset. The page number specifies the page to be accessed, and the page offset indicates the distance from the beginning of the page to the indicated address.

A physical address can also be broken into two parts, a physical page number (also called a *page frame* number) and an offset. This mapping is done at the level of pages, so the page table can be indexed by means of the virtual page number. The page frame number is contained in the page table and is read out during the translation, along with other information about the page. In most implementations the page offset is the same for a virtual address and the physical address to which it is mapped.

The virtual memory hierarchy is different than the cache/main memory hierarchy in a number of respects, resulting primarily from the fact that there is a much greater difference in latency between accesses to the disk and the main memory. While a typical latency ratio for cache and main memory is one order of magnitude (main memory has a latency ten times larger than the cache), the latency ratio between disk and main memory is often four orders of magnitude or more. This large ratio exists because the disk is a mechanical device—with a latency partially determined by velocity and inertia—whereas main memory is limited only by electronic and energy constraints. Because of the much larger penalty for a page miss, many design decisions are affected by the need to minimize the frequency of misses. When a miss does occur, the processor could be idle for a period during which it could execute tens of thousands of instructions. Rather than stall during this time, as may occur upon a cache miss, the processor invokes the operating system and may switch to a different task. Because the operating system is being invoked anyway, it is convenient to rely on the operating system to set up and maintain the page table, unlike cache memory, where it is done entirely in hardware. The fact that this accounting occurs in the operating system enables the system to use virtual memory to enforce protection on the memory. This ensures that no program can corrupt the data in memory that belong to any other program.

Hardware support provided for a virtual memory system generally includes the ability to translate the virtual addresses provided by the processor into the physical addresses needed to access main memory. Thus, only on a virtual address miss is the operating system invoked. An important aspect of a computer implementing virtual memory, however, is the necessity of freezing the processor at the point where a miss occurs, servicing the page table fault, and later returning to continue the execution as if no page fault had occurred. This requirement means either that it must be possible to halt execution at any point—including possibly in the middle of a complex instruction—or it must be possible to guarantee that all memory accesses will be to pages resident in main memory.

As described above, virtual memory requires two memory accesses to fetch a single entry from memory, one into the page table to map the virtual address into the physical address, and the second to fetch the actual data. This process can be sped up in a variety of ways. First, a special-purpose cache memory to store the active portion of the page table can be used to speed up the first access. This special-purpose cache is usually called

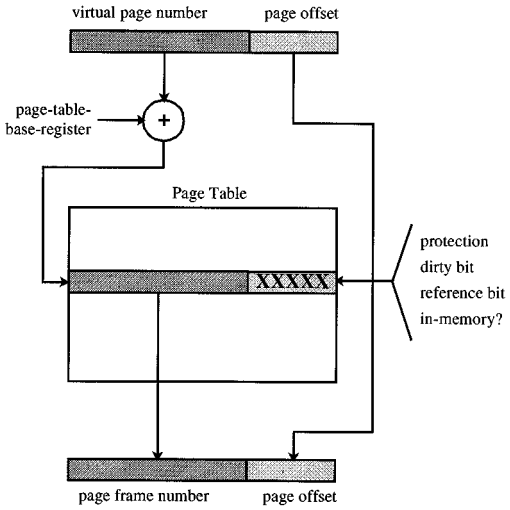


FIGURE 88.7 Virtual-to-real address translation.

a *translation lookaside buffer* (TLB). Second, if the system also employs a cache memory, it may be possible to overlap the access of the cache memory with the access to the TLB, ideally allowing the requested item to be accessed in a single cache access time. The two accesses can be fully overlapped if the virtual address supplies sufficient information to fetch the data from the cache before the virtual-to-physical address translation has been accomplished. This is true for an M-way set-associative cache of capacity C if the following relationship holds:

$$Page\_size \geq \frac{C}{M} \tag{88.7}$$

For such a cache, the index into the cache can be determined strictly from the page offset. Since the virtual page offset is identical to the physical page offset, no translation is necessary, and the cache can be accessed concurrently with the TLB. The physical address must be obtained before the tag can be compared.

An alternative method applicable to a system containing both virtual memory and a cache is to store the virtual address in the tag memory instead of the physical address. This technique introduces consistency problems in virtual memory systems that permit more than a single address space or allow a single physical page to be mapped to more than a single virtual page. This problem is known as the *aliasing* problem.

88.6 Research Issues

Research is occurring on all levels of the memory hierarchy. At the register level, researchers are exploring techniques to provide more registers than are architecturally visible to the compiler. A large volume of work exists (and is occurring) for cache optimizations and alternate cache organizations. For instance, modern processors now commonly split the top level of the cache into separate physical caches, one for instructions (code) and one for program data. Due to the increasing cost of cache misses (in terms of processor cycles), some research trades-off increase the complexity of the cache for reducing the miss rate. Two examples of cache research from opposite ends of the hardware/software spectrum are *blocking* [Lam, 1901] and *skewed-associative caches* [Seznec, 1993]. Blocking is a software technique in which the programmer or compiler reorganizes algorithms to work on subsets of data that are smaller than the cache, instead of streaming entire large data structures repeatedly through the cache. This reorganization greatly improves temporal locality. The skewed-associative cache is one example of a host of hardware techniques that map blocks into the cache differently, with the goal of reducing misses from conflicts within a set. In skewed-associative caches, either one of two

hashing functions may determine where a block should be placed in the cache, as opposed to just the one hashing function (low-order index bits) that traditional caches use. An important cache-related research topic is *prefetching* [Mowry, 1992], in which the processor issues requests for data well before the data are actually needed. Speculative prefetching is also a current research topic. In speculative prefetching, prefetches are issued based on guesses as to which data will be needed soon. Other cache-related research examines placing special structures in parallel with the cache, trying to optimize for workloads that do not lend themselves well to caches. Stream buffers [Jouppi, 1990] are one such example. A stream buffer automatically detects when a linear access through a data structure is occurring. The stream buffer issues multiple sequential prefetches upon detection of a linear array access.

Much of the ongoing research on main memory involves improving the bandwidth from the memory system without greatly increasing the number of banks. Multiple banks are expensive, particularly with the large and growing capacity of modern DRAM chips. Rambus [Rambus Inc., 1992] and Ramlink [IEEE Computer Society, 1993] are two such examples.

Research issues associated with improving the performance of the virtual memory system fall under the domain of operating system research. One proposed strategy for reducing page faults allows each running program to specify its own page replacement algorithm, enabling each program to optimize the choice of page replacements based on its reference pattern [Engler et al., 1995]. Other recent research focuses on improving the performance of the TLB. Two techniques for doing this are the use of a two-level TLB (the motivation is similar to that for a two-level cache), and the use of superpages [Talluri, 1994]. With superpages, each TLB entry may represent a mapping for more than one consecutive page, thus increasing the total address range that a fixed number of TLB entries may cover.

Summary

A computer's memory system is the repository for all the information that the CPU uses and produces. A perfect memory system is one that can immediately supply any datum that the CPU requests. This ideal memory is not implementable, however, as the three factors of memory capacity, speed, and cost are directly in opposition.

By staging smaller, faster memories in front of larger, slower, and cheaper memories, the performance of the memory system may approach that of a perfect memory system—at a reasonable cost. The memory hierarchies of modern general-purpose computers generally contain registers at the top, followed by one or more levels of cache memory, main memory, and virtual memory on a magnetic or optical disk.

Performance of a memory system is measured in terms of latency and bandwidth. The latency of a memory request is how long it takes the memory system to produce the result of the request. The bandwidth of a memory system is the rate at which the memory system can accept requests and produce results. The memory hierarchy improves average latency by quickly returning results that are found in the higher levels of the hierarchy. The memory hierarchy generally reduces bandwidth requirements by intercepting a fraction of the memory requests at higher levels of the hierarchy. Some machines—such as high-performance vector machines—may have fewer levels in the hierarchy, increasing memory cost for better predictability and performance. Some of these machines contain no caches at all, relying on large arrays of main memory banks to supply very high bandwidth, with pipelined accesses to operands that mitigate the adverse performance impact of long latencies.

Cache memories are a general solution for improving the performance of a memory system. Although caches are smaller than typical main memory sizes, they ideally contain the most frequently accessed portions of main memory. By keeping the most heavily used data near the CPU, caches can service a large fraction of the requests without accessing main memory (the fraction serviced is called the hit rate). Caches assume locality of reference to work well transparently—they assume that accessed memory words will be accessed again quickly (temporal locality), and that memory words adjacent to an accessed word will be accessed soon after the access in question (spatial locality). When the CPU issues a request for a datum not in the cache (a cache miss), the cache loads that datum and some number of adjacent data (a cache block) into itself from main memory.

To reduce cache misses, some caches are associative—a cache may place a given block in one of several places, collectively called a set. This set is content-addressable; a block may or may not be accessed based on an address tag, one of which is coupled with each block. When a new block is brought into a set and the set is full, the cache's replacement policy dictates which of the old blocks should be removed from the cache to make room

for the new block. Most caches use an approximation of least-recently-used (LRU) replacement, in which the block last accessed farthest in the past is the one that the cache replaces.

Main memory, or backing store, consists of banks of dense semiconductor memory. Since each memory chip has a small off-chip bandwidth, rows of these chips are placed together to form a bank, and multiple banks are used to increase the total bandwidth from main memory. When a bank is accessed, it remains busy for a period of time, during which the processor may make no other accesses to that bank. By increasing the number of interleaved (parallel) banks, the chance that the processor issues two conflicting requests to the same bank is reduced.

Systems generally require a greater number of memory locations than are available in the main memory (i.e., a larger address space). The entire address space that the CPU uses is kept on large magnetic or optical disks; this is called the virtual address space, or virtual memory. The most frequently used sections of the virtual memory are kept in main memory (physical memory), and are moved back and forth in units called pages. The place at which a virtual address lies in main memory is called its physical address. Since a much larger address space (virtual memory) is mapped onto a much smaller one (physical memory), the CPU must translate the memory addresses issued by a program (virtual addresses) into their corresponding locations in physical memory (physical addresses). This mapping is maintained in a memory structure called the page table. When the CPU attempts to access a virtual address that does not have a corresponding entry in physical memory, a page fault occurs. Since a page fault requires an access to a slow mechanical storage device (such as a disk), the CPU usually switches to a different task while the needed page is read from the disk.

Every memory request issued by the CPU requires an address translation, which in turn requires an access to the page table stored in memory. A translation lookaside buffer (TLB) is used to reduce the number of page table lookups. The most frequent virtual-to-physical mappings are kept in the TLB, which is a small associative memory tightly coupled with the CPU. If the needed mapping is found in the TLB, the translation is performed quickly and no access to the page table needs to be made. Virtual memory allows systems to run larger or more programs than are able to fit in main memory, enhancing the capabilities of the system.

Defining Terms

Bandwidth: The rate at which the memory system can service requests.

Cache memory: A small, fast, redundant memory used to store the most frequently accessed parts of the main memory.

Interleaving: Technique for connecting multiple memory modules together in order to improve the bandwidth of the memory system.

Latency: The time between the initiation of a memory request and its completion.

Memory hierarchy: Successive levels of different types of memory, which attempt to approximate a single large, fast, and cheap memory structure.

Virtual memory: A memory space implemented by storing the most frequently accessed parts in main memory and less frequently accessed parts on disk.

Related Topics

80.1 Integrated Circuits (RAM, ROM) • 80.2 Basic Disk System Architectures

References

P. J. Denning, "Virtual memory," *Computing Surveys*, vol. 2, no. 3, pp. 153–170, Sept. 1970.

D. R. Engler, M. F. Kaashoek, J. O'Toole, Jr., "Exokernel: An operating system architecture for application-level resource management," *Proc. 15th Symposium on Operating Systems Principles,* pp. 251–266, 1995.

J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, San Mateo, Calif.: Morgan Kaufmann Publishers, 1990.

M. D. Hill, "A case for direct-mapped caches," *IEEE Computer*, 21(12), 1988.

IEEE Computer Society, *IEEE Standard for High-Bandwidth Memory Interface Based on SCI Signaling Technology (RamLink),* Draft 1.00 IEEE P1596.4-199X, 1993.

N. Jouppi, "Improving direct-mapped cache performance by the addition of a small fully-associative cache and prefetch buffers," *Proc. 17th Annual International Symposium on Computer Architecture,* pp. 364–373, 1990.
P. M. Kogge, *The Architecture of Pipelined Computers*, New York: McGraw-Hill, 1981.
D. Kroft, "Lockup-Free Instruction Fetch/Prefetch Cache Organization," *Proc. 8th Annual International Symposium on Computer Architecture,* pp. 81–87, 1981.
M. S. Lam, E. E. Rothberg, and M. E. Wolf, "The cache performance and optimizations of blocked algorithms," *Proc. 4th Annual Symposium on Architectural Support for Programming Languages and Operating Systems,* pp. 63–74, 1991.
T. C. Mowry, M. S. Lam, and A. Gupta, "Design and evaluation of a compiler algorithm for prefetching," *Proc. 5th Annual Symposium on Architectural Support for Programming Languages and Operating Systems,* pp. 62–73, 1992.
Rambus, Inc., *Rambus Architectural Overview,* Mountain View, Calif.: Rambus, Inc., 1992.
A. Seznec, "A case for two-way skewed-associative caches," *Proc. 20th International Symposium on Computer Architecture,* pp. 169–178, 1993.
A. J. Smith, "Bibliography and readings on CPU cache memories and related topics," *ACM SIGARCH Computer Architecture News*, 14(1), 22–42, 1986.
A. J. Smith, "Second bibliography on cache memories," in *ACM SIGARCH Computer Architecture News*, 19(4), 154–182, June 1991.
M. Talluri and M. D. Hill, "Surpassing the TLB performance of superpages with less operating system support," *Proc. Sixth International Symposium on Architectural Support for Programming Languages and Operating Systems,* pp. 171–182, 1994.

Further Information

Some general information on the design of memory systems is available in *High-Speed Memory Systems* by A. V. Pohm and O. P. Agrawal.

Computer Architecture: A Quantitative Approach by John Hennessy and David Patterson contains a detailed discussion on the interaction between memory systems and computer architecture.

For information on memory system research, the recent proceedings of the *International Symposium on Computer Architecture* contain annual research papers in computer architecture, many of which focus on the memory system. To obtain copies, contact the IEEE Computer Society Press, 10662 Los Vaqueros Circle, P.O. Box 3014, Los Alamitos, CA 90720-1264.

Sherr, S., Durbeck, R.C., Suryn, W., Veillette, M. "Input and Output"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

89

Input and Output

Solomon Sherr
Westland Electronics

Robert C. Durbeck
IBM Corporation

Witold Suryn
Gemplus

Michel Veillette
Gemplus

89.1 Input Devices
Keyboards • Light Pen • Data Tablet (Graphics, Digitizer) • Mouse • Trackball • Joystick • Touch Input • Scanners • Voice • Summary • Advantages and Disadvantages

89.2 Computer Output Printer Technologies
Classification of Printer Technologies • Page Printer Technologies • Serial Nonimpact Printer Technologies • Impact Printer Technologies

89.3 Smart Cards
Hardware Architecture • Contact ICC, Contactless ICC • Operating Systems • Standards • Applications • Readers • Card-to-System Solutions • Trends

89.1 Input Devices[1]

Solomon Sherr

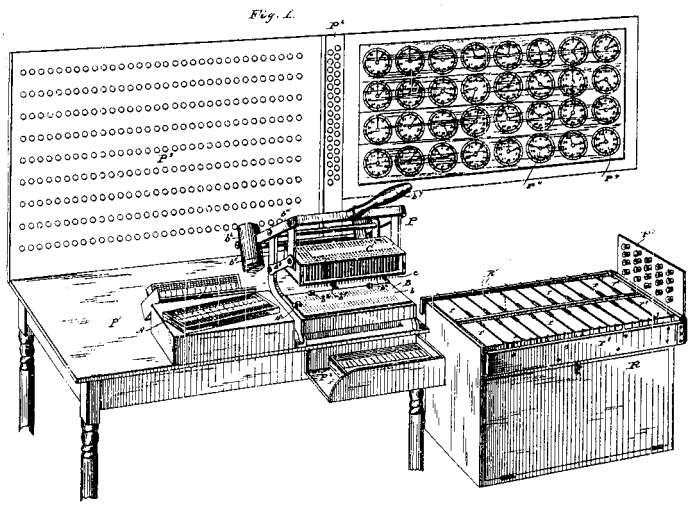
Input devices are those portions of computer, data processing, and information systems that perform the essential function of providing some means for entering commands and data into the system. Therefore, input devices are found in all such systems, but are treated here as a separate equipment group, independent of the total system configuration. However, the place of input devices in a representative computer system may be clarified by reference to Fig. 89.1(a), which shows the interface of the main input device categories in relation to the portions of the generalized system that accept the inputs. The categories and the devices listed in Table 89.1 are the subject of this section.

Keyboards

Keyboards are essentially electromechanical devices, and are still ubiquitous, in spite of the inroads of other input devices. The primary type of keyboard in use as an input device is the alphanumeric (A/N) form, well known in its typewriter application, but with various additions and expansions consisting of numeric and special function keys. This type of keyboard is shown in Fig. 89.1(b) with a standard QWERTY format, so named because of the layout of the top left alpha keys, for the A/N portion, a separate numeric set to the right, and a group of function keys at the top. Other layouts for the A/N portion have been proposed and at least one (Dvorak) accepted by the American National Standards Institute (ANSI), but it has not received much use in spite of its advantages in increased efficiency. At present, the overwhelming majority of system keyboards still use the QWERTY layout, and it is the only one considered here.

As illustrated in Fig. 89.1, a keyboard consists of a number of keyswitches whose exact structure is of prime importance in keyboard design. The relevant characteristics of keyswitch operation are life, actuation force, travel distance, and feedback. Accepted values are shown in Table 89.2 for different keyswitch designs. The elastomer type is preferred to a limited extent over the other two when the electronic audio feedback is included. This indicates that some type of audio feedback is desirable. One form of keyswitch design using an elastomer

[1]The material contained in this section is a shortened version of that which appears in *Electronic Displays,* 2nd ed., by Sol Sherr, Chapter 6, Section 6.1, 1993, published by John Wiley & Sons, Inc., and is reprinted here by permission.



Art of Compiling Statistics

Herman Hollerith
Patented January 8, 1889
#395,781

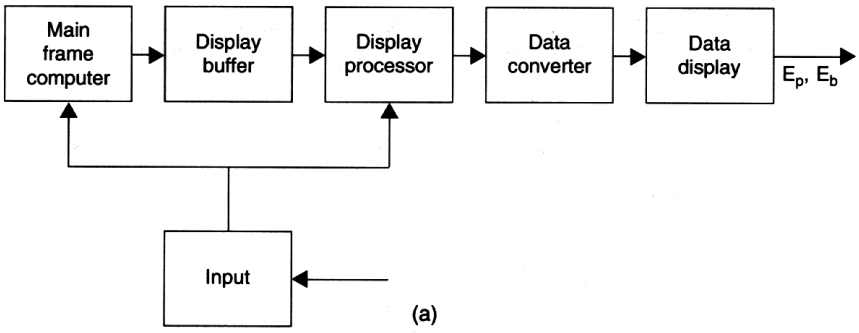
An excerpt from Herman Hollerith's patent application:

Having thus described my invention, what I claim as new is (1) The improvement in the art of compiling statistics, which consists in first preparing a series of separate record-cards, each card representing an individual or subject; second, applying to each card at predetermined intervals circuit-controlling index points arranged according to a fixed plan of distribution, to represent each item or characteristic of the individual or subject, and third, applying said separate record-cards successively to circuit-controlling devices acted upon by the index-points to designate each statistical item represented by one or more of said index-points, substantially as described.

This patent, along with two others, describes a system for tabulating statistical items represented by holes punched in cards. The 1890 U.S. census was completed $5 million under budget and two years ahead of schedule because of Hollerith's system. The punch card system with encoded holes (the code for representing alphanumeric characters with holes was named after Hollerith) was widely used for sorting, counting, and tabulating even into the 1980s. Hollerith's original Tabulating Machine Company was the forerunner to the computer giant, IBM.

TABLE 89.1 List of Input Devices

Category	Designation	Operation Mode
Keyboard	Alphanumeric	Electromechanical
Keyboard	Function	Electromechanical
Pointing	Light pen	Screen pointing
Pointing	Touchscreen	Screen pointing
Pointing	Pen tablet	Tablet pointing
Coordinates	Digitizer	*X-Y* conversion
Coordinates	Data tablet	*X-Y* location
Cursor	Mouse	Movement
Cursor	Trackball	Movement
Cursor	Joystick	Movement
Image	Scanner	Conversion
Verbal	Voice	Conversion



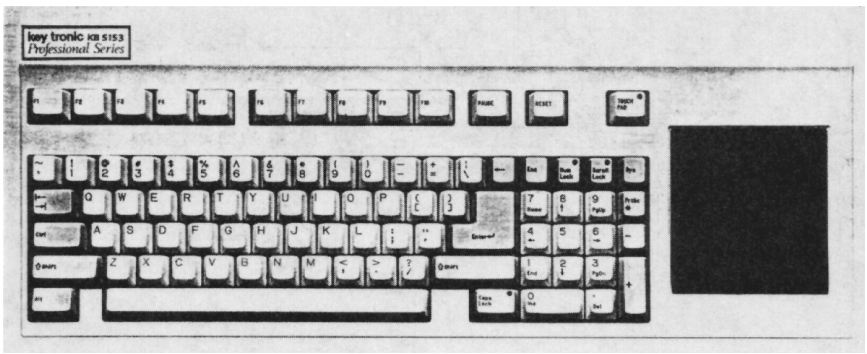


FIGURE 89.1 (a) Generalized display-system block diagram. (Source: After S. Sherr, Electronic Displays, New York: John Wiley & Sons, 1979.With permission.) (b) Alphanumeric keyboard. (Courtesy of Key tronic.)

TABLE 89.2 Keyboard Parameter Values

Parameter	Snap Switch	Elastomer	Foam Pad
Key travel	3.8 mm	3.2 mm	3.8 mm
Force	>60 gm	>50 gm	>30 gm
Life	10 million cycles	10 million cycles	10 million cycles
Feedback	Audio mechanical	Audio electric	Tactile

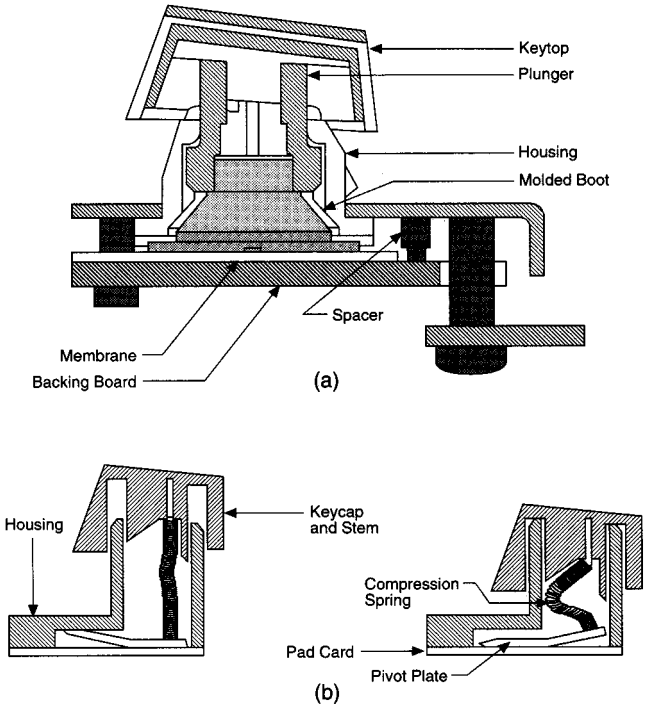


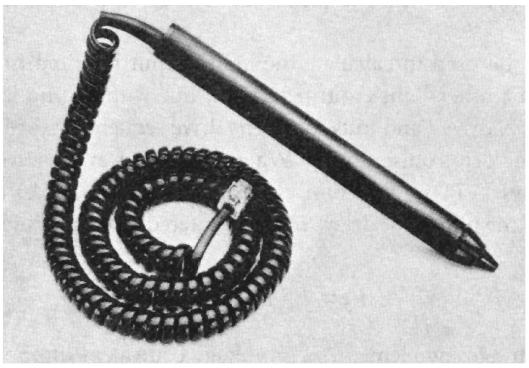
FIGURE 89.2 (a) Elastomer-type keyswitch. (b) Snap switch. (*Source:* After H. Brunner et al., "Effects of key action design on keyboard preference and throughput performance," Micro Switch. With permission.)

or "molded boot" is shown in Fig. 89.2(a), in which the boot consists of two collapsible domes. In this design, the internal movement of the keyswitch is completely silent so that some source of sound must be added to achieve the desired audible feedback. The snap switch design shown in Fig. 89.2(b) has built-in sound and achieves a small reduction in insertion errors over the elastomer design with audio feedback.

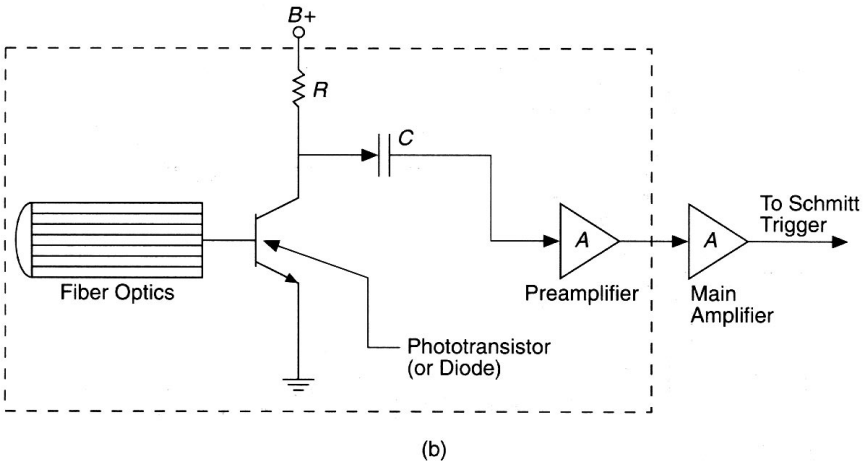
The life requirement is estimated on the basis of workstation users operating at approximately half the accepted rate of 20 million actuations per key used for electronic typewriters. The actual layout and content of the keyboard may vary greatly, ranging from the standard typewriter arrangement, through different combinations of alphanumerics and symbols, to the special-function keyboards that contain legends and symbols specific to the particular application. However, the outputs of each type are the same in that they must contain coded signals that relate the action to be performed by the information system to that defined by the key being operated, in terms of the input code of the system. Thus, many of the keyboards output the ASCII code, and the system is usually designed so that it can accept this type of standard code. Incidentally, ASCII, the acronym for American Standard Code for Information Interchange, is the standard means for encoding alphanumerics and a group of selected symbols for transmission to a display system, among others. It is the standard code used in the United States and most other English-speaking countries and corresponds to the ISO seven-bit code. The seven-bit ASCII is usually used, and it should be noted that for serial data transmission an eighth bit is added for parity. Various keyboard arrangements are possible, and many variants are found in particular applications. The means for coding the key operation may be through magnetic reed relays, solid-state circuits, or more exotic devices such as Hall effect sensors. These device characteristics are only incidental to the operation and beyond the scope of this chapter. Similarly, we do not discuss the human-factors aspects of keyboard design, not because they are not important, but because, apart from the visual considerations, the other factors have to do with tactile and physical features best left to others.

Light Pen

The light pen initially was a very popular means for accomplishing manual input to the random deflection information display systems, but fell out of favor when raster systems became more popular due to its being



(a)



(b)

FIGURE 89.3 (a) Light pen. (Courtesy of FTG Data Systems.) (b) Light pen schematic. (*Source:* After S. Sherr, *Electronic Displays*, New York: John Wiley & Sons, 1979, p. 388. With permission.)

somewhat difficult to use with raster systems. This device goes by a misleading name, as it does not emit light and is not a pen other than being somewhat similar to one in its physical appearance, as shown in Fig. 89.3(a). However, when we consider its functional characteristics, the validity of the term becomes apparent, as it is used to cause the electron beam to "write" patterns on the cathode ray tube (CRT) that are defined by the motion of the light pen on the CRT faceplate.

The light pen operates by sensing the existence or nonexistence of a pulse of light at the point on the screen of the CRT or surface of any other light-emitting device where the point of the pen is placed. This is accomplished by means of the circuit shown in Fig. 89.3(b), where the light pulse is collected and transmitted through the fiber optics to a light-sensitive device that converts the light pulse into an electrical pulse which is shaped by some form of electronics (of which a Schmitt trigger is one example). We need not concern ourselves with the exact form of the electronics except to note that this pulse is then sent to the computer, as shown in Fig. 89.4, and provides a complete, closed-loop system. As the electronic pulse occurs at the time when the light pulse passes under the light pen, the computer is informed of the location at which the designated operation is to be performed and may proceed accordingly. Thus, the light pen is a pointing device that designates a point on the display screen and can be used as an input device. Various light pen programs have been written to expand the capabilities of the original one, and it should be noted that the light pen is coming back into favor as improvements in accuracy, ease of operation, and reliability occur.

There are two characteristics of light pen operation that affect the capabilities of this input device. The first is the sensitivity, given by

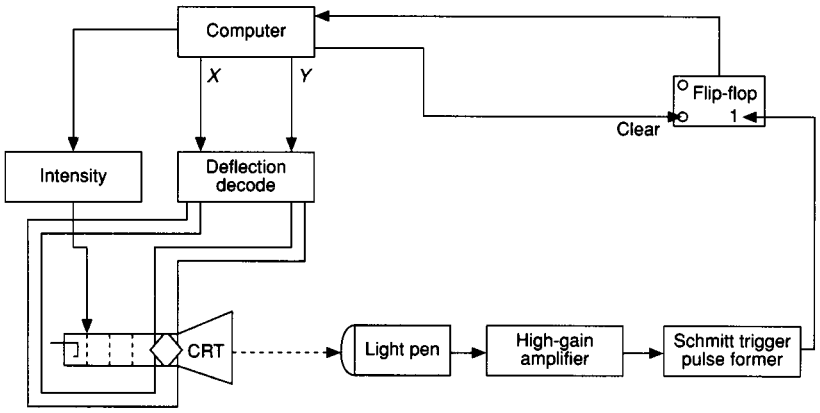


FIGURE 89.4 Block diagram of light pen computer system. (*Source:* S. Sherr, *Electronic Displays*, New York: John Wiley & Sons, 1979, p. 389. With permission.)

$$S = E_L \mu_p A_p A_m \mu_s \mu_f t_L \tag{89.1}$$

where E_L = illuminance at photodetector, μ_p = photodetector sensitivity, A_p = preamplifier gain, A_m = main amplifier gain, μ_s = Schmitt trigger sensitivity, μ_f = flip-flop sensitivity, and t_L = optical loss.

Equation (89.1) may be used to calculate the light output required from the display surface, which may be a CRT or other light-emitting device, but with the limitation that most of the flat panel units are matrix driven and must track the drive sequence in order to know the location of the light pen from the drive pulse timing. When phosphors are involved as for the CRT, vacuum fluorescent displays (VFDs), thin-film electroluminescent (TFEL) units, and color liquid crystal displays (LCDs), the phosphor delays must be entered into the timing, and the total delay is given by

$$E_o = E_i(1 - e - t/\tau) \tag{89.2}$$

where E_o = voltage at triggering element, E_i = voltage equivalent of phosphor light output, t = time, and τ = sum of all delays.

These delays set limits to the positional accuracy, as the computer tracking the signal will be in error by this amount. Other inaccuracies are due to the dimensions of the optical pickup surface, all of which somewhat negate the simplicity of operation. The result is the parameter values shown in Table 89.3.

TABLE 89.3 Light Pen Data

Field of View	Response Time	Sensitivity
0.02–0.08 in.	120–150 ns	0.02–0.04 ft.L

Data Tablet (Graphics, Digitizer)

A very convenient means for data entry, retaining some of the ease of operation of the light pen but with much better accuracy, are the various forms of data tablets available. These tablets differ from the light pen in another significant way in that they do not require a moving spot of light to detect the location of the beam or direct it to a new location. This need for a moving light spot made the light pen difficult to use with the data tablets initially designed to overcome this limitation while still using a device with a pen-like input. The first successful example was the Rand tablet, a digital device that used an *X*–*Y* assembly from which a wand placed above some point on the *X*–*Y* wire matrix could pick up pulse generator output that fed *X* and *Y* electrical pulses into the matrix. By determining the number of pulses in a time period, the location of the wand is established. Another similar device used magnetostrictive rather than electrical signals to accomplish the same result, and this location is converted into display coordinates used to position a cursor on the CRT screen. The cursor may then be

used as a visual feedback element so that the operator can correct the position of the wand until the cursor is properly placed. At this time the information from the tablet may also be transferred to either the host computer or the resident desktop or portable computer, as desired. Since the cursor is not used to signal its position to a pickup device, as is the case with the light pen, it may be used with any type of display system, including the non-light-emitting flat panel displays. Another advantage of the tablet is that it may be used to position cursors in the blank areas of the display, where no light pulses are available unless they are specially generated by the light pen.

There have been numerous improvements and new developments using a variety of technologies that include magnetostrictive, electromagnetic, electrostatic or capacitive, scanned *X–Y* grid, resistive, and sonic. Of these, electromagnetic tablets dominate the digitizer market, and sonic is of interest because it does not require a tablet, but most of the other technologies are essentially restricted to touch input devices covered later. As noted previously, electromagnetic is the most popular technology for high-performance digitizer tablets. Operation is based on transformer principles, whereby a conductor carrying ac creates a magnetic field around it that induces a current in a second conductor. The digitizer tablet uses the amplitude and phase of the induced current to determine digitizing data. The tablet contains an *X–Y* pattern of conductors beneath its surface, in a manner similar to the Rand Tablet, but instead of counting pulses in a time period a circular conductor is used as the pick-up element for the induced current. This coil is placed on the tablet surface, and its position is determined by measuring the phase and amplitude of the current in the coil. Its center is interpolated by sweeping through the *X–Y* grid lines and demodulating the signal in the coil to determine the phase reversal point, or by calculating this point using digitized data fed into a microprocessor. The *X–Y* coordinates may be resolved to better than 0.025 mm using either of these two techniques. Figure 89.5(a) is a photograph of a representative digitizer tablet.

Another digitizer technology is the one that uses the measurement of the time required for sound waves to travel from a source to movable microphone pickups.This sonic technology has the advantage that no special digitizing board is required, and either a stylus or a cursor can be used as the digitizer. Two sonic sources are contained in an *L* frame so that both *X* and *Y* coordinates can be determined by calculating the time it takes for the sound wave to reach the microphones contained in the pickup device. This calculation is made on the basis of sound traveling at 345 m/s at 20°C, and the accuracy is dependent on stable ambient conditions. This tends to limit the resolution to about 300 lpi, and the accuracy to ±0.1%. The device may also be implemented with a single sonic source as the digitizing means and a pair of microphones located outside the digitizing area. In this case the location of the transducer is calculated by triangulation and converted into Cartesian coordinates.

Digitizers are used primarily for inputting accurate coordinate data from maps and engineering drawings. Their high accuracy requirements have led to relatively high prices. Alternative means for inputting data are the data and graphics tablets that meet most input requirements at a lower cost and accuracy. The main technology is still electromagnetic, and the units are essentially the same as the digitizers, but with lower accuracies. However, several of the other technologies have also been used to achieve lower costs. Most successful among them are the capacitive and resistive versions, which may also be used as digitizers. The capacitive units, also termed electrostatic, use capacitive coupling where the coupling between the tablet and the cursor or stylus is determined by the capacitance made up of the tablet surface as one plate and the pickup element as the other. In this case, the capacitance is given by

$$C = f(™A/d) \tag{89.3}$$

where C = capacitance, ™ = permittivity of dielectric, A = relative area of two plates, d = distance between plates, and f = proportionality factor.

A scanned grid approach is used to determine the location of the cursor. As in the electromagnetic tablet, an *X–Y* grid of conductors is embedded in the tablet, with semiconductor switches on each line providing contact on a scanned basis. The charge flowing from each capacitance is summed through a summing amplifier as shown in Fig. 89.5(b). The resultant voltage peaks twice, once for the *X* and once for the *Y* lines, as they are scanned. The peak positions are digitized by means of a counter that starts at the beginning of the scan, and runs at some multiple of the scan rate. The digital values represent the coordinates of the cursor location.

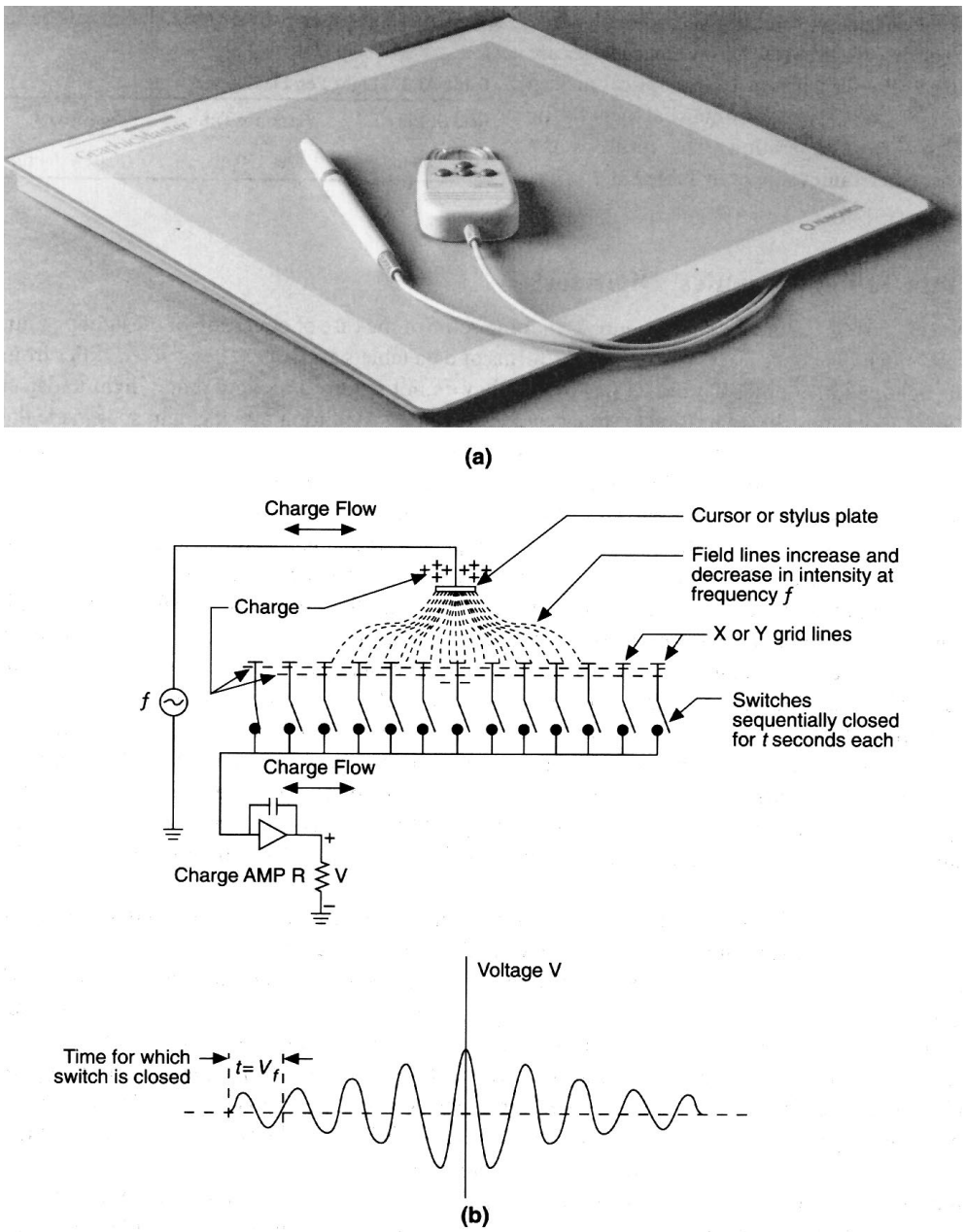


FIGURE 89.5 (a) Digitizer tablet. (Courtesy of Numonics.) (b) Capacitive technology. (*Source:* After T. E. Davies et al., "Digitizers and input tablets," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, p. 186. With permission.)

Mouse

The mouse has gone a long way from its original invention by Engelbart in 1965, through its redesign at Xerox and introduction by Apple as a main input device, and its general acceptance by computer users as an important addition to the group of input devices. It should be noted, in passing, that the mouse is essentially an upside-down trackball, although the latter is now being referred to as an upside-down mouse. However, the trackball came first and is described further in the next section.

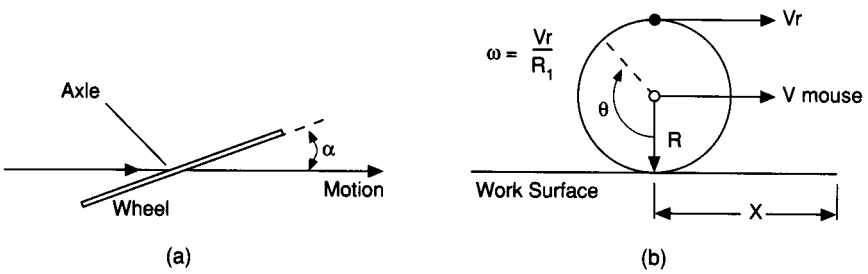


FIGURE 89.6 Wheel showing velocities and slip angle. (*Source:* After C. Goy, "Mice," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, p. 225. With permission.)

Mice contain motion-sensing elements and are operated by moving mechanical or optical elements. One form uses wheels and shafts to drive the sensing elements, as shown schematically in Fig. 89.6. The angular velocity (ω) of the wheel and shaft is given by

$$\omega = V_r/R \quad \text{rad/s} \tag{89.4}$$

where V_r = velocity of wheel and R = wheel radius.

The rotation angle (θ) is given by

$$(\theta) = X/R \quad \text{rad} \tag{89.5}$$

where X = distance moved.

This type of mouse has two sets of wheels and shafts, one for horizontal and the other for vertical motion.

A more popular type of mechanical mouse is the one that uses a ball for the motion sensing device, as shown in Fig. 89.7. Again, the velocity of the ball circumference equals the velocity of the mouse, and the angular velocity is given by

$$\omega = V/R_1 \text{ rad/s} \tag{89.6}$$

where R_1 = shaft radius.

The smaller the shaft the more rapid its rotation for a given mouse velocity. Another form of the ball-and-shaft mouse is the one that uses an optical interrupter, as shown in Fig. 89.8. In this form, the light from the light-emitting diodes (LEDs) is interrupted by the coded disks that are rotated by the shafts, and is then picked up by the phototransistors and converted into the digital signal that represents the disk rotation. An optical interrupter is also used for the optomechanical mouse, and here the interrupter contains a set of slots; as the interrupter rotates quadrature signals are created that correspond to the shaft rotation.

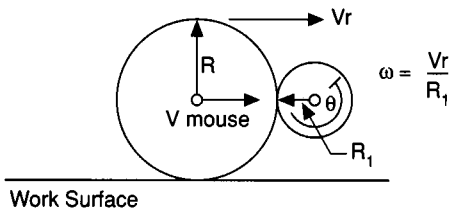


FIGURE 89.7 Ball and shaft. (*Source:* C. Goy, "Mice," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988. With permission.)

In addition to the shaft and optomechanical mice, an early form of mouse used multiturn potentiometers connected to the wheels, and the output voltage that represented the motion varied in direct proportion to the mouse motion. The voltage was then converted by means of an analog-to-digital converter into digital form for input to the computer.

Finally, there are the true optical mice that use a special surface that is printed with a set of geometric shapes, usually a grid of lines or dots, that are illuminated and focused on a light detector. The most common form uses a grid made up of orthogonal lines, with the vertical and horizontal lines printed in different colors. These colors absorb light at different frequencies so that the optical detectors can differentiate between horizontal

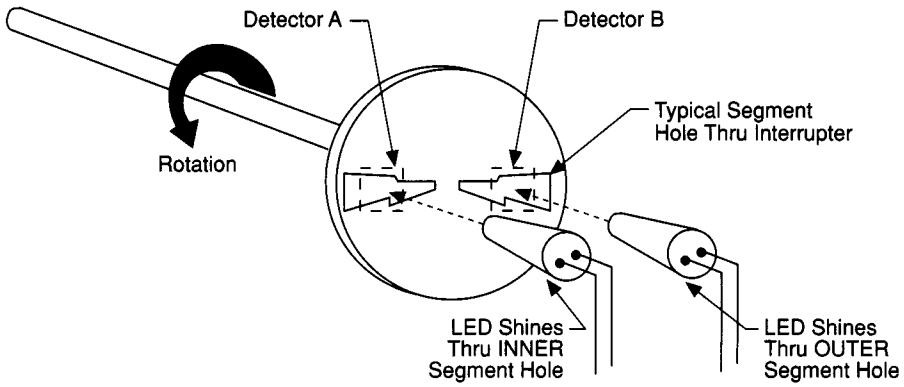


FIGURE 89.8 Optical interrupter. (*Source:* C. Goy, "Mice," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, p. 229. With permission.)

and vertical movement of the mouse. If such a structure is used as the mouse, then the photodetector will pick up a series of light-dark impulses consisting of the reflections from the mirror surface and the grid lines and convert them into square waves. A second LED and photodetector that is mounted orthogonally to the first is used to detect motion in the orthogonal direction, and the combination of the two inks avoids confusion between the two directions of motion. The system then counts the number of impulses created by the mouse motion and converts the result into motion information for the cursor. This type of mouse has the advantage that no mechanical elements are required.

Trackball

As noted previously, the trackball uses technology similar to the mouse, but preceded it as an input device. Thus, the comment that it is an upside-down mouse should be reversed. The movable element is housed in an assembly as is shown in Fig. 89.9, and the assembly remains stationary so that much less desk space is required than for the mouse. In addition, the trackball may be mounted on a keyboard so that very little additional desk space is needed. The movable element can be the same as used in the mouse, and the output can be a set of bits corresponding to the coordinates to which the cursor should be driven, or where the command should be carried out. The output format is essentially equivalent to that used for the mouse, and the same protocols are used.

The typical trackball has an *X* and *Y* optical encoder that generates a pulse for each 0.76 mm of incremental motion of the ball. This means that the pulse train may range from 10 to 2500 pulses per second (pps), depending on how fast the ball is rotated. This is much more rapid than required for satisfactory updates, which need not be greater than about 100 times per second. This can easily be accomodated by the RS-232 protocol using an eight-bit word. Thus, the trackball is an excellent alternative for the mouse, and is rapidly returning to a preferred position as an input device.

Joystick

The joystick has not achieved much acceptance as an input device for electronic display systems, except for video games, although it has been the preferred control for many types of aircraft. However, it can be used to some extent in display systems other than those used in video games, and therefore warrants inclusion in this section. There are two basic types of joysticks, termed "displacement" and "force-operated". A typical displacement joystick is shown in Fig. 89.10, and may have two or three degrees of freedom. The activating means may vary from as few as four switches mounted 90 degrees apart, to full potentiometers for analog output, and optical encoders for digital output. A third axis may be added by allowing the handle to rotate and drive a third potentiometer. Spring forces of 5 to 10 lbs. are usual for the other two axes, and displacements go from 6 to 30 degrees.

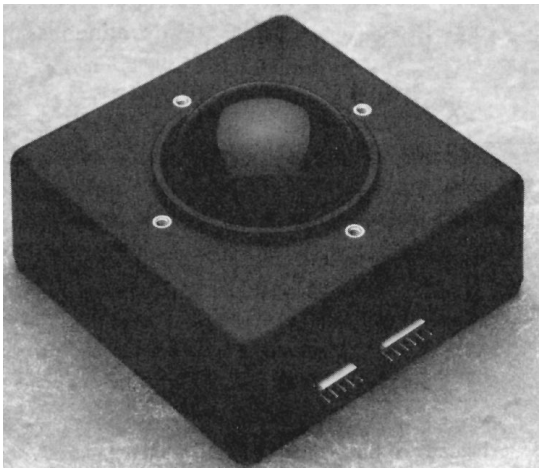


FIGURE 89.9 Trackball. (Courtesy of CH Products,Vista, Calif.)

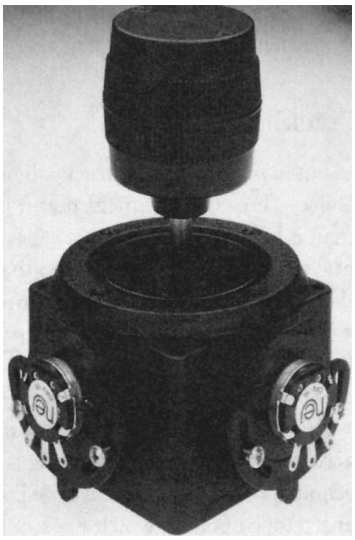


FIGURE 89.10 Three-axis displacement joystick. (Courtesy of CH Products, Vista, Calif.)

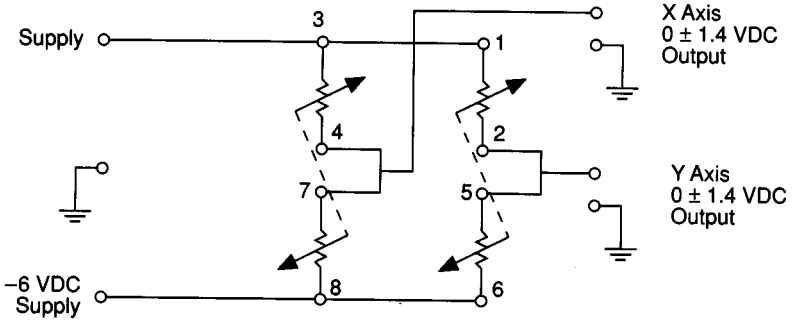


FIGURE 89.11 Schematic connections in a force joystick. (*Source:* After D. Doran, "Trackballs and joysticks," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, p. 260. With permission.)

The force joystick operates by responding to pressure on the handle to generate the *X*–*Y* coordinates. It may be either a two- or three-dimensional version, with the same types of handles as for the displacement joysticks. However, it is difficult to use a rotating handle for the third dimension because some force is usually transmitted to the other dimensions causing crosstalk. Therefore, a separate lever is preferred. The force is detected by means of piezoelectric sensors that are bonded to the handle rod, and a voltage source is applied across the network, as shown in Fig. 89.11. The output is taken from the strain gauge and the analog voltage will be proportional to the amount of force. The same type of protocol and output circuitry may be used as for the displacement unit, and both can generate either position or rate data. An exponential curve with a dead zone threshhold is preferred for pulse rates in order to avoid starting pulse rate uncertainties, with the first pulse starting as soon as the threshhold is exceeded.

Touch Input

Touch input devices come in two basic forms, either placed directly on the display surface, or as a separate panel attached to the computer system. In its second form it is essentially a data tablet differing mainly in that it acts as another display unit with some form of a touch-sensitive surface. In this implementation it is the

same as the *Touchscreen* input device, and this discussion concentrates on the technologies used for Touchscreens. There are five different technologies used for touch input devices, which are capacitive or resistive overlays, piezoelectric, light beam interruption, and surface acoustic wave. The system may be divided into the sensor unit, which senses the location of the pointing element, and the controller that interfaces with the sensor and communicates the location information to the system computer. Since the controller is an electronic device that does not use technology different from the computer it is not covered here. The main differences among the different touch input devices are due to the choice of sensor technology, and the discussion concentrates on these technologies.

Capacitive. Capacitive overlay technology is illustrated in Fig. 89.12 where a transparent metallic coating is placed over the display screen and the finger or stylus capacitance is sensed to determine the touch location. The overlay may consist of a group of separate sections etched into the surface with each separate section connected to the controller, or a continuous surface connected at the four corners. The first form is termed discrete capacitive, and touch location is determined by having each section sequentially connected to an oscillator circuit where the frequency of oscillations is affected by the pointing device. The oscillation frequency is measured and compared to a stored reference frequency. If the frequency difference is large enough then it is recognized as a touch at that location. It is a simple system, but suffers from low resolution and slow response so that it is only practical for menu selection.

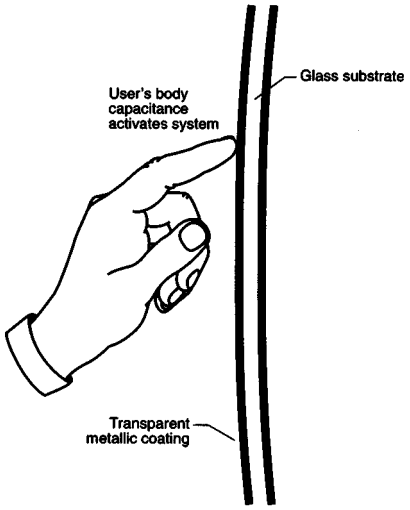


FIGURE 89.12 Capacitive overlay technology. (*Source:* After A. B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lecture Notes*, p. 15.30, 1987. With permission. Courtesy Society for Information Display.)

The analog capacitive system uses the same metallic overlay, but the metallic surface is continuous rather than etched. The connections at the four ends are each connected to a separate oscillator, and the frequency of each is measured and stored. Then when the overlay is touched the change in capacitance will have a different effect on the frequency of each oscillator. These are measured and the differences are used to determine the coordinates of the touch by means of an algorithm. This technique is capable of much higher resolution (250×250) than the digital approach and is preferred for graphics or other high-density displays.

Resistive. Resistive overlay technology requires a more complex assembly consisting of two layers, as illustrated in Fig. 89.13. The layers both contain transparent metallic surfaces and are separated by spacers so that an air gap exists between the layers in the absence of any pressure on the touch panel. The metallic layers face each other and when the outer panel is pressed the metallic layers make contact and form a conductive path at the point of contact. When a voltage is applied between the top of the outer layer and the bottom of the inner layer, the two layers act as a voltage divider, and the voltage at the point of contact may be measured in the X and Y directions by applying the voltage in first one and then the other direction. The measured voltages are then transmitted to the controller where they are converted into coordinates which are then sent to the computer.

The panel may be discrete, in which the conductive coating on the top layer is etched in one direction and that on the bottom layer in the other direction, or analog, where the conductive coatings in both layers are continuous. In the discrete case, the panel then acts as an X–Y matrix, and the resolution is determined by the number of etched lines. The analog configuration requires the addition of linearization networks on each edge of the panel so that a large-area resistor is created with a voltage drop in one direction. Other linearization techniques are also possible, but only the four-element system is described here as shown in Fig. 89.14. In this arrangement, one of the layers acts as the large-area resistor and the other as a voltage probe where either can function in either role. For the Y coordinate value the top layer is the voltage probe, and the voltage is applied by the controller to the bottom layer. Similarly, the X coordinate is found by connecting the voltage to the top layer and making the bottom layer into the voltage probe. In either type of system, the resolution can be very high, but the transmissivity is reduced to under 80% due to the multiple layers.

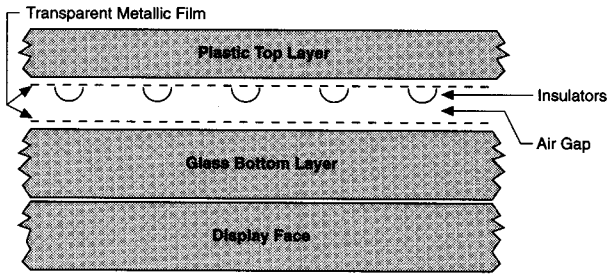


FIGURE 89.13 Resistive overlay technology. (*Source:* After A. B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lecture Notes*, p. 15.31, 1987. With permission. Courtesy Society for Information Display.)

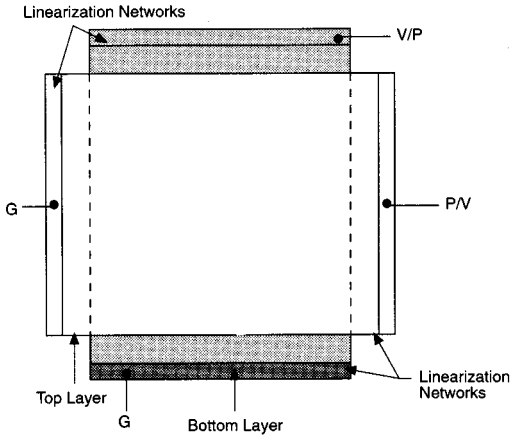


FIGURE 89.14 Four-wire analog resistive. (*Source:* A. B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lecture Notes*, p. 15.32, 1987. With permission. Courtesy Society for Information Display.)

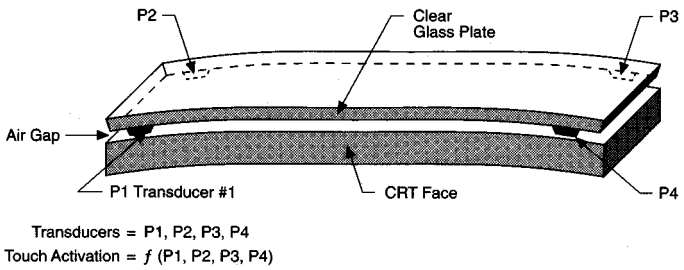


FIGURE 89.15 Piezoelectric technology. (*Source:* A. B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lecture Notes*, p. 15.34, 1987. With permission. Courtesy Society for Information Display.)

Piezoelectric. The piezoelectric technology uses pressure-sensitive transducers as the means for determining the location of the touch, as shown in Fig. 89.15. The sensor is a glass plate with transducers connected to the four corners. Pressure on the plate causes readings to occur at each of the transducers, which depend on the location of the pressure. Thus, the controller can measure the readings and obtain the coordinates by means of a proper algorithm. This technique allows a high-transmissivity plate to be used that can be curved to follow the CRT face plate curvature, but it allows only a limited number of touch points to be used.

Light Beam Interruption. This is a fairly straightforward technology that requires a matrix of light sources and detectors facing each other in the *X* and *Y* directions. When the beams from the *X* and *Y* light sources are interrupted, this is sensed by the facing light detectors and the signals are sent to the controller. The light beams

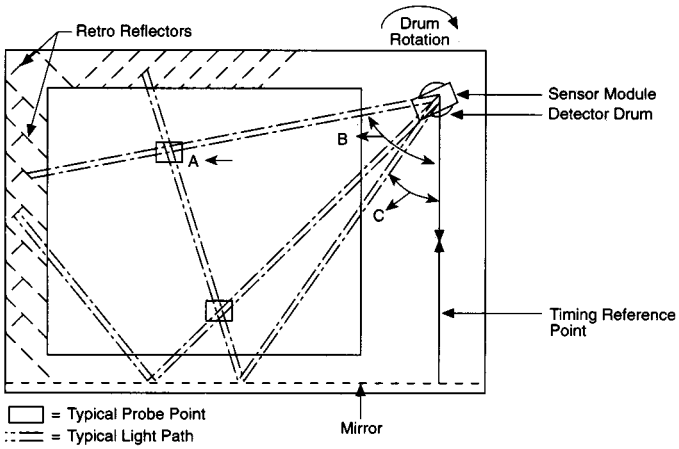


FIGURE 89.16 Rotating infrared beam technology. (*Source:* A. B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lecture Notes*, p. 15.34, 1987. With permission. Courtesy Society for Information Display.)

are turned on sequentially by pulsing the LEDs and thus create a full matrix of light beams without requiring each of them to be on continuously. This system does not reduce the screen transmissivity as there is no obstruction of the screen output, but it is limited in resolution to the number of LED detector pairs that can be placed on the periphery of the screen.

Another approach to light interruption is to use a rotating beam of light, which has the advantage that only one light source and detector pair is required. This technology is depicted in Fig. 89.16 and consists of a LED and a light detector placed inside a rotating drum which has a slit that allows light to be transmitted outside the drum. The light is swept across the surface and strikes the retroreflectors that sends it back directly to the detector. The beam scan is sampled 256 times on each scan, and Fig. 89.16 shows how two angles of interruption are created, angle B by direct interruption, and angle C by mirror reflection interruption. The result is that the location of the interruption can be calculated by comparing the two angles. Again, there is no obstruction of the screen but a moving element must be added, and parallax errors may occur.

Pen-Based Computing. This is an application for touch input devices that is growing at a rapid rate. The input device comes in several forms, each of which can recognize hand printing with the special operating system and software recognizing this type of input. The pen-based input device comes in several forms, of which the one termed TouchPen™ can function both as a digitizer with a touch tablet, and as the touch input device with a touch input pen-based computer system. A second one is that developed by Wacom, Inc., primarily for the GO Systems computer, but used by other pen-based systems as well. Finally, a third unit is that made by Scriptel Corp. and used by Wang Laboratories in its system.

TouchPen™ was developed by Microtouch Systems, Inc., initially for use in GridPad made by the Grid Systems Corp. It is essentially a high-resolution digitizer consisting of an all-glass tablet that can be used with a number of stylus input operating systems to digitize handwriting. It is basically a touch input device using resistive techniques to digitize the handwriting appearing on the display surface of pen-based computer systems. The glass tablet is placed on the display surface and the system pen is used to transmit the digitized data to the computer. As noted previously, the tablet may also be used as a standard touch input device.

The second form of pen-based input device is one that uses electromagnetic technology and consists of a grid of wires that transmit radio waves that are picked up by a tuned circuit in the stylus. This circuit resonates at its own frequency and transmits that signal back to the wires at the grid location it is touching. The pen also transmits its signal to the computer, which turns off the grid transmission, and locates the position of the pen by determining which of the grid wires pick up the pen signal. The pen does not need to actually touch the display surface and does not require any power, which is an advantage somewhat counteracted by the higher cost.

Finally, the Scriptel unit is similar to that made by Microtouch, but differs in that it uses electrostatic technology and is also similar to the capacitive touch panel.

Surface Acoustic Wave (SAW). This technology is more recent than the others and has not received wide acceptance as yet. It is based on the transmission through the glass of SAWs generated by transducers mounted on the glass overlay. These waves are detected by receivers also mounted on the glass, and the time of arrival of the waves at the receivers is known because the wave velocity is known. The placing of a finger on the glass weakens the signal and the location of the finger can be determined by the difference in its effect on the SAW.

There are two types of SAW systems in use, namely those using reflective techniques and those using attenuation as the source of position information. The reflective systems are similar to sonar where the time from the source to the pointing finger and then from the finger to the receiver is measured to arrive at finger location. The attenuation technology is illustrated in Fig. 89.17 and consists of two transducers, two receivers, and four reflector strips, all mounted on a glass substrate. One transducer-receiver pair is used for *X* and the other for *Y* location. Figure 89.17 shows the *X* axis pair, and the transducer transmits a burst of acoustic energy in a horizontal wave. The wave is partially reflected by the top reflector strips and travels down to the bottom strip where the reflectors are at an angle such that it is reflected to the lower left corner receiver. The wave now has a long rectangular shape, and each point in time corresponds to a specific vertical path across the substrate. The *Y* axis is scanned in the same fashion after the *X* wave dies out. Then, when the finger touches the substrate, its water content absorbs some of the energy in the wave, and the wave is attenuated. The dip in the wave amplitude corresponds to the amount of absorbed energy, and the time of the lowest point can be determined, allowing the location of the finger to be calculated. Finally, in addition to the *X* and *Y* coordinates, a *Z* coordinate can be determined, depending on how hard the user presses. This depends on surface contact, which affects the amount of attenuation. The advantages of this system are high resolution, speed of transmission, and the availability of a *Z* axis component. Its main disadvantages are the variation in moisture content in fingers and sensitivity to local moisture on the substrate. However, it is being used in developmental units and should be considered as another input device technology.

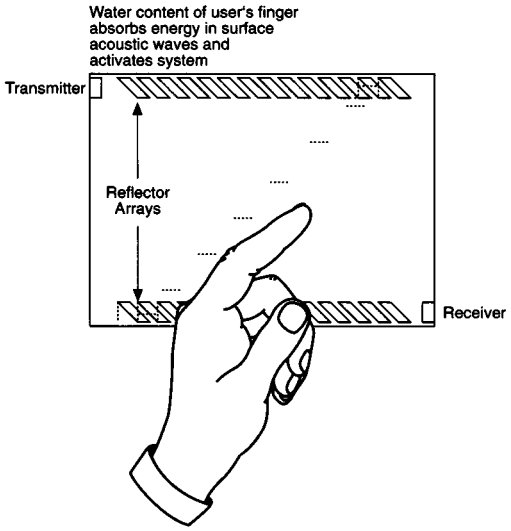


FIGURE 89.17 Attenuation SAW technology. (*Source:* A. B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lecture Notes*, p. 15.35, 1987. With permission. Courtesy Society for Information Display.)

Scanners

Scanners are a means for inputting text and/or images directly into the computer system, thus avoiding the need for retyping and redrawing information contained in other sources. It is a relatively convenient way to avoid repetition if the data to be entered already exist in readable form. This is done by special image-recognition software that accompanies the scanning hardware, and can transfer an entire image containing both text and illustrations, but without the capability to modify the image. However, the addition of optical character recognition (OCR) software allows the entered text to be modified as if it were entered by typewriter. This can greatly simplify entering and editing text from some preexistent source and has resulted in a proliferation of devices that can perform this function.

These devices come in two main forms, *hand-held* and *page* scanners, with or without OCR software in addition to the standard image-recognition software. A typical hand-held scanner is shown in Fig. 89.18 and it consists of a light source, a light-sensitive device such as a charge-coupled device (CCD) array, and the electronics to actuate the elements of the array sequentially under software control. The scanner window is placed over the page, and is moved down or across the page so that the window covers as much of the page as falls within the capability of the software. The light source is reflected from the page to the CCD and the charge in the CCD is modified by the reflectivity of the printed material.

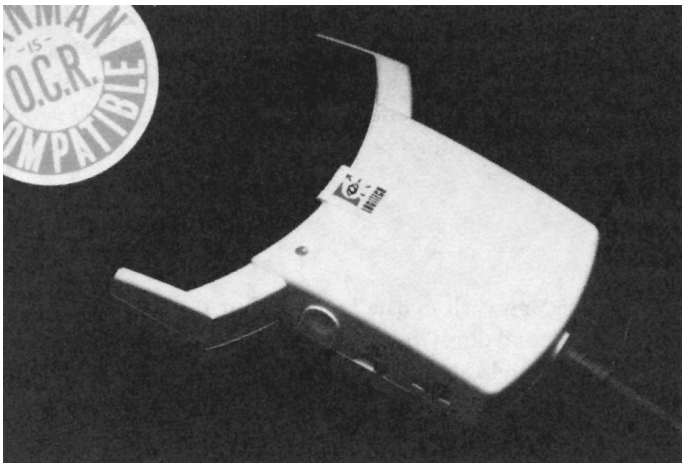


FIGURE 89.18 Hand-held scanner. (Courtesy of Logitech, Freemont, Calif.)

The window area ranges from 4 to 5 in. in width by 0.5 in. in height and may be moved through 14 to 20 in., so that a fairly large area may be covered in a single manual scan. Images wider than the maximum window may be scanned in two passes, and the OCR software can stitch the two scans together into a single image, although this procedure requires considerable care in scanning so that the scans line up properly. Therefore, when images wider than the window of the hand-held scanner are to be scanned, it is advisable to use a *flatbed scanner* of the type shown in Fig. 89.19 which can handle a full 8.5 in. by 11 in. page, or some of the larger scanners than can accept large drawings and input them into the computer system. Resolutions of 400 dpi and higher, with up to 250 levels of gray and 24 bits of color resolution are available. Thus, scanners offer a wide variety of choice and performance capabilities, and are powerful input devices when prepared data in visual form is to be entered into the computer system.

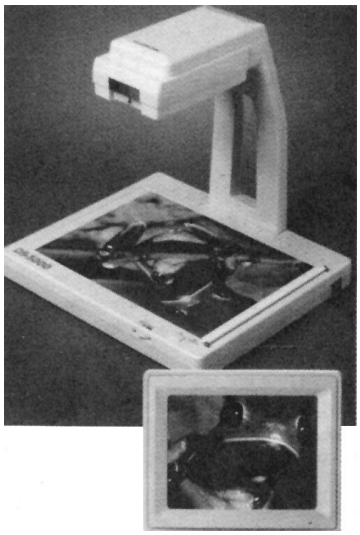


FIGURE 89.19 Page scanner. (Courtesy of Chinon, Torrance, Calif.)

Voice

Voice input is an intriguing approach to data input, with particular attractiveness to managers who want a simple and direct means for inputting data and commands. For many years, this technology tended to promise more than it could achieve, but recent developments have brought it to the point where it can be considered as a viable input means. This has been due to new developments in software that make it possible to minimize the amount of training required and increase the success rate to close to 100%.

One basic approach to speech recognition is represented by the block diagram shown in Fig. 89.20. This is a system that is built around a special chip developed by Texas Instruments. This system uses templates and special algorithms for recognizing the input speech patterns. The system is speaker dependent, with the capability of storing up to 32 word templates and user-defined phrases. The output portion may be superfluous when the system is used only for inputting data and commands, but can be a useful adjunct to the visual response. Other techniques such as speaker-independent and phoneme-recognition systems are also available. Vocabularies range from 50 to 5000 active words, and both isolated and connected words can be recognized, although the larger numbers tend to be associated with isolated word systems. In general, it seems feasible that a combination of speech input and pen-based computing may find a viable market.

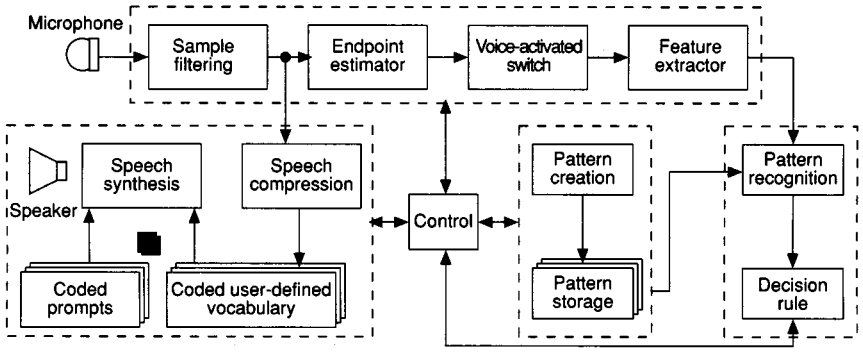


FIGURE 89.20 Block diagram of speech recognition and synthesis chip. (*Source:* After M. Leonard, "Speech poised to join man-machine interface," *Electronic Design*, pp. 43–48, September 26, 1991. With permission.)

TABLE 89.4 Input Device Functional Evaluation

		Function		
Input Device	Control	Data/Text	Data/Graphics	Total
Keyboard	E	E	P	9
Light Pen	G	G	E	10
Tablet	E	G	E	11
Mouse	E	F	E	11
Trackball	E	G	E	11
Joystick	F	F	G	5
Touchscreen	G	F	G	8
Scanner	F	E	G	9
Voice	G	F	P	6
Total	29	23	28	80

E = Excellent = 5; G = Good = 4; F = Fair = 3; P = Poor = 2

Summary

The multiplicity of input devices that are available makes it difficult to determine which is most suitable for any specific set of requirements. However, the limited functional comparison of the input devices covered in this section shown in Table 89.4 may be of some use, and in any event is a starting point in this evaluation. It should be noted that what appears best at one time may become unpopular or obsolete at a later time, as occurred for light pens and trackballs, both of which have come back into favor.

In addition to the generalized evaluation shown in Table 89.4, it is also of interest to examine representative performance parameters. These are shown in Table 89.5 and while representative do not necessarily cover the range of performance parameters offered. More data may be obtained from the vendors of these devices.

Advantages and Disadvantages

Input devices make up one of the functional groups of the display systems, and their technical characteristics are covered in some detail at the beginning of this chapter, with performance information provided in Table 89.5 containing characteristic parameter values for each type, as available. The following material expands somewhat on that information by placing these devices in the context of a full graphics display system and evaluating the functions that the various types of input devices perform in that type of system in terms of their advantages and disadvantages. It is of some interest to compare the advantages and disadvantages of each type at this point, as listed in Table 89.6. This is an imposing list and may be used to aid in choosing the best input devices for specific applications. It also concludes this section on input devices.

TABLE 89.5 Representative Performance Parameters

Input Device	Parameter	Value
Light pen	Response time	150–500 ns
	Spectral response	4200–9500 A
	Luminous sensitivity	0.03–0.7 nts
	Field of view	0.02–0.1 in.
	Ambient rejection	350 nts
Data tablet (digitizers)	Resolution (l/in.)	100–2000
	Accuracy (in.)	0.0005–0.02
	Active area (in.)	12 × 12–60 × 120
	Active height (in.)	0.02–2.5
	Digitizing rate (pps)	100–350
	Transducers	Stylus, puck, cursor
Mouse	Resolution	10–1000 dpi
	Speed	1–20 in./s
	Accuracy	25–1000 dpi
Trackball	Resolution	100–1000 cpi
	Speed	1200–9600 BPS
	Accuracy	100–1000 dpi
	Ball diameter	1.5–2.5 in.
Joystick	Travel	25–30°
	Accuracy	5–10%
	Repeatability	1%
Touchscreen	Resolution	256 × 256–4096 × 4096
	Transmissivity	60–100%
	Viewing area (in.)	3 × 4.5–15 × 20
	Speed	80–200 touch pts./s
Scanner	Resolution (dpi)	75–1600
	Scan rate (in./s)	0.5–2.0
	Scanning width (in.)	4.1–36 gray shades 32–256
	Scan time (s/page)	1–30
Voice	Active vocabulary	13–5000 words

Defining Terms

Data tablet/digitizer: A device consisting of a surface, usually flat, and incorporating means for selecting a specific location on the surface of the device and transmitting the coordinates of this location to a computer or other data processing unit that can use this information for moving a cursor on the screen of the display unit.

Joystick: An input device somewhat in the form of the navigation control device found in early aircraft and operating in a somewhat similar manner by generating series of pulses whose frequency or number depend on how far, with what force, and in what direction the control stick is moved from the central position.

Keyboards: Electromechanical devices consisting of sets of keys labeled with alphanumeric, numeric, and functional designations that enable the user to describe and define the operation to be performed.

Light pen: Neither a pen or a light source but rather an input device in the shape of a pen that operates by sensing the existence or nonexistence of light pulses at specific locations on the surface of a display device and uses this information to signal the computer as to the location of the pen.

Mouse: An input device based on a much older type known as a trackball and fancifully named because it bears only a casual resemblance to a mouse. It consists of a roller ball that is moved on a flat surface and causes orthogonal potentiometers or other types of *X*–*Y*-position signal generators to move and produce electrical signals defining the desired coordinates of the cursor on the screen so that the cursor can be moved to that position.

TABLE 89.6 Input Devices—Advantages and Disadvantages

Device	Advantages	Disadvantages
Keyboard	Simple operation	Requires many keys
	Well known	Requires training
	Standard interface	No graphics
Light pen	Eye-hand coordination	Arm fatigue
	Low cost models	Limited resolution
	No desk space required	May block display
Graphic tablet	Natural hand movements	Eye-hand conflict
	Screen not blocked	Requires desk space
	No parallax	Breakable stylus
	Good for graphics	Poor for A/N entry
Mouse	Small space needed	Some space needed
	Low cost	Slow transmission
	Screen viewing	Low resolution
	Any surface may be used	Grid for optical
	(Optical) noiseless	Mechanical noise
Trackball	High resolution	Poor for A/N input
	Fixed desk space	Slow transmission
	Screen viewing	Mechanical noise
	Tactile feedback	3-D difficult
Joystick	Fixed desk space	Low accuracy
	Low fatigue	Low resolution
	Low cost	No A/N input
Touchscreen	Eye-hand coordination	Arm fatigue
	Minimal training	May block display
	Minimum input errors	Varied resolution
	User acceptance	Parallax
	No special commands	Slow data entry
Scanner	Full A/N page input	Hand scanner width
	Color scan input	High cost for color
	High resolution	Slow input
	OCR software	Compatibility
Voice	Ease of use	Limited words
	Minimal training	Machine training
	No special devices	Graphics difficult

Scanners: Means for converting hard copy into electrical signals that can be entered into a computer or data processing system. The usual means for accomplishing such conversion is to move a light beam over the surface containing the data either by hand or automatically and using arrays of light-sensitive devices to convert the reflected light into electrical pulses.

Touch input: A means for selecting a location on the surface of the display unit using a variety of technologies that can respond to the placing of a finger or other pointing device on the surface. These are essentially data panels placed either on the display surface or between the user and the display surface.

Trackball: The earliest version of an input device using a roller ball, differing from the mouse in that the ball is contained in a unit that can remain in a fixed position while the ball is rotated. It is sometimes referred to as an upside-down mouse, but the reverse is more appropriate as the trackball came first.

Voice: Means for enabling a computer or data processing system to recognize spoken commands and input data and convert them into electrical signals that can be used to cause the system to carry out these commands or accept the data. Various types of algorithms and stored templates are used to achieve this recognition.

Related Topic

89.2 Computer Output Printer Technologies

References

H. Brunner et al., "Effects of key action design on keyboard preference and throughput performance," Micro Switch.

A.B. Carrell and J. Carstedt, "Touch input technology," *SID Sem. Lec. Notes*, pp. 15.30–15.35, 1987.

T.E. Davies et al., "Digitizers and input tablets," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, p. 186.

D. Doran, "Trackballs and joysticks," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, pp. 251–262.

C. Goy, "Mice," in *Input Devices*, S. Sherr, Ed., New York: Academic Press, 1988, pp. 225–232.

M. Leonard, "Speech poised to join man-machine interface," *Elec. Des.*, pp. 43–48, Sept. 26, 1991.

S. Sherr, *Electronic Displays*, New York: Wiley, 1979, pp. 323, 388–389.

Further Information

Electronic Displays, 2nd ed., by Sol Sherr and published by John Wiley & Sons, Inc., contains an extensive and detailed discussion of other aspects of display systems and technology, as well as a somewhat expanded version of this section. In addition, *Input Devices,* edited by Sol Sherr, and *Output Hardcopy Devices,* edited by Robert C. Durbeck and Sol Sherr, both published by Academic Press, include extensive discussions of a wide variety of devices.

The Society for Information Display (SID) sponsors a yearly symposium at which a large amount of information on new developments in information display as well as tutorials and seminars on basic information display topics are presented and made available in published form. In addition, it publishes two journals, namely, *Proceedings of the Society for Information Display* and *Information Display.* Other relevant meetings and publications are those sponsored by the Computer Society and Electron Devices groups of the IEEE, the SIGGRAPH group of the Association for Computing Machines (ACM), and the National Computer Graphics Association (NCGA).

89.2 Computer Output Printer Technologies

Robert C. Durbeck

Electronic printers for computer output represent a very important part of the computer industry. They range from small, inexpensive printers for personal computers and workstations to very large and fast page printers used for bulk printing output for large-scale computer systems. The technologies employed for this wide scope of printing requirements are diverse: some based on "impact" methods to transfer ink from a sheet or ribbon to paper, others based on more sophisticated "nonimpact" methods. Today, there is no single technology which completely dominates. A wide range of user needs has led to the present proliferation of printer technologies. The most prevalent are discussed in the following.

Classification of Printer Technologies

Table 89.7 illustrates the two main classifications of printer technologies and also the wide range of extant technologies. Those technologies listed under **line impact** and **page printing** are used for large computer system printing; by far the most popular are *fully formed character* and *electrophotographic* (so-called laser printers). The most common printing technologies used for personal/workstation computer systems are *electrophotographic, ink jet,* and *serial wire matrix.* Emphasis in the following is directed to these favored technologies, with brief descriptions of the others included.

Page Printer Technologies

By far the most important page printer technology is electrophotography (EP). EP, as well as the much less employed ionographic and magnetographic technologies, uses "powder toning" development of an intermediate

TABLE 89.7 Types of Printer Technologies

Impact	Nonimpact
Line impact	Page printing
Fully formed character	Electrophotographic
Dot band matrix	Ionographic
Shuttle hammer matrix	Magnetographic
	Electrostatic
	Thermal
Serial impact	Serial nonimpact
Fully formed character	Ink-jet
Serial wire matrix	Continuous
	Piezoelectric/impulse
	Thermal/bubble-jet
	Thermal
	Direct (thermal paper)
	Thermal transfer
	Resistive ribbon

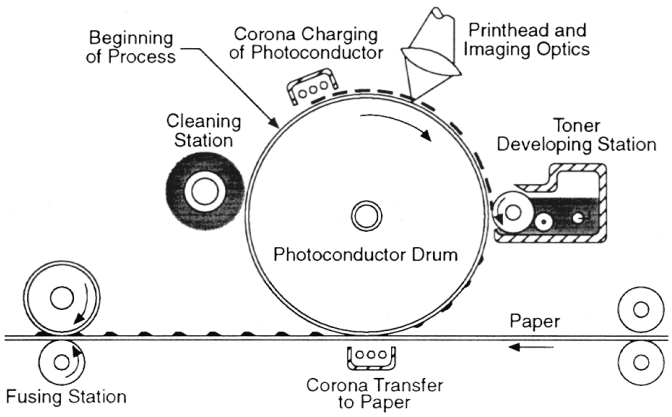


Figure 89.21 The six basic electrophotographic printer process steps: charge, image, develop, transfer, fuse, and clean.

"image" created in the process. Liquid toners are also used to develop electrostatic images. Thermal page printers employ a full-page-width linear array of thin-film heater elements to melt and transfer ink from a ribbon to the paper or to mark special thermal-sensitive paper.

Electrophotographic Printing

EP printers use essentially the same technology found in most "plain paper" copiers, the major exception being the printhead. Instead of using page- or line-imaging optics as in a copier, the printhead utilizes a solid-state laser (usually GaAlAs) or gas laser (typically HeNe) to scan across and expose a photoconductor drum or belt to create a "latent image" (see below). A few EP printers use stitched arrays of light-emitting $GaAs_{1-x}P_x$ diodes (LEDs) with Selfoc™ glass fiber optics or an array of liquid crystal shutters, the latter to modulate light from a bright line light source. Other possibilities are electroluminescent, magnetooptic or electrooptic arrays, but these have not been commercialized to any extent.

There are basically six major steps employed in the EP printing process (see Fig. 89.21): (1) uniform charging of the photoconductor (PC) electrostatically; (2) exposing the PC to the image light pattern, which results in selective discharge of the area charge created in Step 1, creating an electrostatic image; (3) developing the PC by bringing electrostatically charged toner particles (black or colored) to the surface of the PC where they selectively adhere to appropriately charged regions; (4) electrostatically transferring the toned image from the PC to the final medium (usually paper); (5) thermal fusing of the toner to the paper; and (6) cleaning residual toner from the surface of the PC to allow reinitiation of the six step cycle.

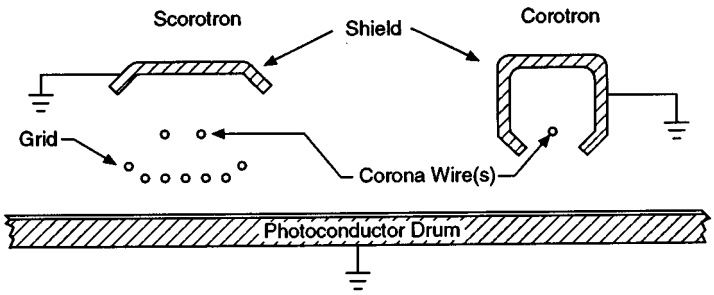


Figure 89.22 Two types of coronas for charging the photoconductor: the scorotron and the corotron.

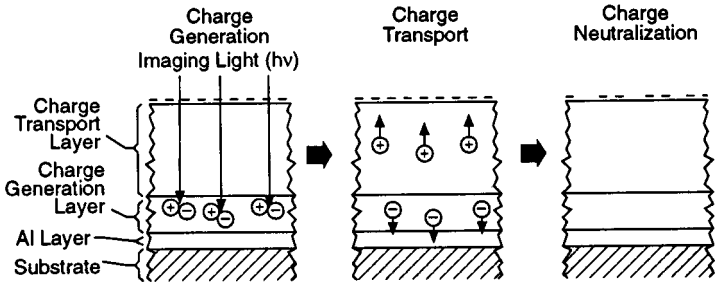


Figure 89.23 Light is used to selectively discharge the photoconductor. Electron-hole pairs are photogenerated in the charge-generation layer, followed by charge transport under a dc bias field and then selective neutralization of the surface charges, thereby creating an electrostatic image.

Step 1—Charging the PC. The most common approach used is corona charging. One or more thin corona wires (typically tungsten) are supported directly above the PC and are energized to 5–8 kV. The resultant high electric field surrounding the wire causes electrons in the immediate region to be accelerated to energies sufficient to ionize local air molecules. Either positive or negative ions are then attracted to the outer surface of the PC (which, when unexposed to light, acts for a short period of time as an insulator) depending on the sign of the potential difference. At the same time, a counter-image charge is formed on the inner side of the PC. The two corona structures most commonly used are the corotron and the scorotron (see Fig. 89.22). The grid on the scorotron is used to more precisely control the resultant voltage charge level on the PC (approximates the grid voltage). Both dc and ac designs are used; the latter usually include a glass sleeve around the corona wire to reduce localized high-emission spots on the wire due to contaminants. To save on cost for very low-cost EP printers and to reduce corona by-products (e.g., ozone), a lower-voltage conductive elastomer charge roll in direct contact with the PC has also been used in place of the corona wire.

Step 2—Exposing the PC. The wavelength of the exposing light source must match the spectral sensitivity of the PC. If the PC is discharged in areas that will be printed white, the overall process is termed charge area development (CAD); if the discharged area will be printed black (or color), the process is called discharge area development (DAD). Both CAD and DAD processes are used in EP printers but CAD is the only common process used in copiers.

Selective discharge of the PC involves two steps: (1) photogeneration of electron-hole pairs and (2) transport of the electrons and holes in opposite directions under the influence of a high-dc bias field, locally dissipating the surface charges created in Step 1 (see Fig. 89.23).

Both organic and inorganic PC materials are used (see Fig. 89.24). A variety of charge-generation and charge-transport material systems have been developed for organic PCs; most use separate layers for charge generation and charge transport. Examples of efficient organic charge-generation materials sensitive at both GaAlAs (7800 Å) and HeNe (6328 Å) wavelengths include squarylium and thiopyridium dyes in an appropriate binder layer (≈0.5 μm thick). The charge-transport layer (CTL) consists of a thicker layer (20–30 μm) of a charge-transport

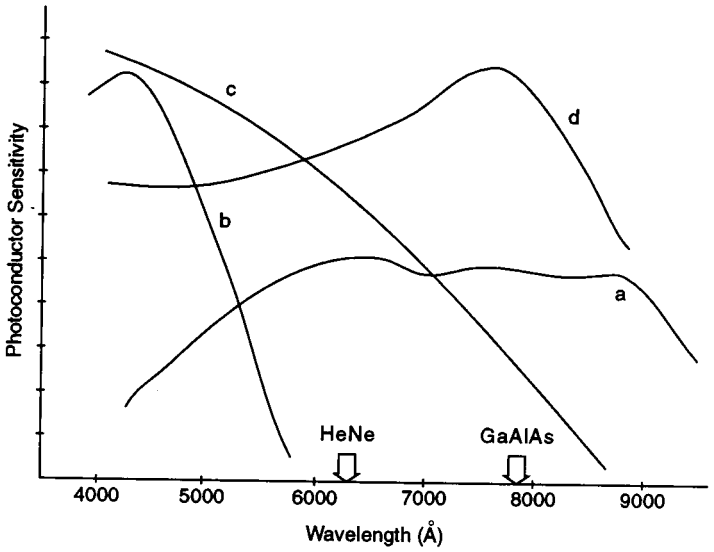


Figure 89.24 Spectral absorption characteristics of several photoconductor materials: (a) squarylium dye, (b) SeTe, (c) As_2Se_3, and (d) a-Si.

molecule dispersed in an inert binder, or simply a good charge-transporting polymer (e.g., polyvinylcarbazole). Since most polymer transport materials are essentially hole-transport materials, this requires that the PC surface charge produced in Step 1 be negative. The CTL must be transparent at the imaging wavelength.

Examples of inorganic PC materials include amorphous chalcogenide alloys such as a-Se, a-As_2Se_3 and Te-doped Se. Most do not have a high sensitivity at GaAlAs wavelength, but a-Si does. a-Si also offers other desirable properties (e.g., durability and lack of fatigue), but is quite expensive to produce.

The laser beam is scanned linearly over the PC in a direction orthogonal to the PC motion; the combined motions covering a "page". The most common gas laser scanning technique employs a high-speed rotating polygon mirror along with beam-expanding optics, an acoustic-optic modulator (not required for solid-state lasers) and an f-θ imaging lens. To produce a quality image, the multifaceted scanning mirror system must be essentially free of facet defects, up-and-down wobble, variations in polygon rotational velocity and lack of synchronization with the **pixel** clock. Some nonplanarity of the facet surfaces can be corrected with anamorphic optics. LED arrays and liquid crystal shutter systems do not have these same technical challenges but, so far, they are more expensive to produce.

Step 3—Developing the PC. There are basically three development techniques: dual component, monocomponent and liquid development. The first two use powder toner. Until the advent of EP printers for personal and workstation computers, the most common method of development was dual component, where polymer-coated magnetic carrier beads are mixed with the toner particles and development is done with a "magnetic brush." This technique is still prevalent in high-end printers. With this approach, the 5–20 μm toner particles (consisting mainly of resin plus carbon particles or colorant) are triboelectrically charged by repeated contacts during mixing with the much larger (60–250 μm) magnetic carrier beads. The toner particles then electrostatically adhere to the opposite-sign charged carrier beads. Charge control agents (e.g., complex organometallic salts) are often included in the toner composition to control the charge level, rate of charging, and consistency of charge. The mix is mechanically directed to a nonmagnetic rotating shell which has fixed magnets located within its core, adjacent to the gap between the PC and the shell (see Fig. 89.25). The gap is typically 0.5 to 6 mm, depending on the specific system. As the shell rotates, the mixture is carried to the gap, and chains of carrier beads (coated with toner—the "magnetic brush"!) form along the local magnetic field lines. These field lines are approximately perpendicular to the shell at the smallest gap. In addition, a development voltage (200–500 V) is applied between the PC and shell. This provides a high field in the gap whose local value is determined by the applied voltage, gap dimension and the electrical properties of the material mix in the gap.

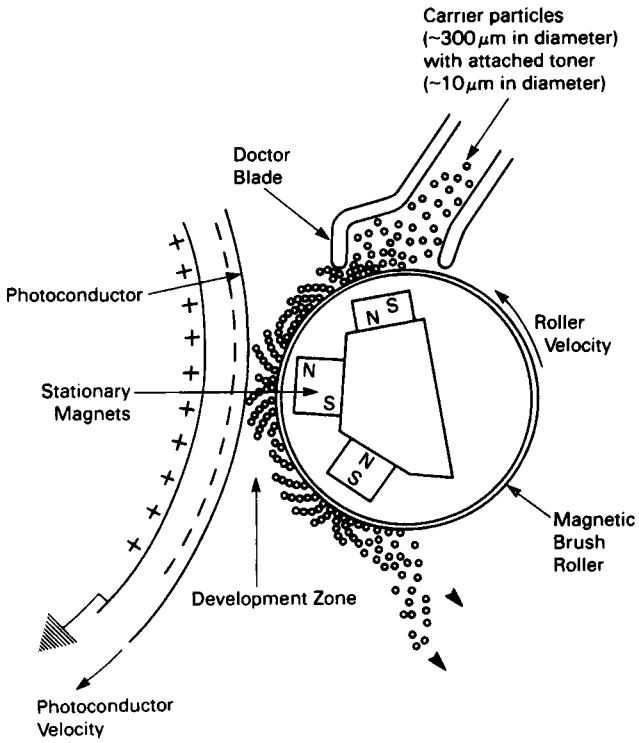


Figure 89.25 Dual-component magnetic brush developer. Toner particles, adhering electrostatically to much larger magnetic carrier beads, are transported into the photoconductor-developer gap to tone the electrostatic image. (*Source:* R. C. Durbeck and S. Sherr, Eds., *Output Hardcopy Devices,* San Diego, Calif.: Academic Press, 1988, p. 242. With permission.)

The electric field at the end of the last carrier bead (next to the PC) in a chain may be up to 50 times the nominal unfilled gap field, the value being greatest with uncoated carrier beads. It can be shown that for coated carrier beads, the mass of toner per unit area developed on the PC is approximately

$$\frac{m}{A} = \frac{\pi \epsilon_0 V}{2 L_g C_m} \left\{ r_c \, / [(r_t / \kappa) + \delta_c] \right\} \cdot \left| v_r / v_p \right| \tag{89.7}$$

where v_r/v_p is the surface speed ratio between the developer roll (shell) and the PC surface, V the voltage across the gap L_g, C_m the toner charge-to-mass ratio, ™$_0$ the permittivity of free space, r_c the carrier bead radius, r_t the toner radius, κ the dielectric constant of the bulk toner, and δ_c the thickness of the carrier polymer coating.

Monocomponent development is used almost exclusively for low-end printers because this process does not require carrier beads, toner concentration sensors or toner replenishment hardware, resulting in much lower manufacturing costs. This approach has also allowed the use of replaceable toner/developer cartridges which, although more costly on a supplies cost-per-page basis, adds greatly to the user-perceived reliability of the system (if it fails—just replace the cartridge!). A rotating donor/development roll with appropriate charging properties is employed to charge the toner by touch-and-rubbing contacts (see Fig. 89.26). The toner electrostatically adheres to the donor roll and is transported to contact the PC at the nip. Here, in the presence of a development bias field, the toner is selectively transferred to those areas on the PC with opposite sign charge.

Liquid development employs a high-resistivity hydrocarbon dispersion of very fine toner particles (<1 μm) that are charged naturally in the solvate. Mechanical means are used to bring the liquid into contact with the PC, and the toner is then electrophoretically transferred to the latent image areas on the PC.

Color can be accomplished by using multiple development stations, one each for the subtractive colors (cyan, yellow and magenta) plus black. Toners are colored by either dyes or pigments. The four-colored images may

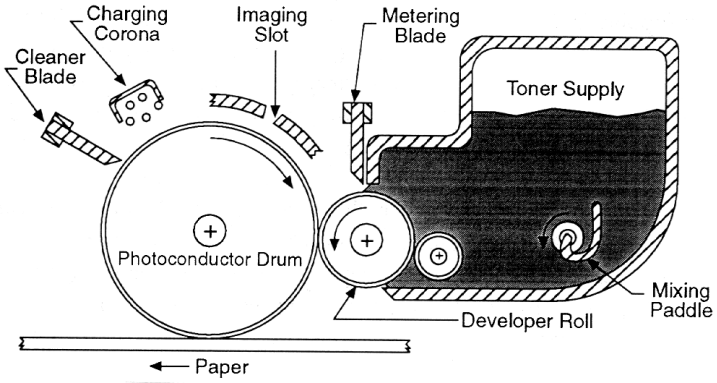


Figure 89.26 Monocomponent developer system.

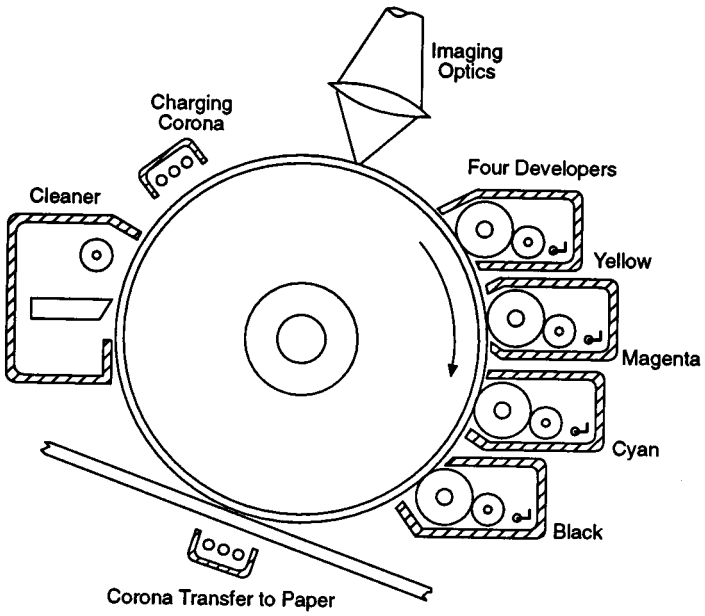


Figure 89.27 Four developers are used to produce color plus black. Four photoconductor drum rotations are needed to accumulate the four-color toner images.

be accumulated on the PC (see Fig. 89.27) or alternatively on an intermediate belt or drum, or even on the paper itself. A wide range of colors and the visual illusion of gray scale can be achieved by the use of laser pulse-width modulation and the use of "super-pixels" consisting of $N \div M$ arrays of binary pixels to provide digital half-toning.

Step 4—Transfer of Toner to Paper. Here a corona is used to charge the back side of the paper and the toner is transferred from the PC to the paper. Typically, some small fraction of the toner usually has a charge of the wrong sign. It is at this step where most of this wrong-sign toner is removed, minimizing "background" on the page, since primarily only toner with the correct sign gets transferred.

Step 5—Fusing Toner on the Paper. After transfer the toner is only loosely held (electrostatically) onto the paper; it must be fixed by fusing. There are several fusing techniques: (1) hot-roll using hot fuser rolls under pressure, (2) cold-roll, (3) solvent vapor, (4) flash lamp, and (5) radiant heating. Hot-roll fusing is predominantly used but all five techniques have been commercialized. The most efficient approach is hot-roll but all

of the thermal approaches require 0.5–2.0 J/cm^2 energy at the paper surface. The energy required is determined by the contact duration (related to the nip compliance and roll speed), pressure, paper water content, and the melt-flow rheological characteristics of the toner. Flash fusing has only been used in mid-range to high-end printers because expensive power supplies and capacitor banks are needed. A xenon flash lamp is typically pulsed for a few milliseconds to provide the required energy. Thermal efficiency improves with very short pulses, but undesired toner degradation volatiles increase with the higher temperatures produced. Care must be taken with radiant fusing to insure that excessive paper overheating does not occur if the paper is stopped under the incandescent lamp.

Step 6—Cleaning the PC. To restart the overall EP process, the PC must be cleaned of residual toner and contaminants. Fiber brushes, scraper blades, and "magnetic-brush" cleaners are used. For low-end printers, scraper blades suffice and are replaced (in some systems) with each new cartridge. Rotating soft-fiber brushes with air flow collection are used commonly in mid-range to high-end printers. Magnetic-brush cleaning is very similar to magnetic-brush development (see above).

Ionographic Printing

Ionographic printer devices also use powder toner technology and the overall technology is very similar to EP. One major difference, however, is the creation of the electrostatic image for subsequent toner development. The ionographic process uses an ion source (high-voltage drive electrode) and a page-wide array of control and screen electrodes to gate ions directed toward a drum coated with a thin-film dielectric material (e.g., Al_2O_3), thus creating a charged image on the drum. In present commercial ionographic printers, the transfer and fuse steps of EP are also replaced with a "transfix" process where the paper sheet is squeezed between the toner-developed dielectric drum and a compliant cold pressure roll. The ionographic technology has two fewer process steps (vis-à-vis EP) but the ion printhead represents significant technical challenges (cost and lifetime).

Magnetographic Printing

Magnetographic printing is also similar to EP but employs magnetic powder toner and a magnetic printhead. The magnetic head is an array of individually addressable magnetic write gaps, each representing a pixel location on a line across the page. The head writes a magnetic "image" on a belt or drum coated with a magnetic material such as γ-Fe_2O_3, Co:P or Co:Cr. Toner is attracted to those areas on the drum where flux reversals (and, hence, magnetic fields external to the the magnetic media) are present. The other process steps are essentially the same as used with EP but toner charge levels must be kept under control to ensure that electrostatic forces do not dominate over magnetic forces.

Electrostatic Printing

There are two basic approaches that have been developed: (1) one using a special dielectric paper, and (2) the other using plain paper. In the first approach, a page-wide linear array of electrode discharge pins are independently pulsed to charge the surface of a moving, conductive-base, dielectric-coated paper. Both powder and liquid toners may be used to develop the image; thermal fusing of the image is required for powder toner. With the plain paper version, either a precharged or uncharged dielectric drum or belt can be used depending on whether the array of electrode pins are used to charge or discharge the media surface. The other process steps are essentially the same as with the nominal EP printing process. Cost savings can be realized by multiplexing the electrode driver lines because there is a process threshold of hundreds of volts. This requires, however, that segmented counter electrodes be positioned behind the receiving surface (drum, belt, or paper).

Thermal Page Printing

Thermal transfer page printing is accomplished by using a stationary page-wide linear array of thermal heating elements coupled with a page-wide transfer ribbon roll (see Serial Thermal Printing below for discussion of thermal printhead technologies). The ribbon is positioned between the printhead and the receiving paper. The ribbon typically consists of a polycarbonate or polyester film substrate (10–20 μm) coated with a waxy ink layer. Heat from the thermal printhead elements must penetrate through the substrate to heat (melt) the dye/wax coating. The melted ink layer is pressed into the paper surface by printhead pressure, and after the paper and ribbon move together away from the printhead, the partially cooled ink layer adheres better to the paper than

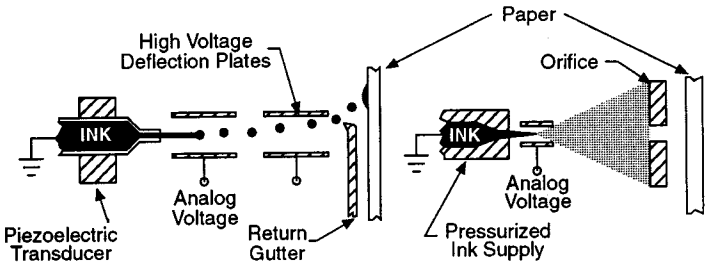


Figure 89.28 (a) Continuous drop ink-jet, and (b) continuous spray ink-jet technologies.

to the ribbon substrate. The ribbon is then peeled away from the paper, leaving the desired image on the paper. Considerable energy must be applied to melt the ink layer (e.g., 2–4 J/cm^2) which does restrict speed. Also, comparably smooth paper (≤50–100 ml/min Sheffield roughness) is required for quality printing with minimal gaps or voids. Print resolution of better than 8 pixels/mm can be achieved with smooth paper.

Sublimable transfer dyes have been used for high-quality colored images since the amount of dye transferred can be controlled somewhat by the energy supplied to the head heater elements. Also, since the sublimable dyes tend to penetrate into the paper sizing and fibers, a less smooth paper can be used than with waxy materials. The energy required for dye transfer is similar to that required for waxy ribbons.

Serial Nonimpact Printer Technologies

Ink-jet and thermal printers represent the two major classes of serial nonimpact technologies used for personal and workstation computer printer output. As shown in Table 89.7, the ink-jet technologies can be subclassified as continuous, piezoelectric, and thermal/bubble-jet. The thermal technologies can be subdivided into direct thermal, thermal transfer, and resistive ribbon.

Serial Ink-Jet Printing

Ink-jet technologies have evolved over the last two decades with continuous ink-jet being the first to be developed, followed by piezoelectric and then thermal or bubble-jet technologies. All three have been commercialized, although bubble-jet is by far the most popular today and continuous ink-jet is used primarily for speciality page printer or high-quality color graphics and image applications. Both piezoelectric and bubble-jet are "drop-on-demand" or "impulse" technologies, i.e., a drop is ejected from the printhead only when desired. With continuous ink-jet, a continuous stream of droplets is generated by the printhead and undesired drops are deflected electrostatically away from the paper (or vice versa). Ink requirements for these technologies are very demanding and the development of appropriate inks is as important (and difficult) as the development of the printer and printhead hardware. Demands on the ink are both extensive and conflicting, and include the requirements of nonclogging in the nozzle but fast dry time on the paper, water-based but water-resistant after drying on the paper, and quality printing on a wide range of papers (requires minimal feathering and controlled spot size). Nonaqueous solid inks have also been used; the ink is solid at room temperature but is liquid at an elevated head temperature.

Continuous Ink-Jet Printing. With this technology, ink is continuously jetted from a small-diameter nozzle under pressure (see Fig. 89.28). Although the resultant jet stream will naturally break into small drops, this phenomenon is assisted and stabilized by the inclusion of a piezoelectric perturbation transducer, driven at the desired drop rate. Lord Rayleigh was the first to determine that the *dimensionless instability factor*

$$\frac{g_r}{(\sigma/\rho d^3)^{1/2}} \tag{89.8}$$

(where g_r is the growth rate of jet instability, σ the surface tension, ρ the fluid density, and d the jet diameter) is maximized for $\lambda/d = 4.51$, where λ is the perturbation wavelength. Operation at this design point produces

very consistent drop breakoff and diameter. Operation at frequencies in excess of 100 kHz is possible with drop velocities of typically 25 m/s with a 50-μm nozzle diameter. Because the drops must be electrostatically charged at the breakoff point, conductive ink must be used. The ink source is typically grounded and a controllable voltage electrode is placed at the breakoff point (usually surrounding the jet stream). The charge level on each drop is then proportional to the applied voltage. Each charged drop can then be deflected by parallel downstream deflection plates with a field of typically 10 kV/cm. This deflection (assuming parallel plates and a uniform electric field between the plates) can be approximated by:

$$d_l = \frac{l_p}{2}\left(\frac{qE}{mv^2}\right)\cdot(2l_z - l_p) \tag{89.9}$$

where d_l is the deflection length, q the drop charge, m the drop mass, E the electric field, v the drop velocity, l_p the deflection plate length, and l_z the distance from the upstream end of the deflection plates to the paper. Drop deflection can be either binary or analog. With the former, the drop either reaches the paper or is directed into a collection gutter. With analog deflection, the drop may be deflected linear (e.g., any position over the height of a printed character).

Smaller satellite drops may be also produced between the primary drops. To eliminate their effect, the excitation system is designed to produce forward-merging satellites as these charge simultaneous with the preceding drop; thus the forward-merging (through "drafting") and subsequent drop coalescence will not alter the charge-to-mass ratio of the augmented drop.

Compensation must be provided for both aerodynamic and electrostatic interactions. Two primary examples are that (1) the first drop in a sequence of drops encounters much greater aerodynamic drag than subsequent drops, and (2) the charge on a drop is influenced by the charge on the previous few drops. One approach to greatly reduce electrostatic drop interaction and to stabilize merging effects is to include noncharged drops between charged drops. This obviously reduces the effective drop rate by a factor of two and requires a design configuration where only charged drops reach the paper. Charging electrode voltage adjustment algorithms based on voltages applied to prior drops are also used to reduce the electrostatic interaction.

An alternate approach is the *continuous spray* design where a smaller nozzle (10–20 μm) is used. Much smaller drops are produced at higher velocities (≈40 m/s). It is often called the Hertz method [see Fig. 89.28(b)]. As with the *continuous drop* approach, a controlled voltage electrode is positioned around the breakoff point of the jet stream, a conductive ink is used, and the smaller droplets are charged proportionally to the applied voltage. The stream of droplets is directed to the paper when no voltage is applied. When a voltage is applied, the resulting electrostatic charge on the droplets produces strong mutual repulsion forces and the stream transforms into a spray, the cone angle of which is determined by the applied voltage. The spray is intercepted by a collecting surface surrounding the collection orifice which allows only uncharged or low-charged droplets through to the paper. With analog voltage control, the amount of spray that passes through the orifice can be varied, thus providing gray-scale capability. This approach (with multiple orifices) has been commercialized for very high-quality color image and graphics applications.

Piezoelectric Ink-Jet Printing. Piezoelectric ceramic transducers are employed with this technology. These materials (e.g., lead zirconate titanate and barium titanate), when polarized, change their physical dimensions when subject to an electric field—usually applied through surface electrodes. Deflections of several angstroms per volt are typical. When the transducer is pulsed with a voltage, the deflection generates a pressure wave in an adjacent ink chamber, resulting in the ejection of a single drop—hence the descriptors, "impulse" or "drop-on demand". Four implementations are shown in Fig. 89.29. Often arrays of these devices are integrated into a serial printhead which allows the printing of a one-character-high-per-head pass across the paper. Color printing can be accomplished by assigning one or more nozzles per color.

Only a very small (100–1000 Å) deflection of the piezoelectric transducer is needed to create the ink chamber pressure wave if the displacement is very rapid (10–100 μs). Hence, these devices are very efficient; only a few microjoules per drop are required for robust operation. Drop ejection rates of over 20 kHz have been demonstrated in the laboratory but commercial devices are typically designed for the 5–10 kHz range. Drops that produce a

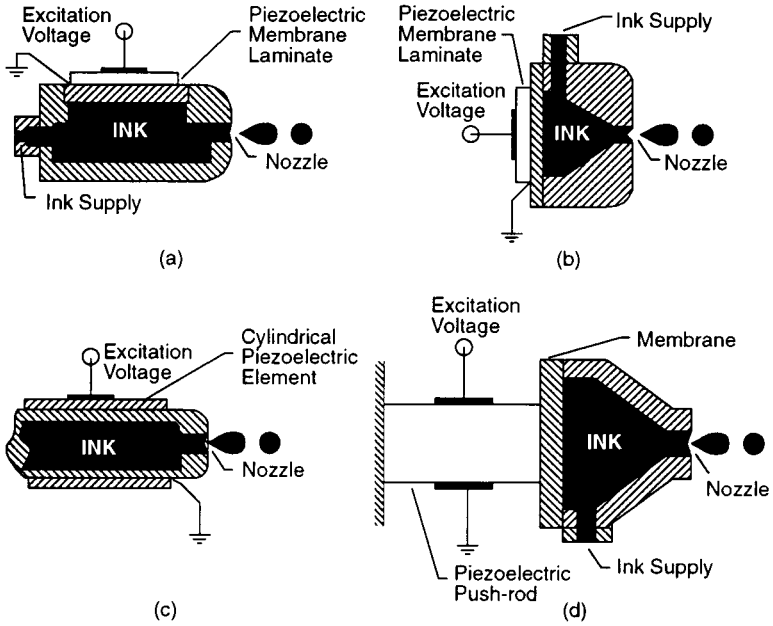


Figure 89.29 Four approaches to piezoelectric drop-on-demand ink-jet technology: (a) piezoelectric/membrane laminate design, (b) "oil-can" version of (a), (c) piezoelectric squeeze-tube approach, and (d) piezoelectric push-rod method.

spot size of 150–200 μm on paper can be achieved with an orifice diameter of about 50 μm. Even greater device efficiency can be obtained by synchronizing the arrival of a direct wave and a reflected wave at the nozzle (e.g., a negatively reflected wave caused by an initial expansion pulse plus the direct wave from a following compression pulse). These two waves can be made to reinforce one another at the time and place of drop ejection.

Figure 89.29(a) shows a piezoelectric/membrane laminate design with a disk-shaped piezoelectric transducer which, upon being pulsed, deforms (assumes a finite radius of curvature) and thus creates a pressure wave in the ink chamber. Figure 89.29(b) illustrates an "oil-can" version of the laminate approach. In Figure 89.29(c), a tubular transducer is used to squeeze the ink chamber. Figure 89.29(d) represents a push-rod or piston design where the piezoelectric material is used in the extensional mode. Attempts have also been made to use modern semiconductor planar/etching processing techniques to create low-cost arrays of devices.

Bubble-Jet Printing. This technology has also been called thermal-jet. With this approach, very small thermal resistors on the ink chamber wall are electrically pulsed. Joule heating of the resistor causes the temperature of the ink adjacent to the heater to rise to 350–400°C (see Fig. 89.30). Because the ink becomes locally superheated, nucleation of tiny bubbles takes place on the surface over the heater. These bubbles coalesce and very rapidly form a single expanding bubble which, by displacement (like a piston), propels a single drop of ink out at the orifice. The electrical pulse must be short (typically 3–6 μs) to insure low conductive heat losses; however, the power density is extremely high (~500 MW/m^2). The energy applied per drop is 30–50 μJ but only a small fraction (a few percent) represents the kinetic energy of the drop. The remaining energy is thermally dissipated in the ink and device structure.

When the thermal energy in the superheated layer is depleted, the bubble begins to collapse. The total cycle (nucleation plus bubble growth and collapse) is normally complete in about 20 μs. Drop rate, however, is typically limited to less than 10 kHz, mainly because of the limits of thermal dissipation. Cavitation damage can occur to the heater structure if the bubble collapse is too violent. Proper design of the ink chamber geometry can provide the necessary damping and minimize this problem. Heater element materials used for this technology include HfB_2, ZrB_2, Ta_2Al and TaN. Passivation over-layers (e.g., SiC, SiO_2, Si_3N_4, plus certain metals such as tantalum) are deposited on top of the thin-film heaters to provide protection from chemical and mechanically enhanced corrosion. Materials, structures and thin-film deposition processes must be carefully

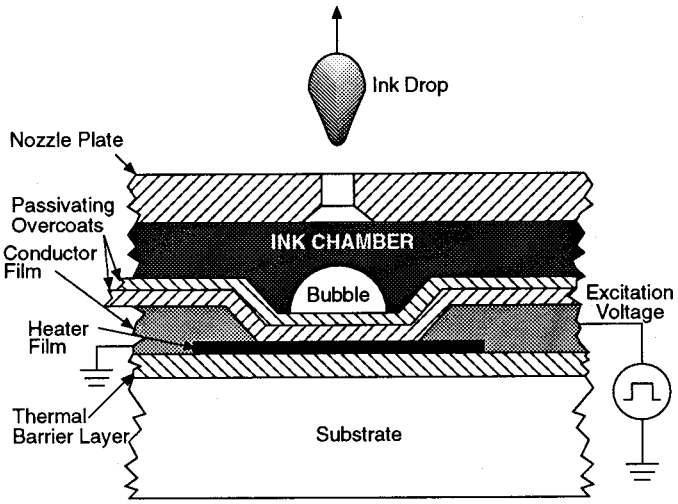


Figure 89.30 Thermal-jet device showing one heater-nozzle package. Joule heating of the thin-film heater causes nucleate boiling in the ink adjacent to the heater. A bubble forms and pushes a drop of ink out of the nozzle.

designed to allow billions of heater pulse cycles without failure at a local power density of 500 MW/m^2. Also, very special ink must be used so that very minimal chemistry occurs at the hot interface layer, e.g., so that the chemicals in the ink do not break down and form a thick carbonaceous film over the heater. These carbonaceous films, when more than a few hundred angstroms thick, destroy the device velocity and thermal efficiency characteristics. The ink must also not etch the overcoats.

Very low-cost, compact, low-power printers have been developed based on this technology because energy requirements are very low and the printhead can be made inexpensively using semiconductor-like planar processing techniques. A proliferation of these printer products (many with color) have appeared in the market over the past five years.

Serial Thermal Printing

Thermal technologies have been used in many low-cost serial printer applications. There are three basic approaches: (1) use of a heat-sensitive special paper, (2) thermal transfer of ink from a ribbon to paper, and (3) a variant on (2) with a special ribbon structure and printhead that improve thermal efficiency and allow printing on a wide range of standard papers (including the relatively rough office bond papers).

Direct Thermal Printing. The key aspects to this technology are the thin-film resistive serial printhead and the special paper required. A typical array structure (see Fig. 89.31) consists of photolithographically defined and deposited heater material (e.g., Ta_2N or TaAl) on a contoured insulating substrate. One or more protective layers (e.g., SiO_2, Ta_2O_5 or SiC) are deposited on top of the heater layer for abrasion resistance and electrical insulation. Contours (raised areas or bumps) on the surface under the heater areas are constructed by a raised glass "glaze" of about 40 μm and provide both improved contact with the paper and short time constant thermal insulation to the substrate. This structure is often called the *thin-film head structure.* Alternates include (1) the *silicon mesa technology,* where a two-dimensional array of silicon mesas is fabricated from monolithic silicon, and (2) the *thick-film technology.* With the former, each mesa contains its own resistor-transistor/diode on the base of the silicon chip. Joule heating occurs when voltage is applied to the resistor via the transistor/diode. With the thick-film approach, a resistor

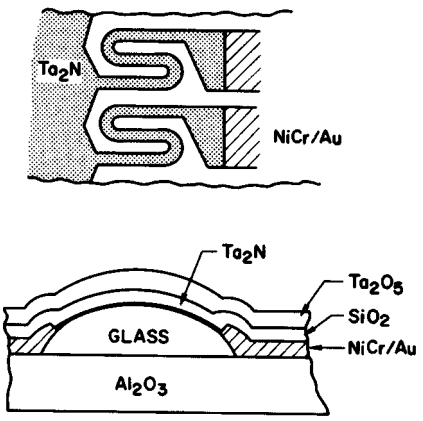


Figure 89.31 Thin-film thermal head structure. (*Source:* R.C. Durbeck and S. Sherr, Eds., *Output Hardcopy Devices*, San Diego, Calif.: Academic Press, 1988, p. 282. With permission.)

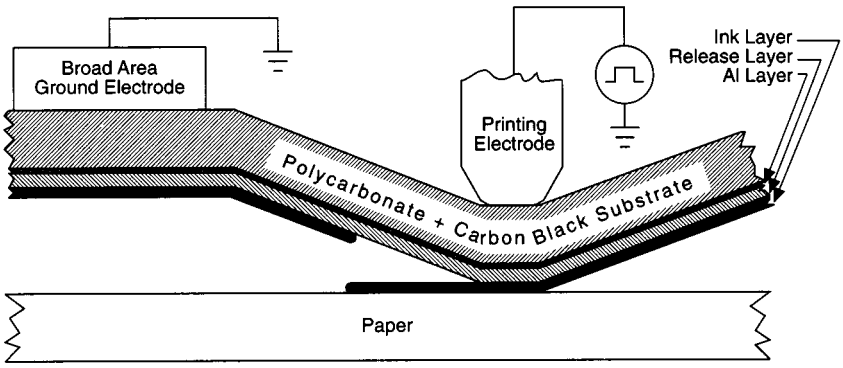


Figure 89.32 Resistive ribbon head and ribbon system.

paste is typically screened onto a ceramic substrate. Materials for the resistor paste include borosilicate glass and lanthanum glass-ruthenium oxide. The elements are typically 25 μm thick and slowly wear down by contact with the paper. Since, with each approach, the resistive elements must be allowed to cool before the head is ready to "write" the next pixel position on the paper, the heating and cooling time constants of the head primarily determine the attainable print speed.

Special thermal paper typically has a coating with a leuco dye plus a phenol developer in a polymer binder which react at the elevated temperature provided by the printhead to form a colored species. The resulting optical density is a function of the temperature reached so that some gray-scale capability is possible. Energy density requirements are in the range of 2–6 J/cm^2.

Thermal Transfer Printing. The desire to eliminate costly special thermal papers even for low-cost serial printers led to the development of the thermal transfer printing technology where a separate transfer ribbon is interposed between the thermal printhead and the paper (see Thermal Page Printing above for discussion of ribbon technology).

Resistive Ribbon Printing. This technology grew from the need to provide quality thermal printing at higher speeds and on a wide range of papers, including standard office bonds. With this technology, the heating function is repositioned from within the printhead into the ribbon itself. The ribbon (see Fig. 89.32) includes an additional aluminum conductive and heating thin-film layer, sandwiched between the "standard" thermoplastic ink layer and the ribbon substrate. The substrate is made conductive (350–900 Ω/□) by the incorporation of sufficient (20–30%) carbon black in the polycarbonate film. Joule heating occurs as current flows from the addressable head pixel-size electrodes through the aluminum layer to the large single-ground return electrode. Current density is maximum directly beneath the addressable electrodes so that temperatures there (but not elsewhere) are sufficient to cause ink layer melt and transfer. In addition, a thin release layer can be added between the ink layer and the aluminum heating layer to provide enhanced and smooth-edged pixels; this provides sharp character edges and allows more detail in images.

This technology allows (1) 2–3 speed increase (vis-à-vis conventional thermal transfer printing) because printhead thermal time constants are much less of a factor, (2) very high-resolution printing (up to 40 pixels/mm), and (3) printing on office-quality bond papers. Also, already deposited ink can be selectively lifted off the paper by reducing the printhead power and using a fresh ribbon area. Offsetting these advantages is the fact that the ribbon is more complex and costly than conventional transfer ribbons.

Impact Printer Technologies

The earliest electronic printing technologies were impact devices employed on the early plug-board programmable accounting machines that used punched-card input. Before that there were typewriters and teleprinters. *Line impact* printers are used on mid-range to high-end computer systems; their continuing appeal is based first on the ability to print multipart forms, and secondly on lower cost of printing and greater reliability (as compared with electrophotographic printers). *Serial impact* printers historically represent the greatest sales

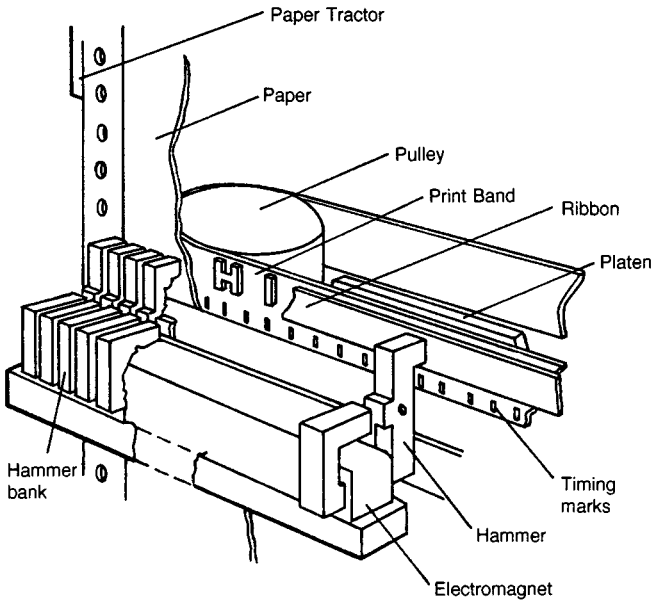


Figure 89.33 Line impact printer mechanism. (*Source:* R.C. Durbeck and S. Sherr, Eds., *Output Hardcopy Devices*, San Diego, Calif.: Academic Press, 1988, p. 130. With permission.)

volume for all computer printer technologies by a wide margin. By far the largest volume has been for personal computer output applications. Their appeal stems from products offering very low cost (device and supplies), high reliability and, again, multipart forms. More recently, however, thermal ink-jet and low-end EP printers have largely supplanted these impact products.

Line Impact Printer Technologies

As shown in Table 89.7, there are three distinct technology approaches for line impact printers: (1) *fully formed character*, (2) *dot band matrix*, and (3) *shuttle hammer matrix*. The first method dominates in high-end computer applications. The other two approaches are primarily used with mid-range computer systems.

Fully Formed Character Printing. Various line printer technologies have evolved over the past but the most popular extant technology makes use of a band of etched engraved characters (plus timing and position marks) in a continuously moving loop configuration (see Fig. 89.33). This band is positioned between a page-wide bank (typically 132) of hammers-actuators and the paper/ribbon set. With timing and position marks, the location of the characters (replicated several times on the band loop) can be easily tracked and the appropriate hammers can be asynchronously fired to provide a full line of print in tens of milliseconds. A limit to the throughput P_r (lines per minute [lpm]) of such a printer may be approximated by:

$$P_r = \frac{V_b}{V_b T_i + P_c N_c} \tag{89.10}$$

where V_b is the band velocity (in./min), T_i the time to increment the paper to the next print line position (min), P_c the character pitch (in.), and N_c the number of characters in the character string. This relation assumes that the hammer settle-out time is less than the paper advance time, the latter normally not the limiting speed factor. The usual limiting speed factor is related to print quality, i.e., well-registered, sharp-edged, high-optical-density characters. In addition, because not all lines on a page are usually printed, with line skipping the actual throughput may be significantly faster than indicated above.

Several factors are important to achieve high-quality print characters. First, the hammer-actuator must provide enough impact force and energy to produce optically dense print. Second, it must impact the ribbon/paper set at

just the right instant to provide good character registration. Third, it must have a short impact time. The impact force and energy are determined by the kinetic energy transferred by the hammer; this is the main factor in determining the amount of ink transfer to the paper. The hammer energy is typically 4–8 mJ. The total time from electronic print impulse to impact with the ribbon/paper set must be closely controlled to achieve good character registration, increasing in importance with band loop speed. The flight time variance must not be more than 1.7 μs to yield a print registration error of no greater than 0.05 mm with a loop speed of 30 m/s. This can best be done by automatically and periodically measuring the time to impact (using a piezoelectric impact bar) and then making microcode/electronic impulse delay adjustments. To minimize "slur" (i.e., the blurring of a character caused by lateral relative motion between the engraved character and the ribbon/paper set during impact), short impact time is important. It may be shown that this impact time is inversely proportional to the square root of the hammer mass (for given hammer kinetic energy), and inversely proportional to the hammer velocity (for given ribbon/paper set thickness and compliance). For a given amount of acceptable slur, the hammer velocity must increase proportional to the band loop velocity, and the hammer mass (for fixed kinetic energy) must decrease—inversely proportional to the square of the band loop velocity.

Hammer-actuator systems have been designed with one, two and even three moving piece components. With the one-piece design, the mass impacting the paper is largest, limiting this design to slower printers (≤650 lpm). To have less mass in the hammer impacting the ribbon/paper set, separate parts are used for the armature and the impacting hammer (two-piece design). In this case, most of the kinetic energy from the pivoting armature is transferred to the hammer without the large mass handicap of the ferromagnetic armature. Of course, the residual energy in the armature must be absorbed and dissipated within the printhead. Further design improvements are possible by interposing a push rod between the the armature and the hammer pieces (see Fig. 89.34). Since the lengths of adjacent push rods can be made alternately short and long, close packing of these actuator assemblies having a single pivot axis is possible. With the most efficient designs, printing speeds of over 5000 lpm are possible.

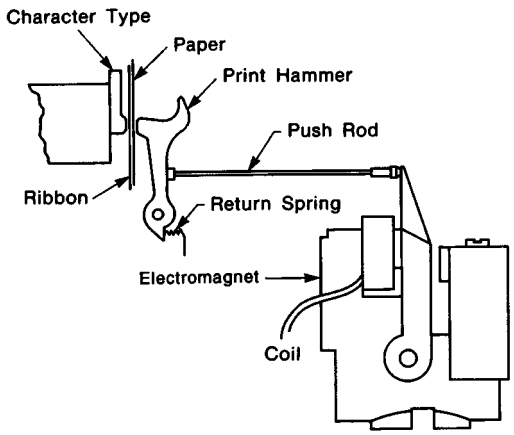


Figure 89.34 Three-piece hammer-actuator design. (*Source:* R.C. Durbeck and S. Sherr, Eds., *Output Hardcop Devices*, San Diego, Calif.: Academic Press, 1988, p. 143. With permission.)

Both moving-coil and stored-energy actuators are used, the latter using a "bucking" coil and a stored-energy flexible spring. The bucking coil, when energized, cancels the magnetic flux from a permanent magnet, holding the actuator in the "cocked" position, and thus converts the stored spring energy to kinetic energy.

Dot Band Matrix Printing. This technology has been employed on some printers used for mid-range computer systems, and, in a sense, is a hybrid technology combining attributes of high-end line band printers with features found in serial wire matrix printers (see below). Present designs also use a moving metal band but incorporate pixel-size raised bumps instead of etched characters. These bumps (typically 120 on a band) are positioned at the apex of chevron-shaped springs etched into the band (see Fig. 89.35). Also, timing/position slots are etched into the band above the spring slots. Instead of one full line of characters being printed in one cycle, one full line of dots is printed with this technology. Also, each hammer covers typically three character positions across the page; thus, 45 hammers can cover a page-width of 135 character positions. Hammer cycle time is typically 1.2 ms and the band speed can be 0.28 m/s or higher. After a line of dots is printed, the paper is advanced *N* times until a full character height is achieved.

This technology is much less expensive than typical engraved band technology, and allows the use of a faster draft mode using fewer dots per character; however, normal throughput (lpm) is generally much less.

Shuttle Hammer Matrix Printing. Transverse shuttling a reduced number of horizontally spaced hammers to cover a full print line is another way to implement matrix line printing. Here, a pixel-size raised bump is

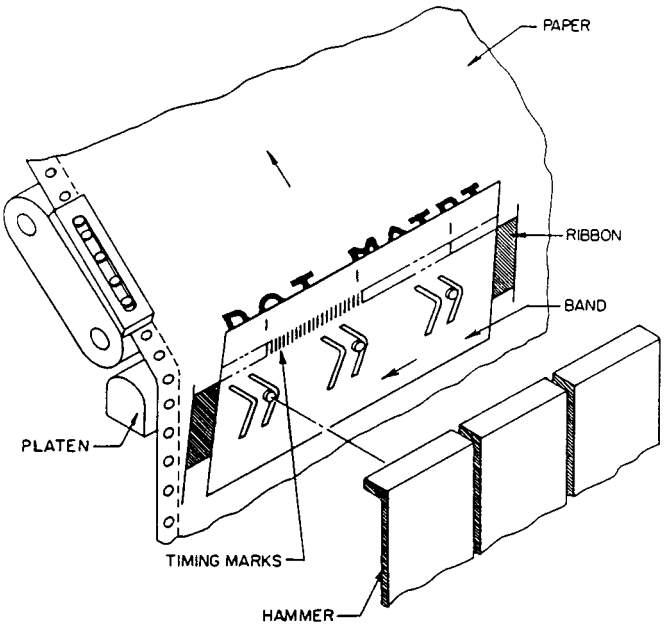


Figure 89.35 Dot band printer technology. (*Source:* R.C. Durbeck and S. Sherr, Eds., *Output Hardcopy Devices*, San Diego, Calif.: Academic Press, 1988, p. 198. With permission.)

incorporated into the strike surface of each hammer. One design includes 33 hammers (each spaced 1 cm from its neighbor) covering 132 character positions. With an oscillatory shuttle motion of 1 cm, all line dot positions are covered, and the shuttle return time can be used for paper advance. Moderate printing speeds (600–900 lpm) can be accomplished with $^1/_4$-cm hammer spacing; 300 lpm is possible with 1-cm spacing. Opposite moving counterweights have been employed to reduce shuttle vibration.

Serial Impact Printing

There are two classes of technologies presently employed for serial impact printing: fully formed character and wire matrix. The former category includes both daisywheel and typeball technologies, the latter now in little use for computer output except for a few dedicated word processing applications. Daisywheel technology is used in many dedicated word processing systems and for some personal computer systems. With these two technologies, font change can be implemented by simply changing a wheel or ball, but all-points-addressable printing for image and graphics is basically not possible. Serial wire matrix technology has for years been the most popular output device for personal computers, but has now lost most of its market share to low-end laser printers and thermal ink-jet.

Daisywheel Printing. This technology makes use of a "petal-like" rotating print wheel with typically 96 spring-like fingers or "petals" radiating outward from a central core. An engraved raised character is present at the end of each petal. The wheel is rotated by a dc servomotor to the desired character. A single hammer-actuator then impacts the back side of the petal, forcing the raised character into the ribbon/paper set. The simplest designs have a constant carriage speed and a fixed time delay to allow for up to 180° rotation of the wheel between strokes. These systems are typically limited to about 30 characters per second (cps). More modern designs have incorporated microcode logic which slow down the carriage motor when the next character position is more than, for example, 30° away. This can increase the net average speed to as high as 60 cps.

Variants on this design use thimble- and cup-like rotary devices with petals bent parallel to the axis of rotation. This lowers the rotational inertia and allows for more petals (and hence more characters—up to 128).

Typeball Printing. Typeball technology was first developed for typewriters in the 1950s, but has also been used for low-speed, correspondence-quality printers. The golf ball-size sphere, rotated and tilted to reach the

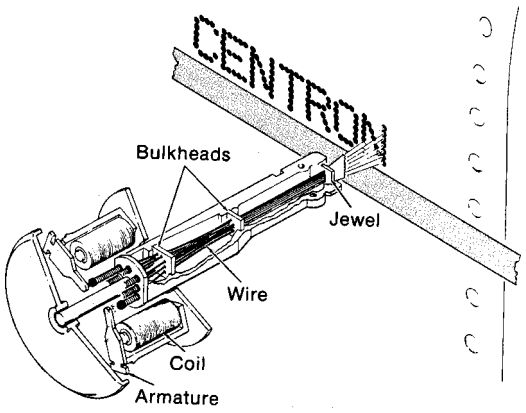


Figure 89.36 Wire matrix serial printhead. (*Source:* R.C. Durbeck and S. Sherr, Eds., *Output Hardcopy Devices*, San Diego, Calif.: Academic Press, 1988, p. 187. With permission.)

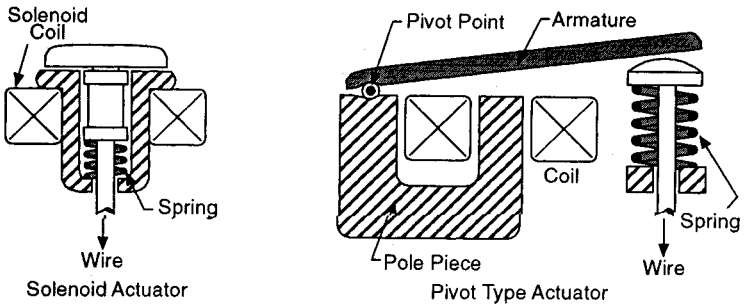


Figure 89.37 Solenoid and pivot-type actuators for serial wire matrix printing.

desired character position, has four rows of 11 characters, repeated on both hemispheres, making a total of 88 character positions. The usual operating speed is 10 cps but up to 15 cps has been achieved.

Serial Wire Matrix Printing. This technology employs an array of guided wires (often tungsten) that are individually driven into the ribbon/paper set (see Fig. 89.36). The array may form a single-row configuration at the plane of ribbon impact, may have two in-line rows, or may have two staggered rows. Nine wire designs became very popular in the 1980s for low-cost personal computer systems, but the demand for higher print quality has moved the "standard" design point to 24 wires. Early technology limited wire cycle repetition rates to about 500 Hz. More advanced designs can perform at 2500 Hz, but most available products operate in the 1000–1500 Hz range.

Actuators to drive these wires are quite robust and can print through many layers (up to 6) of a multipart form. The kinetic energy needed to print through 4–6 layers is normally about 0.5 mJ for 200-μm diameter wires, 1.0 mJ for 300-μm wires. For 2-layer forms, these energies drop to about 0.3 mJ and 0.45 mJ, respectively. There are two common approaches to driving the guided wires. One is a pivot-type actuator and the other is a direct solenoid design (see Fig. 89.37). The magnetic circuit in all cases must be designed to maximize either (1) the moving wire kinetic energy in the case of a free flight wire configuration or (2) the kinetic energy of the combined armature and moving wire in the case where these two components are either permanently connected or remain in contact during the drive cycle.

Other actuator designs use a stored energy approach where a preloaded leaf spring is held by a permanent magnet until released by a "bucking coil" to counteract the static flux. Also, experimental stacked piezoelectric transducers have been investigated using high lever ratios (e.g., 30:1) which have operated at 3 kHz and above. Synthetic ruby or ceramic wire guide holes are sometimes used to combat wear. Several hundred million cycles lifetime operation for each wire is typically required for today's devices.

Higher draft-quality speed (e.g., 200 cps) can be achieved without decreasing wire cycle time by simply increasing the carriage speed. Higher than normal print quality can be accomplished by both slowing down the carriage and interlacing horizontal rows of dots. This near letter-quality printing is usually at greatly reduced speed (e.g., 48 cps). Most printers also print during the carriage return. Color can be produced by shifting four-color ribbons.

Defining Terms

Line printing: A printer prints one full line width of characters or dots at a time. The paper is then moved into the next print line position, ready for the next line of characters or dots. The printer may pause or stop between lines. Printing speed is often given in units of lines per minute (lpm).

Page printing: The information to be printed on a page is electronically composed and stored before shipping to the printer. The printer then prints the full page nonstop. Printing speed is usually given in units of pages per minute (ppm).

Pixel: The nominal printed spot area or "picture element" addressed by a particular printing device. It is sometimes called "pel."

Serial printing: Printing is done one character at a time. The print head must move across the entire page to print a line of characters. The printer may pause or stop between characters. Printing speed is usually given in units of characters per second (cps).

Related Topic

89.1 Input Devices

References

R.C. Durbeck and S. Sherr, Eds., *Output Hard Copy Devices*, San Diego, Calif.: Academic Press, 1988.

J. Heinzel and C.H. Hertz, *Advances in Electronics and Electron Physics,* vol. 65, P.W. Hawkes, Ed., San Diego, Calif.: Academic Press, 1985, pp. 91–285.

L.B. Schein, *Electrophotography and Development Physics*, Berlin: Springer-Verlag, 1988.

J.M. Sturge et al., Eds., *Imaging Processes and Materials,* New York: Van Nostrand Reinhold, 1989.

D. Winkelmann et al., *Ullmann's Encyclopedia of Industrial Chemistry*, vol. A13, Weinheim, Germany: VCH Publishers, 1989, pp. 571–660.

Further Information

Proceedings of the SPSE and IS&T International Congresses on Nonimpact Printing, published by the Society for Imaging Sciences & Technology, 7003 Kilworth Lane, Springfield, VA 22151.

Proceedings of the Society for Information Display, 1980–1996, published by The Society for Information Display, 1526 Brookhollow Drive, Suite 82, Santa Ana, CA 92705–5421.

89.3 Smart Cards

Witold Suryn and Michel Veillette

The smallest laptop in the world.

That is the frequently used expression describing intelligent chip cards. The word "intelligent" plays an important role, as there are also other, very popular chip cards that can hardly be named that way — memory cards containing defined amounts of available memory and some access control logic. One can find them today as the most consumed commodity articles: prepaid services cards. A good example would be a telephone card, preloaded with some electronic "money" when manufactured, debited each time one makes a phone call and, finally, thrown out when emptied. Even if interesting as technology solution this type of card won't be a subject of next chapters as this article is rather dedicated to intelligent members of chip card's family — Smart Cards.

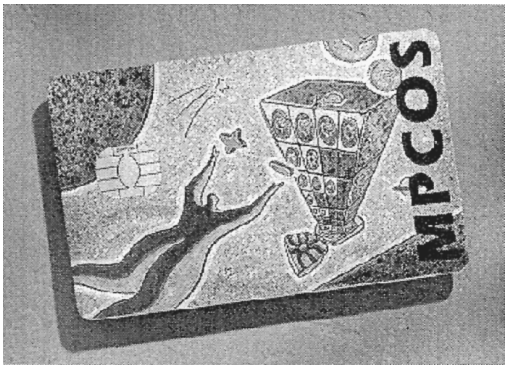


FIGURE 89.38 Pocket computer — Smart Card (courtesy of Gemplus Corporation).

The history of this device is long with plenty of defeats. Invented in 1974 by Roland Moreno in France (there were also earlier announcements: in 1968 in Germany and 1970 in Japan), it was refused any right of recognition as a "usable invention." These years the world was dominated successfully by a still-new fashion — plastic cards with a magnetic stripe.

The very important questions are: did such a status quo change? Where? And why?

Depending on the region of the world, smart cards replaced magnetic-stripe cards more or less successfully. In Europe, you could hardly find a single magnetic stripe card in use, while in America, intelligent chip cards (ICC) still experience their childhood. The reason for such a worldly footprint of ICC technology is due to the market's demand for security measures as well as for diverse functionality. Last, but not least, the costs of on-line telecommunication services should be mentioned. Magnetic stripe offers about 140 bytes of available memory, almost nonexistent security, and in most cases, requires costly on-line service while ICC may be used in both on- and off-line environments, grants a very high, sophisticated level of secure operation and, finally, allows multifunctionality.

Could it be explained more vividly? Yes, if you make comparison between a piece of magnetic tape and a computer. Because *a smart card is a computer.*

In following chapters, we will attempt to present to the reader a complete overview of the world of smart cards. Beginning with internal hardware architecture through card software applications and ending by complete system solutions, we hope to prove the exceptionality of ICC technology both from technical and applicability points of view.

Hardware Architecture

Year 1998 state of the art recognizes two main hardware architectures of the ICC: 8-bit and 32-bit microcontrollers. Why not 16-bit? Because silicon chip technology goes faster than "bits-and-bytes" maturity fights of smart card software developers. 16-bit did not catch its historical moment, however, from time to time there are successful attempts to re-use this technology. This statement may sound strange for nowadays' software programmers used to easy "consumption" of megabytes of memory but for those who began their developer careers in early 1980s, the mastership of putting every required functionality in magic "64k" number of bytes must sound familiar.

8-bit Microcontroller Based ICC

The architecture presented in Fig. 89.39 is one of the most obvious schemas known in computer science. Things become less obvious when put in ICC context. What are the roles of blocks forming ICC when compared to "full-blooded" computer?

- Microprocessor is a typical CISC (Complete Instruction Set Computer) 8-bit unit driven by an external clock of 3.57 MHz frequency you might find in early standalone systems.
- I/O — typical input/output serial communication device.
- RAM — the same memory you find today in any type of a machine.

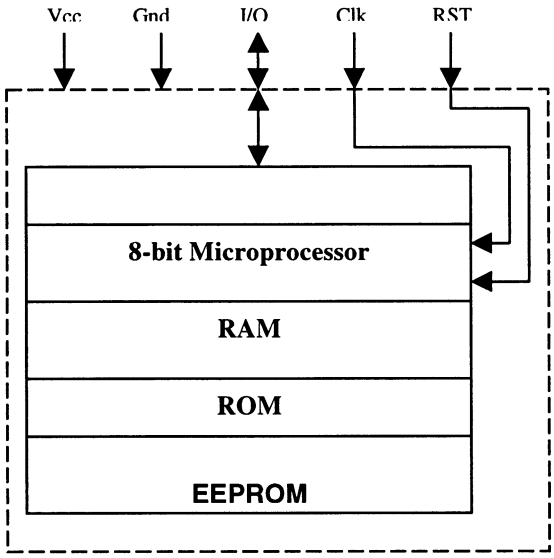


FIGURE 89.39 The general hardware architecture of 8-bit ICC, where:
Vcc — +5V power supply line
Gnd — 0V (ground) line
I/O — bi-directional communication line
Clk — external clock line
RST — external reset signal line
RAM — random access memory
ROM — read only memory
EEPROM — electrically erasable programmable ROM

- ROM — system kernel in ICC, BIOS in PC (observe the difference).
- EEPROM — ICC's hard disk.

It may happen that some readers now feel a bit surprised, but the surprise will become bigger when we look at "numbers."

By default, 8-bit processor maximum memory space is 64 kilobytes (2^{16}), but in most manufactured ICCs that space is not fully used. Why so, if even 64k seems to be today less than nothing? The answer is simple: price. The device being sold yearly in hundreds of millions must be cheap, but memory forms a considerable part of the overall cost of the chip. So, as much software as possible is needed in as little memory as a developer can possibly accept.

Let's then come back to numbers: ROM — average 8k, RAM — 256 to 1024 bytes (no, we did not forget "kilo"), EEPROM — 4 to 32k (we can imagine the ironical smile of the reader — 32 kb of hard disk).

And that's the whole wealth that the developer has for his needs. Now, dear reader you can understand why developers fight for every bit. Sometimes even for a half of it. And why the ICC software world is so specific, closed and exacting.

The last important feature of chips being used is the voltage of power supply (Vcc). The most commonly used voltages are 5V and 3V, however, lower (1V, 0.5V) voltages are being strongly considered.

Let's have a quick look at the industry. Most of manufactured ICCs are based on one of two industrial patterns: Intel's I8051 microcontroller and Motorola's 6805 one. These patterns were broadly accommodated by the world's biggest hardware manufacturers like Siemens, Hitachi, SGS Thomson, Texas Instruments and of course — Motorola. These five names are the leaders of the smart card industry.

32-bit Microcontroller Based ICC

Thirty-two bit microcontrollers are the discovery of the past year. Still new and fresh, they attract card manufacturers' attention by its potential processing power and much larger memory space.

The internal structure of the chip is very similar to this shown in the Fig. 89.39, however, some differences may be observed:

- Microprocessor is 32-bit RISC (Reduced Instruction Set Computer) type.
- EEPROM may be replaced by faster and easier to re-program Flash ROM.
- Two I/O lines.
- In newest releases, ordinary RAM is being replaced by FeRAM (20 times quicker if used in 8-bit chip, several hundred times in 32-bit chip).
- Clock rate is 33MHz (for Vcc = 5V).
- Memory is split into Banks (pages).

Now again numbers: ROM — average 56kb, RAM — 2kb, EEPROM — up to 64kb. The first impression we might have is that such a machine should give the output several tens of times better than its 8-bit sister. Surprise! The recent benchmarking based on Java Card applications (Java Card will be discussed later) have shown that the performances of both cards may be comparable, but not the price, as the 32-bit card is several times more expensive. So why go in this direction?

There are many reasons among which some are crucial. Although the 8-bit solution has almost reached its top performance, that does not mean that it'll vanish overnight. As in the world of cars, there are fans of Rolls Royce, Cadillac, and Geo Metro which exactly reflect the future of 8-bit cards. Many existing and incoming applications will still be satisfied by an 8-bit solution, offering to card issuers and clients something fulfilling their needs and at the same time remaining affordable. The second crucial reason is 32-bit technology maturity being on its learning curve. In certain, software developers as well as chip manufacturers will reach the level of expertise high enough to push prices from the "Cadillac" level to the "Geo Metro" level. The 8-bit card shall survive even if in the "bicycle" position. The market continuously creates a third important factor: multi-application requirement. Today most of us have his/her wallet filled up with dozens of credit and/or debit cards, ID cards, healthcare cards, loyalty program cards, etc. Card issuers try to attract their customers by a vision of merged functionalities installed on one card, requiring a need for more sophisticated technology. The perception today of such a solution directs issuers' as well as card manufacturers' interest toward 32-bit technology. There is still a future for now since there are few 32-bit chip manufacturers known: NEC, Philips, Siemens, and Hitachi, however, the rest of world's biggest players are close to launching their modern products.

Contact ICC, Contactless ICC

The distinction of contact and contactless cards brings us on the next level of smart card technology. "Contact" or "contactless" define the method in which the card communicates with the external world. There are cards combining both methods sometimes called "combi" cards.

Contact ICC (Fig. 89.38) is equipped with a small field of ohmic contacts to which all necessary signals are being applied. To operate such a card, a reader would have to possess the ability of physical touch.

Contacless ICC uses "untouchable" media such as microwave transmission, optical transmission, capacitive coupling and inductive coupling, but — due to physical constraints — the latter one is most commonly accommodated.

Before we proceed further, two facts should be strongly expressed: there is no power supply on board the card and the transmission is unidirectional (terminal to card). In such a case, four problems arise:

- How to power up the card.
- How to transmit clock.
- How to transmit data to the card.
- How to transmit data from the card.

From a purely electrical point of view, the solution seems to be simple: both card and terminal are equipped with coupling loops (electrical coils) allowing transmission of energy similar to a coreless transformer (Fig. 89.40). The card's loop is embedded in a plastic body to avoid getting destroyed or touched.

The main carrier is a sinusoidal magnetic field (f = 4.9152MHz) that transmits the power supply to the card. To transmit the data/clock to the card, the terminal uses frequency modulation of the main carrier, being easily

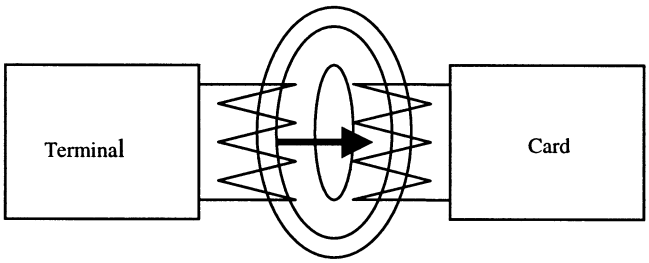


FIGURE 89.40 Card-terminal communication principle.

filtered and decoded by the card's I/O device. Transmission of the data from the card to the terminal is based on the passive amplitude modulation of the main carrier, which is executed by changing the card's I/O circuit impedance differently for "0" and "1." To read the information, a reader must be able to detect a potential drop in the impedance in its internal AC generator.

Inside the contactless type of cards there are two sub-groups recognized: close coupling cards and remote coupling cards. The distinction between the two is based on the distance terminal-card and the power that may be transmitted to the card.

Regardless of which communication carrier is used, the transmission speed between the card and the terminal is, in most cases, one of the following: 9600 bit/s, 19,200 bit/s, or 115 kbit/s.

Operating Systems

Operating system (OS) is the next layer of technology that makes the microprocessor card a *smart* card. The term *operating system* applied to a smart card is in some sense misleading as in many cases even highly experienced programmer/analysts would not be able to find a clear borderline between "system" and "applicative" layers. Due to continuous memory constraints smart card software designers chose the formula of "nested functionalities" that finally resulted in placing real functionality inside OS services (or the opposite, depending on your point of view). Each card manufacturer has its own family of operating systems offering different dedicated functionalities to the user that are characteristic of the specific manufacturer. That type of OS forms a vast group of software products called proprietary OS.

In the mid-1990s, another approach to smart card OS emerged: open OS. By its definition the open OS is quite similar to familiar DOS with the exception of being placed in the card's ROM instead of the computer's hard disk.

For either of them some principles apply:

- The stable part of software application is placed in ROM during the **masking** (chip manufacturing) process. For proprietary OS, it makes all the functions regardless of whether "system" or "applicative" form the basic dedicated platform. For open OS, in most cases ROM keeps, at a minimum, basic OS.
- The changeable part of software application is always loaded to EEPROM. In both cases, EEPROM keeps "add-ins" stored in ROM. For proprietary OS, this is the **"filter"** consisting of either additional functionalities or corrective patches. For open OS, real functional software is stored. EEPROM contents may be loaded to the card during **personalization process** (when a real, customer-dedicated card is manufactured) or in any moment of a card's life cycle.
- RAM is used to store only temporary data during transaction time. No application code is allowed to be loaded to RAM for a card may be withdrawn from the terminal at any moment causing a permanent loss of important code.
- Communication with the external world is processed by means of specific interpreted code structures called **APDU** (Application Protocol Data Unit) and data exchange (for further details see [1–3]).
- The way in which a card performs required functions is based on strictly-defined rules concerning memory file structure, data structure and format, file management, data management, access rights (read/write/execute control), execution control, and atomic (backup) methods (for further details see [1–3]).

- Of the pair, the terminal-card card is always a "slave." Such a situation results from the simple fact that it's the terminal that "wakes up" a card by applying power and reset signals to the card. In response, the card sends back **ATR** (Answer To Reset) **APDU** which identifies itself to a terminal. From this moment on the real session starts. The session may be terminated either by a card or by a terminal.

In the following two sections, we will try to present both types of OS and discuss their pros and cons.

Proprietary Operating Systems

Proprietary OS is always targeted to fulfill a defined scope of dedicated functionalities. The general, however specific, structure of the proprietary OS is shown on the Fig. 89.41.

Such a structure may give the impression that proprietary OS is created in a "fuzzy" manner. This is incorrect. All the memory is perfectly structured, with areas storing data, application code, and basic OS functions as well as native functions (procedures performing most basic functions as I/O protocol, **DES algorithm**, basic math, atomic functions, etc.). What are not that well-structured are *functionalities* performed by the card. Let's take the example of an electronic purse. From the user's point of view, the card should perform at minimum the following functions:

- User authenticate (user proves to the card that he/she is the owner of the card).
- Credit the card within pre-defined limit.
- Debit the card within pre-defined limits.
- Disable any attempt to break into the contents of the card.

In "nested" OS, none of the above functions are standalone. All of them are, in great part or even totally, performed by OS. In most cases, the reader may find in an application area "script(s)" performing needed function(s) by calling necessary subroutines within OS, but rather rarely a standalone block of software performing an independently new function.

What are the pros and cons of such a structure? Let's investigate:

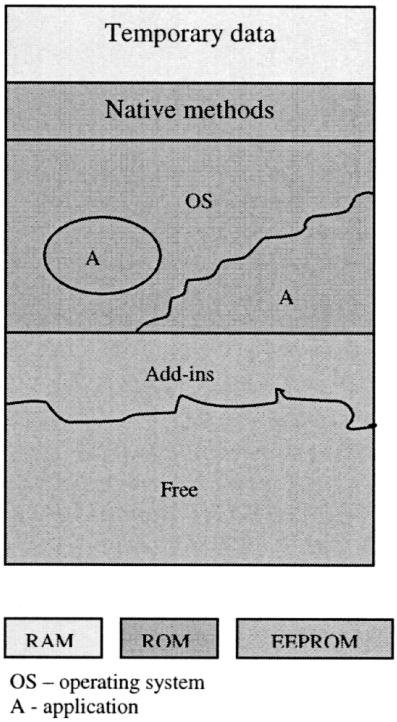


FIGURE 89.41 Proprietary OS structure.

PROS

- High "density" of the software or, in other words, good functionality/memory consumption ratio,
- Good functionality/price ratio,
- Good functionality/performance (speed) ratio,
- Good way to build a stable relationship between the card issuer and card manufacturer (watch the interpretation of this point in *CONS*).

CONS

- Applications are not portable, sometimes even within the same family of operating systems.
- Debugging process requires tremendous efforts as, in the worst case, the entire card software has to be verified.
- There are only two ways to correct found bugs: add the **"filter"** in EEPROM (not always possible, but always temporary) or remask the chip. Specifically, the second possibility in some cases may be called the *disaster* as when the bug is found in the card that has been deployed into the market in 100 million volume.
- Good relationship created by a chain Card Manufacturer → Functionality → Card Issuer is more and more often perceived as a master-slave relationship. Card issuers prefer to be supplier independent.

The conclusion drawn is somewhat obvious: proprietary OS is a handicapped solution, but there are hundreds of millions of such cards issued each year that imply the message: more flexible technologies are still to come.

Open Operating Systems

Smart Card open operating systems (OOS) are being developed continuously since the early 1990s reaching the position today of emerging technology. The basic assumptions for OOS development are:

- Portability of applications.
- Application load/delete function.
- Multi-application ability.
- Upward/downward compatibility of OOS within the family.
- Development environment similar to or being a subset of "full-blooded" computer software development environment.
- High-level programming language.
- High level of security, both inside the application and outside (firewalls).
- Shorter development time (time to market).

The general structure of OOS smart card is presented on Fig. 89.42.

How much clearer is such a structure when compared to proprietary OS? The OOS forms a well-defined group of layers over which the real applications reside. Bottom-up direction shows higher and higher levels of abstraction of consecutive layers:

- Native functions and system resources — all the basic functions that are time critical (I/O subroutines, basic cryptographic functions, atomicity, basic file management) developed in assembly language.
- Virtual Machine — on-card interpreter of bytecodes.
- **API** — application programming interface — set of generic libraries.
- Specific API — nonmandatory API dedicated to specific card functionalities.
- Card Executive — file and application manager (loading/removing applications).
- Applications — developed in high-level languages, stored on the card in bytecode format.

Applications developed for such a type of the card are commonly called applets or, better, cardlets. Usually, the cardlet is developed on the PC using one of several higher-level languages in a not necessarily dedicated development environment. Dedicated environment comes on the stage in the moment of the first compilation

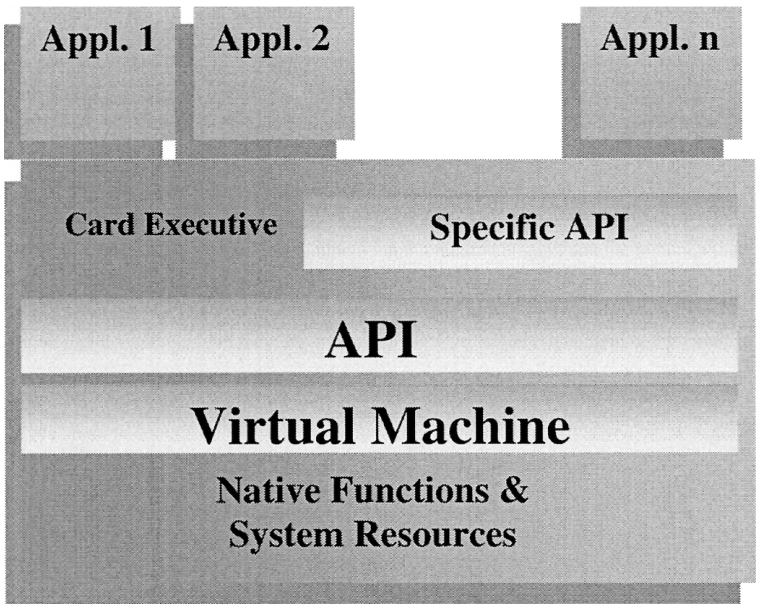


FIGURE 89.42 General structure of OOP Smart Card, where:
Appl1,2,..n — card applications,
API — application programming interface

and debugging. What is important to observe as the difference between PC-dedicated compilation and card-dedicated compilation is the result: for PC you may obtain the executable file (*.exe), for card it's always bytecode (*.cap) as the card *interprets* application code by code. Debugging of the cardlet may be done in a simulated, PC-based environment as well as in a natural card environment.

Ideally, during card life cycle, cardlets may be downloaded, added, or removed changing the profile of the card. All three possibilities generate immediate problems of a different nature:

- Co-existence of two or more cardlets requires high security measures to firewall cardlets against cross-corruption in any form as well as raises the needs of re-usability of common objects.
- Any type of manipulation of the cardlets creates responsibility/legal issues to be solved in the near future such as Who's got the right to manipulate cardlets? Card issuer (bank?), card manufacturer, or card user? How to maintain cardlets and their life cycle, etc.

There are more issues of both a technical and legal nature, to be answered, so those interested in smart card technology may get closer to it by observing the works of international standard bodies (e.g., ISO), industry associations (e.g., SCIA) or international forums (best known is Smart Card Forum).

PROS

- Any developer knowing higher-level programming languages may develop an application. No "bits-and-bytes" fighter knowledge necessary.
- Development environment is PC-like or purely PC.
- Debugging may be off-card as well as on-card.
- Development time is incomparably shorter than this for proprietary OS.
- As applications are loaded by default to EEPROM, any in-the-field bug may be fixed by simply reloading the cardlet.
- Time to market dramatically shorter.
- Portability of applications.
- The customer is supplier independent.

CONS

- Technology still not mastered, both in hardware and software domains.
- Development/debugging tools in a childhood phase.
- Performance still low (due to interpretation mode of operation and slow hardware).
- Due to evolution phase of standards, portability of cardlets merely applicable.
- High cost of cards.

The conclusion might be: first steps are always the heaviest, but to explode, OOS technology needs more maturity rather than technical improvements. Next five years are forecasted as the real startup for OOS smart cards.

Standards

Smart card is one of the best standard-guarded pieces of equipment. The set of ISO 7816 regulations defines all the aspects of smart card technology and life cycle. Most of them are stable, however, due to continuous software and hardware technological improvements all of them are being amended. The complete list of applicable standards is presented below:

- ISO 7816-1 — Identification cards — Integrated circuit(s) cards with contacts — Part 1: Physical characteristics. *Defines physical characteristics of cards as well as tests to be applied.*
- ISO 7816-2 — Identification cards — Integrated circuit(s) cards with contacts — Part 2: Dimensions and location of the contacts. *Defines dimensions and locations of contacts and magnetic stripe as well as tests to be applied.*
- ISO 7816-3 — Identification cards — Integrated circuit(s) cards with contacts — Part 3: Electronic signals and transmission protocols. *Defines basic electronic characteristics of ICCs, power supply, data elements in* ***ATR*** *and transmission protocol T = 0.*

- ISO 7816-4 — Identification cards — Integrated circuit(s) cards with contacts — Part 4: Interindustry command for interchange. *Defines internal software structure of ICCs, security requirements, formats of APDUs, communication codes.*
- ISO 7816-5 — Identification cards — Integrated circuit(s) cards with contacts — Part 5: Numbering system and registration procedure for application identifiers. *Defines system for clear identification of national and international applications in ICC.*
- ISO 7816-6 — Identification cards — Integrated circuit(s) cards with contacts — Part 6: Interindustry data elements. *Defines data elements and structures used in retrieval of data from ICC.*
- ISO 7816-7 — Identification cards — Integrated circuit(s) cards with contacts — Part 7: Enhanced interindustry commands. *Defines additional commands added to Part 4 for SQL access, encryption, secure messaging.*
- ISO 7816-8 — Identification cards — Integrated circuit(s) cards with contacts — Part 8: Interindustry security architecture. *Defines security architecture of ICCs.*
- ISO 7816-9 — Identification cards — Integrated circuit(s) cards with contacts — Part 9: Additional interindustry commands and security attributes. *Describes coding of the life cycle of the cards and related objects, coding of security attributes of card related objects; functions and syntax of interindustry commands not defined in other parts of ISO/IEC 7816; data elements associated with these commands; and a mechanism for initiating card-initiated messages.*
- ISO 7816-10 — Identification cards — Integrated circuit(s) cards with contacts — Part 10: Electronic signals and Answer To Reset for synchronous cards. *Defines basic electronic characteristics, power supply, data elements in **ATR** for synchronous cards.*
- ISO 7816-11 — Identification cards — Integrated circuit(s) cards with contacts — Part 11: Card structure and enhanced functions for multi-application use. *Defines an enhanced logical structure for multi-application cards to support security with applications loaded after issuance, prerequisites for loading of an application after issuance, a mechanism to support an application life cycle independent of the card life cycle, interapplication security issues, including storage partitioning.*

For more details you may refer either directly to International Standards Organization or to ISO Web site: www.iso.org.

Applications

The purpose of this chapter is mostly descriptive, aiming to show the wealth and diversity of existing and upcoming applications deployed all over the world.

The smart card industry is formed by a few big manufacturers covering practically all of the world's needs (Gemplus, Schlumberger, Motorola, Orga, De La Rue, Gisecke&Devrient, and Oberthur). Each of them offers either applications, platforms, or both. The short overview of possible applications will be more persuasive if done by market sectors:

- Banking: credit/debit card, electronic purse, loyalty programs, individual financial programs.
- Retail: loyalty programs, retailers' credit cards.
- Healthcare: personal medical treatment databases (medical files).
- Wireless communication: SIM (Subscriber Identity Module) card, OTA (Over The Air) services, e-mail services.
- IT (Information Technology): security access to IT systems, encrypting/decrypting files, securing e-mail, Internet transactions (e-wallet, e-commerce).
- Universities: multi-functional ID cards (credit, electronic purse).
- Government and agencies: ID cards, internal electronic purse, time control.
- Physical access: ID cards, zone access control.
- Manufacturing: electronic tags.
- Transportation: ticketing, air-miles programs, driving licenses.

Most of the above-mentioned applications are based on a proprietary OS, however, OOS early birds are also observed. The first steps on the way to commercial OOS have been recently done by some of the industry

leaders. The joint efforts of Sun Microsystems, Gemplus, and VISA-created Java-based Visa Open Platform (VOP), Schlumberger has created its Java-based Open Platform — Cyberflex, Gemplus has its Java-based GemXpresso and its own OOS — Nucleos, and last but not least, MULTOS — OOS created by consortium MAOSCO/Mondex. Though still not matured, all of these OOSs have already provoked the chain reaction: creating cardlets' development industry.

Readers

A smart card is of no use if you cannot communicate with it and the reader is the element that will give life to this hand-held device. The market offers several kinds of readers. Readers are categorized in three classes: standalone, peripheral, and contactless.

Standalone Readers

Standalone readers are modules that offer a complete interface to operate and communicate with the smart card. That kind of reader can be a part of a complex machine such as an automatic teller machine or presented as a hand-held device. For an application such as an automated bank teller, the smart card reader is then a programmable device integrated into a larger system with several peripherals such as display, keyboard, real-time clock, and printers. In many cases, the complete system also includes several security mechanisms from protected housings to sensors for detecting attacks and security modules. A security module is a smart card that holds the terminal secret keys, which are used to create card-specific keys. A standalone reader usually has a communication link to operate as a peripheral of a larger transaction server that will verify the user identification and take charge of the information transmission, reconciliation of transactions, etc.

Other applications require a reader that will participate in a secure information exchange transaction between participants. Thus, desktop systems are smaller and have their use in business application to provide secured services. Along with security modules, some readers will accept several cards to authenticate, for instance, the owner and the receiver of a confidential information. In healthcare application, the patient will authorize access to his files while the doctor will provide the proof that he has the right to access this kind of information. Several security modules will ensure security and protection for different applications executed through the same reader, treated separately one from another. Other suitable applications for those readers are payment transactions, electronic purses, and loyalty programs. The integration of the keyboard and the display to the security of the process ensures that the user's PIN and exchanged information are not tapped by another software.

Hand-held systems offer the flexibility of mobile systems. Integrated with telecommunication devices and hand-held computer technology, these systems can present touchscreen interfaces and provide facilities such as gaming, fax, e-mail, telephone, e-purse, loyalty, healthcare or driver's license programs, security, authentication, identification, electronic payment, secure access to networks, secure data transfer over those networks, etc. GSM is gaining more popularity by using the cell phone as a mobile reader to personalize your phone and provide the user with facilities such as e-mail and prepaid services.

Peripheral Readers

A peripheral reader provides a communication interface between a larger system and a smart card. Using standard interfaces, the reader is usually coupled to a control system that will send commands to the reader. A peripheral reader can be integrated into a larger peripheral, such as payphones, a vending machine, or a card personalization device.

Peripheral readers are often attached to a computer to add functionality for card-aware applications. The market offers several types of connections for peripheral devices. Serial connections are suitable for most desktop computers and are standard and affordable. PC Card connections (pcmcia) are mainly for laptop computers. Although slow, keyboard readers, using the PS/2 connection, provide better security, mainly for the PIN management. When entering a pin, the reader intercepts the characters from the keyboard and replaces them with a neutral character that is sent to the system. This way no resident software can intercept the PIN. A keyboard reader can be integrated to the keyboard itself or added to the PS2 connector of the keyboard. **USB** readers are now coming into the market and will compensate for any deficiency of the serial reader. Some

attempts are also done to integrate a reader into a mouse or floppy disk. This last solution offers the user the possibility to use a smart card without adding new hardware to his machine.

Finally, one can find specialized peripheral readers combining the functionality of other device types. For instance, a smart card reader and modem were combined to secure network access in Internet communication, electronic commerce, and home banking. Also, a smart card reader and a television signal processing module were combined for conditional access, electronic program guides, interactive applications, home shopping, and payment systems.

The use of peripheral readers is an emerging market with the recent deployment of application for the electronic commerce, home banking, online gaming, health care, network security on the Internet, and telecommunications. Readers with biometrics interfaces have begun their introduction on the markets. Fingerprints are used instead of a PIN to authenticate the user.

Contactless Reader

A contactless reader (also called proximity reader) needs no physical contact between the card and the reader to exchange data. The card goes immediately into action when it is within a few inches of the reader. Transmission details were described earlier.

Using a hard-wired algorithm and random number generation, it is possible to provide high-level security. The reader and the card use a mutual authentication and a stream ciphering. Such system must also provide anticollision management to operate with multiple cards in the communication field.

Contactless smart card readers are perfectly suited for physical access control and smart tracking applications. The use of Smart Labels (or tags) injected or embedded into a product allows the identification and tracking of product for either authentication or automated manipulation. Since data can also be written into a contactless card, payment functions can be implemented for fare collection in public transportation, for instance. When the distance between communications is a concern, systems using intelligent transponders enable the payment of tolls from equipped vehicles travelling on motorways at normal speed.

Card-to-System Solutions

Naturally, communication with a smart card would be impossible without several software layers that transform the request of the user into a sequence of bits transmitted to the device. At the lowest level, the drivers provide the basic functionality by transferring a sequence of bytes to the reader. Then layers of software are built over the driver to form application program interfaces (APIs). **API** offers to the application developer functions that combine basic operations to hide specificity of a particular device and to ease the development.

Drivers

To communicate with a reader, the basic transmission is handled by three protocol layers: the physical layer, the transport layer, and the command layer. The physical layer is related to the data transmission itself through the physical port (serial, pcmcia, usb, etc.). The transport layer specifies the transmission type, the message addressing, and handles the message validation. It defines how each byte transmitted by the physical layer is interpreted and ensures that it forms a consistent and valid message. Finally, the command layer specifies commands that are transmitted to the reader to execute a set of functions. These functions will allow the communication with a smart card and the management of events such as insertion and removal of the card.

In the communication process, the driver is a piece of software that hides the treatment related to the physical and the transport layers. With documented interfaces, the user can use the command layer to send commands to the reader and to communicate with the card. Usually, the operating system integrates drivers that take charge of the physical layer. The device vendor will then write a **filter driver** that is placed over the other one to wrap up the transport layer, transmitting commands into well-defined packages. The driver will allow basic commands to the reader: setting parameters of the readers (communication speed, voltage levels, card type used), setting communication parameters between reader and card (protocol, speed), power on, power off, transmitting bytes to the card. Moreover, the driver must manage the events coming from the system and control the states of the device. For instance, the insertion or the removal of the reader involves some processes of initialization of the communication and registration of the connection. The driver must also allow the sharing

of the resource by multiple applications or threads. Security issues regarding those accesses must be considered. One must preserve access to the information on the card from the application that was not allowed.

With only the driver, the development of an application requires the knowledge of the command format to send to the card and of the mapping of the smart card. To ease and speed up his development, a programmer needs tools that hide specificity of cards and readers. **API** and middleware are coming to his rescue.

APIs and Middleware

Application program interfaces (API, also called middleware) usually take the form of libraries that provide the user with high-level functions to ease the development and to hide disparities between different technologies. Basic APIs try to hide the command layer to the user and regroup under common interfaces the communication with different readers. At a higher lever, an API hides the smart card operating system particularities, providing the user with card services and with abstract concepts that ease the use of the information on the card. Thus, card services will allow the user to select a file, to read, write, and erase objects on the card, and to authenticate the user. API can also regroup specialized functions such as cryptographic functions, purse functions, etc.

Usually, APIs are proposed by smart card vendors to offer an easy-to-use environment for the development of applications on their products. It offers a wide variety of commands and responses. However, it makes the developer dependent of a product and of the support of these API. Efforts were made to standardize some API. PKCS#11[5] is an example of cryptographic libraries. This standard API is defined by RSA (RSA Data Security, Inc.) and permits the use of data objects in the form of tag-length-value.

Standardizing Approach

With the lack of interoperability between operating systems, API, readers, and smart cards, some efforts were made to offer more compatibility and harmonization. **PC/SC** and the OpenCard Framework are the main stream.

PC/SC

PC/SC[4] is an effort to provide interoperability of the smart card technology in the personal computer environment. The aims of PC/SC is to simplify the development of application by defining high-level programming interfaces that reduce the dependency of application on proprietary implementation. To achieve its goals, PC/SC distinguishes three layers between the hardware device and the application level: The IFDHandler Layer, the resource manager layer, and the service provider layer.

The IFDHandler allows access to the smart card reader. It is the driver level and the hardware manufacturer that will normally provide it. PC/SC defines the interfaces that the IFDHandler must expose to take place in the structure. The IFDHandler is the implementation of the functionality required by the exposed functions to transmit commands to the reader and/or the card. The IFDHandler maps the native capabilities of the smart card reader to the IFDHandler interface. All distinctions between smart cards based on ISO protocol handling, whether synchronous or asynchronous, are hidden. Once installed, the IFDhandler registers itself to the Resource Manager.

The Resource Manager is a key component of the PC/SC architecture. It is responsible for managing the other resources (such as service providers) within the system and for controlling access to the smart card readers and to the card itself. The Resource Manager is provided by the operating system vendor. Thus, the Resource Manager is responsible for identification and tracking of resources. This includes the tracking of the readers, of the smart cards, of the service providers, of the supported interfaces and of the smart card insertion and removal events. Thus, the Resource Manager can list the available resources to each application.

The Resource Manager is also responsible for controlling the allocation of readers and across multiple applications with shared or exclusive modes of operations. Finally, it supports transaction primitives to allow multiple commands to be executed without interruption, ensuring that intermediate state information is not corrupted.

The Service Provider regroups a set of functions exposed by a specific smart card and makes it accessible through high-level programming interfaces. The common functionalities are related to file access, authentication, and cryptographic services. However, these interfaces may be extended to meet the needs of specific domains.

PC/SC divides the Service Provider into two independent components: the Smart Card Service Provider (SCSP) and the Cryptographic Service Provider (CSP). The CSP regroups cryptographic functionality through high-level programming interfaces. This distinction was brought to deal with security issues and with international laws regarding cryptographic devices.

The Answer To Reset (**ATR**) string is used to perform the binding between a card and an interface. Thus, when a service provider is installed on a system, it registers itself to the Resource Manager with the ATR of cards it can serve. An application can define at run-time which Service Provider will be used since the Resource Manager enumerates the list of the available interfaces when the card is inserted.

Although PC/SC is platform neutral, to date, Microsoft did the only implementation of this standard on Windows platform (even now the implementation still lacks of some features: "plug and play" and power management). Card services are exposed to an application through Component Objet Model (COM) interface.

PS/SC v.1.0 supports only T = 0 and T = 1 cards and optionally T = 14 (protocols of communication between a card and a reader. For details see [1–3]). Next release of this standard will consider memory card and multiple interfaces readers such as keyboard, display, multiple cards, and SAM modules.

OpenCard Framework

OpenCard Framework (OCF) [6] objective is to offer to the programmer a transparency with regard to smart card operating systems, card terminals, and card issuers. To achieve this, OCF has defined a provision of functions to support installation, removal, enumeration, selection of applications on the card and of functions to perform name resolution for data files on the card. Several mechanisms allow a user to develop an application without knowing where the card issuer placed his applications. OpenCard Framework based its architecture on two main parts: the CardTerminal layer and the CardService layer.

The CardTerminal layer plays the role of the driver and the reader manufacturer should provide it. It offers an interface that enables the seamless integration of the reader in the OCF environment. With OCF-compliant interfaces, a specific card terminal gives access to the reader and the smart card. A CardTerminal can also act as a bridge to provide a transparent access to a reader through the PCSC interface to take advantage of the existing components.

The CardService layer is defined as a high-level application programming interface that hides the characteristics of a particular provider's components (specific to a type of smart card and/or smart card reader) from the application and service developers. Card services are standard **APIs** regrouping functions to access resources of a card operating system. Thus, a card issuer will provide the users with a card services allowing access to the files (FileAccessCardService), to the signature functions (SignatureCardService), to PKCS#11 functions, to the ISO7816 file system, and to purse functions.

Between those two layers, a specific card service, the application management card service, manages the card resident applications. The application management component lists the applications that a card can support, locates them or even installs or removes them from the card. It is a special card service, also provided by the card issuer, that can manage multiple applications on the same smart card.

OCF essentially was created for the Java programming environment and the world of network computing. Because of this, it claims all the advantages that the Java language provides. Basically, any platform that is capable of running JAVA can exploit OCF right away. Moreover, OCF offers mechanisms to allow one to download from the network missing components for a particular card at hand and plug them into the framework. As it is for Java, OCF targets embedded systems such as automatic teller machines, point-of-sales terminals, and hand-held devices (phone, electronic agenda).

Applications

Developing an application is made easier and more flexible when based over standard API and middleware products. The challenge remains, however, to respond to the new needs of the industry. Applications must integrate security schemes to avoid holes and prevent piracy of a system. To secure their system, applications developers must be aware of the complete structure of the operating system they are using. For instance, the market proposes new software to protect access to a computer and to encrypt files on computers. Such a system requires low level control of the resources of a computer to block unexpected entry. Also, sensitive information must not be exposed. Encryption keys are never kept clearly in memory to prevent software attack.

Applications are also confronted with the Internet challenge. Security systems are facing new offensives and new security schemes that must be put in place. The Internet not only broadcasts the information at the speed of light; it represents to hackers a tremendous power for parallel computing. Competitions are organized on the net to break encryption keys: each participant downloads the required software and adds its contribution to the computing effort. Key size is lengthened, consequently. Secure transactions over the net also bring new problems. Applications are distributed, client-server schemes become more important, the number of transactions are growing and, consequentially, we do not know who is on the other side. For an electronic wallet application, this is crucial because nobody wants to give his credit card number without some credentials. Third party certificates are there to authenticate your invisible partner and mechanisms are set to establish a confidence path between all involved parties of a transaction.

Trends

Trends in smart card industry may, but rather should not be, seen only from an ICC perspective. The reason is amazingly simple: the most powerful, super fast and multi-functional card is nothing more than a piece of plastic and silicon if there is no IT system able to use it. That's why the trends will be discussed in two groups: card technology and complete solutions.

The general challenges driving card technology trends may be defined by four simple words: faster, cheaper, more, easier. Technical solutions to come should allow multi-application (more), flexible functional profile i.e., easy changeable dedicated applications (easier), shorter, nonspecific technology-tied development cycle (faster), shorter time to market (more, cheaper), higher security level (easier). All of the above points to a generally named card open OS solution.

How serious this trend is depends upon industry reactions. Hardware manufacturers design more powerful chips dedicated to operate with higher clock speeds, offering larger memory space, in some cases injecting specific commands into the processor's command list (Siemens's picoJava concept encapsulating specific JavaCard commands in processor design), and adding specialized coprocessors (cryptographic functions). Card manufacturers improve their OOSs to meet multi-application requirements by adding new features up to recently imminent to pure IT (multi-thread, COM, firewalls, add/load/delete), system integrators adapt their solutions to absorb specific card management systems (CMS) allowing control over the live cycle of applications (and cards), card issuers design multi-applicative programs for their customers, and finally, software development companies begin to create competitive, cutting edge competence in smart card OOS-dedicated applications development. The recent entry into the industry includes the software giant — Microsoft with its Smart Card for Windows (SCW). This entry establishes the seriousness of the smart card trend.

To penetrate new markets, the smart card industry must now develop value-added solutions by enhancing the functionality of the conventional card. To be attractive to customers, the same card should be used for access control, securing transactions on the net, and for payment. This trend implies the coexistence of multiple applications on the same card. It also means the customization of the user card. The Java Card and, more recently, the Smart Card for Windows are supporting this stream where developers can download their own application into the card.

Defining Terms

APDU: Application Protocol Data Unit is a set of bytes forming a unitary "container" used to send to or receive data from a card. There are two types of APDUs: command APDU (sent to a card) and response APDU (sent by a card). Formats of both APDUs are different, however, no mismatch is possible as the communication between a card and the reader is ruled by a sequence: command (reader)-response (card).

API: Application Programming Interface. In general terms, API is a set of functionalities (library) loadable or preloaded available for a programmer developing the application. In smart card technology, APIs are found on board the card (e.g., Java Card) as well as on the system side (reader APIs). The goal of creating API is to free the developer from card-related specific knowledge allowing him/her to develop on the higher level of abstraction.

ATR: Answer To Reset is a set of maximum 33 bytes returned by a card to a reader after RESET signal was applied. ATR consists of information necessary to authenticate a card to the system and the reader as well as to negotiate all communication parameters as protocol, transmission speed, etc.

DES algorithm: Data Encryption Standard algorithm is the most popular encryption/decryption mathematical engine. DES is being broadly applied in the security domain for encryption/decryption purposes, secure messaging, and authentication. Best known is 56-bit encryption/decryption key version. For more sophisticated security applications, 3DES (triple DES) is being applied. 3DES is a method using DES three consecutive times with a different encryption key used each time.

Filter: Filter, in smart card technology, defines an additional block of code loaded to the card's EEPROM whose function is to either enhance the existing card's functionality or patch found bugs. In the first case, you can see it as the development "pilot" that allows the manufacturer to check new releases before its industrial launch. In the second case, a "filter" is the easiest way to debug and fix existing application or OS. When checked, verified, and approved, the "filter" is moved to ROM and remasked.

Masking process: The masking process is the essential process of manufacturing the chip to be later embedded onto a card. The masked chip contains all the hardware components as well as all of the software stored in ROM. The remaining EEPROM and RAM are empty and not structured.

Personalizing process: The personalizing process is performed by a card manufacturer. The goal is to make a masked chip card that the customer can use. The process covers loading onto the card all specific customer info (name, primary security keys, serial number) as well as building the internal memory structure (directories, file structures, file types).

PC/SC: Personal Computer/Smart Card (PC/SC) is the "de facto" or industry standard introduced by Microsoft Corporation. The standard defines structure, functionality, and communication entries for smart card readers' drivers being installed on all existing Windows platforms (95, 98, NT4, 2000).

USB: Universal Serial Bus is the newest PC technology applied to serial communication solutions. The technology allows connecting to a PC several serial devices in a chain at the same time. The communication protocol between PC and external serial devices is similar to this used in networking. USB becomes more and more in demand in PC-Smart Card solutions.

References

1. Allen C., Barr W.J. *Smart Card seizing strategic opportunity.* Irvin, 1998.
2. Jurgensen T. *Smart Card developer's kit.* Macmillan, 1998.
3. Rankl W., Effing W. *Smart Card Handbook.* John Wiley & Sons Ltd., 1995.
4. PC/SC Workgroup. *Interoperability Specifications for ICCs and Personal Computer Systems.* Revision 1.0, 1997, available electronically at: http://www.smartcardsys.com.
5. RSA Laboratories. *PKCS#11: Cryptographic Token Interface Standard.* RSA Laboratories technical notes, version 2.01, 1997, available electronically at: http://www.rsa.com.
6. OpenCard Consortium. *OpenCard Framework 1.1 Programmer's guide.* 1998, available electronically at: http://www.opencard.org.

Further Information

Gemplus Web site (www.gemplus.com).
ISO web site (www.iso.ch).
ISO-IEC JTC1-SC17 Web site (www.funkster.com/ossian).
Proceedings of CardTech/SecurTech 1998, CTST 1998.
Smart Card Forum Web site (www.smartcrd.com).
Smart Card Industry Association Web site (www.scia.org).

Argila, C.A., Jones, C., Martin, J.J. "Software Engineering"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

90

Software Engineering

90.1 Tools and Techniques
Approach • Methods • Information Modeling • Essential Modeling • Implementation Modeling • CASE Tools

90.2 Testing, Debugging, and Verification
The Origins and Causes of Software Defects • The Taxonomy and Efficiency of Software Defect Removal • Pre-Test Defect Removal • Testing Software • Selecting an Optimal Series of Defect Prevention and Removal Operations • Post-Release Defect Removal • Recent Industry Trends in Software Quality Control

90.3 Programming Methodology
Analysis of Algorithms • Flow of Control • Abstraction • Modularity • Simple Hierarchical Structuring • Object-Oriented Programming • Program Testing

Carl A. Argila
Software Engineering Consultant

Capers Jones
Software Productivity Research, Inc.

Johannes J. Martin
University of New Orleans

90.1 Tools and Techniques[1]

Carl A. Argila

The last decade has seen a revolution in software engineering tools and techniques. This revolution has been fueled by the ever-increasing complexity of the software component of delivered systems. Although the software component of delivered systems may not be the most expensive component, it is usually, however, "in series" with the hardware component; if the software doesn't work, the hardware is useless.

Traditionally, software engineering has focused primarily on computer programming with ad hoc analysis and design techniques. Each software system was a unique piece of intellectual work; little emphasis was placed on architecture, interchangeability of parts, reusability, etc. These ad hoc software engineering methods resulted in the production of software systems which did not meet user requirements, were usually delivered over budget and beyond schedule, and were extraordinarily difficult to maintain and enhance.

In an attempt to find some solutions to the "software crisis," large governmental and private organizations motivated the development of so-called "waterfall" methods. These methods defined formal requirement definition and analysis phases, which had to be completed before commencing a formal design stage, which in turn had to be completed before beginning a formal implementation phase, etc. Although waterfall methods were usually superior to ad hoc methods, large and complex software systems were still being delivered over budget and beyond schedule, which did not meet user requirements. There were several reasons for this. First, waterfall methods focus on the generation of *work products* rather than "engineering." Simply put, writing documents is not the same as doing good engineering. Second, the waterfall methods do not support the *evolution* of system requirements throughout the development life cycle. Also, the prose English specifications produced within the waterfall methods are not well suited to describing the complex behaviors of software systems.

[1]The material in this article was originally published by CRC Press in *The Electrical Engineering Handbook,* Richard C. Dorf, Editor, 1993.

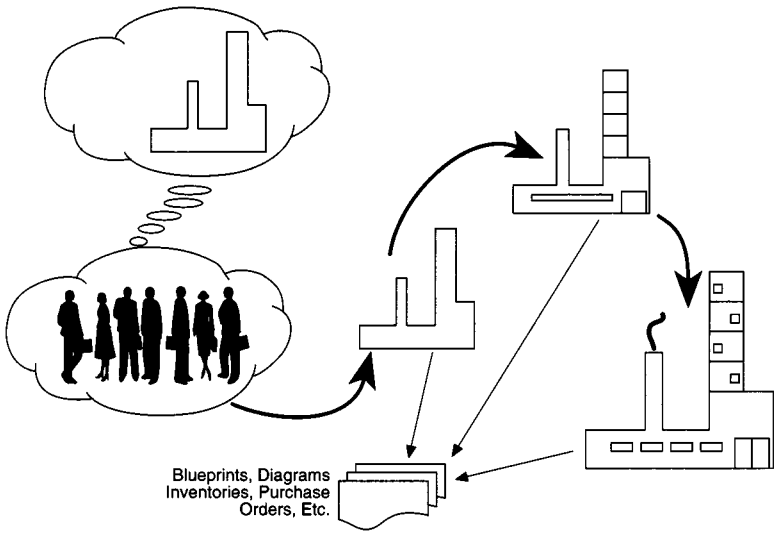


FIGURE 90.1 Model-based software engineering.

The basic, underlying philosophy of how software systems should be developed changed dramatically in 1978 when Tom DeMarco published his truly seminal book, *Structured Analysis and System Specification* [DeMarco, 1979]. DeMarco proposed that software systems should be developed like any large, complex engineering systems—by first building scale models of proposed systems so as to investigate their behavior. This *model-based software engineering* approach is analogous to that used by architects to specify and design large complex buildings (see Fig. 90.1). We build scale models of software systems for the same reason that architects build scale models of houses, so that users can visualize living with the systems of the future. These models serve as vehicles for communication and negotiation between users, developers, sponsors, builders, etc. Model-based software engineering holds considerable promise for enabling large, complex software systems to be developed on budget, within schedule, while meeting user requirements [see Harel, 1992].

As shown in Fig. 90.2, a number of specific software development models may be built as part of the software development process. These models may be built by different communities of users, developers, customers, etc. Most importantly, however, these models are built in an *iterative* fashion. Although work products (documents, milestone reviews, code releases, etc.) may be delivered chronologically, models are built iteratively throughout the software system's development life cycle.

In Fig. 90.3 we illustrate the distinction between *methodology*, *tool*, and *work product*. A number of differing software development methods have evolved, all based on the underlying model-based philosophy. Different methods may in fact be used for the requirements and analysis phases of project development than for design and implementation. These differing methods may or may not integrate well. Tools such as **CASE** may support all, or only a part, of a given method. Work products, such as document production or code generation, may be generated manually or by means of CASE tools.

This article will present a synopsis of various practical software engineering techniques which can be used to construct software development models; these techniques are illustrated within the context of a simple case study system.

Approach

One of the most widely accepted approaches in the software engineering industry is to build two software development models. An **essential model** captures the behavior of a proposed software system, independent of implementation specifics. An essential model of a software system is analogous to the scale model of a house built by an architect; this model is used to negotiate the *essential* requirements of a system between customers and developers. A second model, an **implementation model**, of a software system describes the technical aspects of a proposed system within a particular implementation environment. This model is analogous to the detailed

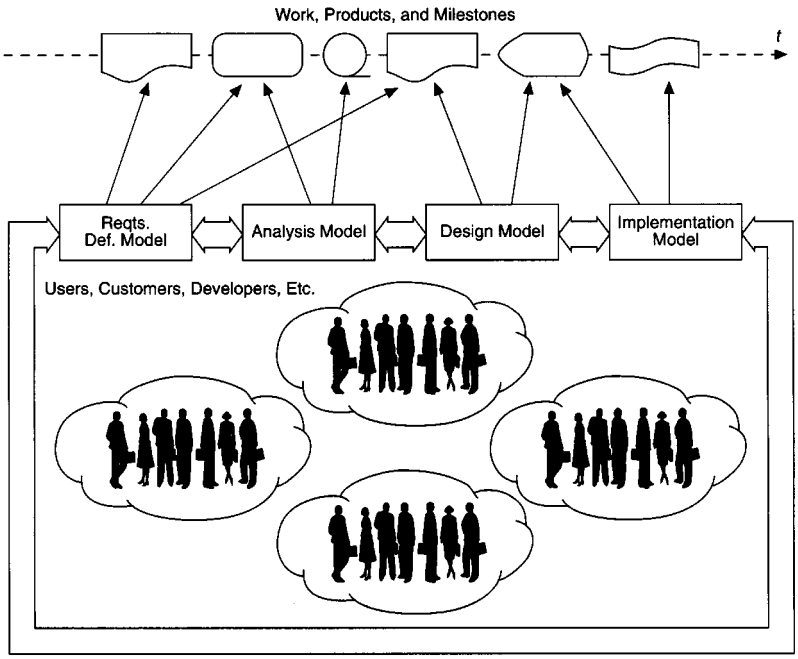


FIGURE 90.2 Modeling life cycle.

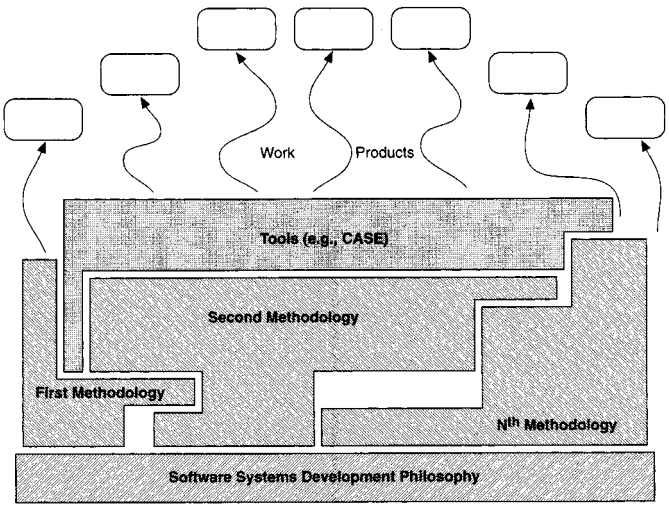


FIGURE 90.3 Methods, tools and work products.

blueprints created by an architect; it specifies the *implementation* aspects of a system to those who will do the construction. These models [described in Argila, 1992] are shown in Fig. 90.4. The essential and implementation models of a proposed software system are built in an iterative fashion.

Methods

The techniques used to build the essential and implementation models of a proposed software system are illustrated by means of a simple case study. The Radio Button System (RBS) is a component of a fully automated, digital automobile sound system. The RBS monitors a set of front-panel *station selection buttons* and performs station selection functions.

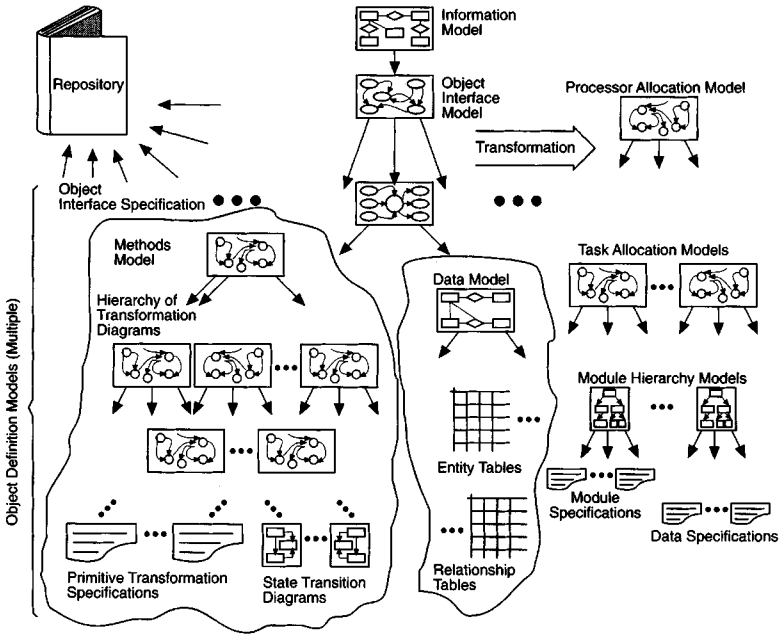


FIGURE 90.4 Software engineering methods overview.

When a station selection button is momentarily depressed, the RBS causes a new station to be selected. This selection is made on the basis of station-setting information stored within the RBS. The RBS can "memorize" new station selections in the following manner: When a given station selection button is depressed longer than "momentarily" (say, for more than 2 seconds), the currently selected station will be "memorized." Future momentary depressions of this button will result in this "memorized" station being selected.

The RBS also performs a muting function. While a station is being selected, the RBS will cause the *audio system* to mute the audio output signal. The RBS will also cause the audio output signal to be muted until a new station selection has been successfully memorized.

The RBS interfaces with the front-panel station selection buttons by "reading" a single-byte memory location. Each bit position of this memory location is associated with a particular front-panel station selection button. The value of 0 in a given bit position indicates that the corresponding button is *not* depressed. The value of 1 in that bit position indicates that the corresponding button *is* depressed. (For example, 0000 0000 indicates no station selection buttons are currently depressed; 0000 0010 indicates that the second button is currently depressed, etc.)

The RBS interfaces with the *tuning system* by means of a common memory location. This single-byte memory location contains a non-negative integer value which represents a station selection. (For example, 0000 0000 might represent 87.9 MHz, 0000 0001 might represent 88.1 MHz, etc.) The RBS may "read" this memory location to "memorize" a current station selection. The RBS may also "write" to this memory location to cause the tuning system to select another station.

Finally, the RBS interfaces with the audio system by sending two signals. The RBS may send a MUTE-ON signal to the audio system causing the audio system to disable the audio output. A MUTE-OFF signal would cause the audio system to enable the audio output.

Information Modeling

The construction of an **information model** is fundamental to so-called object-oriented approaches. An information model captures a "view" of an application domain within which a software system will be built. Information models are based on entity-relationship diagrams and underlying textual information. A sample

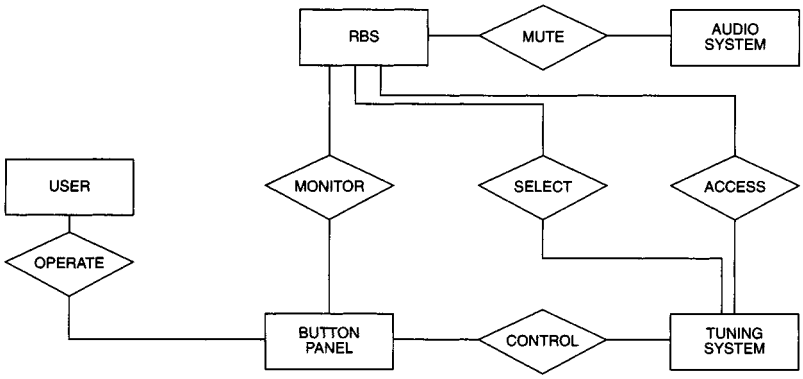
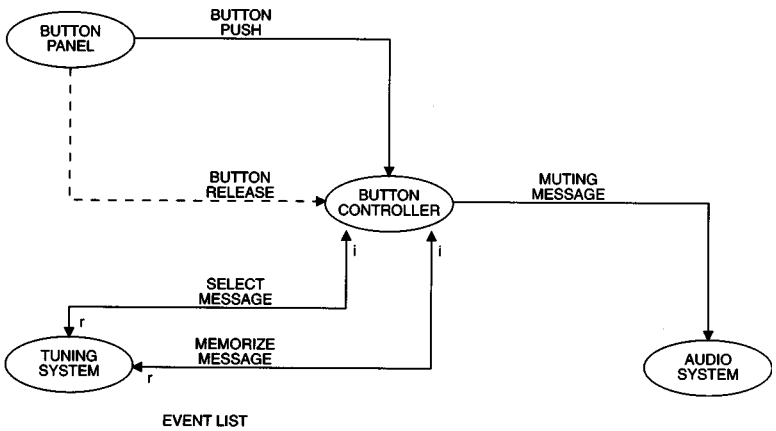


FIGURE 90.5 RBS information model.

information model for the RBS is shown in Fig. 90.5. Entities (shown as rectangles) represent "things" or "**objects**" in the application domain. Entities may be established by considering principal nouns or noun phrases in the application domain. Entities have *attributes* associated with them which express the qualities of the entity. Entities participate in *relationships;* these are shown as diamonds in the entity-relationship diagram. Relationships may be determined by considering principal verbs or verb phrases in the application domain. Relationships have *cardinality* associated with them and entities may participate *conditionally* in relationships. Finally, there are special kinds of relationships which show *hierarchical relationships* between objects.

Essential Modeling

The essential model consists of a number of graphical components with integrated textual information. Figure 90.6 shows the **object collaboration model** for the RBS. This model depicts how a collection of objects or entities can communicate (by exchanging messages) to perform the proposed system functions. An *event list* is part of this model; it shows what responses must be produced for a given external stimulus.



EVENT LIST

EVENT	RESPONSE
1. Selection button depressed.	1. Mute audio system.
2. Button released within two seconds after it was depressed.	2. Select a memorized station. Un-mute audio system.
3. Two seconds elapses after button was depressed while button remained depressed.	3. Memorize current station. Un-mute audio system.

FIGURE 90.6 RBS object collaboration model.

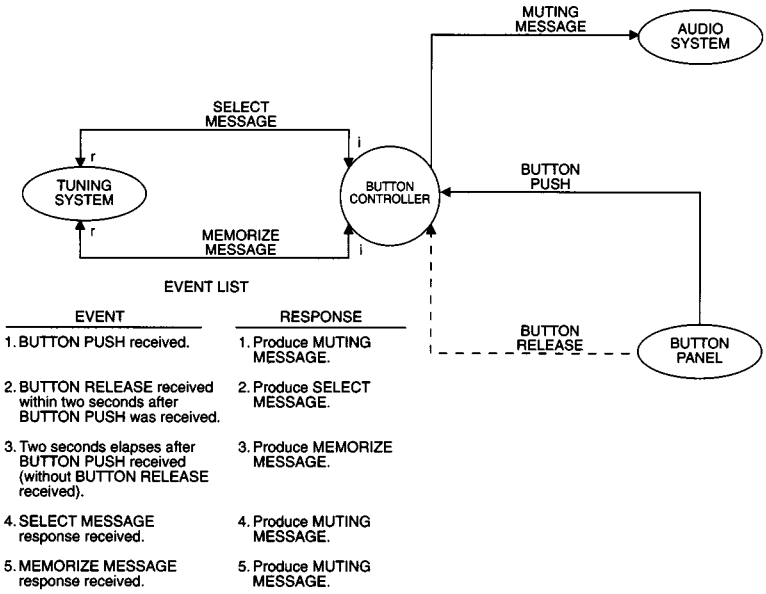


FIGURE 90.7 RBS object interface specification.

For each object there is an **object interface specification** (as shown in Fig. 90.7) which shows the public and private interfaces to an object. An event list is also associated with this specification; it shows how the object will respond to external stimuli. A hierarchy of **transformation diagrams** is associated with each object specification (as shown in Fig. 90.8 for the RBS). This diagram defines all of the functions or "methods" which the object performs. Some behavior may be expressed by means of a **state transition diagram** (Fig. 90.9).

Implementation Modeling

Two principal activities must be accomplished in transitioning from the essential to the implementation model. First, all of the methods and data encapsulated by each object must be mapped to the implementation environment. This process is illustrated in Fig. 90.10. Second, all of the details which were ignored in the essential model (such as user interfaces, communication protocols, hardware limitations, etc.) must now be accounted for.

Each component of the essential model must be allocated to hardware processors. Within each hardware processor, allocation must be continued to the *task* level. Within each task, the computer program controlling that task must be described. This latter description is accomplished by means of a **module structure chart.** As illustrated in Fig. 90.11 for one component of the RBS, the module structure chart is a formal description of each of the computer program units and their interfaces.

CASE Tools

The term *computer-aided software engineering* (CASE) is used to describe a collection of tools which automate all or some of various of the software engineering life cycle phases. These tools may facilitate the capturing, tracking and tracing of requirements, the construction and verification of essential and implementation models and the automatic generation of computer programs. Most CASE tools have an underlying *project repository* which stores project-related information, both textual and graphical, and uses this information for producing reports and work products.

CASE tool features may include:

- Requirements capture, tracing, and tracking
- Maintenance of all project-related information
- Model verification
- Facilitation of model validation

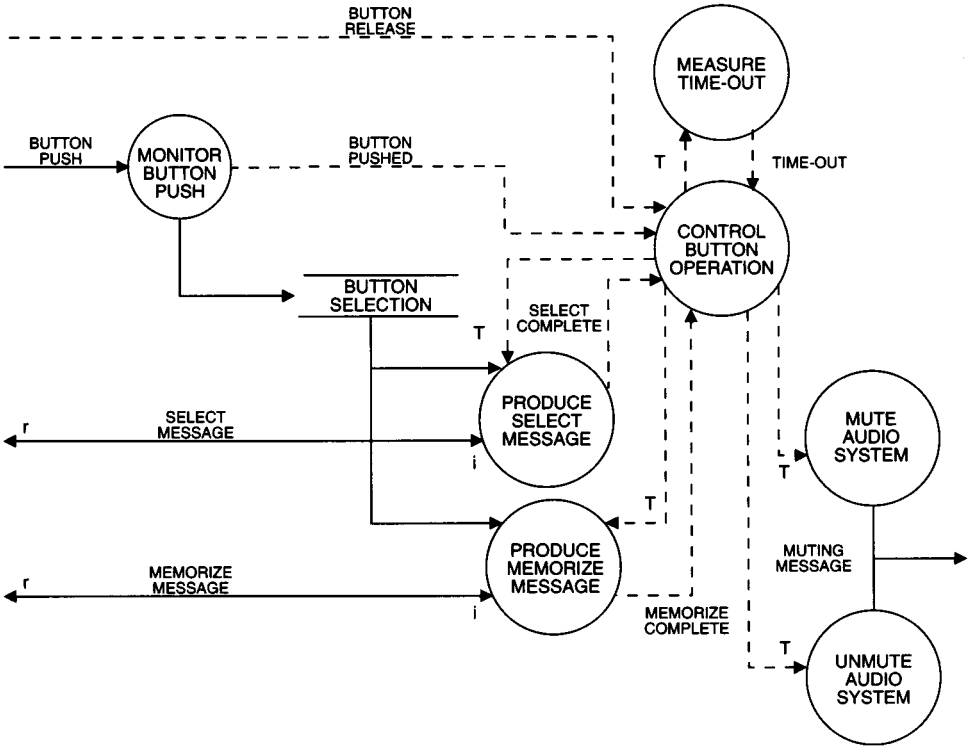


FIGURE 90.8 RBS transformation diagram.

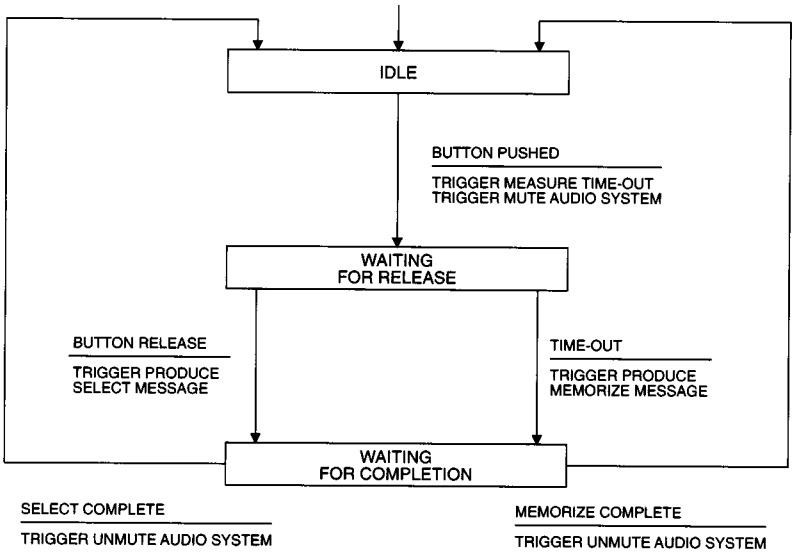


FIGURE 90.9 RBS state transition diagram.

- Document production
- Configuration management
- Collection and reporting of project management data
- CASE data exchange

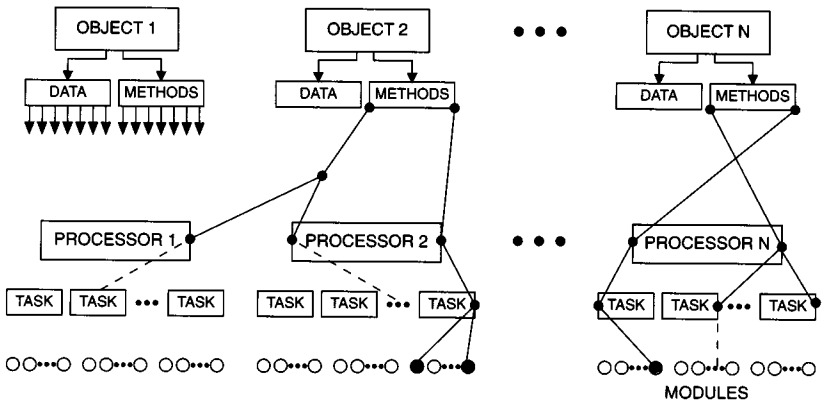


FIGURE 90.10 Implementation modeling.

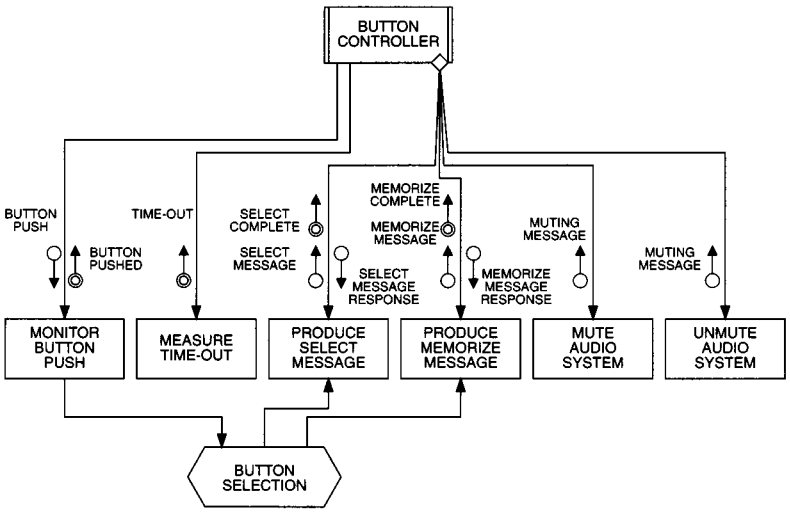


FIGURE 90.11 RBS module structure chart.

Defining Terms

CASE: Computer-aided software engineering. A general term for tools which automate various of the software engineering life cycle phases.

Essential model: A software engineering model which describes the behavior of a proposed software system independent of implementation aspects.

Implementation model: A software engineering model which describes the technical aspects of a proposed system within a particular implementation environment.

Information model: A software engineering model which describes an application domain as a collection of objects and relationships between those objects.

Module structure chart: A component of the implementation model; it describes the architecture of a single computer program.

Object: An "entity" or "thing" within the application domain of a proposed software system.

Object collaboration model: A component of the essential model; it describes how objects exchange messages in order to perform the work specified for a proposed system.

Object interface specification: A component of the essential model; it describes all of the public and private interfaces to an object.
State transition diagram: A component of the essential model; it describes event-response behaviors.
Transformation diagram: A component of the essential model; it describes system functions or "methods."

Related Topic

90.3 Programming Methodology

References

C. Argila, "Object-oriented real-time systems development" (video course notes), Los Angeles: University of Southern California IITV, June 11, 1992.
G. Booch, *Object-Oriented Design with Applications*, Redwood City, Calif.: Benjamin/Cummings, 1991.
P. Coad and E. Yourdon, *Object-Oriented Analysis*, 2nd ed., New York: Prentice-Hall, 1991.
P. Coad and E. Yourdon, *Object-Oriented Design*, New York: Prentice-Hall, 1991.
T. DeMarco, *Structured Analysis and System Specification*, New York: Prentice-Hall, 1979.
D. Harel, "Biting the silver bullet," *Computer*, January 1992.
J. Rumbaugh et al., *Object-Oriented Modeling and Design*, New York: Prentice-Hall, 1991.
S. Shlaer and S. Mellor, *Object-Oriented Systems Analysis: Modeling the World in Data*, New York: Prentice-Hall, 1988.
S. Shlaer and S. Mellor, *Object Life-Cycles: Modeling the World in States*, New York: Prentice-Hall, 1992.
P. Ward and S. Mellor, *Structured Development for Real-Time Systems*, New York: Prentice-Hall, vol. 1, 1985; vol. 2, 1985; vol. 3, 1986.
E. Yourdon and L. Constantine, *Structured Design*, 2nd ed., New York: Prentice-Hall, 1975, 1978.

Further Information

A video course presenting the software engineering techniques described here is available [see Argila, 1992]. The author may be contacted for additional information and comments at (800) 347-6903.

90.2 Testing, Debugging, and Verification

Capers Jones

Achieving acceptable levels of software quality has been one of the most troublesome areas of software engineering since the industry began. It has been so difficult to achieve error-free software that, historically, the cost of finding and fixing **"bugs"** has been the most time-consuming and expensive aspect of large-scale software development.

Software quality control is difficult to achieve, but the results of the careful application of **defect prevention** and **defect removal** techniques are quite good. The software producers who are most effective in quality control discover several derivative benefits as well: software with the highest quality levels also tends to be the least expensive to produce, tends to have minimum schedules during development, and also tends to have the highest levels of post-release user satisfaction.

The topic of defect prevention is outside the primary scope of this article. However, it is important to note that the set of technologies associated with defect prevention must be utilized concurrently with the technologies of defect removal in order to achieve high levels of final quality. The technologies which prevent defects are those concerned with optimizing both clarity and structure and with minimizing ambiguity. Joint application design (JAD) for preventing requirements defects; prototyping; reuse of certified material; clean-room development; any of the standard structured analysis, design, and coding techniques; and quality function deployment (QFD) for evaluating end-user quality demands are examples of the technologies associated with defect prevention. Many aspects of total quality management (TQM) programs are also associated with defect prevention.

TABLE 90.1 Origins and Causes of Software Defects

	Defect Causes		
Defect Origins	Errors of Omission	Errors of Commission	Errors of Clarity or Ambiguity
Requirements defects	Frequent	Seldom	Frequent
Design defects	Frequent	Frequent	Frequent
Coding defects	Frequent	Frequent	Frequent
Document defects	Frequent	Seldom	Frequent
Bad fix defects	Seldom	Frequent	Seldom

The Origins and Causes of Software Defects

Before software defects can be prevented or removed effectively, it is necessary to know where defects originate and what causes them. There are five primary origin points for software defects and three primary causes. The five primary origin points are (1) requirements, (2) design, (3) code, (4) user documentation, and (5) bad fixes, or secondary defects that occur while attempting to repair a primary defect. The three primary causes of software defects are (1) errors of omission, (2) errors of commission, and (3) errors of clarity or ambiguity. Table 90.1 shows the interaction of software defect origins and defect types.

The phrase *errors of omission* refers to problems caused by a failure to include vital information. Such errors are frequently encountered in requirements, specifications, source code, and user documents. An example of such an error became highly visible on February 29, 1992, when it was discovered that the calendar routine for a particular application omitted leap year calculations, thus shutting down the system at midnight on the 28th. From 15% to more than 30% of the post-deployment problems encountered in software can be traced to errors of omission. The probability and frequency of such errors rises with the size of the system.

The phrase *errors of commission* refers to problems caused by an action that is not correct, such as looping through a counter routine one time too often or including conditions in a specification that are mutually contradictory. Such errors are frequently encountered in design, code, and in "bad fixes" or secondary defects created as a by-product of repairing prior defects. From 40 to 65% of the post-deployment problems encountered in software can be traced to errors of commission, thus making it the most common source of software problems.

The phrase *errors of clarity or ambiguity* refers to problems caused by two or more interpretations of the same information. For example, a requirement may contain a phrase that an application should provide "rapid response time" without actually defining what "rapid" means. To the client, sub-second response time may have been intended, but to the development team one-minute response time may have been their interpretation. Such errors are frequently encountered in all software deliverables based on natural language such as English. They are also frequently encountered in source code itself, especially so if the code were not well structured. Also, certain languages such as APL and those using nested parentheses are notorious for being ambiguous. From less than 5% to more than 10% of the post-deployment problems encountered in software can be traced to errors of clarity or ambiguity.

When considering the origins of software defects, it is significant that the requirements and specifications themselves may be the source of as many as 40% of the defects later encountered after deployment. This fact implies that some forms of verification and validation, which assume the requirements and specifications are complete and correct, have hidden logical flaws.

The distribution of defects among the common origin points varies with the size and complexity of the application. Table 90.2 shows the probable defect distributions for four size ranges of software projects, using the language C as the nominal coding language.

The sum of the five defect categories is termed the *defect potential* of a software program or system [Jones, 1986]. The defect potential constitutes the universe of all errors and bugs that might cause an application to either fail to operate or to produce erratic and unacceptable results while operating.

The defect potential of software tends to be alarmingly high. When expressed with the traditional metric "KLOC" (where K stands for 1000 and LOC stands for lines of code) the observed defect potentials have ranged from about 10 bugs per KLOC to more than 100 bugs per KLOC, assuming procedural languages such as C,

TABLE 90.2 Software Defect Origins and Project Size

	Software Project Size (Statements in C Language)			
	1000	10,000	100,000	1,000,000
Requirements	5%	7%	10%	15%
Design	10%	15%	20%	25%
Code	70%	60%	50%	40%
Documents	5%	8%	10%	10%
Bad fixes	10%	10%	10%	10%
Total	100%	100%	100%	100%

Fortran, or Jovial. When expressed with the newer Function Point metric, the range of software defect potentials is from less than 2 to more than 10 bugs per Function Point.

When historical defect data is analyzed, it can easily be seen why defect prevention and defect removal are complementary and synergistic technologies. The defect prevention approaches are used to lower defect potentials, and the defect removal technologies are then used to eliminate residual defects.

Another dimension of concern is the *severity level* of software defects. The severity level scheme used by IBM, and widely adopted by other software producers, is based on a four-point scale. Severity 1 defects are those which stop the software completely or prevent it from being used at all. Severity 2 defects are those where major functions are disabled or unusable. Severity 3 defects are those that can be bypassed or which have a minor impact. Severity 4 defects are very minor problems which do not affect the operation of the software, for example, a spelling error in a text message.

Software **debugging,** testing, and verification methods should be effective against the entire defect potential of software, and not just against coding defects. Defect removal methods should approach 100% in efficiency against severity 1 and severity 2 defects. Further, defect removal methods should deal with errors of omission as well as errors of commission. Software operates at extremely high speed, so defect removal operations must deal with timing and performance problems as well as with structural and logical problems. Finally, and most difficult, defect removal methods should deal with errors of clarity or ambiguity. These are challenging tasks.

The Taxonomy and Efficiency of Software Defect Removal

There is no standard taxonomy that encompasses all of the many forms of defect removal that can be applied to software [Dunn, 1984]. For convenience, we will divide defect removal operations into three broad categories: pre-test defect removal, testing, and post-release defect removal.

Since the goal of defect removal is the elimination of bugs or defects, the primary figure of merit for a defect removal operation is its *efficiency* [Jones, 1991]. **Defect removal efficiency** is defined as the percent of latent defects which a given removal operation will detect. Cumulative defect removal efficiency is the overall efficiency of a complete series of pre-test removal activities combined with all test stages.

Calculating the efficiency of a defect removal operation, or a series, is necessarily a long-range operation. All defects encountered prior to release are enumerated and running totals are kept. After the first year of production, the client-reported defects and the pre-release defects are aggregated as a set, and the efficiencies of pre-release operations are then calibrated. Thus for large systems with multi-year development cycles, it may take up to five years before defect removal efficiency rates are fully known.

Major computer companies, telecommunications manufacturers, defense contractors, and some software vendors have been measuring defect removal efficiencies for more than 20 years. The body of empirical data is large enough to make general observations about the efficiencies of many different kinds of defect removal operation.

Empirical observations and long-range studies of commercial software have demonstrated that most forms of testing are less than 30% efficient. That is, any given stage of testing such as system testing or acceptance testing is likely to find less than 30% of the latent defects that are actually present. However, a carefully planned series of defect removal operations can achieve very respectable levels of cumulative defect removal efficiency. Computer manufacturers, commercial software producers, and defense software producers may utilize as many as 9 to 20 consecutive defect removal operations, and sometimes achieve cumulative defect removal efficiencies that approach the six sigma quality level, i.e., defect removal efficiency rates approaching 99.999999%.

Pre-Test Defect Removal

The set of defect removal activities for software carried out prior to the commencement of testing can be further subdivided into *manual defect removal* and *automated defect removal.*

A fundamental form of manual defect removal is termed *desk checking*, or the private review of a specification, code document, or document by the author. The efficiency of desk checking varies widely in response to both individual talents and to the clarity and structure of the materials being analyzed. However, most humans are not particularly efficient in finding their own mistakes, so the overall efficiency of desk checking is normally less than 35%.

The most widely utilized forms of manual defect removal are reviews, inspections, and walkthroughs. Unfortunately common usage blurs the distinction among these three forms of defect removal. All three are group activities where software deliverables such as the requirements, preliminary or detailed design, or a source code listing are discussed and evaluated by technical team members.

Casual reviews or informal walkthroughs seldom keep adequate records, and so their defect removal efficiencies are not precisely known. A few controlled studies carried out by large companies such as IBM indicate that informal reviews or walkthroughs may average less than 45% in defect removal efficiency.

The term *inspection* is often used to define a very rigorous form of manual analysis [Fagan, 1976]. Ideally, the inspection team will have received formal training before participating in their first inspection. Also, the inspection team will follow formal protocols in terms of preparation, execution, recording of defects encountered, and follow-up of the inspection session. The normal complement for a formal inspection includes a moderator, a recorder, a reader to paraphrase the material being inspected, the developer whose work is undergoing inspection, and often one or more additional inspectors.

Formal inspections of the kind just discussed have the highest measured efficiency of any kind of defect removal operation yet observed. For inspections of plans, specifications, and documents the defect removal efficiency of formal inspections can exceed 65%. Formal inspections of source code listings can exceed 60%.

Formal inspections are rather expensive, but extremely valuable and cost-effective. Not only do inspections achieve high rates of defect removal efficiency, but they are also effective against both errors of omission and errors of clarity or ambiguity, which are extremely difficult to find via testing. Formal inspections also operate as a defect prevention mechanism. Those who participate in the inspection process obviously learn a great deal about the kinds of defects encountered. These observations are kept in memory and usually lead to spontaneous reductions in potential defects within a few months of the adoption of formal inspections.

Military software projects in the United States are governed by various military standards such as DOD 2167A and DOD 1614. These standards call for an extensive series of reviews or inspections that are given the generic name of *verification and validation.* The word *verification* is generally defined as ensuring that each stage in the software process follows the logical intent of its predecessor. The term *validation* is generally defined as ensuring that each delivered feature or function can be traced back to a formal requirement.

U.S. military verification and validation has developed its own argot, and those interested in the full set of U.S. military verification and validation steps should refer to the military standards themselves or to some of the specialized books that deal with this topic. Examples of the forms of verification and validation used on U.S. military projects, and the three-letter abbreviations by which they are commonly known, include system requirements review (SRR), system design review (SDR), preliminary design review (PDR), critical design review (CDR).

The defect removal efficiencies of the various forms of military review have not been published, but there is no reason to doubt that they are equivalent or even superior to similar reviews in the civilian domain. The cumulative defect removal efficiency of the full series of U.S. military defect removal operations is rather good: from about 94% to well over 99% for critical weapons systems.

Large military software projects use a special kind of external review, termed *independent verification and validation* (IV&V). The IV&V activities are often performed by contracting organizations which specialize in this kind of work. The efficiency ranges of IV&V reviews have not been published, but there is no reason to doubt that they would achieve levels similar to those of civilian reviews and inspections, i.e., in the 40 to 60% range.

The most elaborate and formal type of manual defect removal methodology is that of *correctness proofs.* The technique of correctness proofs calls for using various mathematical and logical procedures to validate the assertions of software algorithms. Large-scale studies and data on the efficiency of correctness proofs have not

been published, and anecdotal results indicate efficiency levels of less than 30%. To date there is no empirical evidence that software where correctness proofs were used actually achieves lower defect rates in the field than software not using such proofs. Also, the number of correctness proofs which are themselves in error appears to constitute an alarming percentage, perhaps more than half, of all proofs attempted.

Among commercial and military software producers, a very wide spectrum of manual defect removal activities are performed by *software quality assurance* (SQA) teams [Dunn and Ullman, 1982]. The full set of activities associated with formal software quality assurance is outside the primary scope of this article. However, defect prediction, defect measurement, moderating and participating in formal inspections, test planning, test case execution, and providing training in quality-related topics are all aspects of software quality assurance. Organizations that have formal software quality assurance teams will typically average 10 to 15% higher in cumulative defect removal efficiency than organizations which lack such teams.

The suite of automated pre-test tools that facilitate software defect removal has improved significantly in recent years, but still has a number of gaps that may be filled in the future.

For requirements, specifications, user documents, and other software deliverables based on natural language, the use of word processors and their built-in spelling checkers has minimized the presence of minor spelling errors. Also available, although not used as widely as spelling checkers, are a variety of automated grammar and syntax checkers and even textual complexity analyzers. Errors of clarity and ambiguity, long a bane of software, can be reduced significantly by judicious usage of such tools.

Several categories of specialized tools have recently become available for software projects. Many CASE (computer-aided software engineering) tool suites have integral support for both producing and verifying structural descriptions of software applications. There are also tools that can automatically generate test cases from specifications and tools that can trace test cases back to the original requirements. While such tools do not fully eliminate errors of commission, they do provide welcome assistance.

Errors of omission in requirements, specifications, and user documents are the most resistant to automatic detection and elimination. Studies carried out on operating systems and large switching systems have revealed that in spite of the enormous volume of software specifications (sometimes more than 100 English words per source statement, with many diagrams and tables also occurring) more than 50% of the functionality in the source code could not be found described in the specifications or requirements. Indeed, one of the unsolved challenges of software engineering is to determine if it is even theoretically possible to fully specify large and complex software systems. Assuming that full specification is possible, several practical issues are also unknown: (1) What combination of text, graphics, and other symbols is optimal for software specifications? (2) What is the optimum volume or size of the specifications for an application of any given size? (3) What will be the impact of multi-media software extensions on specifications, debugging, and testing?

The number and utility of debugging tools for programming and programmers have made enormous strides over the past few years, and continued progress can be expected indefinitely. Software development in the 1990s often takes place with the aid of what is termed a *programming environment*. The environment constitutes the set of tools and aids which support various programming activities. For debugging purposes, software syntax checkers, trace routines, trap routines, static analyzers, complexity analyzers, cross-reference analyzers, comparators, and various data recording capabilities are fairly standard. Not every language and not every vendor provides the same level of debugging support, but the best are very good indeed.

Testing Software

There is no standard taxonomy for discussing testing. For convenience, it is useful to consider test planning, test case construction, test case execution, test coverage analysis, and test library control as the major topics encompassing software testing.

For such a mainstream activity as test planning, both the literature and tool suites are surprisingly sparse. The standard reference is Myers [1979]. The basic aspects of test planning are to consider which of the many varieties of testing will be carried out and to specify the number and kind of test cases that will be performed at each step. U.S. military specifications are fairly thorough in defining the contents of test plans. There are also commercial books and courses available, but on the whole the topic of test planning is underreported in the software literature.

A new method for estimating the number of test cases required for software was developed in 1991 and is starting to produce excellent results. A metric termed *Function Points* was invented by A. J. Albrecht of IBM and placed in the public domain in 1978 [Albrecht, 1979]. This metric is derived from the weighted sums of five parameters: the numbers of inputs, outputs, inquiries, logical files, and interfaces that constitute a software application [Garmus, 1991]. The Function Point total of an application can be enumerated during the requirements and design phases.

It is obvious that testing must be carried out for each of the factors utilized by the Function Point metric. Empirical observations indicate that from two to four test cases per Function Point are the normal quantities created by commercial software vendors and computer manufacturers. Thus for an application of 1000 Function Points in size, from 2000 to 4000 test cases may be required to test the user-defined functionality of the application.

Testing can be approached from several directions. Testing which attempts to validate user requirements or the functionality of software, without regard for its inner structure of the application, is termed *black-box* testing.

Black-box test case construction for software has long been a largely manual operation that is both labor-intensive and unreliable. (Empirical observations of operating system test libraries revealed more errors in the test cases than in the product being tested.) Test case generators have been used experimentally since the 1970s and are starting to appear as both stand-alone products and as parts of CASE tool suites. However, in order for test case generation to work effectively, the specifications or written description of the software must be fairly complete, rigorous, and valid in its own right. It is to no purpose to generate automatic test cases for incorrect specifications.

Testing which attempts to exercise the structure, branching, and control flows of software is termed *white-box* or sometimes *glass-box* testing. In this domain, a fairly rich variety of tools has come into existence since 1985 that can analyze the structure and complexity of software. For certain languages such as COBOL and C, tools are available that not only analyze complexity but can restructure or simplify it. In addition, there are tools that can either create or aid in the creation of test cases. Quite a number of tools are available that can monitor the execution of test cases and identify portions of code which have or have not been reached by any given test run.

Although not full test case generators, many supplemental testing tools are available that provide services such as creating matrices of how test cases interact with functions and modules. Also widely used are *record and playback* tools which capture all events during testing. One of the more widely used testing tool classes is that of *test coverage analyzers*. Test coverage analyzers dynamically monitor the code that is being executed while test cases are being run, and then report on any code or paths that may have been missed. Test coverage is not the same as removal efficiency: it is possible to execute 100% of the instructions in a software application without finding 100% of the bugs. However, for areas of code that are not executed at all, the removal efficiency may be zero. Software *defect tracking systems* are exceptionally useful.

Once created, effective test cases will have residual value long after the first release of a software product. Therefore, formal test case libraries are highly desirable, to ensure that future changes to software do not cause regression or damage to existing functionality. Test library tools occur in several CASE tool suites, are available as stand-alone products, and are also often constructed as custom in-house tools by the larger software vendors and computer manufacturers.

Test case execution can be carried out either singly for individual test cases or for entire sets of related test cases using *test scripts*. It is obvious that manually running one test case at a time is too expensive for large software projects. Tools for multi-test execution support, sometimes called a *test harness*, are available in either stand-alone form or as part of several CASE tool suites.

There are no rigorous definitions or standard naming conventions for the kinds and varieties of testing that occur for software. Some of the more common forms of testing include the following.

The testing of an individual module or program by the programmer who created it is normally called *unit test*. Unit testing can include both black-box and white-box test cases. The efficiency of unit test varies with the skill of the programmer and the size and complexity of the unit being tested. However, the observed defect removal efficiency of unit testing seldom exceeds 50% and the average efficiency hovers around 25%. For small projects developed by a single programmer, unit test may be the only test step performed.

For large software projects involving multiple programmers, modules, or components a number of test stages will normally occur that deal with testing multiple facets of the application.

The phrase *new function test* is used to define the testing of capabilities being created or added to an evolving software project. New function testing may consist of testing the aggregated work of several programmers. New function testing may be carried out by the developers themselves, or it may be carried out by a team of testing specialists. The observed efficiency of new function testing is in the 30% range when carried out by developers and in the 35% range when carried out by testing specialists.

The phrase *regression test* is used to define the testing of an evolving software product to ensure that existing capabilities have not been damaged or degraded as a by-product of adding new capabilities. Regression testing is normally carried out using test cases created for earlier releases, or at least created earlier in the development cycle of the current release. The observed removal efficiency of regression testing against the specific class of errors that it targets may exceed 50% when carried out by sophisticated organizations such as computer manufacturers or defense contractors. However, efficiencies in the 20 to 25% range are more common.

The phrase *stress test* is used to define a special kind of testing often carried out for real-time or embedded software. With stress testing, the software is executed at maximum load and under maximum performance levels, to ensure that it can meet critical timing requirements. The observed removal efficiency of stress testing often exceeds 50% against timing and performance-related problems. Stress testing is normally carried out by testing and performance specialists who are supported by fairly sophisticated test tool suites.

The phrase *integration test* is used to define a recurring series of tests that occur when components or modules of an evolving software product are added to the fundamental system. The purpose of integration testing is to ensure that the newer portions of the product can interact and operate safely with the existing portions. The observed removal efficiency of integration testing is normally in the 20 to 35% range. When performed by commercial software producers, integration test is normally carried out by testing specialists supported by fairly powerful tools.

The phrase *system test* is used to define the final, internal testing stage of a complete product before it is released to customers. The test suites that are executed during system test will include both special tests created for system test and also regression tests drawn from the product's test library. The observed removal efficiency of system testing is in the 25 to 35% range when such testing is done by the developers themselves. When system testing is carried out by testing specialists for commercial software, its defect removal efficiency may achieve levels approaching 50%. However, high defect removal efficiency this late in a development cycle often leads to alarming schedule slippages.

The phrase *independent test* is used to define a form of testing by an independent testing contractor or an outside organization. Independent testing is standard for U.S. military software and sometimes used for commercial software as well. Defect removal efficiency of independent test may approach 50%.

The phrase *field test* is used to define testing by early customers of a software product, often under a special agreement with the software vendor. Field testing often uses live data and actual customer usage as the vehicle for finding bugs, although prepared test cases may also occur. The defect removal efficiency of field test fluctuates wildly from customer to customer and product to product, but seldom exceeds 30%.

The phrase *acceptance test* is used to define testing by a specific client, as a precursor to determining whether to accept a software product or not. Acceptance testing may be covered by an actual contract, and if so is a major business factor. The defect removal efficiency of acceptance testing fluctuates from client to client and product to product, but seldom exceeds 30%.

Sometimes the phrases *alpha test* and *beta test* are used as a linked pair of related terms. Alpha testing defines the set of tests run by a development organization prior to release to early customers. Beta testing defines the set of tests and actual usage experiences of a small group of early customers. Unfortunately the term alpha test is so ambiguous that it is not possible to assign a defect removal efficiency rating. Beta testing is better understood, and the observed removal efficiency is usually in the 25% range.

Selecting an Optimal Series of Defect Prevention and Removal Operations

Since most forms of defect removal are less than 60% efficient, it is obvious that more than a single defect removal activity will be necessary for any software project. For mission-critical software systems where it is imperative to achieve a cumulative defect removal efficiency higher than 99%, then at least 10 discrete defect removal activities should be utilized, and careful defect prevention activities should also be planned (i.e., prototyping, use of structured techniques, etc.).

An effective series of defect removal operations utilized by computer manufacturers, telecommunication manufacturers, and defense contractors who wish to approach the six-sigma quality level (99.999999% efficiency) will include from 16 to 20 steps and resemble the following, although there are variances from company to company and project to project:

Pre-Test Defect Removal Activities

1. Requirements inspection
2. Functional design inspection
3. Logical design inspection
4. Test plan inspection
5. User documentation inspection
6. Desk checking and debugging by developers
7. Code inspection
8. Software quality assurance review
9. Independent verification and validation (military projects)

Testing Activities

10. Unit testing by developers
11. New function testing
12. Regression testing
13. Integration testing
14. Stress testing
15. Independent testing (military projects)
16. System testing
17. Field testing
18. Acceptance testing

Post-Release Defect Removal Activities

19. Incoming defect report analysis
20. Defect removal efficiency calibration (after one year of deployment)

The series of defect removal operations just illustrated can achieve cumulative defect removal efficiencies well in excess of 99%. However, these levels of efficiency are normally encountered only for large mission-critical or important commercial-grade software systems, such as operating systems, defense systems, and telecommunication systems.

A rough rule of thumb can predict that number of defect removal operations that are normally carried out for software projects. Using the Function Point total of the application as the starting point, calculate the 0.3rd power, and express the result as an integer. Thus an application of 100 Function Points in size would normally employ a series of 4 defect removal operations. An application of 1000 Function Points in size would normally employ a series of 8 defect removal operations. An application of 10,000 Function Points in size would normally employ 16 defect removal operations.

Post-Release Defect Removal

Defect removal operations do not cease when a software project is released. Indeed, one of the embarrassing facts about software is that post-deployment defect removal must sometimes continue indefinitely.

One of the critical tasks of post-release defect removal is the retroactive calculation of defect removal efficiency levels after software has been in use for one year. For a software project of some nominal size, such as 1000 Function Points, assume that 2000 bugs or defects were found via inspections and tests prior to release. The first year total of user-reported defects might be 50. Dividing the pre-release defect total (2000) by the sum of all defects (2050) indicates a provisional defect removal efficiency of 97.56%.

The defect potential of the application can be retroactively calculated at the same time, and in this case is 2.05 defects per Function Point. The annual rate of incoming user-reported defects for the first year is 0.05 defects per Function Point per year.

In addition to calculating defect removal efficiency, it is also useful to trace each incoming user-reported defect back to its origin, i.e., whether the defect originated in requirements, design, code, documentation, or as a result of a "bad fix." Defect origin analysis also requires a sophisticated tracking system, and full configuration control and traceability tools are useful adjuncts.

Note that the use of Function Points rather than the traditional KLOC metric is preferred for both defect potential and for defect origin analysis. KLOC metrics produce invalid and paradoxical results when applied to requirements, specification, and documentation error classes. Since defects outside the code itself often constitute more than 50% of the total bugs discovered, KLOC metrics are harmful to the long-range understanding of software quality.

If additional bugs are reported during the second or subsequent years, then defect removal efficiency levels should be recalculated as necessary. Removal efficiency analysis should also be performed for each specific review, inspection, and test step utilized on the project.

Calculating post-release defect removal efficiency requires an accurate quality measurement system throughout the life cycle. In addition, rules for counting post-release defects must be established. For example, if the same bug is reported more than once by different users, it should still count as only a single bug. If upon investigation user bug reports should turn out to be usage errors, hardware errors, or errors against some other software package, such bug reports should not be considered in calculating defect removal efficiency rates. If bugs are found by project personnel, rather than users, after release to customers, those bugs should still be considered to be in the post-release category.

The leading computer manufacturers, telecommunication companies, defense contractors, and software vendors utilize powerful defect tracking systems for monitoring both pre-release and post-release software defects. These systems record all incoming defects and the symptoms which were present when the defect was discovered. They offer a variety of supplemental facilities as well. Obviously statistical analysis on defect quantities and severities is a by-product of defect tracking systems. Less obvious, but quite important, is support for routing defects to the appropriate repair center and for notifying users of anticipated repair schedules. The most powerful defect tracking systems can show defect trends by product, by development laboratory, by time period, by country, by state, by city, by industry, and even by specific customer.

Recent Industry Trends in Software Quality Control

Since the previous edition of this handbook, a number of changes have occurred which affect software quality control approaches. Three factors, in particular, are beginning to exert a major influence on the way software is built, inspected, and tested:

1. The explosion of software viruses
2. The on-rushing approach of the Year 2000 problem
3. The marked increase in lawsuits where poor quality is alleged

These three new and emerging topics are likely to stay important software considerations well into the 21st century.

The Explosion of Software Viruses

Over the past 10 years, software viruses have grown from becoming a rare and minor annoyance to becoming major risks for software projects. The rampant increase in the number of viruses and the ever increasing sophistication of viruses has triggered a new subindustry of viral protection tools and software.

For the purpose of this handbook, viruses have also introduced new and "standard" testing phases aimed at eliminating all possibilities of viral contamination from the final delivered versions of software. As of 1996, every major software vendor in the world now includes viral protection screening as a standard activity, and many include multiple kinds of viral protection tools and approaches. This is a serious problem and it is not showing any sign of becoming less serious.

The On-Rushing Year 2000 Software Problem

Another very significant change in recent years has been the sudden recognition that at the end of the century when the calendar changes from 1999 to 2000 AD, many of the world's software applications may stop completely

or begin to put out incorrect data. The reason for the Year 2000 problem (as it has come to be called) is because of the historical practice of storing all dates in two-digit form. Thus, calendar year "1997" would be stored as "97". The obvious failing of the two-digit approach will occur when "99" becomes "00". This will throw off normal collating sequences and cause major errors. This problem is an interesting one because, although obvious for more than 20 years, the problem was never found during testing. The reason for this situation is because the two-digit form of storing dates originated as a requirement and was also included in software designs. Therefore, test cases tended to conform to erroneous requirements rather than actually find the problem. The Year 2000 problem provides a strong lesson to the software community that "quality means conformance to user requirements" may sometimes be dangerous. The Year 2000 problem is going to be the most expensive software problem in history, and it was caused primarily by conforming to a requirement that was dangerous and flawed. Now that the end of the century is rapidly approaching, every corporation and government agency in the world is frantically racing to make software repairs before the problem surfaces at the end of 1999. Some hidden and obscure instances of this problem will almost certainly slip through, causing major errors and probable litigation.

The Emergence of Software Quality as a Litigation Issue

On the topic of software litigation, another major change since the previous edition of this handbook is the increase in lawsuits and litigation involving allegations of poor quality. From observations on half a dozen such lawsuits, Table 90.3 shows the sequence of defect removal operations that are common when clients sue vendors for poor quality. It is an interesting observation that for outsource, military, and systems software that ends up in court for litigation which involves assertions of unacceptable or inadequate quality, the number of testing stages is much smaller, while formal design and code inspections were not utilized at all. Table 90.3 shows the typical patterns of defect removal activities for software projects larger than 1000 function points in size where the client sued the developing organization for producing software with inadequate quality levels. The table simply compares the pattern of defect removal operations observed for reliable software packages with high quality levels to the pattern noted during lawsuits where poor quality and low reliability was part of the litigation. It is interesting and ironic that the "reliable" projects actually had shorter development schedules than the projects that ended up in court with charges of poor quality. The reason for this is because finding and fixing software bugs late in the development cycle is the major cause for slipping delivery dates. High quality and short schedules usually are closely coupled, although few software project managers know this.

It is interesting that during the depositions and testimony of the litigation, the vendor often counter-charges that the short-cuts were made at the direct request of the client. Sometimes the vendors assert that the client

TABLE 90.3 Defect Removal and Testing Stages Noted During Litigation for Poor Quality

	Reliable Software	Software Involved in Litigation for Poor Quality
Formal design inspections	Used	Not used
Formal code inspections	Used	Not used
Subroutine testing	Used	Used
Unit testing	Used	Used
New function testing	Used	Rushed or omitted
Regression testing	Used	Rushed or omitted
Integration testing	Used	Used
System testing	Used	Rushed or omitted
Performance testing	Used	Rushed or omitted
Capacity testing	Used	Rushed or omitted

Note: The phrase "rushed or omitted" indicates that the vendor departed from best standard practices by eliminating a stage of defect removal or by rushing it in order to meet an arbitrary finish date or commitment to the client.

ordered the short-cuts even in the face of warnings that the results might be hazardous. As can be seen, software developed under contractual obligations is at some peril if quality control and testing approaches are not carefully performed.

Summary and Conclusions

Software has been prone to such a wide variety of error conditions that much of the cost and time associated with developing software projects is devoted to defect removal. In spite of the high expense levels and long schedules historically associated with defect removal, deployed software is often unstable and requires continuous maintenance.

Synergistic combinations of defect prevention and defect removal operations, accompanied by careful measurements and the long-range calibration of inspection and test efficiencies, can give software-producing organizations the ability to consistently create stable and reliable software packages. As the state of the art advances, both six-sigma quality levels or even zero-defect quality levels may be achievable. Achieving high quality levels also reduces development schedules and lowers development costs.

Defining Terms

Bug: The generic term for a defect or error in a software program. The term originated with Admiral Grace Hopper in the 1950s, who discovered an actual insect that was blocking a contact in an electromechanical device.

Debugging: The generic term for the process of eliminating bugs, i.e., errors, from software programs. The phrase is often used in a somewhat narrow sense to refer to the defect removal activities of individual programmers prior to the commencement of formal testing.

Defect prevention: The set of technologies which simplify complexity and reduce the probability of making errors when constructing software. Examples of defect prevention technologies include prototypes, structured methods, and reuse of certified components.

Defect removal: The set of activities concerned with finding and eliminating errors in software deliverables. This is a broad term which encompasses both manual and automated activities and which covers errors associated with requirement, design, documentation, code, and bad fixes or secondary defects created as a by-product of eliminating prior bugs.

Defect removal efficiency: The ratio of defects discovered and eliminated to defects present. Removal efficiency is normally expressed as a percent and is calculated based on all defects discovered prior to release and for the first year of deployment.

Related Topic

90.3 Programming Methodology

References

A.J. Albrecht, "Measuring application development productivity," *Proceedings of the Joint SHARE, GUIDE, IBM Application Development Conference,* October 1979, pp. 83–92.

R.H. Dunn, *Software Defect Removal,* New York: McGraw-Hill, 1984.

R.H. Dunn and R. Ullman, *Quality Assurance for Computer Software,* New York: McGraw-Hill, 1982.

M.E. Fagan, "Design and code inspections to reduce errors in program development," *IBM Systems Journal,* vol. 15, no. 3, pp. 182–211, 1976.

D. Garmus, Ed., *Function Point Counting Practices Manual,* Version 4.0. Westerville, Ohio: International Function Point Users Group (IFPUG), 1995.

C. Jones, *Programming Productivity,* New York: McGraw-Hill, 1986.

C. Jones, *Applied Software Measurement,* 2nd ed., New York: McGraw-Hill, 1996.

G. J. Myers, *The Art of Software Testing,* New York: John Wiley, 1979.

Further Information

Both specialized and general-interest journals cover defect removal and software quality control. There are also frequent conferences, seminars, and workshops on these topics. Some of the journals and books include:

ACM Transactions on Software Engineering and Methodology (TOSEM). The Association for Computing Machinery, 11 West 42nd Street, New York, NY 10036.

American Programmer. American Programmer, Inc., 161 West 86th Street, New York, NY 10024-3411.

IEE Software Engineering Journal. Institution of Electrical Engineers and the British Computer Society, Michael Faraday House, Six Hills Way, Stevenage, Herts, SG1 2AY, United Kingdom.

IEEE Transactions on Software Engineering. Institute of Electrical and Electronics Engineers, 345 East 47th Street, New York, NY 10017.

ITEA Journal of Test and Evaluation. International Test and Evaluation Association, 4400 Fair Lakes Court, Fairfax, VA 22033.

Software Maintenance News. Software Maintenance Society, 141 Saint Marks Place, Suite 5F, Staten Island, NY 10301.

T. Gilb and D. Graham, *Software Inspections,* Reading, Mass.: Addison Wesley, 1993.

B. Beizer, *Software Testing Techniques,* International Thomson Computer Press, 1995.

C. Jones, *Software Quality — Analysis and Guidelines for Success,* International Thomson Computer Press, 1996.

90.3 Programming Methodology

Johannes J. Martin

Programming methodology is concerned with the problem of producing and managing large software projects. It relates to programming as the theory of style relates to writing prose. Abiding by its rules does not, in itself, guarantee success, but ignoring them usually creates chaos. Many useful books have been written on the subject and there is a wealth of primary literature. Some of these books, that themselves contain numerous pointers to the primary literature, are listed at the end of this section.

The rules and recommendations of programming methodology are rather independent of particular programming languages; however, significant progress in the understanding of programming methods has always led to the design of new languages or the enhancement of existing ones. Because of the necessity of upward compatibility, enhanced old languages are, in comparison to new designs, less likely to realize progressive programming concepts satisfactorily. Consequently, the year of its design usually determines how well a language supports state-of-the-art programming methods.

Programming methodology promotes program correctness, maintainability, portability, and efficiency.

Program Correctness. For some relatively simple programs specifications and even verifications can be formalized. Unfortunately, however, formal specification methods are not available or not sufficiently developed for most programs of practical interest, and the specifications, given as narratives, are most likely less than complete. For the situations not explicitly covered by the specifications, the resulting program may exhibit bizarre behavior or simply terminate. Furthermore, informal methods of demonstrating correctness may also miss points that were, indeed, addressed by the specifications. The costs of these problems may range from time wasted in the workplace to injuries and loss of lives.

Short of formal methods for proving **program correctness,** simplicity of design gives informal methods a chance to succeed.

Maintainability. Since the programmer who will do the **program maintenance** is usually not the one who also did the original coding (or if he did he cannot be expected to remember the details), proper documentation and straightforward design and implementation of the pertinent algorithms is mandatory.

Portability. A program is called (easily) portable if it can be adapted to a different computer system without much change. One step toward **program portability** is using a high-level language common to both computer systems. The second step is avoiding the use of idiosyncratic features of a system if at all possible or, if such a

feature must be used, to isolate its use to small, well-designed program segments that can easily be rewritten for the new system.

Efficiency. The costs of running a computer program are determined (1) by the time needed to compute the desired result and (2) by the amount of storage space the program requires to run. As the costs of computer hardware have declined dramatically over the past decade, the importance of program efficiency has similarly declined. Yet, there are different dimensions to efficiency. Choosing the right algorithm is still crucial, while local optimization should not be pursued, if it increases the complexity and decreases the clarity of a program and thereby jeopardizes its correctness and maintainability.

Analysis of Algorithms

The solution of a problem as a sequence of computational steps, each taking finite time, is called an algorithm. As an example consider the algorithm that finds the greatest element in a random list of n elements:

1. Give the name A to the first element of the list.
2. For each element x of the list beginning with the second and proceeding to the last, if $x > A$ then give the name A to x.
3. The element named A is the greatest element in the list.

Assume that examining the "next" element, comparing it with A and possibly reassigning the name A to the new element takes a fixed amount of time that does not depend on the number of elements in the list. Then the total time for finding the maximum increases proportional to the length n of the list. We say the *time complexity* of the algorithm is of *order n*, denoted by $O(n)$.

While there are not many algorithms for finding the maximum (or minimum) in an unordered list, other tasks may have many different solutions with distinctly different performances. The task of putting a list in ascending or descending order (sorting), for example, has many different solutions, with time complexities of $O(n^2)$ and $O(n \cdot \log(n))$. Compared with the $O(n \cdot \log(n))$ algorithms, the $O(n^2)$ algorithms are simpler and their individual steps take less time, so that they are the appropriate choice, if n, the number of items to be sorted, is relatively small. However, as n grows, the balance tips decidedly toward the $O(n \cdot \log(n))$ algorithms. In order to compute the break-even point suppose that the individual steps of an $O(n \cdot \log(n))$ and an $O(n^2)$ algorithm take $k \cdot t$ and t units of time, respectively. Thus, for the break-even point we have

$$t \cdot n^2 = k \cdot t \cdot n \cdot \log(n)$$

hence

$$k = n/\log(n)$$

With $k = 5$, for example, the break-even point is $n = 13$. The difference in performance of the two algorithms for large numbers of n is quite dramatic. Again with $k = 5$, the $O(n \cdot \log(n))$ algorithm is 217 times faster if $n =$ 10,000, and it is 14,476 times faster if $n =$ 1,000,000. If the faster algorithm would need, e.g., 10 minutes to sort a million items, the slower one would require about 100 days.

Consequently, while efficiency is no longer a predominant concern, analyzing algorithms and choosing the proper one is still a most important step in the process of program development.

Flow of Control

The term *flow of control* refers to the order in which individual instructions of a program are performed. For the specification of the flow of control, assembly (machine) languages (and older high-level languages) provide conditional and unconditional branch instructions. Programmers discovered that the indiscriminate use of these instructions frequently leads to flow structures too complex to be analyzed at reasonable costs. On the other hand, keeping the flow of control reasonably simple is, indeed, possible since three elementary control

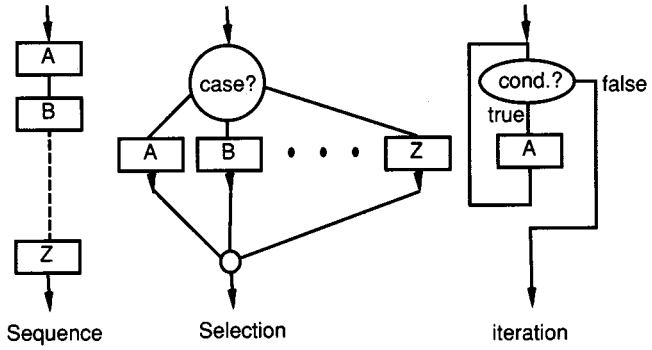


FIGURE 90.12 Basic control constructors.

constructors are sufficient to express arbitrary algorithms. Using only these constructors contributes in an essential way to the manageability of programs. The basic rules are as follows:

1. A proper computational unit has exactly one entry point and one exit.
2. New computational units are formed from existing units A, B, . . . by a.sequencing (perform A, then B, and so forth) b.selection (perform A or B or . . ., depending on some condition) c.iteration (while some condition prevails, repeat A).

Figure 90.12 shows diagrams that illustrate these rules.

High-level languages support these three constructors by (1) juxtaposition for sequencing, (2) if-then-else and case (switch) constructs for selection, and (3) while-do and for loops for iteration. The somewhat rigid restriction on loops sometimes requires the duplication of code. Some languages ease this problem by providing a *break* or *leave* statement that permits leaving a loop prematurely (Fig. 90.13). A typical example for a program that profits from this mechanism is a loop that reads and processes records and terminates when a record with some given property is found. In Fig. 90.13 block B reads and block A processes a record. With a strict while-loop the code of B must be duplicated.

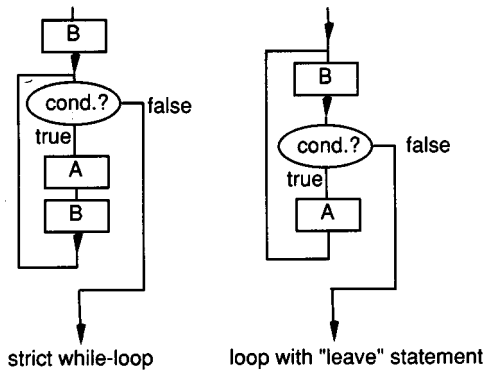


FIGURE 90.13 The use of a leave statement.

Abstraction

In programming, abstraction refers to the separation of what needs to be done from the details of how to do it. As it facilitates the separation of unrelated issues, it is a most powerful tool for making complex structures comprehensible. There are two varieties: *functional (procedural) abstraction* and *data abstraction.*

Almost all programming languages support *functional abstraction* by providing facilities for the definition of *subprograms.* The definition of a subprogram (procedure or function) associates a possibly complex sequence of instructions with an identifier and (usually) a list of parameters. The instruction sequence may then be performed by referring to its associated identifier and providing the required parameters. We could, for example, define a procedure *findRecord* that searches a given list of records for a record with a given key, e.g., an identification number. We could then use the statement

```
findRecord(theRecord, id_number, list);
```

instead of the instruction sequence that actually searches the list for the record. If this operation is needed more than once, then the obvious advantage is saving the replication of instructions. Much more important is a second advantage: When we study the text of a program that *uses findRecord*, we most likely wish to understand how this program processes and updates records or uses the information they contain. As this in itself may be

quite complex, we do not wish to be burdened simultaneously with the details of how to *find* a record, since this second problem has absolutely nothing to do with the process that we are trying to understand. Functional abstraction allows us to separate these different concerns and, thus, make each comprehensible.

The essential property of a data object is not its representation (is it an array, a record or something else?) but the operations that can be applied to it. The separation of these (abstract) operations from the subprograms that implement them, as well as from the representation of the object, is called *data abstraction*. This separation is possible because only the implementations of the operations need to know about the actual representation of the object. A system of data objects and operations defined by data abstraction is called an *abstract data type*. See Chapter 81.3 on *data types and data structures* for more details.

Modularity

With the tool of abstraction programs are made modular, that is, they are broken into fairly small, manageable parts. Each part, called a module, consists of one or more procedures or functions and addresses a particular detail of a project. A module may consist of a sorting program, a package for matrix or time and date calculations, a team of programs that preprocess input or format output, and the like.

A program should be broken up into smaller parts for one of two reasons: (1) if it addresses more than one problem and (2) if it is physically too large.

First, a procedure (function) should address a single problem. If more than one problem is addressed by the same procedure, a reader of the program text experiences a distracting shift in subject, which makes the comprehension of the text unnecessarily difficult. For example, a program involved in interest calculations should not also solve calendar problems (such as how many days are between March 15 and June 27) but use a calendar module that provides the abstract data type *Date* to obtain the desired results. Similarly, a sorting procedure should not be burdened with the details of comparing records. Instead, a different function should be written that contains the details of the comparison, and this function should then be invoked by the sorting program. The program fragments in Example 1, written in the programming language C, illustrate the separation of these levels of concern.

Considering Example 2, the advantage of separating levels of concern may appear to be subtle. The reason is the small size of the original sort program. A quote by B. Stroustrup [1991] emphasizes this point, "You can make a small program (less than 1000 lines) work through brute force even when breaking every rule of good style. For a larger program, this is simply not so. . . ." Advantages that do not seem very significant for small programs become invaluable for large ones. Writing large programs is an inherently difficult activity that demands a high level of concentration, helped by separating and hindered by mixing unrelated levels of concern.

Second, as a rule of thumb, the text of a procedure should fit onto one page (or into a window). A single page of text can be studied exclusively by eye movement and, thus, comprehended more easily than text that spans over several pages. That is not to say that longer procedures are necessarily an indication of poor programming methods, but there should be compelling reasons if the limit of one page is to be exceeded.

Simple Hierarchical Structuring

The relation among modules should be hierarchical, i.e., modules should be the vertices of a directed acyclic graph whose edges describe how modules refer to other modules. In such a structure, called a top-down design, each module can be understood in terms of the modules to which it refers directly. Yet, hierarchical structuring by itself does not yield simple and comprehensible programs. In addition, interfaces between modules must be kept narrow, that is, (1) the use of nonlocal variables must be avoided and (2) the number of parameters passed to a subprogram should be held to a minimum.

First, the scope rules of many programming languages permit subprograms to access entities of the program in which they are defined (static scoping) or of the calling program (dynamic scoping). These entities are called nonlocal in contrast to those defined within the subprogram itself. *Nonvariable* entities can and should be broadcast by means of nonlocal, preferably global definitions (i.e., definitions available to all modules and subprograms within an entire project). For these entities—types, procedures, and constant values—global definitions are very beneficial, since changes that may become necessary need to be made only at the place of

Example 1. A program that addresses more than one problem:

```
void sort (recType *table, int size) /* recType is a structure with fields
                                        suitable for a personnel record */
{
   int i, j, result;
   recType temp;
  for (i = 0; i < size-1; i++)
    for (j = i+1; j<size; j++){
        if ((result = strcmp(table[i].last, table[j].last)) == 0)
           if ((result = strcmp(table[i].f irst, table[j].f irst)) == 0)
              if ((result = strcmp(table[i].midl, table[j].midl)) == 0)
                 if ((result = datecmp(table[i].birthdate, table[j].birthdate)) == 0)
                     result = addresscmp(table[i].address, table[j].address);
        if (result > 0){
             temp = table[i];
             table[i] = table[j];
             table[j] = temp;
         }
   }
}
```

Example 2. Programs tackling one problem at a time:

```
BOOL isGreater (recType a, recType b);
void swap (recType *a, recType *b);

void sort (recType *table, int size)
{
   int i, j;
  for (i = 0; i < size-1; i++)
    for (j = i+1; j<size; j++)
        if (isGreater(table[i], table[j])) swap(&table[i], &table[j]);
}

BOOL isGreater (recType a, recType b)
{
   int result;
  if ((result = strcmp(a.last, b.last)) == 0)
     if ((result = strcmp(a.f irst, b.f irst)) == 0)
       if ((result = strcmp(a.midl, b.midl)) == 0)
          if ((result = datecmp(a.birthdate, b.birthdate)) == 0)
              result = addresscmp(a.address, b.address);
  return result > 0;
}

void swap (recType *a, recType *b)
{
   recType temp;
  temp = * a;
  *a = *b;
  *b = temp;
}
```

definition, not at every place of usage. Moreover, global definitions that replace numeric or alphanumeric constants by descriptive identifiers improve the readability of a program. Nonlocal *variables,* on the other hand, are to be avoided. For reasons of simplicity and clarity, the task performed by a subprogram should be determined *exclusively* by the subprogram itself and the values of its parameters. As a result, all information needed to understand why a subprogram may behave differently for different invocations is provided at each point of invocation. This rule, related to the concept of *referential transparency,* is clearly violated if a subprogram

uses nonlocal variables: in order to understand the behavior of the subprogram, information contained in the nonlocal variables must be consulted. These variables may appear nowhere near the point of invocation.

Similarly, because their return values are clearly identified, functions that do not change their parameters are preferred over those that do and over procedures. As functions are not always appropriate, a programmer choosing to use a procedure should aid the reader's comprehension by adopting some documented standard of consistently changing either the procedure's first or last parameter(s).

Second, the number of parameters should be kept to a minimum. If more than four or five parameters seem to be required, programmers should seriously consider packaging some of them into a record or an array. Of course, only values that in some logical sense belong together should be packaged, while parameters that serve different functions should be kept separate. For example, suppose the operation *updateRecord* is to find a personnel record with a given identification number in a file or a list and then change certain fields such as the job title, rate of pay, etc. A programmer may be tempted to define

```
updateRecord (Rectype myRecord);
```

yet the form

```
updateRecord (Idtype idNumber, Packtype attributes)
```

is better, since it suggests that the record with the key *idNumber* is the one to be updated by modifying the *attributes*.

The term *top-down design* (also called step-wise refinement) is frequently misunderstood as exclusively describing the *method* of design rather than the *structure* of the *finished product*. However, the creative process of designing software (or anything else) frequently proceeds in a rather erratic fashion, and there is no fault in this as long as the final product through analyses and reorganizations has a simple hierarchical structure. This is not to say that the process of design should not be guided by a top-down analysis; however, neither should a programmer's creativity be hampered by the straightjacket of a formalistic design principle, nor can a poorly designed program be defended with reference to the superior (top-down design) method used for its creation.

Object-Oriented Programming

In languages that support object-oriented programming, *classes* (i.e., data types) of objects are defined by specifying (1) the variables that each object will own as *instance variables* and (2) operations, called *methods*, applicable to the objects of the class. As a difference in style, these methods are not invoked like functions or procedures, but are *sent* to an object as a *message*. The expression [*window moveTo* : *x* : *y*], for example, is a message in the programming language *Objective C*, a dialect of C. Here the object *window*, which may represent a window on the screen, is instructed to apply to itself the method *moveTo* using the parameters *x* and *y*.

New objects of a class are created—usually dynamically—by *factory methods* addressed to the *class* itself. These methods allocate the equivalent of a record whose fields are the instance variables of the object and return a reference to this record, which represents the new object. After its creation, an object can receive messages from other objects.

To data abstraction, object-oriented programming adds the concept of inheritance: From an existing class new (sub)classes can be derived by adding additional instance variables and/or methods. Each subclass inherits the instance variables and methods of its superclass. This encourages the use of existing code for new purposes.

With object-oriented programming, classes become the modules of a project. For the most part, the rules of hierarchical structuring apply, interfaces should be kept narrow, and nonlocal variables ought to be restricted to the instance variables of an object. There are, however, at least two situations, both occurring with modern graphical user interfaces, where the strict hierarchical structure is replaced by a structure of mutual referencing.

Objects—instances of classes—frequently represent visible objects on the screen such as windows, panels, menus, or parts thereof. It may now happen that actions (such as mouse clicks) applied to one object influence another object and vice versa. For example, the selection of an object, such as a button or a text field, may launch or modify an inspector panel that, in turn, allows the modification of the appearance or function of the button or the text field (mutual referencing).

In conventional programming, library routines are, with rare exceptions, at the bottom of the hierarchy, i.e., they are called by subprograms written by the user of the library (later referred to as the "user") but they do not call subprograms that the user has written.

With object-oriented systems, library programs may have user-written *delegates* whose methods are used by the library object. For example, a window—a library object—may send the message "windowWillClose" (written by the user) to its delegate in response to the operators clicking of the window's close button. In response, the delegate may now interrogate the window in order to determine whether its contents have been changed and possibly require saving (mutual referencing).

Furthermore, buttons, menus, and the like are library objects. Yet, when operated, they usually activate user methods. Again, the user program may then interrogate the button about its state or request that the button highlight itself, etc. (mutual referencing).

While in the cases discussed, mutual referencing seems to be natural and appropriate, it should, in general, be used very sparingly.

Program Testing

Aside from keeping the structure of a program as simple as possible, a top-down design can be tested in a divide-and-conquer fashion. There are basically two possible strategies: bottom-up testing and top-down testing.

Bottom-up testing proceeds by first testing all those modules that do not depend on other modules (except library modules). One then proceeds always testing those modules that use only modules already tested earlier. Testing of each (sub)program must exercise all statements of the program, and it must ensure that the program handles all exceptional responses from supporting subprograms correctly. For each module a driver program must be written that invokes the module's functions and procedures in an appropriate way. If a program is developed in a top-down fashion, this method cannot be used until a substantial part of the project is completed. Thus design flaws may not be detected until corrections have become difficult and costly.

Top-down testing begins with the modules that are not used by any other module. Since the supporting modules have not been tested yet (or do not even exist yet), simple stand-in programs must be written that simulate the actual programs. This can be done by replacing the computation of the actual program with the programmer, who enters the required results from the keyboard in response to the parameter values displayed on the screen. This method is nontrivial, especially if the supporting program deals with complex objects.

In practice both methods are being used. Frequently, developing the procedure for top-down testing by itself leads to the discovery of errors and can prevent major design flaws.

Defining Terms

Program correctness: A program's conformation with its specifications.
Program maintenance: Modifications and (late) corrections of programs already released for use.
Program portability: A program is called (easily) portable if it can be adapted to a different computer system without much change.

Related Topic

87.3 Data Types and Data Structures

References

B.W. Kernighan and D.M. Ritchie, *The C Programming Language*, Englewood Cliffs, N.J.: Prentice-Hall, 1988.
J.J. Martin, *Data Types and Data Structures*, C.A.R. Hoare, Series Ed., New York: Prentice-Hall, 1986.
NeXT Step Concepts, *Objective C*, NeXT Developers' Library, NeXT, Inc., 1991, chap 3.
J.W. Rozenblit, *Codesign: Software/Hardware Engineering*, IEEE Press, 1995.
J.C. Reynolds, *The Craft of Programmings*, C.A.R. Hoare, Series Ed., New York: Prentice-Hall, 1983.

Proceedings of 1996 International Conference on Software Maintenance, Los Alamitos, Calif.: IEEE Computer Society, 1996.
B. Stroustrup, *The C++ Programming Language*, Reading, Mass.: Addison-Wesley, 1991.
J. Welsh et al., *Sequential Program Structures*, C.A.R. Hoare, Series Ed., New York: Prentice-Hall, 1984.
N. Wirth, *Algorithms + Data Structures = Programs,* Berlin: Springer-Verlag, 1984.

Further Information

Textbooks on programming usually address the subject of good programming style in great detail. More information can be found in articles on object-oriented programming and design and on software engineering as published, for example, in the proceedings of the annual conferences on *Object Oriented Programming Systems, Languages and Applications (OOPSLA)* sponsored by the Association for Computing Machinery (ACM) and on *Computer Software and Applications (CompSac)* sponsored by the Institute of Electrical and Electronics Engineers, Inc. (IEEE). Articles on the subject are also found in periodicals such as *IEEE Transactions on Software Engineering*, *IEEE Transactions on Computers*, *ACM Transactions on Software Engineering and Methodology, Software Engineering Notes* published by the ACM Special Interest Group on Software Engineering, *ACM Transactions on Programming Languages and Systems*, the *Communications of the ACM*, or *Acta Informatica.*

Rozanski, E.P. "Computer Graphics"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

91

Computer Graphics

91.1 Introduction

91.2 Graphics Hardware
Hard Copy Technologies • Display Technologies • Standard CRT • Other Display Technologies

91.3 Graphics Software
Engineering Software Packages • General Purpose Libraries and Packages • Solid Modeling Packages • Object-Oriented Programming • Plotting and Page Description Languages • Interaction

91.4 Conclusion

Evelyn P. Rozanski
Rochester Institute of Technology

91.1 Introduction

The term **computer graphics** refers to the generation, representation, manipulation, processing, and display of data by a computer. Computer-generated images may be real or imagined, animated or still, two-dimensional (2-D) or three-dimensional (3-D). Today most computers, particularly those in the PC, Macintosh, or workstation categories, have graphics capability. Their central components are a graphical display device, usually a cathode ray tube (CRT), and one or more input devices (e.g., keyboard, mouse, digitizer, data glove). Output devices include laser printers or video or such other displays as goggles or "eyephones" as in the case of some **virtual reality** systems.

Computer graphics encompasses a wide variety of applications. It has expanded its scope from the mundane business/presentation graphics to placing desktop publishing at everyone's fingertips. Highly interactive real-time systems are used in flight simulators where the display represents changes in the scene or landscape. In engineering, computer-aided design (CAD) systems allow users to create, store, manipulate, and test objects and designs. Fully integrated systems allow standard component parts libraries to be incorporated into a product. Product design and drafting information is fed into manufacturing operations via numerical control interfaces. Other engineering applications that make extensive use of graphics include very large scale integration (VLSI) and solid modeling.

Graphics has emerged as the vehicle for visualizing physical phenomena and the volume visualization of complex datasets [Purgathofer and Schonhut, 1989; Vince, 1990; Kaufmen et al., 1996]. Some examples include the medical modeling of the anatomy and MRIs [Kaufmen et al., 1996]. One application simulates laboratory testing of a new friction material for disc brakes and visualizes temperature distribution of the brakes' ability to conduct or absorb heat [Purgathofer and Schonhut, 1989]. In mathematics, B. B. Mandelbrot defined the geometry of **fractals**. Fractals, geometrical self-similar objects with fractional dimension, form a powerful tool for generating objects that resemble natural phenomena such as mountains, trees, and coastlines [de Ruiter, 1988; Mandelbrot, 1982].

In the world of animation, the computer has taken the drudgery out of transforming and redrawing objects. It has enhanced cell animation as well as produced glitzy Hollywood special effects such as morphing, a process of letting the computer transform one image to another by generating all the in-between images.

One of the most spectacular uses of graphics is in the area of virtual reality (VR). This technology, which uses high-resolution graphics terminals and head-mounted displays (HMD) or eyephones, provides the user

with a stereo view of a virtual world and an ability to navigate through it. These systems have a tracking device to determine the position of the user and devices, such as data gloves, for inputting commands [Thomas and Stuart, 1992]. Applications include simulation and architecture.

Research in the area of computer graphics has centered on all aspects of hardware, software, and algorithm development. Some of these areas are

1. Object-oriented environments: Design of programming languages, tools, databases, user interfaces, and animation [Purgathofer and Schonhut, 1989; Cunningham et al., 1992; de Ruiter, 1988].
2. Virtual reality: The design of system architecture, the creation and integration of component hardwares, the creation of software, the building of virtual environments, the development of real-world applications, and the study of philosophical and human perceptual issues [Stuart, 1992].
3. **Scientific visualization:** Graphics software solutions, practical implementations, user interfaces, high-resolution hard copy, data representation and metafiles [Purgathofer and Schonhut, 1989].
4. Algorithmic design: Ray tracing [Straber, 1987].
5. Hardware design: Workstation architectures, support for geometric modeling [Straber, 1987].
6. Color models and manipulation [Purgathofer and Schonhut, 1989].
7. Page description languages (PDLs): PostScript interpreters [Purgathofer and Schonhut, 1989].
8. CAD and solid modeling: VLSI, data exchange, geometric modeling [de Ruiter, 1988; Purgathofer and Schonhut, 1989].

91.2 Graphics Hardware

Computer graphics systems comprise several different output components in which to display computer-generated images. These components are classified into two groups: (1) hard copy technologies and (2) display technologies.

Hard Copy Technologies

Hard copy technologies include printers, pen plotters, electrostatic plotters, laser printers, ink-jet plotters, thermal transfer plotters, and film recorders [Foley et al., 1996]. These devices use either a raster or vector style of drawing. The raster style uses discrete dots, and the vector style uses a continuous drawing motion. Each display device is distinguished by its dot size and the number of dots per inch, known as *addressability*. The closer the dots, the smoother the image. The smaller the dot, the finer the detail. *Resolution* is related to dot size and is the number of distinguishable lines per inch. This may vary in the horizontal and vertical directions. High-resolution devices have fine detail, smooth lines, and crisp images.

Color may be achieved in several ways, depending on the device. Some devices use multicolored ribbons with single print heads, multiple print heads with different ribbons, or overstriking to combine colors. Other devices use color pens, spray (e.g., ink jet), toner (e.g., laser printer, electrostatic plotters), or pigment from colored wax paper (e.g., thermal transfer).

The hard copy devices vary in color and intensity levels, addressability, dot size, cost, image quality, and speed. The laser printer is becoming the most common, high-quality output device in this category [Foley et al., 1996].

Display Technologies

Displays are, for the most part, characterized by their responsiveness to a changing image. As with the hard copy technologies, display technologies vary greatly with respect to performance and cost. Guidelines for comparisons are based on the following characteristics: power consumption, screen size, depth, weight, ruggedness, brightness, addressability, contrast, intensity levels per dot, viewing angle, color capability, and relative cost.

Standard CRT

The most common component of graphics displays has been the CRT, which is used in televisions. The CRT is composed of five parts: (1) the electron gun, which when heated emits electrons at an appropriate rate; (2) the control grid, which regulates the flow of electrons; (3) the focusing system, which concentrates the beam

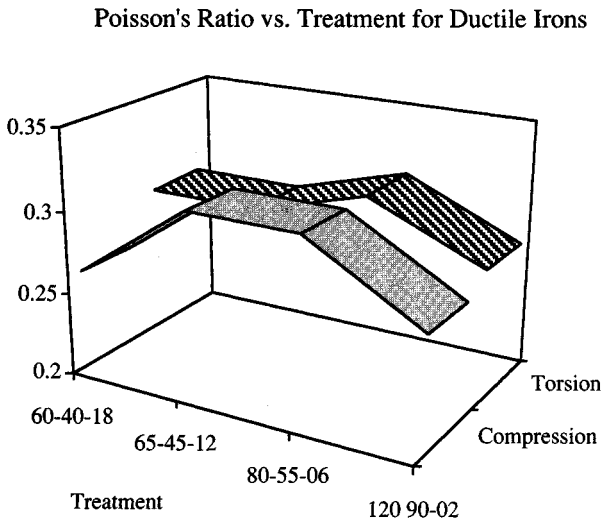


FIGURE 91.1 An example of a figure generated on the Macintosh with Microsoft Excel 3.0, showing the effect of four different treatments on two different measured variables. Although this information could be presented in two dimensions, the 3-D illustration can be more intuitive and interesting.

into a fine point; (4) the deflection system, which directs the beam to the appropriate location; and (5) the phosphor screen, which glows when bombarded with the electron beam. The *persistence* of the phosphor is defined as the time from the removal of excitation to when the phosphorescence has decayed to 10% of the initial light output [Foley et al., 1996]. Depending on the persistence of the phosphor used, the screen will need to be continually *refreshed*, or redrawn. Color is produced by laying triads of red-green-blue (RGB) phosphors on the screen and using three electron guns, one for each color, to excite the corresponding phosphor.

The raster CRT scans the image, one row at a time, from a matrix whose elements correspond to a *pixel,* or point on the screen. This matrix is referred to as the *frame buffer* and allows for a constant refresh rate, usually 60 times per second. Systems may also have more than one frame buffer (double buffer) to facilitate faster image generation. These displays include high resolution (1024×11280), SVGA (768×1024), NTSC ($\sim 350 \times 480$) and HDTV (720×1280 and 1080×1920) [Baily et al., 1996]. In vector CRT displays, the picture is generated in a continuous sweep, much like tracing an image on paper. The refresh rate on the vector displays is a function of the complexity of the image. The result may be a noticeable flicker on the screen.

Other Display Technologies

1. Direct view storage tubes (DVST): These devices were the primary displays used in earlier systems. These vector drawing devices stored their images on a grid, which was continually bombarded with electrons in order to transfer the image to the screen. The advantage was that once the image was drawn, the refresh process took place independently of the complexity of the image, thereby producing a constant image on the screen. The disadvantage of these systems was that no part of the image could be selectively erased without erasing the entire grid and resending the modified image to the display.
2. Liquid crystal display (LCD): This device uses matrix addressing and refreshes the display one row at a time. Appropriate voltages are applied to the crystals, causing them to line up. They remain polarized, not allowing light to pass through; light is absorbed, causing dark spots on the display. These devices are light in weight, rugged, and have a low power consumption, fair intensity, and low cost.
3. Plasma panels: These devices have an array of neon bulbs between glass plates, which may be turned on or off. While color is possible, it has not been commercially available. These devices excel in screen size, weight, ruggedness, and brightness characteristics but are generally high in cost.
4. Electroluminescent displays: These devices also use a grid-like structure for addressing elements. The light-emitting material, a zinc sulfide doped with manganese, is available in color. These devices have excellent brightness characteristics but are high in cost.

91.3 Graphics Software

Software for scientific and engineering applications has changed dramatically in the past several years. In the 1970s and early 1980s, there were few graphics software tools available. Most of the engineering packages were in the CAD area. Many specific engineering applications required users to develop and implement programs to solve their problems. These programs were written in the Fortran or C programming languages using low-level graphical commands or calls to some standard or quasi-standard (e.g., the CORE package) graphical routines. Most of these systems were developed for a mainframe computer environment. A trend begun in the late 1980s resulted in a change in computing hardware environments as well as in software approaches. Predominantly, the hardware platforms are PCs, microcomputers, and powerful Unix workstations, with most of these machines having excellent graphics capabilities. Software moved from code generation to customized stand-alone scientific and engineering software tools. Software development uses standard languages and graphical user interfaces for CH, C, Fortran, and Pascal, as well as more sophisticated languages such as JAVA, HyperText, Unix X.11, Microsoft's Windows, and PostScript. The technical community is relying more and more on the increased power of computers to easily support software packages that manipulate complex data and represent them in a visual manner.

Engineering Software Packages

Several commercial scientific and engineering software packages have graphics functionality. It is difficult to distinguish graphics or visualization capabilities without discussing some of these packages. An excellent reference is found in the *IEEE Spectrum Focus Report: Software*.

These graphical application software packages fall into five categories:

1. Logic simulation for application-specific integrated circuits (ASICs). Software in this area might display a schematic of a multigate ASIC from large functional building blocks. These blocks could represent a finite-state machine with several states and gates. Representative packages are Mentor Graphics' Auto-Logic, Cadence Design Systems' HDL Synthesizer and Optimizer, and Teradyne's Frenchip. HDL is a hardware description language.
2. Electromagnetic design and simulation. Software in this area might simulate a printed-circuit board for a 32-bit-wide, 8-bit-byte reversal network. Multilayers of a board are displayed, with colors indicating current densities in lines. Representative systems are Hewlett-Packard's High Frequency Structure Simulator (HFSS), a finite-element-based product having animation of field plots and conductor loss and 3-D full-wave solution and S-parameter output; Sonnet Software's "em" package with animation of conductor currents; and Compact Software's Microwave Explorer with X-Windows and OSF Motif graphical interfaces.
3. Data acquisition, analysis, display, and technical reporting. Systems in this area have compute-intensive analysis routines and enhanced visualization of data which capitalize on sharper display resolutions. These packages could produce plots and graphs based on acquired data that are displayed in several windows at once; changes to one window could result in recalculation and updating of corresponding windows. Packages in this area frequently have support for standard languages and graphical user interfaces for C and Fortran as well as the Unix X.11 interface or Microsoft's Windows. Representative packages are HP's VEE-Test; Design Science's MathType; DSP Development's DADiSP; National Instruments' LabWindows; Speakeasy Computing's Speakeasy Zeta, which features user-tailored graphical user interface and PostScript output; and Mihalisin Associates' Temple-Graph, which produces a color PostScript output link to *Mathematica*.
4. Mathematical calculations and graphics for visualization. Applications for these packages would be curve fitting, evaluation of integrals, statistical analysis, signal processing, and numerical analysis. Features include programmability in languages such as C, Fortran, and Pascal and 2-D and 3-D representations. The leading package in this area is *Mathematica* by Wolfram Research, which is a general system and programming language for numerical, symbolic, and graphical computations in engineering, research, science, financial analysis, and education [Wolfram, 1991]. Other packages are Amtec Engineering's Tecplot, Integrated Systems' Xmath, MathWorks' Mathlab, Jandel Scientific's SigmaPlot, and NAG's Axiom.

FIGURE 91.2 Examples of rendered 3-D figures generated on the Macintosh 7100 computer with 2MB RAM using Strata Studio Pro software, showing the effects of different rendering, coloring, and lighting parameters on a 3-D scene. Computer artist: (top) Joel Rosen, (bottom) Alex Dao, M.F.A. in Computer Graphics Design program students, Rochester Institute of Technology.

5. Digital signal processors for embedded systems. The tools available in this area let the engineer focus on the application rather than the programming details. The Audio Frequency Fourier Analyzer by National Instruments is a combination of graphical programming with development software for the Macintosh environment. Signals can be analyzed, manipulated, and displayed using custom graphics software. Some packages allow for programming in C and user interfaces. Representative packages include

FIGURE 91.3 Examples of 3-D figures generated on the Macintosh 7100 computer with 2MB RAM using Strata Studio Pro software, showing the effects of different rendering, coloring, and lighting paramaters on a 3-D scene. Computer artist: (top) Hyung-Joo Lee, (bottom) Jennifer Cisney, MFA in Computer Graphics Design Program students, Rochester Institute of Technology.

Signal Technology's N!Power, which has **object-oriented programming** and linkage to X-Windows, and Bitware Research Systems' DsqHq with real-time graphics and algorithm design.

General Purpose Libraries and Packages

Traditionally, graphical software systems are developed as a result of programming in high-level languages with interfaces to standard or quasi-standard software packages. These packages attempt to address the issues of device independence and application program portability by allowing systems to drive a wide variety of display devices as well as isolating the programmer from machine-specific graphics commands. Portability of programs is enhanced by allowing the user to move an application from one system to another. The primary programming languages include C, Fortran, and Pascal.

The quasi-standard graphical package is ACM/SIGGRAPH's Core system developed in 1977 and revised in 1979. While it was not a formally recognized standard, it did fulfill a role as a baseline specification for graphical systems [Foley et al., 1996]. The two official standards are GKS-3D, the 3-D Graphical Kernel System; and

PHIGS and PHIGS+, the Programmer's Hierarchical Interactive Graphics System. Both systems support graphics primitives, such as lines, polygons, and character strings, and their attributes. The GKS system allows for groupings into segments with no nesting capabilities. PHIGS supports geometrical transformations (i.e., scaling, translating, and rotating) and a database structure that allows for selective editing and manipulation of the picture. PHIGS runs best when there is hardware support for the transformation, clipping, and **rendering** features. Other software include the cross-platform OpenGL, which is a low-level graphics rendering and imaging library, and Inventor, which is object-oriented and built on top of Open GL [Reynolds and Danielson, 1996].

In traditional graphical systems development, image data are stored either as Cartesian coordinates or as vectors. These data are manipulated through the geometrical transformations of scaling, translating, and rotating in a reference system known as the *world coordinate system* (WCS). The units of the WCS system might be inches, millimeters, or miles. Physical devices use their own coordinate systems known as *screen coordinate systems* (SCS). In order to ready the image for display, a *viewing transformation* takes place, which changes the image data in the WCS to its corresponding device-specific screen coordinates in SCS. A *window* or portion of the world picture is chosen to be shown in an area of the display known as the *viewport*. Because some of the data in the world could be outside the window, a *clipping* operation is necessary. Clipping will eliminate any data points outside the window. These values are then converted to an intermediate coordinate system known as the *normalized device coordinate system* (NDC). Values in this system are in the range of 0 to 1. Because a viewport may be any portion of the display area and the image could be displayed on more than one device, the NDC values are easily adjusted to screen coordinates. In 3-D, the clipping volume uses the viewing transformation which must take into account the *view reference point* (i.e., the position from which an object is to be viewed) and the *perspective* or *parallel projection* (i.e., the conversion from the object's 3-D coordinates to the screen's 2-D coordinates).

Solid Modeling Packages

Feature-based systems such as solid or geometric modeling rather than mathematical-based systems form the basis of some CAD systems. **Solid modeling** (SM) systems use constructive solid geometry to build complicated objects. These systems have a descriptive language which uses a database of 3-D primitive objects such as block, cylinder, sphere, wedge, cone, and torus. These solids are combined to form other solids using the set operators of union, intersection, and difference. The resultant object can then be named, saved, and positioned into a picture or drawing. Attributes stored with the objects allow them to be displayed in wire-frame format or as a completely rendered image. Representative SM systems are MAGI (Mathematical Applications Group, Inc.), Synthavision, PADL-2 (Production Automation Project), GM Solid (a proprietary package of General Motors), and McDonnell Douglas's UNISOLID [Teicholz, 1985].

Object-Oriented Programming

Object-oriented programming is the paradigm for designing and implementing software and is particularly important in computer graphics. An engineering approach, these languages allow software to be constructed from reusable, interchangeable, and extensible parts [Cunningham et al., 1992]. *Class* libraries of graphical objects are being developed. Classes of objects are defined in terms of what an object can do (i.e., what actions and reactions it might produce) and communicate via messages. Subclasses *inherit* actions or characteristics of the superclass. For example, a robot could be constructed from instances of such classes as legs, arms, and head. Each class would have actions defined for it (e.g., a head would be able to nod up and down or shake from side to side). An instance of a head in the object robot would preserve these characteristics. Representative object-oriented languages are Smalltalk, C++, Objective-C, Actor, and Object Pascal.

Plotting and Page Description Languages

Plotting packages, such as ISSCO's DISSPLA and Precision Visuals' DI-3000, consist of routines that are callable from a high-level program. These packages handle 2-D and 3-D images and generally display them in a wire-frame format.

Page description languages are desktop publishing formats that produce graphical output on a printer, display, or other output device. They are used in application programs such as composition systems and illustrators where text, graphical shapes, and sampled images are to be combined into a single document. The dominant language in this category is PostScript, which is a simple interpretive programming language with powerful graphics capabilities. It communicates the description of a document to a printing system in a high level, device-independent manner. PostScript features construction of arbitrary shapes, which may self-intersect, be painted, transformed, cropped, or rendered. The commands are embedded in a general purpose programming language. PostScript programs can be created, transmitted, and interpreted in the form of ASCII source text. The resultant representations will allow for document interchange [Adobe, 1990].

Interaction

The power of computer graphics is to be able to input commands or data in a manner that is appropriate for an application and to have the program react in a timely fashion. These interactions may involve typing words or labels, pointing to items or commands, specifying values or directions for movement, or choosing picture parts displayed on the screen.

Some of the input devices that are available include mouse, special purpose keyboards using buttons or dials, data gloves and other VR devices, touch panels and screens, light pens, graphics tablets, joysticks, 3-D digitizers, trackballs, and voice systems. Each of these devices is capable of sending appropriate values to the graphics program for action [Hearn and Baker, 1997].

Graphics software packages categorize input devices as one of the following logical devices:

1. Locator: a device for specifying a coordinate position (x,y) or orientation (e.g., tablet)
2. Valuator: a device for specifying scalar values (e.g., dials)
3. Keyboard: a device for specifying text input
4. Pick: a device for selecting displayed entities (e.g., mouse)
5. Choice/button: a device for selecting among alternatives (e.g., function keys)

In some systems, an input device might be used for more than one operation. For example, in the Macintosh computer, the mouse is used as a locator, valuator, and pick device [Foley et al., 1996].

91.4 Conclusion

The field of computer graphics has changed dramatically over the past decade. Scientific and engineering applications have expanded from the CAD systems to scientific visualization of complex systems, enhanced solid modeling systems, real-time animated simulations, and now to another dimension, virtual reality.

On the hardware side, we have seen a movement from large, costly systems to putting the power and speed of computers with advanced graphics capabilities on a desktop. PCs, microcomputers, and professional workstations have provided cost-effective platforms that are within the reach of every engineer.

Interaction with a system has been simplified. In most cases, the user has been relieved of the task of keying in and remembering commands. By merely pointing to menu items, the user is led through a system.

Advances in hardware have driven the software development side. Gone are the days of tediously programming and interfacing with low-level graphics commands. Off-the-shelf and vendor-supplied applications packages that incorporate sophisticated graphics abound. These systems are characterized by user-friendly interfaces and high-quality output capabilities.

When programming is necessary, high-level picture constructs through object-oriented environments make manipulation of graphical images more natural. Other support allows for high-level interfaces to X-Windows, Windows, and PostScript by providing the programmer with more graphical development tools.

Overall, scientists and engineers will find the visual dimension for their applications an integral and common component of their tool kit.

Defining Terms

Computer graphics: The generation, representation, manipulation, processing, and display of data by a computer.
Fractals: Geometrical self-similar objects with fractional dimension.
Object-oriented programming: An engineering approach that uses software constructs that are reusable, interchangeable, and extensible.
Rendering: The preparation of the representation of an image to include illumination, shading, depth cueing, coloring, texture, and reflection.
Scientific visualization: The use of computer graphics techniques to represent complex physical phenomena and multidimensional data in order to aid in its understanding and interpretation.
Solid modeling: The use of constructive geometry to build complicated 3-D objects.
Virtual reality (VR): Three or more dimensionality of computer-generated images, which give the user a sense of presence (i.e., a first-person experience) in the scene.
Volume visualization: A method of extracting information from datasets with interactive graphics and imaging; it is concerned with the representation, manipulation, and rendering of volumetric data [Kaufmen et al., 1996].

Related Topics

87.2 High-Level Languages • 89.2 Computer Output Printer Technologies

References

Adobe Systems Incorporated, *PostScript Language Reference Manual*, 2nd ed., Reading, Mass.: Addison-Wesley, 1990.
L. Ammeraal, *Programming Principles in Computer Graphics*, New York: John Wiley, 1986.
M. Bailey, A. Glassner, and P. Wenner, Introduction to Computer Graphics, Course Notes, 23rd International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96, New Orleans, LA, August 1996.
M. Brown, *Understanding PHIGS*, TEMPLATE, San Diego: Megatek Corporation, 1985.
S. Cunningham, N.K. Craighill, M.W. Fong, and J.R. Brown, Eds., *Computer Graphics Using Object-Oriented Programming*, New York: John Wiley, 1992.
M.M. de Ruiter, Ed., *Advances in Computer Graphics III*, New York: Springer-Verlag, 1988.
J.D. Foley, A. van Dam, S.K. Feiner, and J.F. Hughes, *Computer Graphics: Principles and Practice,* 2nd ed. in C, Reading, Mass.: Addison-Wesley, 1996.
S. Harrington, *Computer Graphics: A Programming Approach*, New York: McGraw-Hill, 2nd ed., 1997.
D. Hearn and M.P. Baker, *Computer Graphics,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
A. Kaufman, R. Avila, B. Lorensen, L. Sobierajski, and R. Yagel, Volume Visualization: Principles and Practice, Course Notes, 23rd International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96, New Orleans, LA, August, 1996.
IEEE Spectrum Focus Report: Software, vol. 28, no. 11, November 1991.
B.B. Mandelbrot, *The Fractal Geometry of Nature*, San Francisco: W.H. Freeman, 1982.
W. Purgathofer and J. Schonhut, Eds., *Advances in Computer Graphics V,* New York: Springer-Verlag, 1989.
T. Reynolds and K. Danielson, Programming with OpenGL: An Introduction, Course Notes, 23rd International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96, New Orleans, LA, August, 1996.
W. Straber, Ed., *Advances in Computer Graphics Hardware I*, New York: Springer-Verlag, 1987.
R. Stuart, "Virtual reality: directions in research and development," Interactive Learning Int., vol, 8, pp. 95–100, 1992.
E. Teicholz, Ed., *CAD/CAM Handbook*, New York: McGraw-Hill, 1985.
J.C. Thomas and R. Stuart, "Virtual reality and human factors," Proc. Human Factors Society, 36th Annual Meeting, 1992.
J. Vince, *The Language of Computer Graphics,* New York: Van Nostrand Reinhold, 1990.
S. Wolfram, *Mathematica: A System for Doing Mathematics by Computer*, 2nd ed., Redwood City, Calif.: Addison-Wesley, 1991.

Further Information

Two professional computing organizations publish periodicals that are specifically devoted to the field of computer graphics and provide an excellent forum for current research and techniques. The Association for Computing Machinery (ACM) publishes *ACM Transactions on Graphics* and the IEEE publishes *IEEE Computer Graphics and Applications*.

SIGGRAPH, ACM's special interest group on graphics, sponsors an annual conference and exhibit as well as offering a variety of tutorials and course notes. Other major conferences are sponsored by the National Computer Graphics Association and Eurographics.

Robertazzi, T.G. "Computer Networks"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

92

Computer Networks

92.1 Introduction
92.2 Local Area Networks
Carrier Sense Buses • Token Ring • Token Bus • Wireless LANs • Asyncronous Transfer Mode (ATM) LANs Private Branch Exchange
92.3 Metropolitan Area Networks
FDDI • DQDB
92.4 Wide Area Networks
92.5 The Future

Thomas G. Robertazzi
State University of New York, Stony Brook

92.1 Introduction

Computer networks are geographically distributed collections of communication links and switching processors, the purpose of which is to transport data between computers, workstations, and terminals. In general the elements of a computer network must follow compatible rules of operation together to function effectively. These rules of operation are known as protocols.

There are three broad categories of computer networks, distinguished by geographical extent. Local **area networks** (LANs) connect computer equipment in a single building or floor of a building. Metropolitan area networks (MANs) interconnect network users over a campus or metropolitan-sized region. Finally, wide area networks (WANs) interconnect users on a national or an international scale. In the following, key features of these types of networks will be outlined.

92.2 Local Area Networks

There are six main types of local network architectures that have been commercially produced to date: carrier sense multiple-access buses with collision detection, token rings, token buses, wireless LANs, ATM LANs, and private branch exchanges. The first four have been standardized in the IEEE 802 series standards.

Carrier Sense Buses

The **IEEE 802.3 standard** deals with a network architecture and protocol first constructed at Xerox in the 1970s and termed *Ethernet*. All stations in an Ethernet can be connected, through interfaces, to a **coaxial cable** that is usually run through the ceiling near each user's computer equipment.

The coaxial cable essentially acts as a private radio channel for the users. An interesting protocol called carrier sense multiple-access with collision detection (CSMA/CD) is used in such a network. Each station constantly monitors the cable and can detect when it is idle (no user transmitting), when one user is transmitting (successfully), or when more than one user is simultaneously transmitting (resulting in an unsuccessful collision on the channel). The cable basically acts as a broadcast bus. Any station can transmit on the cable if the station detects it to be idle. Once a station transmits, other stations will not interrupt the transmission. As there is no central control in the network, occasionally two or more stations may attempt to transmit at about the same

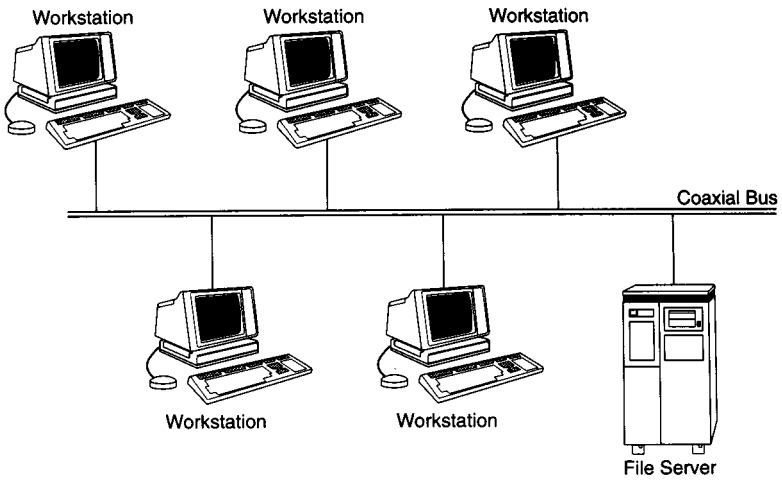


FIGURE 92.1 Bus-type local area network.

time. The transmissions will overlap and be unintelligible (collision). The transmitting stations will detect such a situation and each will retransmit at a randomly chosen later time.

Ethernet and 802.3 networks have raw speeds of up to 10 million bits per second (Mbps). Idle time and collisions, however, can reduce the useful information throughput significantly. The maximum length of these networks is limited by signal propagation delay. An 802.3 coaxial bus differs from an internal computer bus in size and in the lack of a bus controller. Ethernet connections can also be made over unshielded twisted pair or, for long runs, over fiber optics.

Since the introduction of ethernet more than a dozen years ago, desktop computer capabilities and networking requirements have increased significantly. To meet these demands a 100-Mb/s (fast) ethernet has been developed by a consortium of companies. This will be standardized within IEEE 802.3. Media options for the 100-Mb/s ethernet include shielded or unshielded twisted pair as well as fiber optics. Wiring distances of up to 100 m between an end system and wiring closet can be supported. Adapters as well as repeaters that can operate at either 10 or 100 Mb/s will be available. In 1994 there were 50 million ethernet nodes with 15 million new nodes being added each year.

Token Ring

Token ring LANs were developed by IBM in the early 1980s. Topologically, stations are arranged in a circle with point-to-point links between neighbors. Transmissions flow in only one direction (clockwise or counterclockwise). A message transmitted is relayed over the point-to-point links to the receiving station and then forwarded around the rest of the ring and back to the sender to serve as an acknowledgment.

Only a station possessing a special digital code word known as a *token* may transmit. When a station is finished transmitting, it passes the token to its downstream neighbor. Thus, there are no collisions in a token ring, and utilization can approach 100% under heavy loads.

Because of the use of point-to-point links, token rings can use various transmission media such as twisted-pair wire or fiber-optic cables. The transmission speed of a token ring can range from 1 to 16 Mbps, depending on the type of point-to-point links used. Token rings are often wired in star configurations for ease of installation. Token rings are covered by the **IEEE 802.5 standard**. In 1994 there were 10 million token ring nodes with 3–4 million new nodes being added each year.

Token Bus

A token bus uses a coaxial cable along with the token concept to produce a LAN with improved throughput compared to the 802.3 protocol. That is, stations pass a token from one to another to determine which station currently has permission to transmit. Also, in a token bus (and in a token ring), response times can be

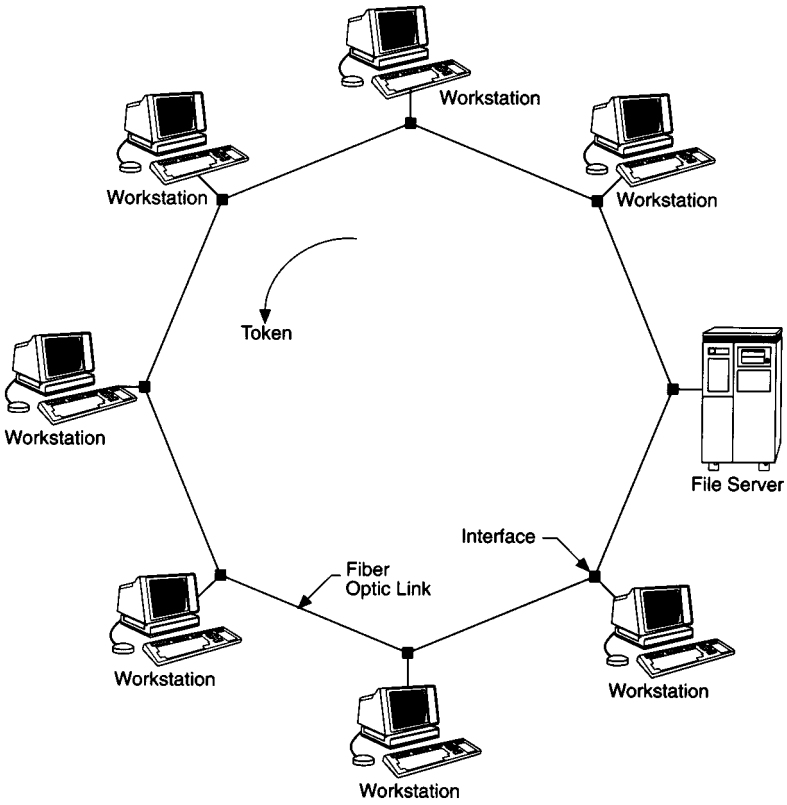


FIGURE 92.2 Token ring local area network.

deterministically bounded. This is important in factory automation, where commands to machines must be received by set times. By way of comparison, response times in an Ethernet-like network can only be probabilistically defined. For this reason, General Motors' Manufacturing Automation Protocol makes use of the token bus. Token buses can operate at 1, 5 and 10 Mbps. Token bus operation is standardized in the **IEEE 802.4 standard**.

Wireless Lans[1]

Wireless LANs use a common radio channel to provide LAN connectivity without any physical wiring. Protocols for wireless LANs are currently being standardized as the IEEE 802.11 standard. Although, one might consider the use of a CSMA/CD protocol in this environment, stations using radio technology are unable to listen to the same channel on which they are transmitting. Thus, it is not possible to implement the collision detection (CD) part of CSMA/CD in a radio environment. Therefore, a modified protocol known as carrier sense multiple access with collision avoidance (CSMA/CA) can be used.

In this variant on CSMA contention for the channel at the end of a successful transmission is mitigated by computing a stochastic idle time for each station during which a station puts off its transmission to see if the channel remains free. A second method of channel access in the form of polling from a master station can be provided for traffic with time delay constraints.

Wireless LANs have aggregate capacities from several hundred kilobit per second to the low megabit per second range. The future of wireless LANs is unclear since it is possible that this capacity will not be sufficient to meet new demands for services.

[1]This material was previously published in *The Mobile Communications Handbook*, J. D. Gibson, Ed., Boca Raton, Fla.: CRC Press, 1996.

One possible way around the problem of the limited spectrum available for wireless LANs is to use infrared light as the transmission medium. There are both direct infrared systems (range: 1–3 mi) and nondirect systems (bounced off walls or ceilings). For small areas data rates are consistent with those of existing ethernet and token ring networks with 100-Mb/s systems on the horizon.

Asyncronous Transfer Mode (ATM) LANs[2]

ATM LANs are relatively new to the LAN marketplace. This is a packet switching technology utilizing relatively short, fixed length packets to provide networking services. ATM was originally seen as a way to develop the next generation wide area telephone network using packet switching, rather than the more conventional circuit switching technology. It was envisioned as a way to transport video, voice, and data in an integrated network. A short packet size was chosen to meet several requirements including minimizing real-time queueing delay.

While progress on the original goal of using ATM technology in wide area telephone networks proceeded slowly because of the complexity of the challenge and the large investments involved, a number of smaller companies introduced ATM local area network products using much the same technology.

An ATM LAN consists of a switch into which are wired end users in a star-type topology. There are several possibilities for the internal architecture of the switch. A low-cost switch may essentially be a computer bus in a box. More sophisticated switches may use switching fabrics. These are very large-scale integrated (VLSI) implementations of patterned networks of simple switching elements, sometimes referred to as space division switching. A great deal of effort has gone into producing cost-effective ATM switches over the past 12 years. It should be pointed out that many of the issues that are not yet resolved for wide area network ATM (i.e., traffic policing, billing) are more tractable in the private ATM LAN environment.

ATM LANs can support a relatively small number of users at high data access rates (low megabit per second). Although ATM is good at handling mixed media traffic at high speeds, it remains to be seen if enough applications are developed requiring its high-bandwidth capability to make it a success. The cost effectiveness of ATM technology is another issue awaiting resolution.

Private Branch Exchange

Historically, private branch exchanges (PBXs) were privately owned telephone switching computers that would be placed in the basement of a building and serve to interconnect phones in the building and provide access to outside lines provided by common carriers. However, PBXs are now available that offer both telephone and data service. In a typical system a phone may have a data socket for terminals or workstations. PBXs are wired in a star topology with the PBX at the center of the star and each user wired directly to it.

92.3 Metropolitan Area Networks

While several network architectures have been proposed for use as MANs, the two that are closest to widespread commercial implementation are fiber-distributed data interface (FDDI) and distributed queue dual bus (DQDB) interface. A key feature of a MAN is the ability to interconnect LANs. This is a problem because of the high data rates at which LANs operate.

FDDI

The FDDI is similar to a token ring LAN except that two rings, instead of one, may be used. Stations needing high-reliability communication are connected to both rings. In the case of a break in the rings the network can be automatically reconfigured. FDDI rings operate at 100 Mbps with a maximum of 500 nodes and a maximum fiber length of 200 km. In fact, most actual FDDI installations have only a small number of nodes (such as routers). There is an American National Standards Institute (ANSI) standard for FDDI.

[2]This material was previously published in *The Mobile Communications Handbook,* J. D. Gibson, Ed., Boca Raton, Fla.: CRC Press, 1996.

DQDB

The DQDB forms the basis of the **IEEE 802.6 standard** for MANs. DQDB is descended from the earlier QPSX, which was developed at the University of Western Australia and Telecom Australia. DQDB uses two unidirectional linear fiber-optic buses. Stations are connected to both buses. Through the clever use of counters the DQDB protocol provides approximate first in, first out (FIFO) service to arriving packets. There are no collisions in DQDB, so utilization can approach 100%. Bus speeds of 150 Mbps are possible.

92.4 Wide Area Networks

Data are generally transmitted over long distances by wide area packet networks. These networks generally lease telephone lines from telecommunications carriers that are used to carry data exclusively. Packet switching technology was first used on a large scale in the ARPANET beginning in the 1960s. The Internet (which replaced the earlier ARPANET) serves to connect universities, industrial and government research centers, and private users.

One problem area unique to wide area packet networks is that of routing. Unlike the previously mentioned networks, there are usually multiple routes available between sources and destinations. Distributed routing algorithms have been developed that route based on current traffic conditions.

92.5 The Future

The future is likely to see an increase in data rates as fiber-optic cables are widely deployed. This will spur the development of faster switching nodes through the use of parallel processing and VLSI implementation. Protocols will have to be simplified to increase processor throughput. New forms of traffic such as video and graphics will become more important. Computer networks will proliferate throughout the world, making possible the ubiquitous transport of data between any two points. These networks are likely to consist of both private networks and new service offerings from telecommunications companies.

Defining Terms

Area networks: LAN, within single building; MAN, metropolitan-sized region; WAN, national/international region.

Coaxial cable: A shielded cable that conducts electrical signals and is used in bus-type local area networks.

Fiber-optic cable: A glass fiber cable that conducts light signals and can be used in token ring local area networks and metropolitan area networks. Fiber optics can provide higher data rates than coaxial cable. They are also immune to electrical interference.

IEEE standards: 802.3, CSMA/CD bus; 802.4, token bus; 802.5, token ring; 802.6, DQDB MAN.

Related Topics

72.3 Local-Area Networks • 75.3 Stochastic Processes

References

U. Black, *Data Networks: Concepts, Theory and Practice,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

M. De Prycker, *Asynchronous Transfer Mode: Solution for Broadband ISDN,* New York: Simon and Schuster, 1991.

A. De Simone and S. Nanda, "Wireless data: Systems, standards, services," *Wireless Networks,* 1(3):241–253, 1995.

N. J. Muller, *Wireless Data Networking,* Boston, Mass: Artech House, 1995.

L. Peterson and B. Davie, *Computer Networks: A System Approach,* San Francisco, Calif.: Morgan Kaufman, 1995.

T. G. Robertazzi, *Performance Evaluation of High Speed Switching Fabrics and Networks: ATM, Broadband ISDN and MAN Technolgoy,* Piscataway N.J.: IEEE Press, 1993.

P. Scherer, The 100 Mbps Ethernet Standard. In Distinguished Lecture Series (IX) (videotape), Stanford, Calif.: Univ. Video Communications, 1994.

J. D. Spragins, J. L. Hammond, and K. Pawlikowski, *Telecommunications Protocols and Design,* Reading, Mass.: Addison–Wesley, 1991.
J. N. D. Walrand, *Communication Networks: A First Course,* Boston, Mass.: Aksen Associates, Inc., and Richard Irwin, Inc. 1991.

Further Information

The following are the IEEE 802 series of standards related to local area networks.

IEEE 802.3: CSMA/CD bus protocol standard.
IEEE 802.4: Token bus standard.
IEEE 802.5: Token ring standard.
IEEE 802.6: DQDB metropolitan area network standard.
IEEE 802.11: CSMA/CA wireless LAN standard.

Tutorial articles on LANs appear in *IEEE Communications Magazine and IEEE Network* magazine. Technical articles on LANs appear in *IEEE Transactions on Networking, IEEE Transactions on Communications, IEEE Journal on Selected Areas in Communications,* and the journal *Wireless Networks.*

Johnson, B.W. "Fault Tolerance"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

93

Fault Tolerance

93.1 Introduction
93.2 Hardware Redundancy
93.3 Information Redundancy
93.4 Time Redundancy
93.5 Software Redundancy
93.6 Dependability Evaluation

Barry W. Johnson
University of Virginia

93.1 Introduction

Fault tolerance is the ability of a system to continue correct performance of its tasks after the occurrence of hardware or software faults. A **fault** is simply any physical defect, imperfection, or flaw that occurs in hardware or software. Applications of fault-tolerant computing can be categorized broadly into four primary areas: long-life, critical computations, maintenance postponement, and high availability. The most common examples of long-life applications are unmanned space flight and satellites. Examples of critical-computation applications include aircraft flight control systems, military systems, and certain types of industrial controllers. Maintenance postponement applications appear most frequently when maintenance operations are extremely costly, inconvenient, or difficult to perform. Remote processing stations and certain space applications are good examples. Banking and other time-shared systems are good examples of high-availability applications. Fault tolerance can be achieved in systems by incorporating various forms of redundancy, including hardware, information, time, and software redundancy [Johnson, 1989].

93.2 Hardware Redundancy

The physical replication of hardware is perhaps the most common form of fault tolerance used in systems. As semiconductor components have become smaller and less expensive, the concept of hardware redundancy has become more common and more practical. There are three basic forms of hardware redundancy. First, *passive* techniques use the concept of fault masking to hide the occurrence of faults and prevent the faults from resulting in **errors**. Passive approaches are designed to achieve fault tolerance without requiring any action on the part of the system or an operator. Passive techniques, in their most basic form, do not provide for the detection of faults but simply mask the faults. An example of a passive approach is triple modular redundancy (TMR), which is illustrated in Fig. 93.1. In the TMR system three identical units perform identical functions, and a majority vote is performed on the output.

The second form of hardware redundancy is the *active* approach, which is sometimes called the *dynamic* method. Active methods achieve fault tolerance by detecting the existence of faults and performing some action to remove the faulty hardware from the system. In other words, active techniques require that the system perform reconfiguration to tolerate faults. Active hardware redundancy uses fault detection, fault location, and fault recovery in an attempt to achieve fault tolerance. An example of an active approach to hardware redundancy is standby sparing, which is illustrated in Fig. 93.2. In standby sparing one or more units operate as spares and replace the primary unit when it fails.

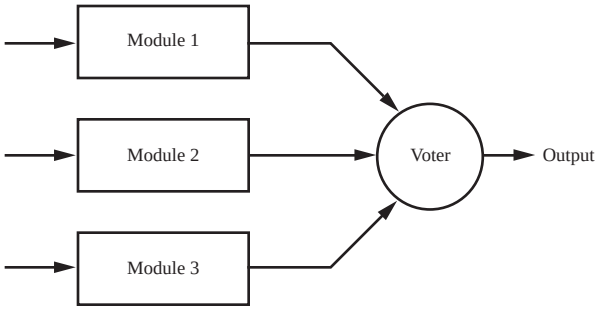


FIGURE 93.1 Fault masking using triple modular redundancy (TMR).

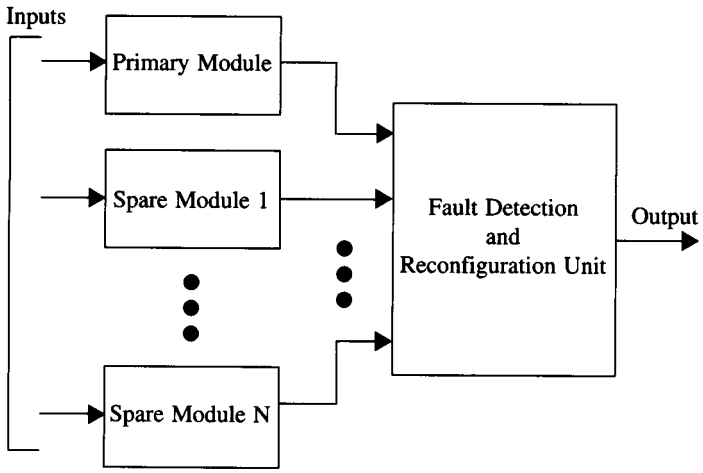


FIGURE 93.2 General concept of standby sparing.

The final form of hardware redundancy is the *hybrid* approach. Hybrid techniques combine the attractive features of both the passive and active approaches. Fault masking is used in hybrid systems to prevent erroneous results from being generated. Fault detection, fault location, and fault recovery are also used in the hybrid approaches to improve fault tolerance by removing faulty hardware and replacing it with spares. Providing spares is one form of providing redundancy in a system. Hybrid methods are most often used in the critical-computation applications where fault masking is required to prevent momentary errors, and high reliability must be achieved. The basic concept of the hybrid approach is illustrated in Fig. 93.3.

93.3 Information Redundancy

Another approach to fault tolerance is to employ redundancy of information. Information redundancy is simply the addition of redundant information to data to allow fault detection, fault masking, or possibly fault tolerance. Good examples of information redundancy are error detecting and error correcting codes, formed by the addition of redundant information to data words or by the mapping of data words into new representations containing redundant information [Lin and Costello, 1983].

In general, a *code* is a means of representing information, or data, using a well-defined set of rules. A *code word* is a collection of symbols, often called digits if the symbols are numbers, used to represent a particular piece of data based upon a specified code. A *binary code* is one in which the symbols forming each code word consist of only the digits 0 and 1. A code word is said to be *valid* if the code word adheres to all of the rules that define the code; otherwise, the code word is said to be *invalid.*

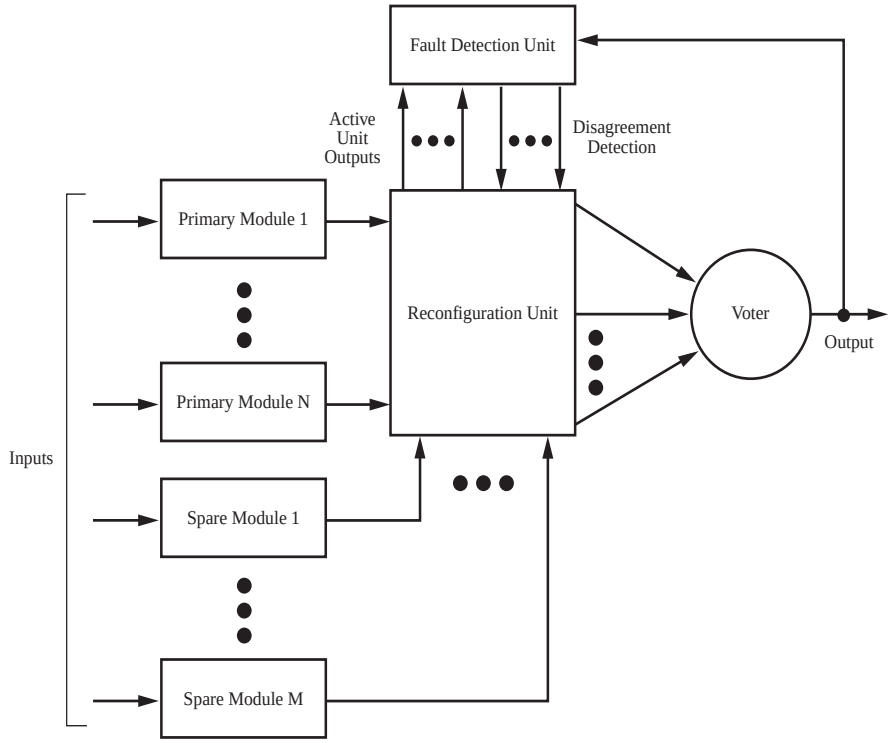


FIGURE 93.3 Hybrid redundancy approach.

The *encoding operation* is the process of determining the corresponding code word for a particular data item. In other words, the encoding process takes an original data item and represents it as a code word using the rules of the code. The *decoding operation* is the process of recovering the original data from the code word. In other words, the decoding process takes a code word and determines the data that it represents.

It is possible to create a binary code for which the valid code words are a subset of the total number of possible combinations of 1s and 0s. If the code words are formed correctly, errors introduced into a code word will force it to lie in the range of illegal, or invalid, code words, and the error can be detected. This is the basic concept of the *error detecting codes.* The basic concept of the *error correcting code* is that the code word is structured such that it is possible to determine the correct code word from the corrupted, or erroneous, code word.

A fundamental concept in the characterization of codes is the *Hamming distance* [Hamming, 1950]. The *Hamming distance* between any two binary words is the number of bit positions in which the two words differ. For example, the binary words 0000 and 0001 differ in only one position and therefore have a Hamming distance of 1. The binary words 0000 and 0101, however, differ in two positions; consequently, their Hamming distance is 2. Clearly, if two words have a Hamming distance of 1, it is possible to change one word into the other simply by modifying one bit in one of the words. If, however, two words differ in two bit positions, it is impossible to transform one word into the other by changing one bit in one of the words.

The Hamming distance gives insight into the requirements of error detecting codes and error correcting codes. We define the *distance* of a code as the minimum Hamming distance between any two valid code words. If a binary code has a distance of two, then any single-bit error introduced into a code word will result in the erroneous word being an invalid code word because all valid code words differ in at least two bit positions. If a code has a distance of 3, then any single-bit error or any double-bit error will result in the erroneous word being an invalid code word because all valid code words differ in at least three positions. However, a code distance of 3 allows any single-bit error to be corrected, if it is desired to do so, because the erroneous word with a single-bit error will be a Hamming distance of 1 from the correct code word and at least a Hamming

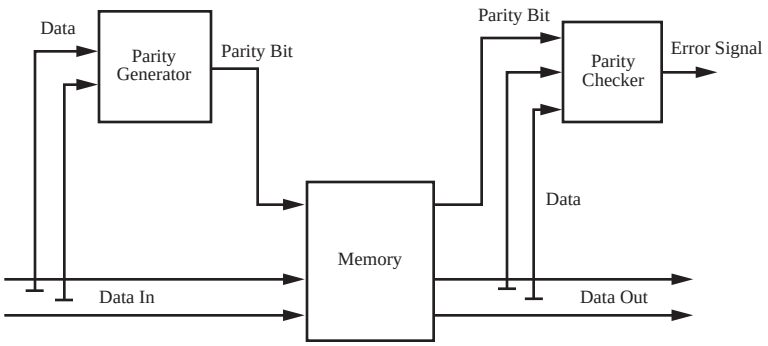


FIGURE 93.4 Use of parity coding in a memory application.

distance of 2 from all others. Consequently, the correct code word can be identified from the corrupted code word.

In general, a binary code can correct up to c bit errors and detect an additional d bit errors if and only if

$$2c - d - 1 \leq H_d$$

where H_d is the distance of the code [Nelson and Carroll, 1986]. For example, a code with a distance of 2 cannot provide any error correction but can detect single-bit errors. Similarly, a code with a distance of 3 can correct single-bit errors or detect a double-bit error.

A second fundamental concept of codes is *separability.* A *separable code* is one in which the original information is appended with new information to form the code word, thus allowing the decoding process to consist of simply removing the additional information and keeping the original data. In other words, the original data is obtained from the code word by stripping away extra bits, called the code bits or check bits, and retaining only those associated with the original information. A *nonseparable code* does not possess the property of separability and, consequently, requires more complicated decoding procedures.

Perhaps the simplest form of a code is the parity code. The basic concept of parity is very straightforward, but there are variations on the fundamental idea. Single-bit parity codes require the addition of an extra bit to a binary word such that the resulting code word has either an even number of 1s or an odd number of 1s. If the extra bit results in the total number of 1s in the code word being odd, the code is referred to as *odd parity.* If the resulting number of 1s in the code word is even, the code is called *even parity.* If a code word with odd parity experiences a change in one of its bits, the parity will become even. Likewise, if a code word with even parity encounters a single-bit change, the parity will become odd. Consequently, a single-bit error can be detected by checking the number of ls in the code words. The single-bit parity code (either odd or even) has a distance of 2, therefore allowing any single-bit error to be detected but not corrected. Figure 93.4 illustrates the use of parity coding in a simple memory application.

Arithmetic codes are very useful when it is desired to check arithmetic operations such as addition, multiplication, and division [Avizienis, 1971]. The basic concept is the same as all coding techniques. The data presented to the arithmetic operation is encoded before the operations are performed. After completing the arithmetic operations, the resulting code words are checked to make sure that they are valid code words. If the resulting code words are not valid, an error condition is signaled. An arithmetic code must be invariant to a set of arithmetic operations. An arithmetic code, A, has the property that $A(b^*c) = A(b)^*A(c)$, where b and c are operands, * is some arithmetic operation, and $A(b)$ and $A(c)$ are the arithmetic code words for the operands b and c, respectively. Stated verbally, the performance of the arithmetic operation on two arithmetic code words will produce the arithmetic code word of the result of the arithmetic operation. To completely define an arithmetic code, the method of encoding and the arithmetic operations for which the code is invariant must be specified. The most common examples of arithmetic codes are the *AN* codes, residue codes, and the inverse residue codes.

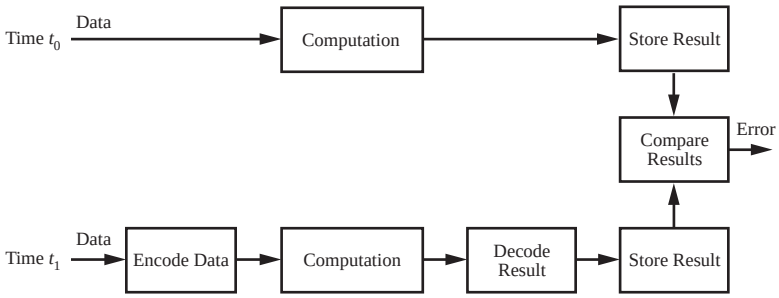


FIGURE 93.5 Time redundancy concept.

93.4 Time Redundancy

Time redundancy methods attempt to reduce the amount of extra hardware at the expense of using additional time. In many applications, the time is of much less importance than the hardware because hardware is a physical entity that impacts weight, size, power consumption, and cost. Time, on the other hand, may be readily available in some applications. The basic concept of time redundancy is the repetition of computations in ways that allow faults to be detected. Time redundancy can function in a system in several ways. The fundamental concept is to perform the same computation two or more times and compare the results to determine if a discrepancy exists. If an error is detected, the computations can be performed again to see if the disagreement remains or disappears. Such approaches are often good for detecting errors resulting from transient faults but cannot provide protection against errors resulting from permanent faults.

The main problem with many time redundancy techniques is assuring that the system has the same data to manipulate each time it redundantly performs a computation. If a transient fault has occurred, a system's data may be completely corrupted, making it difficult to repeat a given computation. Time redundancy has been used primarily to detect transients in systems. One of the biggest potentials of time redundancy, however, now appears to be the ability to detect permanent faults while using a minimum of extra hardware. The fundamental concept is illustrated in Fig. 93.5. During the first computation or transmission, the operands are used as presented and the results are stored in a register. Prior to the second computation or transmission, the operands are encoded in some fashion using an encoding function. After the operations have been performed on the encoded data, the results are then decoded and compared to those obtained during the first operation. The selection of the encoding function is made so as to allow faults in the hardware to be detected. Example encoding functions might include the complementation operator and an arithmetic shift.

93.5 Software Redundancy

Software faults are unusual entities. Software does not break as hardware does, but instead software faults are the result of incorrect software designs or coding mistakes. Therefore, any technique that detects faults in software must detect design flaws. A simple duplication and comparison procedure will not detect software faults if the duplicated software modules are identical, because the design mistakes will appear in both modules.

The concept of N self-checking programming is to first write N unique versions of the program and to develop a set of acceptance tests for each version. The acceptance tests are essentially checks performed on the results produced by the program and may be created using consistency checks and capability checks, for example. Selection logic, which may be a program itself, chooses the results from one of the programs that passes the acceptance tests. This approach is analogous to the hardware technique known as hot standby sparing. Since each program is running simultaneously, the reconfiguration process can be very fast. Provided that the software faults in each version of the program are independent and the faults are detected as they occur by the acceptance tests, then this approach can tolerate $N - 1$ faults. It is important to note that the assumptions of fault independence and perfect fault coverage are very big assumptions to make in almost all applications.

The concept of *N*-version programming was developed to allow certain design flaws in software modules to be tolerated [Chen and Avizienis, 1978]. The basic concept of *N*-version programming is to design and code the software module *N* times and to vote on the *N* results produced by these modules. Each of the *N* modules is designed and coded by a separate group of programmers. Each group designs the software from the same set of specifications such that each of the *N* modules performs the same function. However, it is hoped that by performing the *N* designs independently, the same mistakes will not be made by the different groups. Therefore, when a fault occurs, the fault will either not occur in all modules or it will occur differently in each module, so that the results generated by the modules will differ. Assuming that the faults are independent the approach can tolerate $(N-1)/2$ faults where N is odd.

The recovery block approach to software fault tolerance is analogous to the active approaches to hardware fault tolerance, specifically the cold standby sparing approach. *N* versions of a program are provided, and a single set of acceptance tests is used. One version of the program is designated as the primary version, and the remaining $N-1$ versions are designated as spares, or secondary versions. The primary version of the software is always used unless it fails to pass the acceptance tests. If the acceptance tests are failed by the primary version, then the first secondary version is tried. This process continues until one version passes the acceptance tests or the system fails because none of the versions can pass the tests.

93.6 Dependability Evaluation

Dependability is defined as the quality of service provided by a system [Laprie, 1985]. Perhaps the most important measures of dependability are reliability and availability. Fundamental to reliability calculations is the concept of failure rate. Intuitively, the *failure rate is* the expected number of **failures** of a type of device or system per a given time period [Shooman, 1968]. The failure rate is typically denoted as λ when it is assumed to have a constant value. To more clearly understand the mathematical basis for the concept of a failure rate, first consider the definition of the reliability function. The **reliability** $R(t)$ of a component, or a system, is the conditional probability that the component operates correctly throughout the interval $[t_0, t]$ given that it was operating correctly at the time t_0.

There are a number of different ways in which the failure rate function can be expressed. For example, the failure rate function $z(t)$ can be written strictly in terms of the reliability function $R(t)$ as

$$z(t) = \left(-\frac{dR(t)/dt}{R(t)} \right)$$

Similarly, $z(t)$ can be written in terms of the unreliability $Q(t)$ as

$$z(t) = -\frac{dR(t)/dt}{R(t)} = \frac{dQ(t)/dt}{1-Q(t)}$$

where $Q(t) = 1 - R(t)$. The derivative of the unreliability, $dQ(t)/dt$, is called the *failure density function.*

The failure rate function is clearly dependent upon time; however, experience has shown that the failure rate function for electronic components does have a period where the value of $z(t)$ is approximately constant. The commonly accepted relationship between the failure rate function and time for electronic components is called the bathtub curve and is illustrated in Fig. 93.6. The bathtub curve assumes that during the early life of systems, failures occur frequently due to substandard or weak components. The decreasing part of the bathtub curve is called the early-life or infant mortality region. At the opposite end of the curve is the wear-out region where systems have been functional for a long period of time and are beginning to experience failures due to the physical wearing of electronic or mechanical components. During the intermediate region, the failure rate function is assumed to be a constant. The constant portion of the bathtub curve is called the useful-life phase

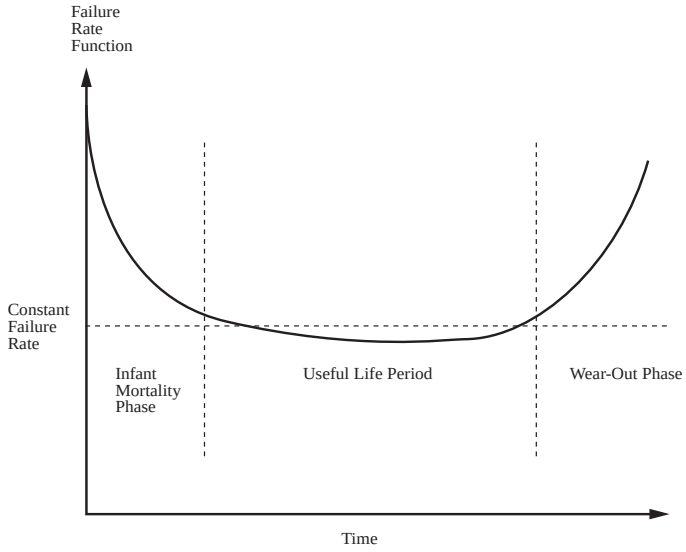


FIGURE 93.6 Bathtub curve relationship between the failure rate function and time.

of the system, and the failure rate function is assumed to have a value of λ during that period. λ is referred to as the failure rate and is normally expressed in units of failures per hour.

The reliability can be expressed in terms of the failure rate function as a differential equation of the form

$$\frac{dR(t)}{dt} = -z(t)R(t)$$

The general solution of this differential equation is given by

$$R(t) = e^{-\int z(t)dt}$$

If we assume that the system is in the useful-life stage where the failure rate function has a constant value of λ, the solution to the differential equation is an exponential function of the parameter λ given by

$$R(t) = e^{-\lambda t}$$

where λ is the constant failure rate. The exponential relationship between the reliability and time is known as the *exponential failure law,* which states that for a constant failure rate function, the reliability varies exponentially as a function of time.

In addition to the failure rate, the mean time to failure (MTTF) is a useful parameter to specify the quality of a system. The MTTF is the expected time that a system will operate before the first failure occurs. The MTTF can be calculated by finding the expected value of the time of failure.

From probability theory, we know that the expected value of a random variable, X, is

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

where $f(x)$ is the probability density function. In reliability analysis we are interested in the expected value of the time of failure (MTTF), so

$$\text{MTTF} = \int_0^\infty t f(t) dt$$

where $f(t)$ is the failure density function, and the integral runs from 0 to ∞ because the failure density function is undefined for times less than 0. We know, however, that the failure density function is

$$f(t) = \frac{dQ(t)}{dt}$$

so, the MTTF can be written as

$$\text{MTTF} = \int_0^\infty t \frac{dQ(t)}{dt} dt$$

Using integration by parts and the fact that $dQ(t)/dt = -dR(t)/dt$ we can show that

$$\text{MTTF} = \int_0^\infty t \frac{dQ(t)}{dt} dt = -\int_0^\infty t \frac{dR(t)}{dt} dt = \left[-tR(t) + \int R(t) dt\right]\Big|_0^\infty = \int_0^\infty R(t) dt$$

Consequently, the MTTF is defined in terms of the reliability function as

$$\text{MTTF} = \int_0^\infty R(t) dt$$

which is valid for any reliability function that satisfies $R(\infty) = 0$.

The mean time to repair (MTTR) is simply the average time required to repair a system. The MTTR is extremely difficult to estimate and is often determined experimentally by injecting a set of faults, one at a time, into a system and measuring the time required to repair the system in each case. The MTTR is normally specified in terms of a repair rate, μ, which is the average number of repairs that occur per time period. The units of the repair rate are normally number of repairs per hour. The MTTR and the rate, μ, are related by

$$\text{MTTR} = \frac{1}{\mu}$$

It is very important to understand the difference between the MTTF and the mean time between failure (MTBF). Unfortunately, these two terms are often used interchangeably. While the numerical difference is small in many cases, the conceptual difference is very important. The MTTF is the average time until the first failure of a system, while the MTBF is the average time between failures of a system. If we assume that all repairs to a system make the system perfect once again just as it was when it was new, the relationship between the MTTF and the MTBF can be determined easily. Once successfully placed into operation, a system will operate, on the average, a time corresponding to the MTTF before encountering the first failure. The system will then require

some time, MTTR, to repair the system and place it back into operation once again. The system will then be perfect once again and will operate for a time corresponding to the MTTF before encountering its next failure. The time between the two failures is the sum of the MTTF and the MTTR and is the MTBF. Thus, the difference between the MTTF and the MTBF is the MTTR. Specifically, the MTBF is given by

$$\text{MTBF} = \text{MTTF} + \text{MTTR}$$

In most practical applications the MTTR is a small fraction of the MTTF, so the approximation that the MTBF and MTTF are equal is often quite good. Conceptually, however, it is crucial to understand the difference between the MTBF and the MTTF.

An extremely important parameter in the design and analysis of fault-tolerant systems is fault coverage. The fault coverage available in a system can have a tremendous impact on the reliability, safety, and other attributes of the system. Fault coverage is mathematically defined as the conditional probability that, given the existence of a fault, the system recovers [Bouricius et al., 1969]. The fundamental problem with fault coverage is that it is extremely difficult to calculate. Probably the most common approach to estimating fault coverage is to develop a list all of the faults that can occur in a system and to form, from that list, a list of faults from which the system can recover. The fault coverage factor is then calculated appropriately.

Reliability is perhaps one of the most important attributes of systems. The reliability of a system is generally derived in terms of the reliabilities of the individual components of the system. The two models of systems that are most common in practice are the series and the parallel. In a series system, each element of the system is required to operate correctly for the system to operate correctly. In a parallel system, on the other hand, only one of several elements must be operational for the system to perform its functions correctly.

The series system is best thought of as a system that contains no redundancy; that is, each element of the system is needed to make the system function correctly. In general, a system may contain N elements, and in a series system each of the N elements is required for the system to function correctly. The reliability of the series system can be calculated as the probability that none of the elements will fail. Another way to look at this is that the reliability of the series system is the probability that all of the elements are working properly. The reliability of a series system is given by

$$R_{\text{series}}(t) = R_1(t)R_2(t) \ldots R_N(t)$$

or

$$R_{\text{series}}(t) = \prod_{i=1}^{N} R_i(t)$$

An interesting relationship exists in a series system if each individual component satisfies the exponential failure law. Suppose that we have a series system made up of N components, and each component, i, has a constant failure rate of λ_i. Also assume that each component satisfies the exponential failure law. The reliability of the series system is given by

$$R_{\text{series}}(t) = e^{-\lambda_1 t} e^{-\lambda_2 t} \ldots e^{-\lambda_N t}$$

$$R_{\text{series}}(t) = e^{-\sum_{i=1}^{N} \lambda_i t}$$

The distinguishing feature of the basic parallel system is that only one of N identical elements is required for the system to function. The reliability of the parallel system can be written as

$$R_{\text{parallel}}(t) = 1.0 - Q_{\text{parallel}}(t) = 1.0 - \prod_{i=1}^{N} Q_i(t) = 1.0 - \prod_{i=1}^{N} (1.0 - R_i(t))$$

It should be noted that the equations for the parallel system assume that the failures of the individual elements that make up the parallel system are independent.

M-of-*N* systems are a generalization of the ideal parallel system. In the ideal parallel system, only one of *N* modules is required to work for the system to work. In the *M*-of-*N* system, however, *M* of the total of *N* identical modules are required to function for the system to function. A good example is the TMR configuration where two of the three modules must work for the majority voting mechanism to function properly. Therefore, the TMR system is a 2-of-3 system.

In general, if there are *N* identical modules and *M* of those are required for the system to function properly, then the system can tolerate $N - M$ module failures. The expression for the reliability of an *M*-of-*N* system can be written as

$$R_{M\text{-of-}N}(t) = \sum_{i=0}^{N-M} \binom{N}{i} R^{N-i}(t)(1.0 - R(t))^i$$

where

$$\binom{N}{i} = \frac{N!}{(N-i)!\,i!}$$

The **availability**, $A(t)$, of a system is defined as the probability that a system will be available to perform its tasks at the instant of time *t*. Intuitively, we can see that the availability can be approximated as the total time that a system has been operational divided by the total time elapsed since the system was initially placed into operation. In other words, the availability is the percentage of time that the system is available to perform its expected tasks. Suppose that we place a system into operation at time $t = 0$. As time moves along, the system will perform its functions, perhaps fail, and hopefully be repaired. At some time $t = t_{\text{current}}$, suppose that the system has operated correctly for a total of t_{op} hours and has been in the process of repair or waiting for repair to begin for a total of t_{repair} hours. The time t_{current} is then the sum of t_{op} and t_{repair}. The availability can be determined as

$$A(t_{\text{current}}) = \frac{t_{\text{op}}}{t_{\text{op}} + t_{\text{repair}}}$$

where $A(t_{\text{current}})$ is the availability at time t_{current}.

If the average system experiences *N* failures during its lifetime, the total time that the system will be operational is *N*(MTTF) hours. Likewise, the total time that the system is down for repairs is *N*(MTTR) hours. In other words, the operational time, t_{op}, is *N*(MTTF) hours and the downtime, t_{repair}, is *N*(MTTR) hours. The average, or steady-state, availability is

$$A_{SS} = \frac{N(\text{MTTF})}{N(\text{MTTF}) + N(\text{MTTR})}$$

We know, however, that the MTTF and the MTTR are related to the failure rate and the repair rate, respectively, for simplex systems, as

$$\text{MTTF} = \frac{1}{\lambda}$$

$$\text{MTTR} = \frac{1}{\mu}$$

Therefore, the steady-state availability is given by

$$A_{SS} = \frac{1/\lambda}{1/\lambda + 1/\mu} = \frac{1}{1 + \lambda/\mu}$$

Defining Terms

Availability, $A(t)$: The probability that a system is operating correctly and is available to perform its functions at the instant of time t.

Dependability: The quality of service provided by a particular system.

Error: The occurrence of an incorrect value in some unit of information within a system.

Failure: A deviation in the expected performance of a system.

Fault: A physical defect, imperfection, or flaw that occurs in hardware or software.

Fault avoidance: A technique that attempts to prevent the occurrence of faults.

Fault tolerance: The ability to continue the correct performance of functions in the presence of faults.

Maintainability, $M(t)$: The probability that an inoperable system will be restored to an operational state within the time t.

Performability, $P(L,t)$: The probability that a system is performing at or above some level of performance, L, at the instant of time t.

Reliability, $R(t)$: The conditional probability that a system has functioned correctly throughout an interval of time, $[t_0,t]$, given that the system was performing correctly at time t_0.

Safety, $S(t)$: The probability that a system will either perform its functions correctly or will discontinue its functions in a well-defined, safe manner.

Related Topics

98.1 Introduction • 98.4 Relationship between Reliability and Failure Rate

References

A. Avizienis, "Arithmetic error codes: Cost and effectiveness studies for application in digital system design," *IEEE Transactions on Computers,* vol. C-20, no. 11, pp. 1322–1331, November 1971.

W. G. Bouricius, W. C. Carter, and P. R. Schneider, "Reliability modeling techniques for self-repairing computer systems," in *Proceedings of the 24th ACM Annual Conference,* pp. 295–309, 1969.

L. Chen and A. Avizienis, "N-version programming: A fault tolerant approach to reliability of software operation," in *Proceedings of the International Symposium on Fault Tolerant Computing,* pp. 3–9, 1978.

R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal,* vol. 26, no. 2, pp. 147–160, April 1950.

B. W. Johnson, *Design and Analysis of Fault-Tolerant Digital Systems,* Reading, Mass.: Addison-Wesley, 1989.

J-C. Laprie, "Dependable computing and fault tolerance: Concepts and terminology," in *Proceedings of the 15th Annual International Symposium on Fault-Tolerant Computing,* Ann Arbor, Mich.: pp. 2–11, June 19–21, 1985.

S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications,* Englewood Cliffs, N.J.: Prentice-Hall, 1983.

V. P. Nelson and B. D. Carroll, *Tutorial: Fault-Tolerant Computing,* Washington, D.C.: IEEE Computer Society Press, 1986.

M. L. Shooman, *Probabilistic Reliability: An Engineering Approach,* New York: McGraw-Hill, 1968.

Further Information

The *IEEE Transactions on Computers, IEEE Computer* magazine, and the *Proceedings of the IEEE* have published numerous special issues dealing exclusively with fault tolerance technology. Also, the IEEE International Symposium on Fault-Tolerant Computing has been held each year since 1971. Finally, the following textbooks are available, in addition to those referenced above:

P. K. Lala, *Fault Tolerant and Fault Testable Hardware,* Englewood Cliffs, N.J.: Prentice-Hall, 1985.
D. K. Pradhan, *Fault-Tolerant Computing: Theory and Techniques,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
D. P. Siewiorek and R. S. Swarz, *The Theory and Practice of Reliable Systems Design,* 2nd ed., Bedford, Mass.: Digital Press, 1992.

Abdelguerfi, M., Eskicioglu, R., Liebowitz, J. "Knowledge Engineering"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

94

Knowledge Engineering

94.1 Databases
Database Abstraction • Data Models • Relational Databases • Hierarchical Databases • Network Databases • Architecture of a DBMS • Data Integrity and Security • Emerging Trends

94.2 Rule-Based Expert Systems
Problem Selection • Knowledge Acquisition • Knowledge Representation • Knowledge Encoding • Knowledge Testing and Evaluation • Implementation and Maintenance

M. Abdelguerfi and R. Eskicioglu
University of New Orleans

Jay Liebowitz
George Washington University

94.1 Databases

M. Abdelguerfi and R. Eskicioglu

In the past, file processing techniques were used to design information systems. These systems usually consist of a set of files and a collection of application programs. Permanent records are stored in the files, and application programs are used to update and query the files. The application programs were in general developed individually to meet the needs of different groups of users. In many cases, this approach leads to a duplication of data among the files of different users. Also, the lack of coordination between files belonging to different users often leads to a lack of data consistency. In addition, changes to the underlying data requirements usually necessitate major changes to existing application programs. Among other major problems that arise with the use of file processing techniques are lack of data sharing, reduced programming productivity, and increased program maintenance. Because of their inherent difficulties and lack of flexibility, file processing techniques have lost a great deal of their popularity and are being replaced by **database management systems (DBMS).**

A DBMS is designed to efficiently manage a shared pool of interrelated data (**database**). This includes the existence of features such as a *data definition language* for the definition of the logical structure of the database (*database schema*), a *data manipulation language* to query and update the database, a *concurrency control* mechanism to keep the database consistent when shared by several users, a *crash recovery* strategy to avoid any loss of information after a system crash, and *safety* mechanisms against any unauthorized access.

Database Abstraction

A DBMS is expected to provide for *data independence*, i.e., user requests are made at a *logical level* without any need for the knowledge of how the data is stored in actual files. This implies that the internal file structure could be modified without any change to the user's perception of the database. To achieve data independence, the Standards Planning and Requirements Committee (SPARC) of the American National Standards Institute (ANSI) in its 1977 report recommended three levels of database abstraction (see Fig. 94.1). The lowest level in the abstraction is the internal level. Here, the database is viewed as a collection of files organized according to one of several possible internal data organizations (e.g., B^+-tree data organization). In the conceptual level, the database is viewed at an abstract level. The user at this level is shielded from the internal storage details. At the external level, each group of users has their own perception or *view* of the database. Each view is derived from

the conceptual database and is designed to meet the needs of a particular group of users. Such a group can only have access to the data specified by its particular view. This, of course, ensures both privacy and security.

The mapping between the three levels of abstraction is the task of the DBMS. When changes to the internal level (such as a change in file organization) do not affect the conceptual and external levels, the system is said to provide for *physical data independence*. *Logical data independence* prevents changes to the conceptual level to affect users' views. Both types of data independence are desired features in a database system.

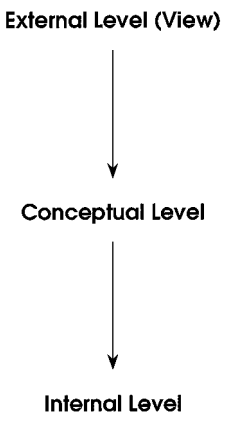


FIGURE 94.1 Data abstraction.

Data Models

A **data model** refers to an integrated set of tools used to describe the data and its structure, data relationships, and data constraints. Some data models provide a set of operators that is used to update and query the database. Data models can be classified in two main categories: *record-based* and *object-based*. Both classes are used to describe the database at the conceptual and external levels. With object-based data models, constraints on the data can be specified more explicitly.

There are three main record-based data models: the *relational*, *network*, and *hierarchical* models. In the relational model, data at the conceptual level is represented as a collection of interrelated tables. These tables are *normalized* so as to minimize data redundancy and update anomalies. In this model, data relationships are implicit and are derived by matching columns in tables. In the hierarchical and network models, the data is represented as a collection of records and data relationships are explicit and are represented by *links*. The difference between the last two models is that in the hierarchical model, data is represented as a tree structure, while it is represented as a generalized graph in the network model.

In hierarchical and network models, the existence of physical pointers (links) to link related records allows an application program to retrieve a single record at a time by following the pointer's chain. The process of following the pointer's chain and selecting one record at a time is referred to as *navigation*. In nonnavigational models such as the relational model, records are not related through pointer's chains, but relationships are established by matching columns in different tables.

The hierarchical and network models require the application programmer to be aware of the internal structure of the database. The relational model, on the other hand, allows for a high degree of physical and logical data independence. Earlier DBMSs were for the most part navigational systems. Because of its simplicity and strong theoretical foundations, the relational model has since received wide acceptance. Today, most DBMSs are based on the relational model.

Other data models include a popular high level conceptual data model, known as the *Entity-Relationship* (ER) model. The ER model is mainly used for the conceptual design of databases and their applications. The ER model describes data as entities, attributes, and relationships.

An *entity* is an "object" in the real world with an independent existence. Each entity has a set of properties, called *attributes,* that describes it. A *relationship* is an association between entities. For example, a professor entity may be described by its name, age, and salary and can be associated with a department entity by the relationship "works for".

With the advent of advanced database applications, the ER modeling concepts became insufficient. This has led to the enhancement of the ER model with additional concepts, such as generalization, categories, and inheritance, leading to the *Enhanced-ER* or *EER* model.

Relational Databases

The relational model was introduced by E. F. Codd [1970]. Since the theoretical underpinnings of the relational model have been well defined, it has become the focus of most commercial DBMSs.

In the relational model, the data is represented as a collection of relations. To a large extent, each relation can be thought of as a table. The example of Fig. 94.2 shows part of a university database composed of two

FAC_INFO (*lname, social_sec#, street, city, dept*)

Hosch	383909164	Esplanade	Kenner	CS
Loggins	482233364	Bonnabel	Metairie	EE
Martin	399254402	Williams	Kenner	CH
Krad	100995678	Bourbon	New Orleans	EE
Hanoura	400919945	Bonnabel	Metairie	CH
Prados	388998800	Severn	Matairie	EE
Abdel	389390164	St Charles	New Orleans	CS

DEPT_CHAIR (*dept, chair*)

CS	Hosch
EE	Prados
CH	Martin

FIGURE 94.2 An example of two relations: FAC_INFO and DEP_CHAIR.

relations. FAC_INFO gives personal information (last name, social security, street and city of residence, and department) of a faculty. DEP_CHAIR gives the last name of the chairman of each department. A faculty is not allowed to belong to two departments. Each row in a relation is referred to as a *tuple*. A column name is called an *attribute name*. The data type of each attribute name is known as its *domain*. A *relation scheme* is a set of attribute names. For instance, the relation scheme (or scheme for short) of the relation FAC_INFO is (lname, social_sec#, street, city, dept). A *key* is a set of attribute names whose composite value is distinct for all tuples. In addition, no proper subset of the key is allowed to have this property. It is not unusual for a scheme to have several possible keys. In FAC_INFO, both lname and social_sec# are possible keys. In this case, each possible key is known as a *candidate key,* and the one selected to act as the relation's key, say, lname, is referred to as the *primary key.* A *superkey* is a key with the exception that there is no requirement for minimality. In a relation, an attribute name (or a set of attribute names) is referred to as a *foreign key,* if it is the primary key of another relation. In FAC_INFO, the attribute name dept is a foreign key, since the same attribute is a key in DEP_CHAIR. Because of updates to the database, the content of a relation is dynamic. For this reason, the data in a relation at a given time instant is called an *instance* of the relation.

There are three integrity constraints that are usually imposed on each instance of a relation: primary key integrity, entity integrity, and referential integrity. The key integrity constraint requires that no two tuples of a relation have the same key value. The entity integrity constraint specifies that the key value of each tuple should have a known value (i.e., no *null* values are allowed for primary keys). The referential integrity constraint specifies that if a relation r_1 contains a foreign key that matches the primary key of a relation r_2, then each value of the foreign key in r_1 must either match a value of the primary key in r_2 or must be null. For the database of Fig. 94.2 to be consistent, each value of dept in FAC_INFO must match a value of dept in DEP_CHAIR.

Relational Database Design

The relational database design [Maier, 1983] refers to the process of generating a set of relation schemes that minimizes data redundancy and removes update anomalies. One of the most popular approaches is the use of the *normalization theory*. The normalization theory is based on the notion of *functional dependencies*.

Functional dependencies are constraints imposed on a database. The notion of superkey, introduced in the previous section, can be formulated as follows: A subset of a relation scheme is a superkey if, in any instance of the relation, no two distinct tuples have the same superkey value. If $r(R)$ is used to denote a relation r on a schema R, $K \subseteq R$ a superkey, and $t(k)$ the K-value of tuple t, then no two tuples t_1 and t_2 in $r(R)$ are such that $t_1(K) = t_2(K)$.

The notion of a functional dependency can be seen as a generalization of the notion of superkey. Let X and Y be two subsets of R; the functional dependency $X \rightarrow Y$ exists in $r(R)$ if whenever two tuples in $r(R)$ have the same X-value, their Y-value is also the same. That is, if $t_1(X) = t_2(X)$, then $t_1(Y) = t_2(Y)$. Using functional dependencies, one can define the notion of a key more precisely. A key k of a relation $r(R)$ is such that $k \rightarrow R$ and no proper subset of k has this property. Note that if the schema R is composed of attribute names $\{A_1, A_2, \ldots, A_n\}$, then each attribute name A_i is functionally determined by the key k, i.e., $k \rightarrow A_i$, $i = 1, \ldots, n$. An

PRODUCT *(supplier_name,*	*product_name,*	*price,*	*location)*
Martin	sofa	500.99	Kenner
Martin	bed	100.99	Kenner
Martin	desk	150.99	Kenner
Evans	sofa	600.99	Metairie
Evans	desk	250.99	Metairie
Rudd	bed	110.99	Metairie

FIGURE 94.3 Instance of PRODUCT (supplier_name, product_name, price, quantity).

attribute name that is part of a key is referred to as a *prime attribute*. In the example of Fig. 94.2, both attribute names street and city are nonprime attributes.

The normalization process can be thought of as the process of decomposing a scheme with update anomalies and data redundancy into smaller schemes in which these undesirable properties are to a large extent eliminated. Depending on the severity of these undesirable properties, schemes are classified into *normal forms*. Originally, Codd defined three normal forms: *first normal form* (1NF), *second normal form* (2NF), and *third normal form* (3NF). Thereafter, a stronger version of the 3NF, known as Boyce-Codd normal form (BCNF), was suggested. These four normal forms are based on the concept of functional dependencies.

The 1NF requires that attribute name values be *atomic*. That is, composite values for attribute names are not allowed. A 2NF scheme is a 1NF scheme in which all nonprime attributes are fully dependent on the key. Consider the relation of Fig. 94.3. Each tuple in PRODUCT gives the name of a supplier, a product name, its price, and the supplier's location. The scheme (supplier_name, product_name, price, quantity) is in 1NF since each attribute name is atomic. It is assumed that many products can be supplied by a single supplier, that a given product can be supplied by more than one supplier, and that a supplier has only one location. So, (supplier_name, product_name) is the relation's key and the functional dependency supplier_name $\rightarrow$ location should hold for any instance of PRODUCT.

The structure of the relation of Fig. 94.3 does not allow a supplier to appear in the relation unless it offers at least one product. Even the use of null values is not of much help in this case as product_name is part of a key and therefore cannot be assigned a null value. Another anomaly can be encountered during the deletion process. For instance, deleting the last tuple in the relation results in the loss of the information that Rudd is a supplier located in Metairie. It is seen that the relation PRODUCT suffers from insertion and deletion anomalies.

Modifications can also be a problem in the relation PRODUCT. Suppose that the location of the supplier Martin is moved from Kenner to Slidell. In order not to violate the functional dependency supplier_name $\rightarrow$ location, the location attribute name of all tuples where the supplier is Martin needs to be changed from Kenner to Slidell. This modification anomaly has a negative effect on performance.

In addition, the relation PRODUCT suffers from data redundancy. For example, although Martin has only one location "Kenner", such a location appears in all three tuples where the supplier_name is Martin.

The update anomalies and data redundancy encountered in PRODUCT are all due to the functional dependency supplier_name $\rightarrow$ location. The right-hand side of this dependency "location" is a nonprime attribute, and the left-hand side represents part of the key. Therefore, we have a nonprime attribute that is only partially dependent on the key (supplier_name, product_name). As a consequence, the schema (supplier_name, product_name, price, location) is not in 2NF. The removal of the partial dependency supplier_name $\rightarrow$ location will eliminate all the above anomalies. The removal of the partial dependency is achieved by decomposing the scheme (supplier_name, product_name, price, quantity) into two 2NF schemes: (supplier_name, product_name, price), and (supplier_name, location). This decomposition results in relations PRO_INFO and SUP_LOC shown in Fig. 94.4. The keys of PRO_INFO and SUP_LOC are (supplier_name, product_name), and supplier_name, respectively.

Normalizing schemes into 2NF removes all update anomalies due to nonprime attributes being partially dependent on keys. Anomalies of a different nature, however, are still possible.

Update anomalies and data redundancy can originate from *transitive dependencies*. A nonprime attribute A_i is said to be transitively dependent on a key k via attribute name A_j, if $k \rightarrow A_j$, $A_j \rightarrow A_i$, and A_j does not functionally determine A_k. A 3NF is a 1NF where no nonprime attribute is transitively dependent on a key.

PRO_INFO *(supplier_name,*	*product_name,*	*price)*
Martin	sofa	500.99
Martin	bed	100.99
Martin	desk	150.99
Evans	sofa	600.99
Evans	desk	250.99
Rudd	bed	110.99

SUP_LOC *(supplier_name,*	*location)*
Martin	Kenner
Evans	Metairie
Rudd	Metairie

FIGURE 94.4 Decomposition of PRODUCT into PRO_INFO and SUP_LOC.

SUPPLIES *(client_name,*	*supplier_name,*	*location)*
Hosch	Martin	Kenner
Krad	Martin	Kenner
Shengru	Evans	Metairie
Tillis	Rudd	Metairie
Greene	Evans	Metairie

FIGURE 94.5 Instance of SUPPLIES.

SUP_CLI *(client_name,*	*supplier_name)*
Hosch	Martin
Krad	Martin
Shengru	Evans
Tillis	Rudd
Greene	Evans

SUP_LOC *(supplier_name,*	*location)*
Martin	Kenner
Evans	Metairie
Rudd	Metairie

FIGURE 94.6 Decomposition of SUPPLIES into SUP_CLI and SUP_LOC.

The relation of Fig. 94.5, which is in 2NF, highlights update anomalies and data redundancy due to the transitive dependency of a nonprime attribute on a key. The relation gives the name of a client (client_name), the corresponding supplier (supplier_name), and the supplier's location. Each client is assumed to have one supplier. The relation's key is client_name, and each supplier has only one location. A supplier and his location cannot be inserted in SUPPLIES unless the supplier has at least one client. In addition, the relation has a deletion anomaly since if Tillis is no longer a client of Rudd, the information about Rudd as a supplier and his location is lost. A change to a supplier's location may require updating the location attribute name of several tuples in the relation. Also, although each supplier has only one location, such a location is sometimes repeated several time unnecessarily, leading to data redundancy.

The relation exhibits the following transitive dependency: client_name $\rightarrow$ supplier_name, supplier_name $\rightarrow$ location (but not the inverse). The relation CLIENT is clearly in 2NF, but because of the transitive dependency of the nonprime attribute location on the key, it is not in 3NF. This is the cause of the anomalies mentioned above. Eliminating this transitive dependency by splitting the schema into two components will remove these anomalies. Clearly, the resulting two relations SUP_CLI and SUP_LOC are in 3NF (see Fig. 94.6).

Each partial dependency of a nonprime attribute on a key can be expressed as a transitive dependency of a nonprime attribute on a key. Therefore, a scheme in 3NF is also in 2NF.

BCNF is a stricter form of 3NF, where a relation r on a schema R is in BCNF if whenever a functional dependency $X \rightarrow Y$ exists in $r(R)$, then X is a superkey of R. The condition of 3NF, which allows Y to be prime if X is not a superkey, does not exist in BCNF. Thus, every scheme in BCNF is also in 3NF, but the opposite is not always true.

A detailed discussion of higher level normalizations, such as 4NF and 5NF, which are based on other forms of dependencies, can be found in [Elmasri and Navathe, 1994].

Data Definition and Manipulation in Relational Databases

Upon completion of the relational database design, a descriptive language, usually referred to as Data Definition Language (DDL), is used to define the designed schemes and their relationships. The DDL can be used to create new schemes or modify existing ones, but it cannot be used to query the database. Once DDL statements are compiled, they are stored in the *data dictionary*. A data dictionary is a repository where information about database schemas, such as attribute names, indexes, and integrity constraints are stored. Data dictionaries also contain other information about databases, such as design decisions, usage standards, application program descriptions, and user information. During the processing of a query, the DBMS usually checks the data dictionary. The data dictionary can be seen as a relational database of its own. As a result, data manipulation languages that are used to manipulate databases can also be used to query the data dictionary.

An important function of a DBMS is to provide a Data Manipulation Language (DML) with which a user can retrieve, change, insert, and delete data from the database. DMLs are classified into two types: *procedural* and *nonprocedural*. The main difference between the two types is that in procedural DMLs, a user has to specify the desired data and how to obtain it, while in nonprocedural DMLs, a user has only to describe the desired data. Because they impose less burden on the user, nonprocedural DMLs are normally easier to learn and use.

The component of a DML that deals with data retrieval is referred to as *query language*. A query language can be used interactively in a stand-alone manner, or it can be embedded in a general-purpose programming language such as C and Cobol.

One of the most popular query languages is SQL (Structured Query Language). SQL is a query language based to a large extent on Codd's *relational algebra*. SQL has additional features for data definition and update. Therefore, SQL is a comprehensive relational database language that includes both a DDL and DML.

SQL includes the following commands for data definition: CREATE TABLE, DROP TABLE, and ALTER TABLE. The CREATE TABLE is used to create and describe a new relation. The two relations of Fig. 94.4 can be created in the following manner:

```
CREATE TABLE PRO_INFO   (supplier_name   VARCHAR(12)     NOT NULL,
                         product_name    VARCHAR(8)      NOT NULL,
                         price           DECIMAL(6,2));
CREATE TABLE SUP_LOC    (supplier_name   VARCHAR(12)     NOT NULL,
                         location        VARCHAR(10));
```

The CREATE TABLE command specifies all the attribute names of a relation and their data types (e.g., INTEGER, DECIMAL, fixed length character "CHAR", variable length character "VARCHAR", DATE). The constraint NOT NULL is usually specified for those attributes that cannot have null values. The primary key of each relation in the database is usually required to have a nonnull value.

If a relation is created incorrectly, it can be deleted using the DROP TABLE command. The command is DROP TABLE followed by the name of the relation to be deleted. A variation of DROP command, DROP SCHEMA, is used if the whole schema is no longer needed.

The ALTER TABLE is used to add new attribute names to an existing relation, as follows:

```
ALTER TABLE SUP_LOC ADD zip_code CHAR(5);
```

The SUP_LOC relation now contains an extra attribute name, zip_code. In most DBMSs, the zip_code value of existing tuples will automatically be assigned a null value. Other DBMSs allow for the assignment of an initial value to a newly added attribute name. Also, definitions of attributes can be changed and new constraints can be added, or current constraints can be dropped.

The DML component of SQL has one basic query statement, sometimes called a mapping, that has the following structure:

```
SELECT    <attribute_name list>
FROM      <relation_list>
WHERE     <restriction>
```

PROFESSOR *(faculty,*	*department,*	*salary)*
Smith	Electrical Eng.	$39,000
Joe	Mechanical Eng.	$35,000
Susan	Computer Sci.	$36,000
Erick	Electrical Eng.	$38,000
Paul	Electrical Eng.	$37,000
Johannes	Computer Sci.	$65,000
Rick	Computer Sci.	$32,000
Gerard	Computer Sci.	$43,000
Kenneth	Mechanical Eng.	$40,000

FIGURE 94.7 Instance of the relation PROFESSOR.

In the above statement, the SELECT clause specifies the attribute names that are to be retrieved, FROM gives the list of the relations involved, and WHERE is a Boolean predicate that completely specifies the tuples to be retrieved.

Consider the database of Fig. 94.4, and suppose that we want the name of all suppliers that supply either beds or desks. In SQL, this query can be expressed as:

```
SELECT    supplier_name
FROM      PRO_INFO
WHERE     product_name = "bed" OR product_name = "sofa"
```

The result of an SQL command may contain duplicate values and is therefore not always a true relation. In fact, the result of the above query, shown below, has duplicate entries.

```
supplier_name
Martin
Martin
Rudd
```

The entry Martin appears twice in the result, because the supplier Martin supplies both beds and sofas. Removal of duplicates is usually a computationally intensive operation. As a result, duplicate entries are not automatically removed by SQL. To ensure uniqueness, the command DISTINCT should be used. In the above query, if we want the supplier names to be listed only once, the above query should be modified as follows:

```
SELECT DISTINCT   supplier_name
FROM              PRO_INFO
WHERE             product_name = "bed" OR product_name = "sofa"
```

In SQL, a query can involve more than one relation. Suppose that we want the list of all suppliers from Metairie who supply beds. Such a query, shown below, involves both PRO_INFO and SUP_LOC.

```
SELECT    supplier_name
FROM      PRO_INFO, SUP_LOC
WHERE     PRO_INFO.supplier_name = SUP_LOC.supplier_name
                                AND product_name = "bed"
```

When an SQL expression, such as the one above, involves more than one relation, it is sometimes necessary to qualify attribute names, that is, to precede an attribute name by the relation (a period is placed between the two) it belongs to. Such a qualification removes possible ambiguities.

In SQL, it is possible to have several levels of query nesting; this is done by including a SELECT query statement within the WHERE clause.

The output data can be presented in sorted order by using the SQL ORDER BY clause followed by the attribute name(s) according to which the output is to be sorted.

In database management applications it is often desirable to categorize the tuples of a relation by the values of a set of attributes and extract an aggregated characteristic of each category. Such database management tasks are referred to as *aggregation functions*. For instance, SQL includes the following built-in aggregation functions: SUM, COUNT, AVERAGE, MIN, MAX. The attribute names used for the categorization are referred to as

GROUP BY columns. Consider the relation PROFESSOR of Fig. 94.7. Each tuple of the above relation gives the name of a faculty and his department and academic year salary.

Suppose that we want to know the number of faculty in each department and the result to be ordered by department. This query requests for each department a count of the number of faculty. Faculty are therefore categorized according to the attribute name department. As a result, department is referred to as a GROUP BY attribute. In SQL, the above query is formulated as follows:

```
SELECT      department, COUNT (faculty)
FROM        PROFESSOR
GROUP BY    department
ORDER BY    department
```

The result of applying the COUNT aggregation function is a new relation with two attribute names. They are a GROUP BY attribute (department in this case) and a new attribute called COUNT. The tuples are ordered lexicographically in ascending order according to the ORDER BY attribute, which is department in this case:

```
department          COUNT (faculty)
Computer Sc.          4
Electrical Eng.       3
Mechanical Eng.       2
```

The relations created through the CREATE TABLE command are known as *base relations*. A base relation exists physically and is stored as a file by the DBMS. SQL can be used to create views using the CREATE VIEW command. In contrast to base relations, the creation of a view results in a *virtual relation*, that is, one that does not necessarily correspond to a physical file. Consider the database of Fig. 94.4, and suppose that we want to create a view giving the name of all suppliers located in Metairie, the products each one provides, and the corresponding prices. Such a view, called METAIRIE_SUPPLIER, can be created as follows:

```
CREATE VIEW     METAIRIE_SUPPLIER
AS SELECT       PRO_INFO.supplier_name, product_name, price
FROM            PRO_INFO, SUP_LOC
WHERE           PRO_INFO.supplier_name = SUP_LOC.supplier_name
                                  AND location = "Metairie"
```

Because a view is a virtual relation that can be constructed from one or more relations, updating a view may lead to ambiguities. As a result, when a view is generated from more than one relation, there are, in general, restrictions on updating such a view.

Hierarchical Databases

The hierarchical data model [Elmasri and Navathe, 1994] uses a tree data structure to conceptualize associations between different record types. In this model, record types are represented as nodes and associations as links. Each record type, except the root, has only one parent; that is, only parent-child (or one-to-many) relationships are allowed. This restriction gives hierarchical databases their simplicity. Since links are only one way, from a parent to a child, the design of hierarchical database management systems is made simpler, and only a small set of data manipulation commands are needed.

Because only parent-child relationships are allowed, the hierarchical model cannot efficiently represent two main types of relationships: many-to-many relationships and the case where a record type is a child in more than one *hierarchical schema*. These two restrictions can be handled by allowing redundant *record instances*. However, such a duplication requires that all the copies of the same record should be kept consistent at all times.

The example of Fig. 94.8 shows a hierarchical schema. The schema gives the relationship between a DEPARTMENT, its employees (D_EMPLOYEE), the projects (D_PROJECT) handled by the different departments, and how employees are assigned to these projects. It is assumed that an employee belongs to only one department, a project is handled by only one department, and an employee can be assigned to several projects. Notice that since a project has several employees assigned to it, and an employee can be assigned to more than one project, the relationship between D_PROJECT and D_EMPLOYEE is many-to-many. To model this relationship multiple instances of the same record type D-EMPLOYEE may appear under different projects.

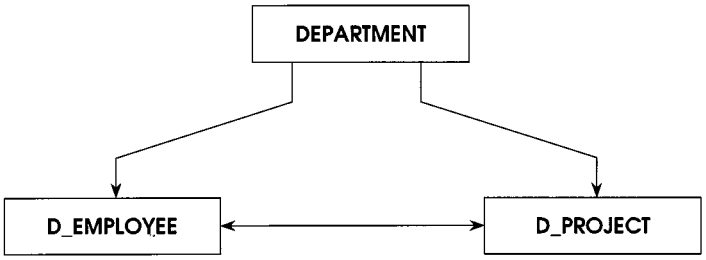


FIGURE 94.8 A hierarchical schema.

Such redundancies can be reduced to a large extent through the use of *logical links*. A logical link associates a *virtual record* from a hierarchical schema with an actual record from either the same schema or another schema. The redundant copy of the actual record is therefore replaced by a virtual record, which is nothing more than a pointer to the actual one.

Hierarchical DLLs are used by a designer to declare the different hierarchical schemas, record types, and logical links. Furthermore, a root node must be declared for each hierarchical schema, and each record type declaration must also specify the parent record type.

Unlike relational DMLs, hierarchical DMLs such as DL/1 are record at-a-time languages. DL/1 is used by IBM's IMS hierarchical DBMS. In DL/1 a tree traversal is based on a preorder algorithm, and within each tree, the last record accessed through a DL/1 command can be located through a *currency indicator*.

Retrieval commands are of three types:

GET UNIQUE <record type> **WHERE** <restrictions>
Such a command retrieves the leftmost record that meets the imposed restrictions. The search always starts at the root of the tree pointed to by the currency indicator.

GET NEXT [<record type> **WHERE** <restrictions>]
Starting from the current position, this command uses the preodrer algorithm to retrieve the next record that satisfies the restrictions. The clause enclosed between brackets is optional. GET NEXT is used to retrieve the next (preorder) record from the current position.

GET NEXT WITHIN PARENT [<record type> **WHERE** <restrictions>]
It retrieves all records that have the same parent and that satisfy the restrictions. The parent is assumed to have been selected through a previous GET command.
Four commands are used for record updates:

INSERT
Stores a new record and links it to a parent. The parent has been already selected through a GET command.

REPLACE
The current record (selected through a previous GET command) is modified.

DELETE
The current record and all its descendants are deleted.

GET HOLD
Locks the current record while it is being modified.

The DL/1 commands are usually embedded in a general-purpose (host) language. In this case, a record accessed through a DL/1 command is assigned to a program variable.

Network Databases

In the network model [Elmasri and Navathe, 1994] associations between record types are less restrictive than with the hierarchy model. Here, associations among record types are represented as graphs.

One-to-one and one-to-many relationships are described using the notion of *set type*. Each set type has an *owner* record type and a *member* record type. In the example of Fig. 94.8, the relationship between DEPARTMENT and

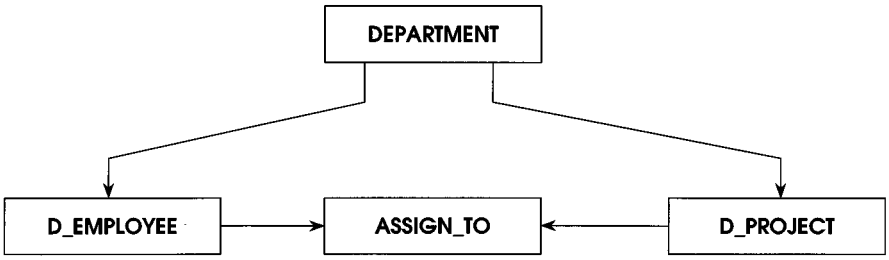


FIGURE 94.9 Representing many-to-many relationships in the network model.

employee (D_EMPLOYEE) is one-to-many. This relationship defines a set type where the owner record type is DEPARTMENT and the member record type is D_EMPLOYEE. Each instance of an owner record type along with all the corresponding member records represents a set instance of the underlying set type. In practice, a set is commonly implemented using a circular-linked list which allows an owner record to be linked to all its member records. The pointer associated with the owner record is known as the FIRST pointer, and the one associated with a member record is known as a NEXT pointer.

In general, a record type cannot be both the owner and a member of the same set type. Also, a record cannot exist in more than one instance of a specific set type. The latter requirement implies that many-to-many relationships are not directly implemented in the network data model.

The relationship between D_PROJECT and D-EMPLOYEE is many-to-many. In the network model, this relationship is represented by two set types and an intermediate record type. The new record type could be named ASSIGNED_TO (see Fig. 94.9). One set has D_EMPLOYEE as owner and ASSIGNED_TO as member record type, and the other has D_PROJECT as owner and ASSIGNED_TO as member record type.

Standards for the network model's DDL and DML were originally proposed by the CODASYL (Conference On Data SYstems Languages) committee in 1971. Several revisions to the original proposal were made later.

In a network DDL, such as that of the IDMS database management system, a set declaration specifies the name of the set, its owner record type, and its member record type. The insertion mode for the set members needs to be specified using combinations of the following four commands:

AUTOMATIC
An inserted record is automatically connected to the appropriate set instance.

MANUAL
In this case, records are inserted into the appropriate set instance by an application program.

OPTIONAL
A member record does not have to be a member of a set instance. The member record can be connected to or disconnected from a set instance using DML commands.

MANDATORY
A member record needs to be connected to a set instance. A member record can be moved to another set instance using the network's DML.

FIXED
A member record needs to be connected to a set instance. A member record cannot be moved to another set instance.

The network's DDL allows member records to be ordered in several ways. Member records can be sorted in ascending or descending order according to one or more fields. Alternatively, a new member record can be inserted next (prior) to the current record (pointed to by the currency indicator) in the set instance. A newly inserted member record can also be placed first (or last) in the set instance. This will lead to a chronological (or reverse chronological) order among member records.

As with the hierarchy model, network DMLs are record-at-a-time languages, and currency indicators are necessary for navigation through the network database. For example, the IDMS main data manipulation commands can be summarized as follows:

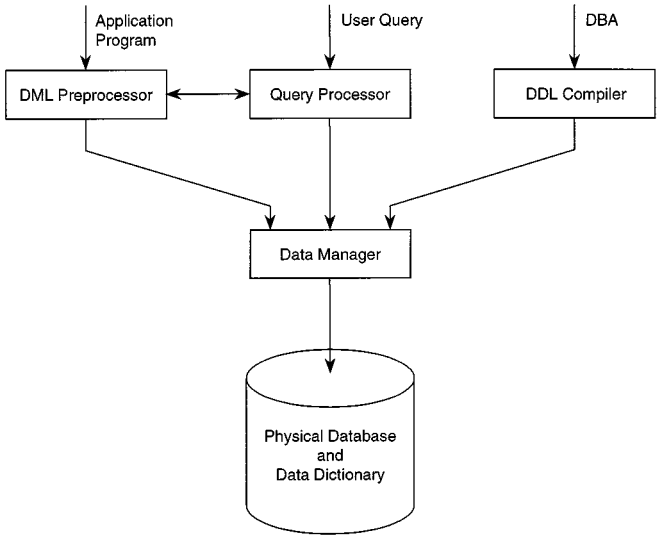


FIGURE 94.10 Simplified architecture of a DBMS.

`CONNECT`
Connects a member record to the specified set instance.

`DISCONNECT`
A member record is disconnected from a set instance (set membership must be manual in this case).

`STORE, MODIFY,` and **`DELETE`**
These commands are used for data storage, modification, and deletion.

`FIND`
Retrieval command based on set membership.

`GET`
Retrieval command based on key values.

Architecture of a DBMS

A DBMS is a complicated software structure that includes several components (see Fig. 94.10). The DBMS has to interact with the operating system for secondary storage access. The *data manager* is usually the interface between the DBMS and the operating system. The *DDL compiler* converts schema definitions, expressed using DDL statements, into a collection of metadata tables that are stored in the data dictionary. The design of the schemas is the function of the *database administrator* (DBA). The DBA is also responsible for specifying the data storage structure and access methodology and granting and revoking access authorizations. The *query processor* converts high-level DML statements into low-level instructions that the database manager can interpret. The *DML preprocessor* separates embedded DML statements from the rest of an application program. The resulting DML commands are processed by a DML compiler, and the rest of the application program is compiled by a host compiler. The object codes of the two components are then linked.

Data Integrity and Security

Data Integrity

In general, during the design of a database schema several integrity constraints are identified. These constraints may include the uniqueness of a key value, restrictions on the domain of an attribute name, and the ability of an attribute to have a null value. A DBMS includes mechanisms with which integrity constraints can be specified. Constraints such as key uniqueness and the admissibility of null values can be specified during schema definition. Also, more elaborate integrity constraints can be specified. For example, constraints can be imposed

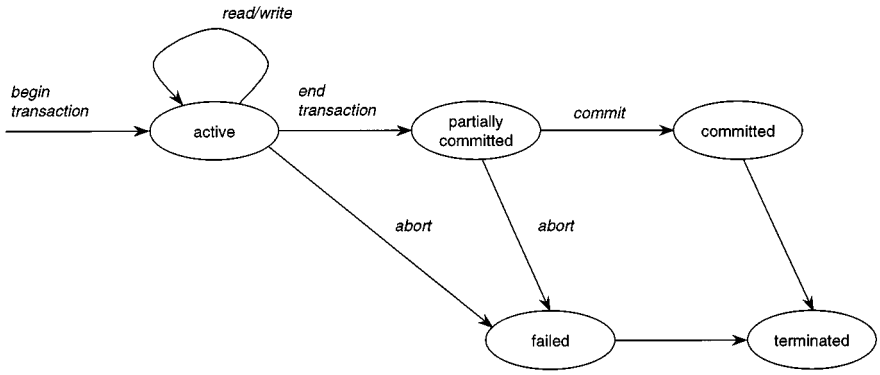


FIGURE 94.11 State transition diagram for transaction execution.

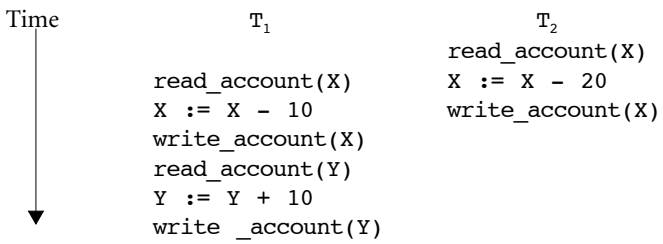
on the domain of an attribute name, and any transaction that violates the imposed constraints is aborted. In some cases, it is useful to specify that the system take some actions, rather than just have the transaction responsible for the constraint violation being aborted. A mechanism called *trigger* can be used for that purpose. A trigger specifies a condition and an action to be taken when the condition is met.

Transactions and Data Integrity

In a multiuser DBMS, the database is a shared resource that can be accessed concurrently by many users. A *transaction* usually refers to the execution of a retrieval or an update program. A transaction performs a single logical operation in a database application. Therefore, it is an *atomic* unit of processing. That is, a transaction is either performed in its entirety or is not performed at all. Basically, a transaction may be in one of the following states (Fig. 94.11):

- active — where read and write operations are performed.
- partially committed — when the transaction ends and various checks are made to ensure that the transaction did not interfere with other transactions.
- failed — when one of the checks failed or the transaction is aborted during the active state.
- committed — when the execution was successfully completed.
- terminated — when the transaction leaves the system.

Transactions originating from different users may be aimed at the same database records. This situation, if not carefully monitored, may cause the database to become *inconsistent*. Starting from a database in a consistent state, it is obvious that if all transactions are executed one after the other, then the database will remain in a consistent state. In a multiuser DBMS, serial execution of transactions is wasteful of system resources. In this case, the solution is to interleave the execution of the transactions. However, the interleaving of transactions has to be performed in a way that prevents the database from becoming inconsistent. Suppose that two transactions T_1 and T_2 proceed in the following way:



The first transaction transfers $10 from bank account X to bank account Y. The second transaction withdraws $20 from bank account X. Assume that initially there was $200 in X and $100 in Y. When the two transactions are performed serially, the final amounts in X and Y are $170 and $110, respectively. However, if the two transactions are interleaved as shown, then after the completion of both transactions, there will be $190 in X and $110 in Y. The database is now in an inconsistent state.

It is therefore important to ensure that the interleaving of the execution of transactions leaves the database in a consistent state. One way of preserving data consistency is to ensure that the interleaved execution of transactions is equivalent to their serial execution. This is referred to as *serializable* execution. Therefore, an interleaved execution of transactions is said to be serializable if it is equivalent to a serial execution.

Locking is one of the most popular approachs to achieving serializability. Locking is the process of ensuring that some actions are not performed on a data item. Therefore, a transaction may request a lock on a data item to prevent it from being either accessed or modified by other transactions. There are two basic types of locks. A *shared lock* allows other transactions to read but not write to the data item. An *exclusive lock* allows only a single transaction to read and write a data item. To achieve a high degree of concurrency, the locked data item size must be as small as possible. A data item can range from the whole database to a particular field in a record. Large data items limit concurrency, while small data items result in a large storage overhead and a greater number of lock and unlock operations that the system will have to handle.

Transactions scheduling based on locking achieves serializability in two phases. This is known as *two-phase locking*. During the first phase, the growing phase, a transaction can only lock new data items, but it cannot release any locked ones. During the second phase, the shrinking phase, existing locks can be released, but no new data item can be locked. The two-phase locking scheme guarantees the serializability of a schedule.

Because of its simplicity, the above scheduling method is very practical. However, it may lead to a *deadlock*. A deadlock occurs when two transactions are waiting for each other to release locks and both cannot proceed. A deadlock prevention (or detection) strategy is needed to handle the situation. For example, this can be achieved by requiring that a transactions locks all data items it needs for its execution before it can proceed; when the transaction finds that a needed data item is already locked, then it releases all locks.

If a transaction fails for whatever reason after (partially committed) or (active) while updating the database, it may be necessary to bring the database to its previous (original) state by undoing the transaction. This operation is called *roll-back*. A roll-back operation requires some information about the changes made on the data items during a transaction. Such information is usually kept outside the database in a system log. Generally, roll-back operations are part of the techniques used to recover from transaction failures.

Database Security

A database needs to be protected against unauthorized access. It is the responsibility of the DBA to create account numbers and passwords for legitimate users. The DBA can also specify the type of privileges a particular account has. In relational databases, this includes the privilege to create base relations, create views, alter relations by adding or dropping a column, and delete relations. The DBA can also revoke privileges that were granted previously. In SQL, the command GRANT is used to grant privileges and the REVOKE command to revoke privileges that have been granted.

The concept of views can serve as a convenient security mechanism. Consider a relation EMPLOYEE that gives the name of an employee, date of birth, the department worked for, address, phone number, and salary. A database user who is not allowed to have access to the salary of employees from his own department can have this portion of the database hidden from him. This can be achieved by limiting his access to a view obtained from the relation EMPLOYEE by selecting only those tuples where the department attribute is different from his.

Database security can be enhanced by using *data encryption*. The idea here is to encrypt the data using some coding technique. An unauthorized user will have difficulty deciphering the encrypted data. Only authorized users are provided with keys to decipher the encoded data.

Emerging Trends

Object-Oriented Databases

Object-oriented database systems (OODBMSs) [Brown, 1991] are one of the latest trends in database technology. The emergence of OODBMS is in response to the requirements of advanced applications. In general, traditional commercial and administrative applications can be effectively modeled using one of the three record-based data models. These applications are characterized by simple data types. Furthermore, for such applications, access and relationships are based on data values. Advanced database applications such as those found in engineering CAD/CAM require complex data structures. When these applications are modeled using the relational model, they require an excessive number of relations. In addition, a large number of complex

operations are usually needed to produce an answer. This leads, in most cases, to unacceptable performance levels.

The notion of "object" is central to OODBMS. An object can be seen as being an entity consisting of its own *private memory* and *external interface* (or protocol). The private memory is used to store the state of the object, and the external interface consists of a set of operations that can be performed on the object. An object communicates with other objects through messages sent to its external interface. When an object receives a message, it responds by using its own procedures, known as *methods*. The methods are responsible for processing the data in the object's private memory and sending messages to other objects to perform specific tasks and possibly send back appropriate results.

The object-oriented approach provides for a high level of *abstraction*. In addition, this model has constructs that can be used to define new data types and specialized operators that can be applied to them. This feature is known as *encapsulation*.

An object is usually a member of a class. The class specifies the internal structure and the external interface of an object. New object classes can be defined as a *specialization* of existing ones. For example, in a university environment, the object type "faculty" can be seen as a specialization of the object type "employee." Since a faculty is a university employee, it has all the properties of a university employee plus some of its own. For example, some of the general operations that can be performed on an employee could be "raise_salary," "fire_employee," "transfer_employee." For a faculty, specialized operations such as "faculty_tenure" could be defined. Faculty can be viewed as a subclass of employee. As a result, faculty (the subclass) will respond to the same messages as employee (the superclass) in addition to those defined specifically for faculty. This technique is known as *inheritance*. A subclass is said to inherit the behavior of its superclass.

Opponents to the object-oriented paradigm point to the fact that while this model has greater modeling capability, it lacks the simplicity and the strong theoretical foundations of the relational model. Also, the reappearance of the navigational approach is seen by many as a step backward.

Supporters of the object-oriented approach believe that a navigational approach is a necessity in several applications. They point to the rich modeling capability of the model, its high level of abstraction, and its suitability for modular design.

Distributed Databases

A **distributed database** [Ozsu and Valdurez, 1991] is a collection of interrelated databases spread over the nodes of a computer network. The management of the distributed database is the responsibility of a software system usually known as distributed DBMS (DDBMS). One of the tasks of the DDBMS is to make the distributed nature of the database transparent to the user. A distributed database usually reflects the distributed nature of some applications. For example, a bank may have branches in different cities. A database used by such an organization is usually distributed over all these sites. The different sites are connected by a computer network. A user may access data stored locally or access data stored at other sites through the network.

Distributed databases have several advantages. In distributed databases, the effect of a site failure or data loss at a particular node can be minimized through data replication. However, data replication reduces security and makes the process of keeping the database consistent more complicated.

In distributed databases, data is decomposed into fragments that are allocated to the different sites. A fragment is allocated to a site in a way that maximizes local use. This allocation scheme, which is known as *data localization*, reduces the frequency of remote access. In addition, since each site deals with only a portion of the database, local query processing is expected to exhibit increased performance.

A distributed database is inherently well suited for parallel processing at both interquery and intraquery levels. Parallel processing at the interquery level is the ability to have multiple queries executed concurrently. Parallelism at the intraquery level results from the possibility of a single query being simultaneously handled by many sites, each site acting on a different portion of the database.

The data distribution increases the complexity of DDBMS over a centralized DBMS. In fact, in distributed databases, several research issues in distributed query processing, distributed database design, and distributed transaction processing remain to be solved. It is only then that the potential of distributed databases can be fully appreciated.

Parallel Database Systems

There has been a continuing increase in the amount of data handled by database management systems (DBMSs) in recent years. Indeed, it is no longer unusual for a DBMS to manage databases ranging in sizes from hundreds of gigabytes to terabytes. This massive increase in database sizes is coupled with a growing need for DBMSs to exhibit more sophisticated functionality such as the support of object-oriented, deductive, and multimedia applications. In many cases, these new requirements have rendered existing DBMSs unable to provide the necessary system performance, especially given that many mainframe DBMSs already have difficulty meeting the I/O and CPU performance requirements of traditional information systems that service large numbers of concurrent users and/or handle massive amounts of data [DeWitt and Gray, 1992].

To achieve the required performance levels, database systems have been increasingly required to make use of parallelism. Two approaches were suggested to provide parallelism in database systems [Abdelguerfi and Lavington, 1995]. The first approach uses massively parallel general-purpose hardware platforms. Commercial systems, such as Intel's nCube and IBM's SP2 follow this approach and support Oracle's Parallel Server. The second approach makes use of arrays of off-the-shelf components to form custom massively parallel systems. Usually, these hardware systems are based on MIMD parallel architectures. The NCR 3700 and the Super Database Computer II (SDC-II) are two such systems. The NCR 3700 now supports parallel version of Sybase relational DBMS.

The number of general purpose or dedicated parallel database computers is increasing each year. It is not unrealistic to envisage that most high performance database management systems in the year 2000 will support parallel processing. The high potential of parallel databases in the future urges both the database vendors and practitioners to understand the concept of parallel database system in depth.

It is noteworthy that in recent years, popularity of the client/server architecture has increased. This architecture is practically a derivative of shared-nothing case. In this model, clients' nodes access data through one or more servers. This approach derives its strength from an attractive price/performance ratio, a high level of scalability, and the ease with which additional remote hosts can be integrated into the system. Another driving force of the client/server approach is the current trend toward corporate downsizing.

Multimedia

Yet another new generation database application is multimedia, where non-text forms of data, such as voice, video, and image, are accessed via some form of a user interface. Hypermedia interfaces are becoming the primary delivery system for the multimedia applications. These interfaces, such as Mosaic, allow users to browse through an information base consisting of many different types of data. The basis of hypermedia is the hypertext, where some text based information is accessed in a non-sequential manner. Hypermedia is an extension of hypertext paradigm into multimedia.

Defining Terms

Database: A shared pool of interrelated data.

Database computer: A special hardware and software configuration aimed primarily at handling large databases and answering complex queries.

Database management system (DBMS): A software system that allows for the definition, construction, and manipulation of a database.

Data model: An integrated set of tools to describe the data and its structure, data relationships, and data constraints.

Distributed database: A collection of multiple, logically interrelated databases distributed over a computer network.

Related Topic

87.3 Data Types and Data Structures

References

M. Abdelguerfi and A. K. Sood, Eds., Special Issue on Database Computers, *IEEE Micro,* December 1991.

M. Abdelguerfi and S. Lavingston, Eds., *Emerging Trends in Database and Knowledge Base Machines,* IEEE Computer Science Press, 1995.

A. Brown, *Object-Oriented Databases: Applications in Software Engineering,* New York: McGraw-Hill, 1991.

E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM,* pp. 377–387, June 1970.

D. DeWitt and J. Gray, "Parallel database systems: The future of high performance database systems", *Communications of the ACM,* pp. 85-98, June 1992.

R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems,* Redwood City, Calif.: Benjamin/ Cummings, 2 ed., 1994.

D. Maier, *The Theory of Relational Databases,* New York: Computer Science Press, 1983.

M. T. Ozsu and P. Valdurez, *Principles of Distributed Database Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1991.

94.2 Rule-Based Expert Systems

Jay Liebowitz

Expert systems is probably the most practical application of artificial intelligence (AI). Artificial intelligence, as a field, has two major thrusts: (1) to supplement human brain power with intelligent computer power and (2) to better understand how we think, learn, and reason. Expert systems are one application of AI, and they are being developed and used throughout the world [Feigenbaum et al., 1988; Liebowitz, 1990]. Other major applications of AI are robotics, speech understanding, natural-language understanding, computer vision, and neural networks.

Expert systems are computer programs that emulate the behavior of a human expert in a well-bounded domain of knowledge [Liebowitz, 1988]. They have been used in a number of tasks, ranging from sheep reproduction management in Australia, hurricane damage assessment in the Caribbean, boiler plant operation in Japan, computer configuration in the United States, to strategic management consulting in Europe [Liebowitz, 1991b]. Expert systems technology has been around since the late 1950s, but it has been only since 1980–1981 that the commercialization of expert systems has emerged [Turban, 1992].

An expert system typically has three major components: the dialog structure, inference engine, and knowledge base [Liebowitz and DeSalvo, 1989]. The dialog structure is the user interface that allows the user to interact with the expert system. Most expert systems are able to explain their reasoning, in the same manner that one would want human experts to explain their decisions. The inference engine is the control structure within the expert system that houses the search strategies to allow the expert system to arrive at various conclusions. The third component is the **knowledge base,** which is the set of facts and heuristics (rules of thumb) about the specific domain task. The knowledge principle says that the power of the expert system lies in its knowledge base. Expert system shells have been developed and are widely used on various platforms to help one build an expert system and concentrate on the knowledge base construction. Most operational expert systems are integrated with existing databases, spreadsheets, optimization modules, or information systems [Mockler and Dologite, 1992].

The most successful type of expert system is the rule-based, or production, system. This type of expert system is chiefly composed of IF-THEN (condition-action) rules. For example, the infamous MYCIN expert system, developed at Stanford University for diagnosing bacterial infections in the blood (meningitis), is rule-based, consisting of 450–500 rules. XCON, the expert system at Digital Equipment Corporation used for configuring VAX computer systems, is probably the largest rule-based expert system, consisting of over 11,000 rules. There are other types of expert systems that represent knowledge in ways other than rules or in conjunction with rules. Frames, scripts, and semantic networks are popular knowledge representation methods that could be used in expert systems.

The development of rule-based systems is typically called **knowledge engineering.** The knowledge engineer is the individual involved in the development and deployment of the expert system. Knowledge engineering, in rule-based systems, refers primarily to the construction of the knowledge base. As such, there are six major steps in this process, namely (1) problem selection, (2) knowledge acquisition, (3) knowledge representation, (4) knowledge encoding, (5) knowledge testing and evaluation, and (6) implementation and maintenance. The knowledge engineering process typically uses a rapid prototyping approach (build a little, test a little). Each of the six steps in the knowledge engineering process will be briefly discussed in turn.

Problem Selection

In selecting an appropriate application for expert systems technology, there are a few guidelines to follow:

- Pick a problem that is causing a large number of people a fair amount of grief.
- Select a "doable," well-bounded problem (i.e., task takes a few minutes to a few hours to solve)—this is especially important for the first expert system project for winning management's support of the technology.
- Select a task that is performed frequently.
- Choose an application where there is a consensus on the solution of the problem.
- Pick a task that utilizes primarily symbolic knowledge.
- Choose an application where an expert exists and is willing to cooperate in the expert systems development.
- Make sure the expert is articulate and available and a backup expert exists.
- Have the financial and moral support from management.

The problem selection and scoping are critical to the success of the expert systems project. As with any information systems project, the systems analysis stage is an essential and crucial part of the development process. With expert systems technology, if the problem domain is not carefully selected, then difficulties will ensue later in the development process.

Knowledge Acquisition

After the problem is carefully selected and scoped, the next step is knowledge acquisition. Knowledge acquisition involves eliciting knowledge from an expert or multiple experts and also using available documentation, regulations, manuals, and other written reports to facilitate the knowledge acquisition process. The biggest bottleneck in expert systems development has, thus far, been in the ability to acquire knowledge. Various automated knowledge acquisition tools, such as Boeing Computer Services' AQUINAS, have been developed to assist in this process, but there are very few knowledge acquisition tools on the market. The most commonly used approaches for acquiring/eliciting knowledge include: interviewing (structured and unstructured), protocol analysis, questionnaires (structured and open-ended), observation, learning by example/analogy, and other various techniques (Delphi technique, statistical methods).

To aid the knowledge acquisition process, some helpful guidelines are:

- Before interviewing the expert, make sure that you (as the knowledge engineer) are familiar/ comfortable with the domain.
- The first session with the expert should be an introductory lecture on the task at hand.
- The knowledge engineer should have a systematic approach to acquiring knowledge.
- Incorporate the input and feedback from the expert (and users) into the system—get the expert and users enthusiastic about the project.
- Pick up manuals and documentation on the subject material.
- Tape the knowledge acquisition sessions, if allowed.

Knowledge Representation

After acquiring the knowledge, the next step is to represent the knowledge. In a rule-based expert system, the IF-THEN (condition-action) rules are used. Rules are typically used to represent knowledge if the preexisting knowledge can best be naturally represented as rules, if the knowledge is procedural, if the knowledge is mostly context-independent, and if the knowledge is mostly categorical ("yes-no" type of answers). Frames, scripts, and semantic networks are used as knowledge representation schemes for more descriptive, declarative knowledge. In selecting an appropriate knowledge representation scheme, try to use the representation method which most closely resembles the way the expert is thinking and expressing his/her knowledge.

Knowledge Encoding

Once the knowledge is represented, the next step is to encode the knowledge. Many knowledge engineers use expert system shells to help develop the expert system prototypes. Other developers may build the expert system from scratch, using such languages as Lisp, Prolog, C, and others. The following general guidelines may be useful in encoding the knowledge:

- Remember that for every shell there is a perfect task, but for every task there is NOT a perfect shell.
- Consider using an expert system shell for prototyping/proof-of-concept purposes—remember to first determine the requirements of the task, instead of force-fitting a shell to a task.
- Try to develop the knowledge base in a modular format for ease of updating.
- Concentrate on the user interface and human factors features, as well as the knowledge base.
- Use an incremental, iterative approach.
- Consider whether uncertainty should play a part in the expert system.
- Consider if the expert reasons in a data-driven manner (forward chaining) or a goal-directed manner (backward chaining), or both.

Knowledge Testing and Evaluation

Once the knowledge is encoded in the system, testing and evaluation need to be conducted. Verification and validation refers to checking for the consistency of the knowledge/logic and checking the quality/accuracy of advice reached by the expert system. Various approaches to testing can be used, such as: performing "backcasting" by running the expert system (using a representative set of test cases) against documented cases and comparing the expert system-generated results with the historical results, using blind verification tests (modified Turing test), having the expert and other experts test the system, using statistical methods for testing, and others. In evaluating the expert system, the users should evaluate the design of the human factors in the system (i.e., instructions, free-text comments, ease of updating, exiting capabilities, response time, display and presentation of conclusions, ability to restart, ability for user to offer degree of certainty, graphics, utility of the system, etc.).

Implementation and Maintenance

Once the system is ready to be deployed within the organization, the knowledge engineer must be cognizant of various institutionalization factors [Liebowitz, 1991a; Turban and Liebowitz, 1992]. Institutionalization refers to implementing and transitioning the expert system into the organization. Frequently, the technology is not the limiting factor—the *management* of the technology is often the culprit. An expert system may be accurate and a technical success, but without careful attention to management and institutionalization considerations, the expert system may be a technology transfer failure. There are several useful guidelines for proper institutionalization of expert systems:

- Know the corporate culture in which the expert system is deployed.
- Planning for the institutionalization process must be thought out well in advance, as early as the requirements analysis stage.

- Through user training, help desks, good documentation, hotlines, etc., the manager can provide mechanisms to reduce "resistance to change."
- Solicit and incorporate users' comments during the analysis, design, development, and implementation stages of the expert system.
- Make sure there is a team/individual empowered to maintain the expert system.
- Be cognizant of possible legal problems resulting from the use and misuse of the expert system.
- During the planning stages, determine how the expert system will be distributed.
- Keep the company's awareness of expert systems at a high level throughout the system's development and implementation, and even after its institutionalization.

Defining Terms

Expert systems: A computer program that emulates a human expert in a well-bounded domain of knowledge.
Knowledge base: The set of facts and rules of thumb (heuristics) on the domain task.
Knowledge engineering: The process of developing an expert system.

References

E.A. Feigenbaum, P. McCorduck, and P. Nii, *The Rise of the Expert Company,* New York: Times Books, 1988.

J.K. Lee, J. Liebowitz, and Y.M. Chae, Eds., Proceedings of the Third World Congress on Expert Systems, New York: Cognizant Communication Corp., 1996.

J. Liebowitz, *Introduction to Expert Systems,* New York: Mitchell/McGraw-Hill Publishing, 1988.

J. Liebowitz, Ed., *Expert Systems for Business and Management,* Englewood Cliffs, N.J.: Prentice-Hall, 1990.

J. Liebowitz, *Institutionalizing Expert Systems: A Handbook for Managers,* Englewood Cliffs, N.J.: Prentice-Hall, 1991a.

J. Liebowitz, Ed., *Operational Expert System Applications in the United States,* New York: Pergamon Press, 1991b.

J. Liebowitz, and D. DeSalvo, Eds., *Structuring Expert Systems: Domain, Design, and Development,* Englewood Cliffs, N.J.: Prentice-Hall, 1989.

R. Mockler and D. Dologite, *An Introduction to Expert Systems,* New York: Macmillan Publishing, 1992.

E. Turban, *Expert Systems and Applied Artificial Intelligence,* New York: Macmillan Publishing, 1992.

E. Turban and J. Liebowitz, Eds., *Managing Expert Systems,* Harrisburg, Pa.: Idea Group Publishing, 1992.

Further Information

There are several journals and magazines specializing in expert systems that should be consulted:
Expert Systems with Applications: An International Journal, New York/Oxford: Pergamon Press, Elsevier.
Expert Systems, Medford, N.J.: Learned Information, Inc.
IEEE Expert, Los Alamitos, Calif.: IEEE Computer Society Press.
AI Expert, San Francisco: Miller Freeman Publications.
Intelligent Systems Report, Atlanta: AI Week, Inc.

Feng, T. "Parallel Processors"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

95

Parallel Processors

95.1 Introduction
95.2 Classifications
95.3 Types of Parallel Processors
Ensemble Processors • Array Processors • Associative Processor
95.4 System Utilization

Tse-yun Feng
The Pennsylvania State University

95.1 Introduction

A computer usually consists of four major components: the arithmetic-logic unit (ALU), the main memory unit (MU), the input/output unit (I/O), and the control unit (CU). Such a computer is known as a uniprocessor since the processing is achieved by operating on one word or word pair at a time. In order to increase the computer performance, we may improve the device technology to reduce the switching (gate delay) time. Indeed, for the past half century we have seen switching speeds improve from 200 to 300 ms for relays to present-day subnanosecond very large scale integration (VLSI) circuits. As the switching speeds of computer devices approach a limit, however, any further significant improvement in performance is more likely to be in increasing the number of words or word pairs that can be processed simultaneously. For example, we may use one ALU to compute N sets of additions N times in a uniprocessor, or we may design a computer system with N ALUs to add all N sets once. Conceptually, such a computer system may still consist of the four major components mentioned previously except that there are N ALUs. An organization with multiple ALUs under the control of a single CU is called a **parallel processor.** To make a parallel processor more efficient and cost-effective, a fifth major component, called **interconnection networks,** is usually required to facilitate the interprocessor and processor-memory communications. In addition, each ALU requires not only its own registers but also network interfaces; the expanded ALU is then called a **processing element** (PE). Figure 95.1 shows a block diagram of a parallel processor.

95.2 Classifications

Flynn has classified computer systems according to the multiplicity of instruction and data streams, where computers are partitioned into four groups [Flynn, 1966]:

1. Single instruction stream, single data stream (SISD): The conventional, word-sequential architecture including pipelined computers (usually with parallel ALU)
2. Single instruction stream, multiple data stream (SIMD): The multiple ALU-type architectures (e.g., parallel/array processor). The ALU may be either bit-serial or bit-parallel.
3. Multiple instruction stream, single data stream (MISD): Not as practical as the other classes.
4. Multiple instruction stream, multiple data stream (MIMD): The multiprocessor system.

As a general rule, one could conclude that SISD and SIMD machines are single CU systems, whereas MIMD machines are multiple CU systems. Flynn's classification does not address the interactions among the processing

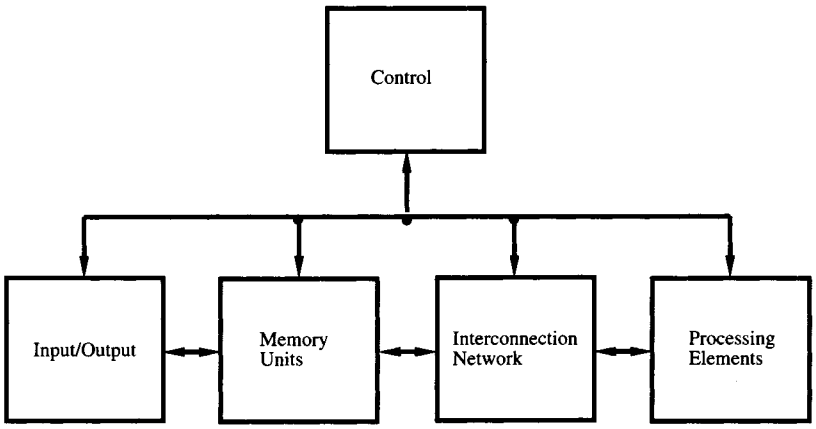


FIGURE 95.1 A basic parallel processor organization.

modules and the methods by which processing modules in concurrent system are controlled. As a result, one can classify both uniprocessors and pipelined computers as SISD machines, because both instructions and data are provided sequentially.

We may also classify computer systems according to the number of bits or bit pairs a computer executes at any instant [Feng, 1972]. For example, a computer may perform operations on one bit or bit pair at a time through the use of a simple serial ALU. For an M-bit word or operand, the operation repeats M times (Point A in Fig. 95.2). To speed up the processing, a parallel ALU is usually used so that all bits of a word can be operated on simultaneously. This is how a conventional word-sequential computer executes on its operands (Point B in Fig. 95.2). In a parallel processor, it may execute either (a) all the ith bits of N operands or operand pairs (i.e., bit slice or bis) or (b) all N M-bit operands or operand pairs simultaneously (Points C and D in Fig. 95.2, respectively). Figure 95.2 also shows some of the systems in this classification. It is seen from this classification that the performance of a computer is proportional to the total number of bits or bit pairs it can execute simultaneously.

Feng's classification [Hwang and Briggs, 1984] was originally intended for parallel processors, and as a result, the number of CUs in a computer system was not specified. Händler extended Feng's classification by adding a third dimension, namely, the number of CUs. Pipelined systems are also included in this classification through additional parameters [Händler, 1977].

95.3 Types of Parallel Processors

Ensemble Processors

An ensemble system is an extension of the conventional uniprocessor systems. It is a collection of N PEs (a PE here consists of an ALU, a set of local registers, and limited local control capability) and N MUs, under the control of a single CU. Thus, the organization of an **ensemble processor** is similar to that shown in Fig. 95.1 except that there are no direct interprocessor and processor-memory communications, i.e., no interconnection networks. When the need for communication arises, it is done through the CU. This slows down the system for applications requiring extensive interprocessor and processor-memory communications. For example, the sum of two matrices A and B can be executed in one step, if R^2 PEs are available in an ensemble processor, where R is the rank of the matrices. On the other hand, the product of the same two matrices requires extensive data alignment between the elements of A and B. As a result, it is ineffective for performing matrix multiplications with an ensemble processor. Therefore, while the ensemble processors are capable of executing up to N identical jobs simultaneously, they have very limited applications. Parallel element processing ensemble (PEPE) [Evensen and Troy, 1973] is an example of such parallel processors.

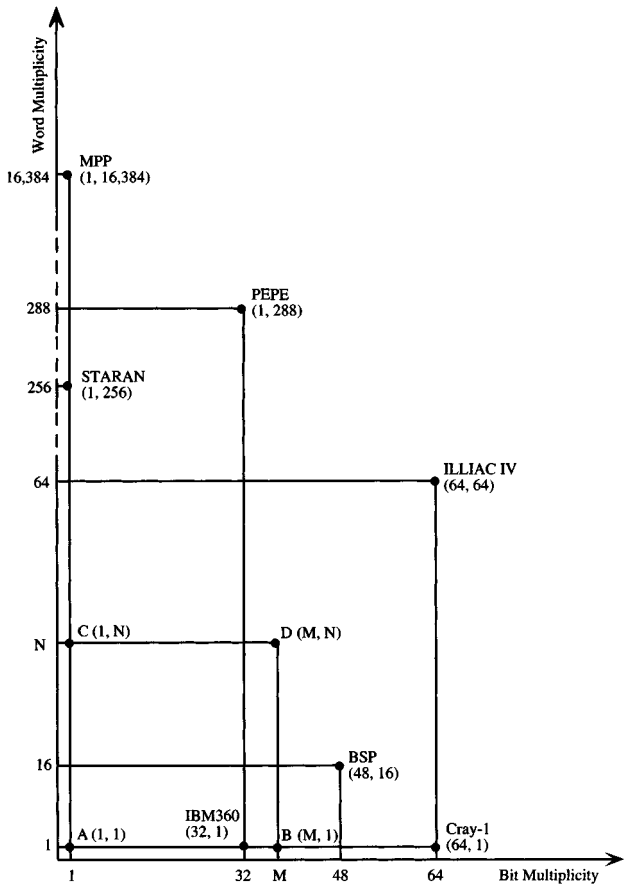


FIGURE 95.2 Feng's classification.

Array Processors

Because of the need for interprocessor and processor-memory communication for most applications, a parallel processor usually has one or more circuits (known as interconnection networks) to support various applications for efficient processing. In general, an **array processor** may consist of N identical PEs under the control of a single CU and a number of MUs. Within each PE there are circuits for network interface as well as its own local memories. The PEs and MUs communicate with each other through an interconnection network. A typical array processor organization is shown in Fig. 95.3. Depending on the design, each PE may perform serial-by-bit (as in MPP) or parallel-by-bit (as in ILLIAC IV) operations.

As can be seen from Fig. 95.3, the interconnection networks play a very important role in parallel processors. The network usually provides a uniform interconnection among PEs on one hand and PEs and MUs on the other. Different array processor organizations might use different interconnection networks. In general, the interconnection networks can be classified into two categories: static and dynamic, as shown in Fig. 95.4.

ILLIAC IV [Barnes et al., 1968] and MPP [Batcher, 1979] are examples of parallel processors using static interconnections, while STARAN [Batcher, 1973] and BSP [Kuck and Stokes, 1982] are examples using dynamic interconnections.

The CU usually has its own high-speed registers, local memory, and arithmetic unit. Thus, in many cases, it is a conventional computer and the instructions are stored in a main memory, together with data. However, in some machines such as ILLIAC IV, programs are distributed among the local memories of the PEs. Hence, the instructions are fetched from the processors' local memories into an instruction buffer in the CU. Each

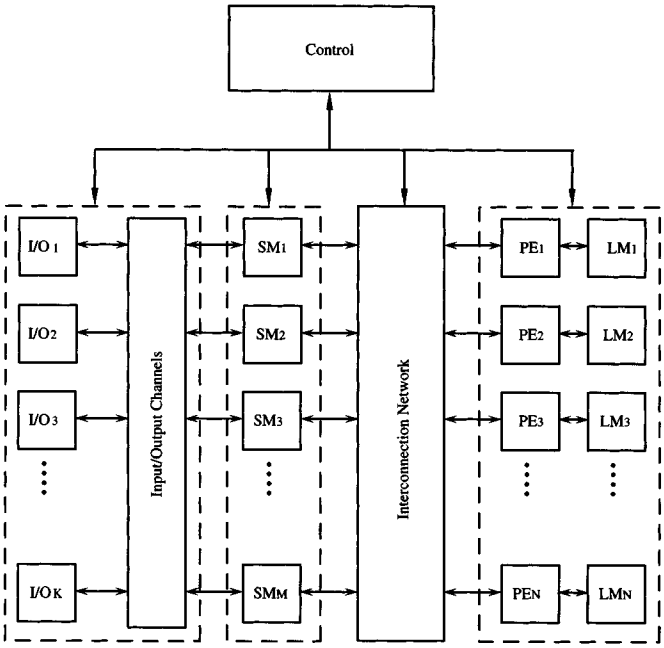


FIGURE 95.3 An array processor organization. I/O, input/output devices; LM, local memory; PE, processing element; SM, shared memory.

instruction is either a local type instruction, where it is executed entirely within the CU, or it is a parallel instruction and is executed in the processing array. The primary function of the CU is to examine each instruction as it is to be executed and to determine where the execution should take place.

Associative Processor

Associative memories, also known as content-addressable memories, retrieve information on the basis of data content rather than addresses. An associative memory performs comparison (i.e., exclusive-OR or equivalence) operations at its bit level. The results of the comparison on a group of bits in a word for all words in the memory are transmitted to a register called a response register or flag. In addition, there are circuits such as multiple match resolver, enable/disable register, a number of temporary registers, as well as appropriate logic gates for resolving multiple responses and information retrieval. For **associative processors,** arithmetic capabilities are added to this unit. The unit can be viewed as consisting of a number of bit-serial PEs. Furthermore, the bit-level logic is moved out of the memory so that the memory part of the processor consists of a number of random-access memories called word modules. A typical associative processor is shown in Fig. 95.5. STARAN and MPP (Fig. 95.2) are representative of this bit-serial, word-parallel SIMD organization. In Fig. 95.5 the common register is where the common operand is stored and the mask register defines the bit positions requiring operation. The enable/disable register provides local control of individual PEs. Because of its simplicity in design the per-PE cost of an associative processor is much lower, but the bit-serial operations slow down the system drastically. To compensate for this, these systems are useful only for applications requiring a large number of PEs.

95.4 System Utilization

As discussed previously, for any computer there is a maximum number of bits or bit pairs that can be processed concurrently, whether it is under single-instruction or multiple-instruction control [Feng, 1972, 1973]. This maximum degree of concurrency, or maximum concurrency (C_m), is an indication of the computer-processing capability. The actual utilization of this capability is indicated by the average concurrency defined to be

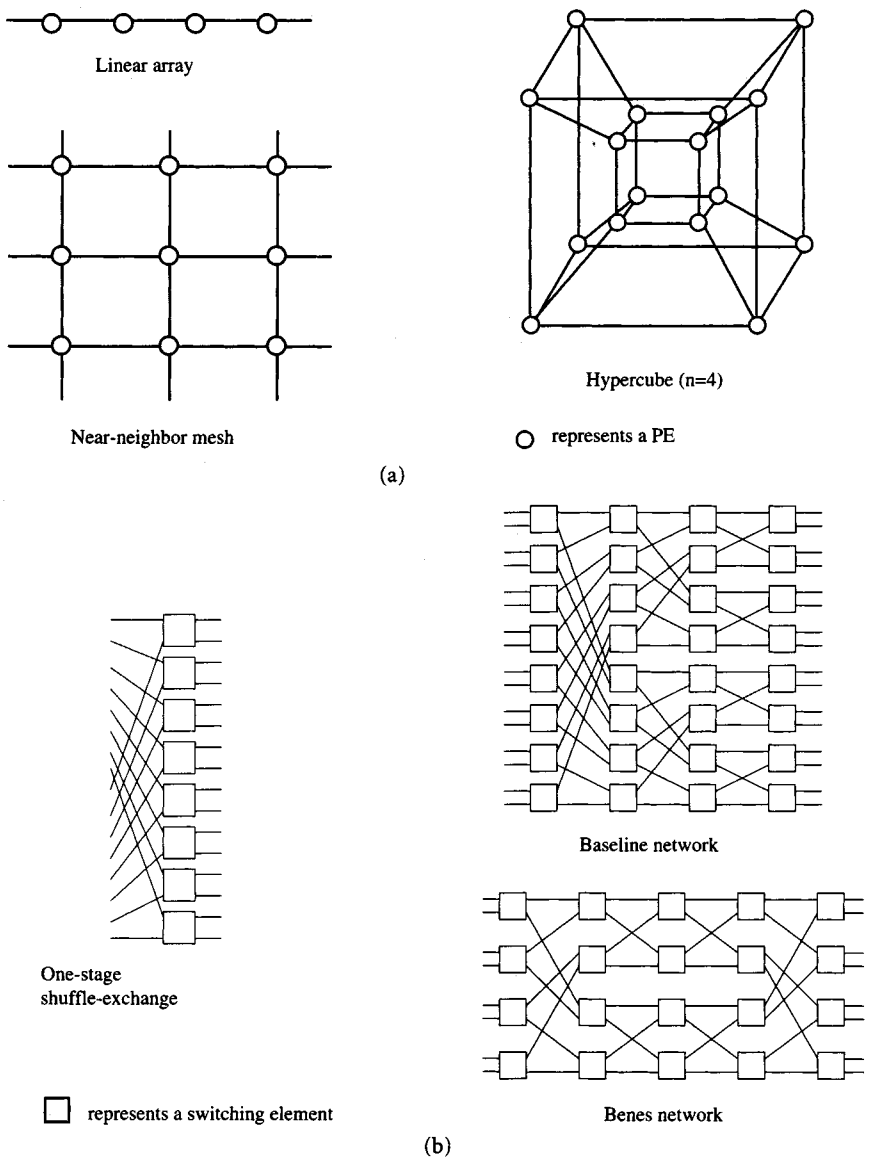


FIGURE 95.4 Some examples of static (a) and dynamic (b) interconnection networks.

$$C_a = \frac{\Sigma c_i \Delta t_i}{\Sigma \Delta t_i}$$

where c_i is the concurrency at Δt_i. If Δt_i is set to one time unit, then the average concurrency over a period of T time units is

$$C_a = \frac{\sum_{i=1}^{T} c_i}{T}$$

The average hardware utilization is then

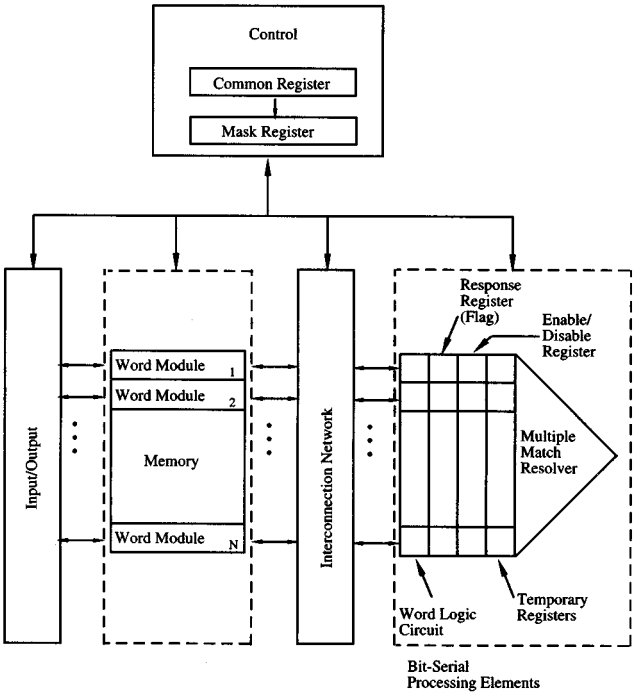


FIGURE 95.5 An associative processor organization.

$$\mu = \frac{C_a}{C_m} = \frac{\sum_{i=1}^{T} c_i}{TC_m} = \frac{1}{T}\sum_{i=1}^{T} \sigma_i$$

where σ_i is the hardware utilization at time *i*. Whereas C_m is determined by the hardware design, C_a or μ is highly dependent on the software and applications. A general-purpose computer should achieve a high μ for as many applications as possible, whereas a special-purpose computer would yield a high μ for at least the intended applications. In either case, maximizing the value of μ for a computer design is important. This equation can also be used to evaluate the relative effectiveness of machine designs.

For a parallel processor, the degree of concurrency is called the degree of parallelism. A similar discussion can be used to define the average hardware utilization of a parallel processor. The maximum parallelism is then P_m, and the average parallelism is

$$P_a = \frac{\Sigma p_i \Delta t_i}{\Sigma \Delta t_i}$$

or

$$P_a = \frac{\sum_{i=1}^{T} p_i}{T}$$

for T time units. The average hardware utilization becomes

$$\upsilon = \frac{P_a}{P_m} = \frac{\sum_{i=1}^{T} p_i}{TP_m} = \frac{1}{T}\sum_{i=1}^{T} \rho_i$$

where ρ_i is the hardware utilization for parallel processors at time i. With appropriate instrumentation, the average hardware utilization of a system can be determined.

In practice, however, it is not always true that every bit or bit pair that is being processed would be productive. Some of the bits produce only repetitious (superfluous) or even meaningless results. This happens more often and more severely in a parallel processor than in a word-sequential processor. Consider, for example, performing a maximum search operation in a mesh-connected parallel processor (such as ILLIAC IV). For N operands, it takes $(N/2)\log_2 N$ comparisons ($N/2$ comparisons for each of $\log_2 N$ iterations) instead of the usual $N-1$ comparisons in word-sequential machines. Thus, in effect there are

$$\left(\frac{N}{2}\log_2 N\right) - (N-1) = \frac{N}{2}(\log_2 N - 2) + 1$$

comparisons that are nonproductive. If we let $\hat{P}_a$ be the effective parallelism over a period of T time units and $\hat{\upsilon}$, $\hat{p}_i$, and $\hat{\rho}_i$ be the corresponding effective values, the effective hardware utilization is then

$$\hat{\upsilon} = \frac{\hat{P}_a}{P_m} = \frac{\sum_{i=1}^{T} \hat{p}_i}{TP_m} = \frac{1}{T}\sum_{i=1}^{T} \hat{\rho}_i$$

A successful parallel processor design should yield a high $\hat{\upsilon}$, as well as the required throughput for, at least, the intended applications. This not only involves a proper hardware and software design but also the development of efficient parallel algorithms for these applications.

Suppose T_u is the execution time of an application program using a conventional word-sequential machine, and T_c is the execution time of the same program using a concurrent system; the speed-up ratio is then defined as

$$S = \frac{T_u}{T_c}$$

Naturally, for a specific parallel organization, the speed-up ratio determines how well an application program can utilize the hardware resources. Supporting software has a direct effect on the speed-up ratio.

Defining Terms

Array processor: A parallel processor consisting of a number of processing elements, memory modules, and input/output devices as well as interconnection networks under a single control unit.

Associative processor: A parallel processor consisting of a number of processing elements, memory modules, and input/output devices under a single control unit. The capability of the processing elements is usually limited to the bit-serial operations.

Ensemble processor: A parallel processor consisting of a number of processing elements, memory modules, and input/output devices under a single control unit. It has no interconnection network to provide interprocessor or processor-memory communications.

Interconnection network: A network of interconnections providing interprocessor and processor-memory communications. It may be static or dynamic, distributed, or centralized.

Parallel processor: A computing system consisting of a number of processors, memory modules, input/output devices, and other components under the control of a single control unit. It is known to be a single-instruction-stream, multiple-data-stream (SIMD) machine.

Processing element: A basic processor consisting of an arithmetic-logic unit, a number of registers, network interfaces, and some local control facilities.

Related Topics

18.1 Special Architectures • 96.5 Parallel Processing

References

G.H. Barnes, R.M. Brown, M. Kato, D.J. Kuck, D.L. Slotnick, and R.A. Stokes, "The ILLIAC IV computer," *IEEE Trans. Comput.,* vol. C-7, pp. 746–757, 1968.

K.E. Batcher, "STARAN/RADCAP hardware architecture," *Proc. Sagamore Computer Conf. on Parallel Processing,* pp. 147–152, 1973.

K.E. Batcher, "MPP—A massively parallel processor," *Proc. Int. Conf. on Parallel Processing,* p. 249, 1979.

A.J. Evensen and J.L. Troy, "Introduction to the architecture of a 288-element PEPE," *Proc. Sagamore Computer Conference*, pp. 162–169, 1973.

T. Feng, "An overview of parallel processing systems," *1972 WESCON Tech. Papers,* Session 1 — "Parallel Processing Systems," pp. 1–2, 1972.

T. Feng, Parallel Processing Characteristics & Implementation of Data Manipulating Functions, Technical Report RADC-TR-73-189, July 1973.

M.J. Flynn, "Very high speed computing systems," *Proc. IEEE,* vol. 54(12), pp. 1901–1909, 1966.

W. Händler, "The impact of classification schemes on computer architecture," *Proc. Int. Conf. on Parallel Processing*, pp. 7–15, 1977.

K. Hwang and F.A. Briggs, *Computer Architecture and Parallel Processing*, New York: McGraw-Hill, 1984.

D. J. Kuck and R.A. Stokes, "The Borroughs Scientific Processor (BSP)," *IEEE Trans. Comput.,* vol. C-31(5), pp. 363–376, 1982.

Further Information

Proceedings of International Conference on Parallel Processing: An annual conference held since 1972. Recent proceedings published by CRCPress.

IEEE Transactions on Parallel and Distributed Systems: Started in 1990 as a quarterly, now a monthly, published by the IEEE Computer Society.

Journal of Parallel and Distributed Computing: A monthly published by Academic Press.

Computer Architecture and Parallel Processing: A book by K. Hwang and F. A. Briggs published by McGraw-Hill.

Boykin, J. "Operating Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

96

Operating Systems

96.1 Introduction
96.2 Types of Operating Systems
96.3 Distributed Computing Systems
96.4 Fault-Tolerant Systems
96.5 Parallel Processing
96.6 Real-Time Systems
96.7 Operating System Structure
96.8 Industry Standards
96.9 Conclusions

Joseph Boykin
Clarion Advanced Storage

96.1 Introduction

An operating system is just another program running on a computer. It is unlike any other program, however. An operating system's primary function is the management of all hardware and software resources. It manages processors, memory, I/O devices, and networks. It enforces policies such as *protection* of one program from another and *fairness* to ensure that users have equal access to system resources. It is privileged in that it is the only program that can perform specialized hardware operations. The operating system is the primary program upon which all other programs rely.

To understand modern operating systems we must begin with some history [Boykin and LoVerso, 1990]. The modern digital computer is only about 40 years old. The first machines were giant monoliths housed in special rooms, and access to them was carefully controlled. To program one of these systems the user scheduled access time well in advance, for in those days the user had sole access to the machine. The program such a user ran was the *only* program running on the machine.

It did not take long to recognize the need for better control over computer resources. This began in the mid-1950s with the dawn of batch processing and early operating systems that did little more than load programs and manage I/O devices.

In the 1960s we saw more general-purpose systems. New operating systems that provided time-sharing and **real-time computing** were developed. This was the time when the foundation for all modern operating systems was laid.

Today's operating systems are sophisticated pieces of software. They may contain millions of lines of code and provide such services as distributed file access, security, **fault tolerance**, and **real-time** scheduling. In this chapter we examine many of these features of modern operating systems and their use to the practicing engineer.

96.2 Types of Operating Systems

Different operating systems (OS) provide a wide range of functionality. Some are designed as single-user systems and some for multiple users. The operating system, with appropriate hardware support, can protect one executing program from malicious or inadvertent attempts of another to modify or examine its memory. When connected to a storage device such as a disk drive, the OS implements a **file system** to permit storage of files.

The file system often includes security features to protect against file access by unauthorized users. The system may be connected to other computers via a *network* and thus provide access to remote system resources.

Operating systems are often categorized by the major functionality they provide. This functionality includes **distributed computing,** *fault tolerance,* **parallel processing,** *real-time,* and *security.* While no operating system incorporates all of these capabilities, many have characteristics from each category.

An operating system does not need to contain every modern feature to be useful. For example, MS-DOS[1] is a single-user system with few of the features now common in other systems. Indeed, this system is little more than a program loader reminiscent of operating systems from the early 1960s. Unlike those vintage systems, there are numerous applications that run under MS-DOS. It is the abundance of programs that solve problems from word processing to spreadsheets to graphics that has made MS-DOS popular. The simplicity of these systems is exactly what makes them popular for the average person.

Systems capable of supporting multiple users are termed *time-sharing* systems; the system is shared among all users, with each user having the view that he or she has all system resources available. Multiuser operating systems provide protection for both the file system and the contents of main memory. The operating system must also mediate access to peripheral devices. For example, only one user may have access to a tape drive at a time.

Fault-tolerant systems rely on both hardware and software to ensure that the failure of any single hardware component, or even multiple components, does not cause the system to cease operation. To build such a system requires that each critical hardware component be replicated at least once. The operating system must be able to dynamically determine which resources are available and, if a resource fails, move a running program to an operational unit.

Security has become more important during recent years. Theft of data and unauthorized access to data are prevented in secure systems. Within the United States, levels of security are defined by a government-produced document known as the *Orange Book.* This document defines seven levels of security, denoted from lowest to highest as *D, C1, C2, B1, B2, B3,* and *A1.* Many operating systems provide no security and are labeled D. Most time-sharing systems are secure enough that they could be classified at the C1 level. The C2 and B1 levels are similar, and this is where most secure operating systems are currently classified. During the 1990s B2 and B3 systems will become readily available from vendors. The A1 level is extremely difficult to achieve, although several such systems are being worked on.

In the next several sections we expand upon the topics of distributed computing, fault-tolerant systems, parallel processing, and real-time systems.

96.3 Distributed Computing Systems

The ability to connect multiple computers through a communications network has existed for many years. Initially, computer-to-computer communication consisted of a small number of systems performing bulk file transfers. The 1980s brought the invention of high-speed *local area networks,* or LANs. A LAN allows hundreds of machines to be connected together. New capabilities began to emerge, such as *virtual terminals* that allowed a user to log on to a computer without being physically connected to that system. Networks were used to provide remote access to printers, disks, and other peripherals. The drawback to these systems was the software; it was not sophisticated enough to provide a totally integrated environment. Only small, well-defined interactions among machines were permitted.

Distributed systems provide the view that *all* resources from every computer on the network are available to the user. What's more, access to resources on a remote computer is viewed in the same way as access to resources on the local computer. For example, a file system that implements a directory hierarchy, such as UNIX,[2] may have some directories on a local disk while one or more directories are on a remote system. Figure 96.1 illustrates how much of the directory hierarchy would be on the local system, while user directories (shaded directories) could be on a remote system.

[1]MS-DOS is a trademark of Microsoft, Inc.

[2]UNIX is a trademark of UNIX Software Laboratories (USL).

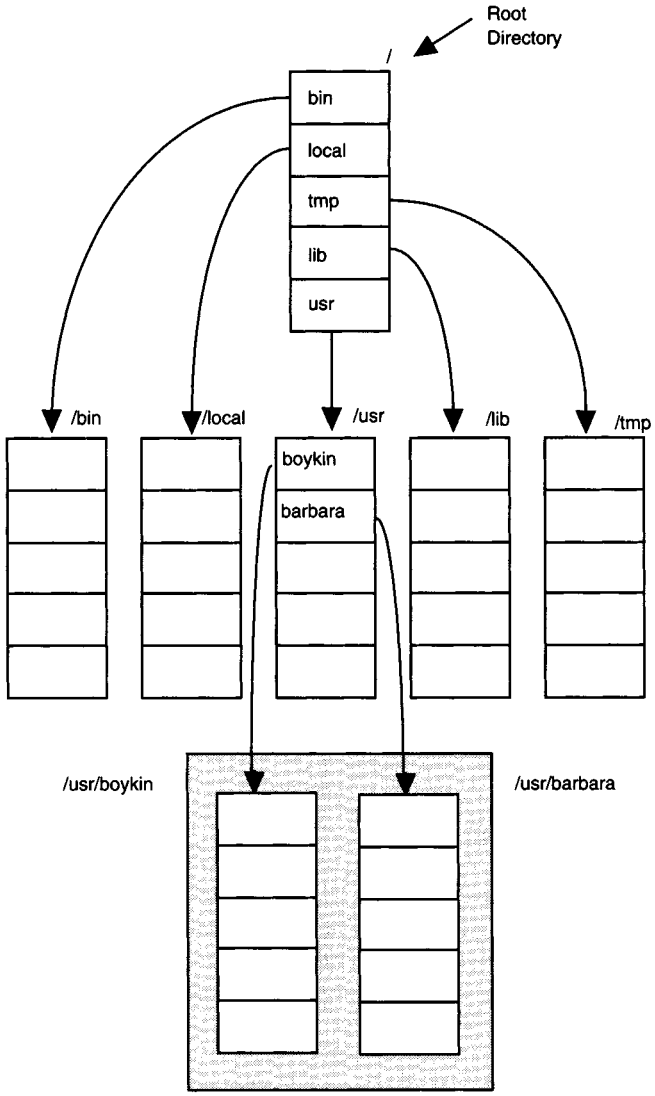


FIGURE 96.1 UNIX file system hierarchy in a distributed environment.

There are many advantages of distributed systems. Advantages over centralized systems include [Tanenbaum, 1992]:

- *Economics:* Microprocessors offer a better price/performance than mainframes.
- *Speed:* A distributed system may have more total computing power than a mainframe.
- *Reliability:* If one machine crashes, the system as a whole can still survive.
- *Incremental growth:* Computing power can be added in small increments.

Advantages over nonnetworked personal computers include [Tanenbaum, 1992]:

- *Data sharing:* Allow many users access to a common database.
- *Device sharing:* Allow many users to share expensive peripherals like color printers.
- *Communication:* Make human-to-human communication easier, for example, by electronic mail.
- *Flexibility:* Spread the workload over the available machines in the most cost effective way.

While there are many advantages to distributed systems, there are also several disadvantages. The primary difficulty is that software for implementing distributed systems is large and complex. Small personal computers could not effectively run modern distributed applications. Software development tools for this environment are not well advanced. Thus, application developers are having a difficult time working in this environment.

An additional problem is network speed. Most office networks are currently based on IEEE standard 802.3 [IEEE, 1985], commonly (although erroneously) called Ethernet, which operates at 10 Mb/s (ten million bits per second). With this limited bandwidth, it is easy to saturate the network. While higher-speed networks such as FDDI[1] and ATM[2] networks do exist, they are not yet in common use. While distributed computing has many advantages, we must also understand that without appropriate safeguards, our data may not be secure.

Security is a difficult problem in a distributed environment. Whom do you trust when there are potentially thousands of users with access to your local system? A network is subject to security attack by a number of mechanisms. It is possible to monitor all packets going across the network; hence, unencrypted data are easily obtained by an unauthorized user. A malicious user may cause a *denial-of-service* attack by flooding the network with packets, making all systems inaccessible to legitimate users.

Finally, we must deal with the problem of scale. To connect a few dozen or even a few hundred computers together may not cause a problem with current software. However, global networks of computers are now being installed. Scaling our current software to work with tens of thousands of computers running across large geographic boundaries with many different types of networks is a challenge that has not yet been met.

96.4 Fault-Tolerant Systems

Most computers simply stop running when they break. We take this as a given. There are many environments, however, where it is not acceptable for the computer to stop working. The space shuttle is a good example. There are other environments where you would simply prefer if the system continued to operate. A business using a computer for order entry can continue to operate if the computer breaks, but the cost and inconvenience may be high. Fault-tolerant systems are composed of specially designed hardware and software that are capable of continuous operation.

To build a fault-tolerant system requires both hardware and software modifications. Let's take a look at an example of a small problem that illustrates the type of changes that must be made. Remember, the goal of such a system is to achieve continuous operation. That means we can never purposely shut the computer off. How then do we repair the system if we cannot shut it off? First, the hardware must be capable of having circuit boards plugged and unplugged while the system is running; this is not possible on most computers. Second, removing a board must be detected by the hardware and reported to the operating system. The operating system, the manager of resources, must then discontinue use of that resource.

Each component of the computer system, both hardware and software, must be specially built to handle failures. It should also be obvious that a fault-tolerant system must have redundant hardware. If, for example, a disk controller should fail, there must be another controller communicating with the disks that can take over.

One problem with implementing a fault-tolerant system is knowing when something has failed. If a circuit board totally ceases operation, we can determine the failure by its lack of response to commands. Another failure mode exists where the failing component appears to work but is operating incorrectly. A common approach to detect this problem is a *voting* mechanism. By implementing three hardware replicas the system can detect when any one has failed by its producing output inconsistent with the other two. In that case, the output of the two components in agreement is used.

The operating system must be capable of restarting a program from a known point when a component on which the program was running has failed. The system can use *checkpoints* for this purpose. When an application program reaches a known state, such as when it completes a transaction, it stores the current state of the

[1]Fiber distributed data interface. The FDDI standard specifies an optical fiber ring with a data rate of 100 Mb/s.

[2]Asynchronous transfer mode. A packet-oriented transfer mode moving data in fixed-size packets called *cells*. There is no fixed speed for ATM. Typical speed is currently 155 Mb/s, although there are implementations running at 2 Gb/s.

program and all I/O operations; this is known as a checkpoint. Should a component on which this program is running fail, the operating system can restart the program from the most recent checkpoint.

While the advantage of fault-tolerant systems is obvious, they come at a price. Redundant hardware is expensive, and software capable of recovering from faults runs more slowly. As with many other systems, the price may be more than offset by the advantage of continuous computing.

96.5 Parallel Processing

No matter how fast computers become, it seems they are never fast enough. Manufacturers make faster computers by decreasing the amount of time it takes to do each operation. An alternative is to build a computer that performs several operations simultaneously. A parallel computer, also called a multiprocessor, is one that contains more than one CPU.[1]

The advantage of a parallel computer is that it can run more than one program simultaneously. In a general-purpose time-sharing environment, parallel computers can greatly enhance overall system throughput. A program shares a CPU with fewer programs. This approach is similar to having several computers connected on a network but has the advantage that all resources are more easily shared.

To take full advantage of a parallel computer will require changes to the operating system [Boykin and Langerman, 1990] and application programs. Most programs are easily divided into pieces that can each run at the same time. If each of these pieces is a separate thread of control, they could run simultaneously on a parallel computer. By so dividing the application, the program may run in less time than it would on a single-processor (uniprocessor) computer.

Within the application program, each thread runs as if it were the only thread of control. It may call functions, manipulate memory, perform I/O operations, etc. If the threads do not interact with each other, then, to the application programmer, there is little change other than determining how to subdivide the program. However, it would be unusual for these threads not to interact. It is this interaction that makes parallel programming more complex.

In principle, the solution is rather simple. Whenever a thread will manipulate memory or perform an I/O operation, it must ensure that it is the *only* thread that will modify that memory location or do I/O to that file until it has completed the operation. To do so, the programmer uses a *lock*. A lock is a mechanism that allows only a single thread to execute a given code segment at a time. Consider an application with several threads of control. Each thread performs an action and writes the result to a file—the same file. Within each thread we might have code that looks as follows:

```
thread()
{
        dowork();
        writeresult();
}
writeresult()
{
        lock();
        write(logfid, result, 512);
        unlock();
}
```

In this example the **writeresult** function calls the **lock** function before it writes the result and calls **unlock** afterward. Other threads simultaneously calling **writeresult** will wait at the call to **lock** until the thread that currently holds the lock calls the **unlock** function.

While this approach is simple in principle, in practice it is more difficult. It takes experience to determine how a program may be divided. Even with appropriate experience, it is more difficult to debug a multithreaded

[1]Central processing unit, the hardware component that does all arithmetic and logical operations.

application. With several threads of control operating simultaneously, it is not simply a matter of stepping through the program line by line to find a mistake. Most often, it is the interaction between threads that is the problem.

Multithreading a program may not be a trivial matter. As with most types of programming, however, experience makes the process easier. The benefit is significantly enhanced performance.

96.6 Real-Time Systems

Real-time systems are those that guarantee that the system will respond in a predetermined amount of time. We use real-time systems when, for example, computers control an assembly line or run a flight simulator. In such an environment we define an action that must occur and a deadline by which we wish that action to take place. On an assembly line an event may occur, such as a part arriving at a station, and an action, such as painting that part. The deadline we impose will be based on the speed of the assembly line. Obviously, we must paint the part before it passes to the next station. This is called a *hard* real-time system because the system must meet a strict deadline.

Another class of system is termed *soft* real-time. These are environments in which response time is important, but the consequences are not as serious as, for example, on an assembly line. Airline reservation systems are in this category. Rapid response time to an event, such as an agent attempting to book a ticket, is important and must be considered when the system performs other activities.

One way of distinguishing hard and soft real-time systems is by examining the *value* of a response over time. For example, if a computer was controlling a nuclear reactor and the reactor began to overheat, the command to open the cooling valves has extremely high value until a deadline, when the reactor explodes. After that deadline, there is no value in opening the valves (see Fig. 96.2).

Relatively few events require that type of responsiveness. Most events have a deadline, but there continues to be value in responding to that event even past the deadline. In our airline reservation example, the airline may wish to respond to a customer request within, say, 10 seconds. However, if the response comes in 11 seconds, there is still value in the response. The value is lessened because the customer has become upset. As time increases, the customer becomes more and more upset and the value of responding decreases. We illustrate this in Fig. 96.3.

96.7 Operating System Structure

Operating systems are large, complex pieces of software. They must handle asynchronous events such as interrupts from I/O devices, control hardware memory management units (MMUs) to implement virtual memory, support multiple simultaneous users, implement complex network protocols, and much more. As with any software of this magnitude, an operating system is logically divided into smaller pieces. The structure of a typical modern operating system is depicted in Fig. 96.4.

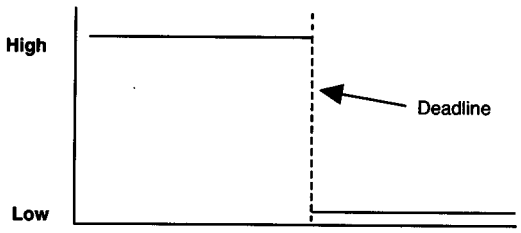


FIGURE 96.2 Relative value of a response over time in a critical situation.

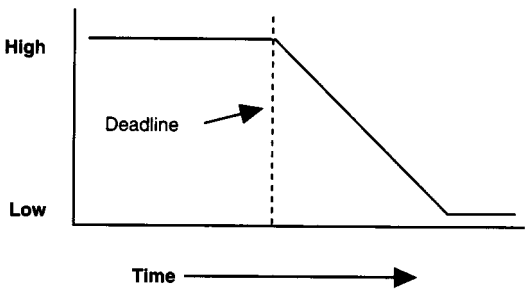


FIGURE 96.3 Relative value of a response over time in a noncritical situation.

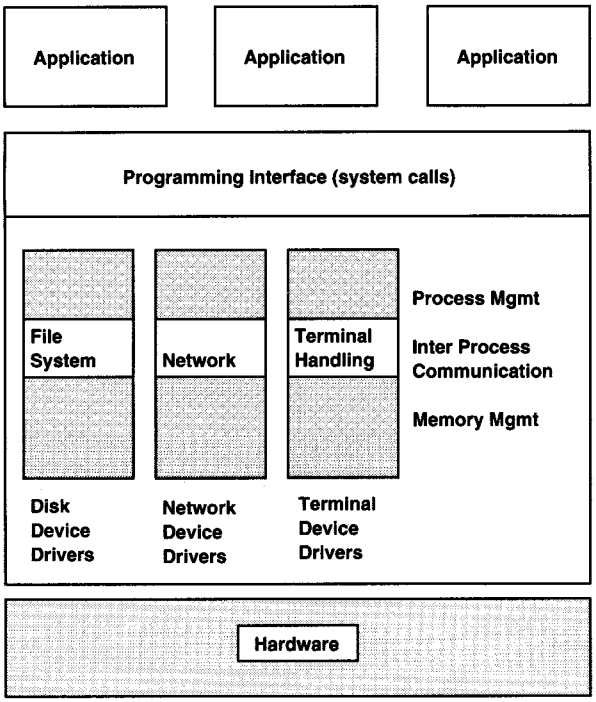


FIGURE 96.4 The structure of a modern operating system.

From the user's standpoint, the operating system is a collection of system calls—the *programmers' interface.* Sometimes this is termed an *application program interface,* or API. System calls provide the mechanism for an application program to obtain services from the system. System calls exist to perform file operations such as *create, open, close, read,* and *write.* For terminals, system calls would perform such functions as changing the baud rate and number of parity bits. Network connections may be established or network protocol options, such as the size of network buffers, are also controlled through system calls.

While every operating system provides a system call interface, there is little uniformity to the appearance of that interface. Some systems provide an interface that appears as a simple function call. For example, to open a file under the UNIX operating system, we use the following system call:

```
open("/home/boykin/crc-press/oschapter", O_RDONLY);
```

Other operating systems require a user to fill in complex data structures for various operations. For example, the following code fragment illustrates how to send an IPC message using the Mach operating system's **interprocess communication** (IPC) facility [Boykin et al., 1993]:

```
msg_header_t header;

header.msg_simple = TRUE;
header.msg_size = sizeof(header);
header.msg_type = MSG_TYPE_NORMAL;
header.msg_local_port = PORT_NULL;
header.msg_remote_port = remote_port;
header.msg_id = 100;
```

Regardless of the interface format, a programmer should become familiar with the parameters, options, and return codes from each system call to use the system proficiently.

Beneath the programming interface lies the heart of the operating system. We can divide the system into two major sections. The first section directly implements the system calls. This includes the file system, terminal

handling, etc. The second section provides basic capabilities upon which the rest of the system is built. Interprocess communication, memory management, and **process** management are all examples of these basic capabilities. A brief explanation of each of these sections will be given shortly.

The lowest level of the operating system interfaces directly with the computer hardware. For each physical device, such as a disk, tape, or serial line, a device driver must exist to communicate with the hardware. Device drivers accept requests to read or write data or determine the status of the device. They may do polled I/O or be interrupt driven, although polled I/O is usually only done on small personal computers. Writing a device driver requires a thorough knowledge of the hardware as well as the interface to the operating system.

In addition to I/O devices, the system must also manipulate such hardware as counters, timers and memory management units. Timers are used to satisfy user requests such as terminating an operation after a specified length of time. MMUs provide the ability to protect memory. Each time a program is run, the operating system programs the MMU with the physical memory addresses the program may access. Any attempt to access other memory is not allowed by the MMU.

An MMU is also required to implement *virtual memory*. Virtual memory allows a program to use more memory than is physically present on the machine. The operating system implements virtual memory by using an external device, typically a disk, to store portions of the program that are not currently in use. When a program attempts to access memory temporarily stored on disk, the MMU traps[1] to the operating system, which reads the memory from disk and restarts the program.

In recent years the structure depicted here has been changing. A new concept, called the *microkernel*, has begun to emerge. The idea behind a micro-kernel is to dramatically reduce the size of the operating system by placing most OS subsystems in the application layer. A micro-kernel would not be a usable system by itself. A number of programs would be run on top of the micro-kernel to provide such services as a file system and network protocols.

In the micro-kernel architecture shown in Fig. 96.5, notice that subsystems traditionally within the operating system are now at the same level as an application program. An application program wishing to, for example, open a file makes its request to the file system program, rather than the micro-kernel. The file system may call upon other OS subsystems or on the micro-kernel to perform an operation.

From the user standpoint, there is no programming difference between a micro-kernel structure and the traditional structure. There are two advantages of the micro-kernel approach. The first is that programming

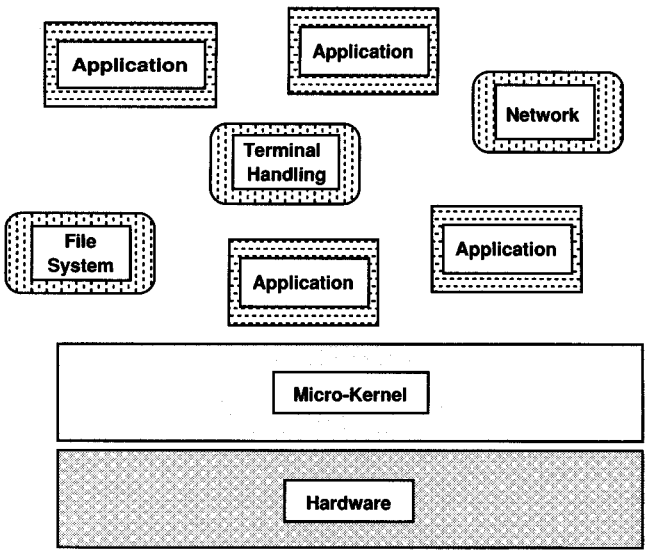


FIGURE 96.5 Micro-kernel structure.

[1]A *trap* is a hardware signal that is received by the operating system. It is very similar to an interrupt from an I/O device

and debugging at the application layer is inherently simpler than programming at the OS layer. The benefit here is to the OS designer and implementors who can now write and debug OS code faster and easier than before. This benefits the user by having an operating system that is more reliable.

The second advantage stems from the ability to incorporate several different OS environments on top of the same micro-kernel. In this way, the computer acts as though it is running several operating systems. For example, if both MS-DOS and UNIX coexisted on the same micro-kernel, the user could choose to run an MS-DOS spreadsheet or word processor and communicate using UNIX network commands. The user has gained increased flexibility.

96.8 Industry Standards

As computer technologies come into widespread use, users begin to desire standardization. Standardization allows a user to know that a program written to a standard will work without concern for which vendor supplies the programming environment. Operating systems are no exception to this general rule, and there are several standards, both industry standards and *de facto* standards, that apply. Porting software from one system to another, often an expensive proposition, becomes a trivial task.

Perhaps the most notable OS standard is POSIX, standard number 1003 [IEEE, 1990], sponsored by the IEEE Computer Society's Technical Committee on Operating Systems. POSIX is a family of standards based on the UNIX operating system that includes the system call interface, user-level commands, real-time extensions, and networking extensions. The POSIX system call interface, 1003.1, was adopted by the U.S. government verbatim as a Federal Information Processing Standard, FIPS 151. Many vendors conform to POSIX; thus, a program that conforms to this standard can be ported to many system platforms without change.

An example of a *de facto* standard is the X/Open Portability Guide (XPG) [X/Open, 1989]. X/Open is not a standards-setting body but is a joint initiative by members of the business community to adopt and adapt existing standards into a consistent environment. The X/Open system interface and headers are based on POSIX 1003.1 but also include extensions to POSIX-defined interfaces as well as additional interfaces.

The importance of such standards is evidenced by the strong support of such organizations as the Open Software Foundation. OSF's OSF/1 operating system conforms to various POSIX standards. Where not superseded by POSIX, it also conforms to XPG and AT&T's System V Interface Definition (SVID) [AT&T, 1985]. Conforming to these standards is considered critical for the success of OSF/1.

Some might consider an operating system such as MS-DOS to be a *de facto* standard. While MS-DOS is in common use, however, it is proprietary software subject to change without notice. Defining a standard implies an open system on which vendors and users agree.

96.9 Conclusions

I have been hearing for the past 15 years about the demise of the operating system. It has been said over and over that the role of the OS will go away. So far, the only change has been to *expand* on the role the operating system plays. One must remember that the operating system is not the user interface it portrays or the applications that run on it. It is, as it always has been, the manager of all resources on a computer system.

While the interface to computers has changed and the use to which we apply computer technology has changed, there will always be the need for an operating system. Without question, the OS will change as well. We have already seen micro-kernel architectures begin to emerge from the research labs into commercial operating systems. Distributed computing will become more widespread and force additional changes to the operating system. Regardless of the changes that come, it will always be the operating system on which all other programs rely.

Defining Terms

Distributed computing: An environment in which multiple computers are networked together and the resources from more than one computer are available to a user. Those resources are accessed in a manner identical to accessing resources on a local computer system.

Fault-tolerant systems: A computer system with both hardware and software that are capable of continuous operation even in the event hardware components fail.

File system: The logical organization of files on a storage device, typically a disk drive. The file system may support a hierarchical structure with directories and subdirectories (sometimes called folders).

Interprocess communication: The transfer of information between two cooperating programs. Communication may take the form of a *signal* (the arrival of an event) or the transfer of data.

Parallel processing: A parallel computer is one that contains more than one CPU. Parallel processing is when a program is divided into multiple threads of control, each of which is capable of running simultaneously. On a parallel computer, multiple threads could be running at the same time, thus resulting in better performance than on a uniprocessor system.

Process: A single executable program. A process is the context in which an operating system places a running program. It contains the program itself as well as allocated memory, open files, network connections, etc.

Real-time computing: Support for environments in which response time to an event must occur within a predetermined amount of time. Real-time systems may be categorized into *hard* and *soft* real-time.

Related Topics

90.3 Programming Methodology • 95.2 Classifications

References

AT&T, *System V Interface Definition,* Spring 1985, Issue 1, AT&T Customer Information Center, Indianapolis, Indiana.

J. Boykin, D. Kirschen, A. Langerman, and S. LoVerso, *Programming Under Mach,* Reading, Mass.: Addison-Wesley, 1993.

J. Boykin and A. Langerman, "Mach/4.3BSD: Parallelization without reimplementation," *Computing Systems Journal,* vol. 3, no. 1, 1990.

J. Boykin and S. LoVerso, "Recent developments in operating systems," *Computer,* vol. 23, no. 5, 1990.

H.M. Dietel, *Operating Systems,* 2nd ed., Reading, Mass.: Addison-Wesley, 1990.

IEEE, *Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, American National Standard ANSI/IEEE Std. 802.3, 1985.

IEEE, *Information Technology—Portable Operating System Interface (POSIX) Part 1: System Application Program Interface (API) [C Language],* New York: IEEE, 1990.

A. Silberschatz, J.L. Peterson, and P.B. Galvin, *Operating Systems Concepts*, 3rd ed., Reading, Mass.: Addison-Wesley, 1991.

A.S. Tanenbaum, *Modern Operating Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1992.

X/Open Portability Guide, X/Open Company Ltd., Englewood Cliffs, N.J.: Prentice-Hall, 1989.

Further Information

Many textbooks describe operating system concepts. The three cited in the reference section [Dietel, 1990; Silberschatz et al., 1991; and Tanenbaum, 1992] are excellent. The IEEE Computer Society has a number of tutorials on operating system related topics such as fault tolerance, real-time, local area networks and distributed processing. Readers should contact the Computer Society Press office at 10662 Los Vaqueros Circle, Los Alamitos, Calif. 90720. Phone: 714-821-8380.

For those interested in learning more about the implementation of specific operating systems, M.J. Bach, *The Design of the UNIX Operating System*, Prentice-Hall, 1986, describes the implementation of AT&T System V. The 4.3BSD operating system is described in Leffler et al., *The Design and Implementation of the 4.3BSD UNIX Operating System,* Addison-Wesley, 1990.

notion of breaking it. Specifically, the attacker is allowed to obtain signatures to any message of his choice. One may argue that in many applications, such a general attack is not possible (as messages to be signed must have a specific format). Yet, our view is that it is impossible to define a general (i.e., application-independent) notion of admissible messages, and thus a general/robust definition of an attack seems to have to be formulated, as suggested here. (Note that, at worst, our approach is overly cautious.) Likewise, the adversary is said to be successful if he can produce a valid signature to any message for which he has not asked for a signature during his attack. Again, this refers to the ability to form signatures to possibly "nonsensical" messages as a breaking of the scheme. Yet, again, we see no way to have a general (i.e., application-independent) notion of "meaningful" messages (so that only forging signatures to them will be consider a breaking of the scheme).

A *chosen message attack* is a process that, on input, a verification key can obtain signatures (relative to the corresponding signing key) to messages of its choice. Such an attack is said to *succeed* (in existential forgery) if it outputs a valid signature to a message for which it has *not* requested a signature during the attack. A signature scheme is *secure* (or unforgeable) if every *feasible* chosen message attack succeeds with, at most, negligible probability.

We stress that *plain* RSA (similar to plain versions of Rabin's [1979] scheme and DSS) is not secure under the above definition. However, it may be secure if the message is "randomized" before RSA (or the other schemes) is applied. Thus, the randomization paradigm seems pivotal here too.

Message Authentication Schemes

Message authentication is a task related to the setting considered for encryption schemes; that is, communication over an insecure channel. This time, we consider an active adversary that is monitoring the channel and may alter the messages sent on it. The parties communicating through this insecure channel wish to authenticate the messages they send so that their counterpart can tell an original message (sent by the sender) from a modified one (i.e., modified by the adversary). Loosely speaking, a *scheme for message authentication* requires that:

- Each of the communicating parties can *efficiently produce an authentication tag* to any message of his choice.
- Each of the communicating parties can *efficiently verify* whether a given string is an authentication tag of a given message.
- But, *it is infeasible for an external adversary* (i.e., a party other than the communicating parties) to *produce authentication tags* to messages not sent by the communicating parties.

Note that in contrast to the specification of signature schemes, we do not require universal verification. That is, only the receiver is required to be able to verify the authentication tags; and the fact that the receiver can also produce such tags is of no concern. Thus, schemes for message authentication can be viewed as a private-key version of signature schemes. The difference between the two is that, in the setting of message authentication, the ability to verify tags can be linked to the ability to authenticate messages; whereas in the setting of signature schemes, these abilities are separated (i.e., everybody can verify signatures but only the holder of the signing key can produce valid signatures). Hence, digital signatures provide a solution to the message authentication problem, but message authentication schemes do not necessarily constitute digital signature schemes.

Constructions

Message authentication schemes can be constructed using pseudorandom functions. However, as noted by Bellare, Canetti, and Krawczyk [1996], *extensive* usage of pseudorandom functions seems an overkill for achieving message authentication, and more efficient schemes can be obtained based on other cryptographic primitives.

Three central paradigms in the construction of *signature schemes* are the "refreshing" of the "effective" signing-key, the usage of an "authentication tree," and the "hashing paradigm." The first paradigm is aimed at limiting the potential dangers of a chosen message attack by signing the given message using a newly generated (random) instance of the signature scheme, and authenticating this random instance relative to the fixed public key. A natural way of carrying on the authentication of the many newly generated keys is by using an authentication tree [Merkle, 1980]. Finally, the hashing paradigm refers to the common practice of signing real documents

via a two-stage process: first, the document is hashed into a (relatively) short bit string, and then the basic signature scheme is applied to the resulting string. This practice, as well as other usages of the paradigm, is sound provided the hashing function belongs to a family of *Universal One-Way Hash Functions* (cf., [Naor and Yung, 1989]).

Cryptographic Protocols

A general framework for casting n-party cryptographic problems consists of specifying a random process that maps n inputs to n outputs. The inputs to the process are to be thought of as local inputs of n parties, and the n outputs are their corresponding local outputs. The random process describes the desired functionality. That is, if the n parties were to trust each other (or trust some outside party), then they could each send their local input to the trusted party, who would compute the outcome of the process and send each party the corresponding output. The question addressed in this section is: to what extent can such a trusted party be "emulated" by the mutually distrustful parties themselves? We consider two general settings: two-party protocols and multi-party protocols in which the majority of the players are honest. We start with the latter case.

Multi-Party Protocols with Honest Majority

Consider any multi-party protocol. We first observe that each party can change its local input before even entering the execution of the protocol. However, this is also unavoidable when the parties utilize a trusted party. In general, the basic paradigm underlying the definitions of *secure multi-party computations* amounts to saying that situations that can occur in the real protocol, can be simulated in an ideal model (where the parties can employ a trusted party). Thus, the effective malfunctioning of parties in secure protocols is restricted to what is postulated in the corresponding ideal model.

Here, we consider an ideal model in which any minority group (of the parties) may collude as follows. First, this minority shares its original inputs and decides together on replaced inputs to be sent to the trusted party. (The other parties send their respective original inputs to the trusted party.) When the trusted party returns the output, each majority player outputs it locally, whereas the colluding minority can compute outputs based on all they know. A *secure multi-party computation with honest majority* is required to simulate this ideal model. That is, the effect of any feasible adversary that controls a minority of the players in the actual protocol, can be essentially simulated by a (different) feasible adversary that controls the corresponding players in the ideal model. This means that in a secure protocol, the effect of each minority group is "essentially restricted" to replacing its own local inputs (independently of the local inputs of the majority players) before the protocol starts, and replacing its own local outputs (depending only on its local inputs and outputs) after the protocol terminates. (We stress that in the real execution, the minority players do obtain additional pieces of information; yet in a secure protocol, they gain nothing from these additional pieces of information, as they can actually reproduce those by themselves.)

It turns out that efficient and secure multi-party computation of any functionality is possible, provided that a majority of the parties are honest and that trapdoor permutations do exist (e.g., factoring is intractable) [Goldreich, Micali, and Wigderson, 1987].

Two-Party Protocols

Secure multi-party protocols with honest majority can even tolerate a situation in which a minority of the parties abort the execution. This cannot be expected to happen when there is no honest majority (e.g., in a two-party computation). In light of this fact, we consider an ideal model in which each of the two parties can "shut down" the trusted (third) party at any point in time. In particular, this can happen after the trusted party has supplied the outcome of the computation to one party, but before it has supplied it to the second. A *secure two-party computation allowing abort* is required to simulate this ideal model. That is, each party's "effective malfunctioning" in a secure protocol is restricted to supplying an initial input of its choice and aborting the computation at any point in time. We stress that, as above, the choice of the initial input of each party may *not* depend on the input of the other party. Again, it turns out that efficient and secure two-party computation (allowing abort) of any functionality is possible, provided that trapdoor permutations do exist [Yao, 1986].

Defining Terms

Message authentication scheme: A scheme to generate keys so that signing messages and authenticating the validity of signatures can be executed efficiently when knowing the key; but without the key, it is infeasible to falsely authenticate a new message even when given authentication tags to messages of one's choice.

NP-sets: The set S is in NP if there exists a polynomial-time recognizable binary relation R_S and a polynomial ℓ so that $x \in S$ if and only if there exists y so that $|y| \leq \ell(|x|)$ and $(x, y) \in R_S$. Such a y is called an *NP-witness* to $x \in S$.

One-way functions: Functions that are easy to evaluate but difficult (on average) to invert.

Private-key encryption scheme: A scheme to generate keys so that encryption and decryption of messages can be executed efficiently when knowing the key, but it is infeasible to gain any information of the message given only the encrypted message (but not the key).

Pseudorandom functions: Efficient deterministic programs that, given a random seed, present an input-output behavior that is indistinguishable from the one of a truly random function.

Pseudorandom generators: Efficient deterministic programs that stretch short random seeds into longer sequences that are computationally indistinguishable from truly random sequences.

Public-key encryption scheme: A scheme to generate pairs of encryption-decryption keys to that encryption and decryption of messages can be executed efficiently when knowing the corresponding key, but it is infeasible to gain any information of the message given only the encrypted message and the encryption key (but not the decryption key).

Signature scheme: A scheme to generate pairs of signing-verifying keys so that signing messages and verifying the validity of signatures can be executed efficiently when knowing the corresponding key; but without the signing key, it is infeasible to forge a signature to a new message even when given signatures to messages of one's choice.

Zero-knowledge proofs: Two-party protocols by which one party can convince the other party of the validity to an assertion without yielding anything else. That is, when executing such a protocol, the verifier is essentially in the same situation as he would have been if he was granted by a trusted party that the assertion is valid.

References

W. Alexi, B. Chor, O. Goldreich, and C.P. Schnorr. RSA/Rabin functions: certain parts are as hard as the whole, *SIAM Journal on Computing*, 17, 194–209, 1988.

M. Bellare, R. Canetti, and H. Krawczyk. Keying hash functions for message authentication, in *Crypto96*, Springer Lecture Notes in Computer Science, 1109, 1–15, 1996.

M. Blum and S. Goldwasser. An efficient probabilistic public-key encryption scheme which hides all partial information, in *Crypto84*, Lecture Notes in Computer Science, 196, 289–302, 1984.

M. Blum and S. Micali. How to generate cryptographically strong sequences of pseudorandom bits, *SIAM Journal on Computing*, 13, 850–864, 1984. Preliminary version in *23rd IEEE Symposium on Foundations of Computer Science*, 1982.

W. Diffie and M.E. Hellman. New directions in cryptography, *IEEE Trans. on Info. Theory*, IT-22, 644–654, 1976.

D. Dolev, C. Dwork, and M. Naor. Non-malleable cryptography, in *23rd ACM Symposium on the Theory of Computing*, 542–552, 1991.

O. Goldreich, S. Goldwasser, and S. Micali. How to construct random functions, *Journal of the ACM*, 33, 792—807, 1986.

O. Goldreich and L.A. Levin. Hard-core predicates for any one-way function, in *21st ACM Symposium on the Theory of Computing*, 25–32, 1989.

O. Goldreich, S. Micali, and A. Wigderson. Proofs that yield nothing but their validity or all languages in NP have zero-knowledge proof systems, *Journal of the ACM*, 38, 691–729, 1991. Preliminary version in *27th IEEE Symposium on Foundations of Computer Science*, 1986.

O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game — A completeness theorem for protocols with honest majority, in *19th ACM Symposium on the Theory of Computing*, 218–229, 1987.

S. Goldwasser and S. Micali. Probabilistic encryption, *Journal of Computer and System Science*, 28, 270–299, 1984. Preliminary version in *14th ACM Symposium on the Theory of Computing*, 1982.

S. Goldwasser, S. Micali, and C. Rackoff. The knowledge complexity of interactive proof systems, *SIAM Journal on Computing*, 18, 186–208, 1989. Preliminary version in *17th ACM Symposium on the Theory of Computing*, 1985.

S. Goldwasser, S. Micali, and R.L. Rivest. A digital signature scheme secure against adaptive chosen-message attacks, *SIAM Journal on Computing*, 281–308, 1988.

D.E. Knuth. *The Art of Computer Programming, Vol. 2 (Seminumerical Algorithms)*. Addison-Wesley Publishing Company, 1969 (first edition) and 1981 (second edition).

R.C. Merkle. Protocols for public key cryptosystems, in *Proc. of the 1980 Symposium on Security and Privacy*, 1980.

M. Naor and M. Yung. Universal one-way hash functions and their cryptographic application, *21st ACM Symposium on the Theory of Computing*, 33–43, 1989.

M.O. Rabin. Digitalized signatures, in *Foundations of Secure Computation* (R.A. DeMillo et al., eds.), Academic Press, 1977.

M.O. Rabin. Digitalized signatures and public key functions as intractable as factoring, MIT/LCS/TR-212, 1979.

R. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public key cryptosystems, *CACM*, 21, 120–126, 1978.

C.E. Shannon. Communication theory of secrecy systems, *Bell Sys. Tech. J.*, 28, 656–715, 1949.

A.C. Yao. Theory and application of trapdoor functions, in *23rd IEEE Symposium on Foundations of Computer Science*, 80–91, 1982.

A.C. Yao. How to generate and exchange secrets, in *27th IEEE Symposium on Foundations of Computer Science*, 162–167, 1986.

Further Information

O. Goldreich. *Foundation of Cryptography — Fragments of a Book*. February 1995. Available from http://theory.lcs.mit.edu/~oded/frag.html.

O. Goldreich. *Modern Cryptography, Probabilistic Proofs and Pseudorandomness*, Algorithms and Combinatorics series (Vol. 17), Springer, 1998.

O. Goldreich. *Secure Multi-Party Computation*. June 1998. Available from http://theory.lcs.mit.edu.

A.J. Menezes, P.C. van Oorschot, and S.A. Vanstone. *Handbook of Applied Cryptography*, CRC Press, 1996.

Guy, C.G. "Computer Reliability"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

98

Computer Reliability

98.1 Introduction
98.2 Definitions of Failure, Fault, and Error
98.3 Failure Rate and Reliability
98.4 Relationship Between Reliability and Failure Rate
98.5 Mean Time to Failure
98.6 Mean Time to Repair
98.7 Mean Time Between Failures
98.8 Availability
98.9 Calculation of Computer System Reliability
98.10 Markov Modeling
98.11 Software Reliability
98.12 Reliability Calculations for Real Systems

Chris G. Guy
University of Reading

98.1 Introduction

This chapter outlines the knowledge needed to estimate the **reliability** of any electronic system or subsystem within a computer. The word *estimate* was used in the first sentence to emphasize that the following calculations, even if carried out perfectly correctly, can provide no guarantee that a particular example of a piece of electronic equipment will work for any length of time. However, they can provide a reasonable guide to the probability that something will function as expected over a given time period. The first step in estimating the reliability of a computer system is to determine the likelihood of failure of each of the individual components, such as resistors, capacitors, integrated circuits, and connectors, that make up the system. This information can then be used in a full system analysis.

98.2 Definitions of Failure, Fault, and Error

A *failure* occurs when a system or component does not perform as expected. Examples of failures at the component level could be a base-emitter short in a transistor somewhere within a large integrated circuit or a solder joint going open circuit because of vibrations. If a component experiences a failure, it may cause a fault, leading to an error, which may lead to a system failure.

A *fault* may be either the outward manifestation of a component failure or a design fault. Component failure may be caused by internal physical phenomena or by external environmental effects such as electromagnetic fields or power supply variations. *Design faults* may be divided into two classes. The first class of design fault is caused by using components outside their rated specification. It should be possible to eliminate this class of faults by careful design checking. The second class, which is characteristic of large digital circuits such as those found in computer systems, is caused by the designer not taking into account every logical condition that could occur during system operation. All computer systems have a software component as an integral part of their operation, and software is especially prone to this kind of design fault.

A fault may be permanent or transitory. Examples of *permanent faults* are short or open circuits within a component caused by physical failures. *Transitory faults* can be subdivided further into two classes. The first, usually called *transient faults,* are caused by such things as alpha-particle radiation or power supply variations. Large random access memory circuits are particularly prone to this kind of fault. By definition, a transient fault is not caused by physical damage to the hardware. The second class is usually called *intermittent faults.* These faults are temporary but reoccur in an unpredictable manner. They are caused by loose physical connections between components or by components used at the limits of their specification. Intermittent faults often become permanent faults after a period of time. A fault may be *active* or *inactive.* For example, if a fault causes the output of a digital component to be stuck at logic 1, and the desired output is logic 1, then this would be classed as an inactive fault. Once the desired output becomes logic 0, then the fault becomes active.

The consequence for the system operation of a fault is an error. As the error may be caused by a permanent or by a transitory fault, it may be classed as a *hard error* or a *soft error.* An error in an individual subsystem may be due to a fault in that subsystem or to the propagation of an error from another part of the overall system.

The terms *fault* and *error* are sometimes interchanged. The term *failure* is often used to mean anything covered by these definitions. The definitions given here are those in most common usage.

Physical faults within a component can be characterized by their external electrical effects. These effects are commonly classified into *fault models.* The intention of any fault model is to take into account every possible failure mechanism, so that the effects on the system can be worked out. The manifestation of faults in a system can be classified according to the likely effects, producing an *error model.* The purpose of error models is to try to establish what kinds of corrective action need be taken in order to effect repairs.

98.3 Failure Rate and Reliability

An individual component may fail after a random time, so it is impossible to predict any pattern of failure from one example. It is possible, however, to estimate the rate at which members of a group of identical components will fail. This rate can be determined by experimental means using accelerated life tests. In a normal operating environment, the time for a statistically significant number of failures to have occurred in a group of modern digital components could be tens or even hundreds of years. Consequently, the manufacturers must make the environment for the tests extremely unfavorable in order to produce failures in a few hours or days and then extrapolate back to produce the likely number of failures in a normal environment. The **failure rate** is then defined as the number of failures per unit time, in a given environment, compared with the number of surviving components. It is usually expressed as a number of failures per million hours.

If $f(t)$ is the number of components that have failed up to time t, and $s(t)$ is the number of components that have survived, then $z(t)$, the *failure rate* or *hazard rate,* is defined as

$$z(t) = \frac{1}{s(t)} \cdot \frac{df(t)}{d(t)} \tag{98.1}$$

Most electronic components will exhibit a variation of failure rate with time. Many studies have shown that this variation can often be approximated to the pattern shown in Fig. 98.1. For obvious reasons this is known as a *bathtub* curve. The first phase, where the failure rate starts high but is decreasing with time, is where the components are suffering infant mortality; in other words, those that had manufacturing defects are failing. This is often called the *burn-in* phase. The second part, where the failure rate is roughly constant, is the useful life period of operation for the component. The final part, where the failure rate is increasing with time, is where the components are starting to wear out.

Using the same nomenclature as before, if:

$$s(t) + f(t) = N \tag{98.2}$$

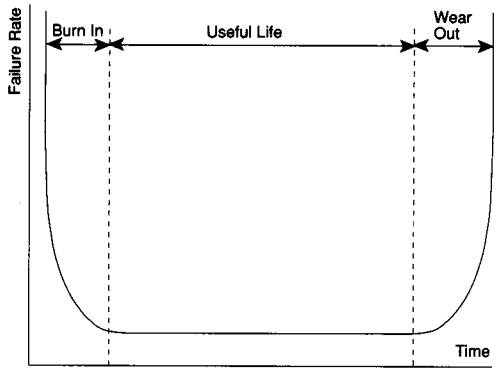


FIGURE 98.1 Variation of failure rate with time.

i.e., N is the total number of components in the test, then the *reliability* $r(t)$ is defined as

$$r(t) = \frac{s(t)}{N} \tag{98.3}$$

or in words, and using the definition from the *IEEE Standard Dictionary of Electrical and Electronic Terms,* reliability is the probability that a device will function without failure over a specified time period or amount of usage, under stated conditions.

98.4 Relationship Between Reliability and Failure Rate

Using Eqs. (98.1), (98.2), and (98.3) then

$$z(t) = -\frac{N}{s(t)} \cdot \frac{dr(t)}{d(t)} \tag{98.4}$$

λ is commonly used as the symbol for the failure rate $z(t)$ in the period where it is a constant, i.e., the useful life of the component. Consequently, we may write Eq. (98.4) as

$$\lambda = -\frac{1}{r(t)} \cdot \frac{dr(t)}{d(t)} \tag{98.5}$$

Rewriting, integrating, and using the limits of integration as $r(t) = 1$ at t=0 and $r(t) = 0$ at $t = \infty$ gives the result:

$$r(t) = e^{-\lambda t} \tag{98.6}$$

This result is true only for the period of operation where the failure rate is a constant. For most common components, real failure rates can be obtained from such handbooks as the American military MIL-HDBK-217E, as explained in Section 98.12.

It must also be borne in mind that the calculated reliability is a probability function based on lifetime tests. There can be no guarantee that any batch of components will exhibit the same failure rate and hence reliability as those predicted because of variations in manufacturing conditions. Even if the components were made at

the same factory as those tested, the process used might have been slightly different and the equipment will be older. Quality assurance standards are imposed on companies to try to guarantee that they meet minimum manufacturing standards, but some cases in the United States have shown that even the largest plants can fall short of these standards.

98.5 Mean Time to Failure

A figure that is commonly quoted because it gives a readier feel for the system performance is the **mean time to failure** or MTTF. This is defined as

$$\text{MTTF} = \int_0^\infty r(t)\, dt \tag{98.7}$$

Hence, for the period where the failure rate is constant:

$$\text{MTTF} = \frac{1}{\lambda} \tag{98.8}$$

98.6 Mean Time to Repair

For many computer systems it is possible to define a **mean time to repair** (MTTR). This will be a function of a number of things, including the time taken to detect the failure, the time taken to isolate and replace the faulty component, and the time taken to verify that the system is operating correctly again. While the MTTF is a function of the system design and the operating environment, the MTTR is often a function of unpredictable human factors and, hence, is difficult to quantify. Figures used for MTTR for a given system in a fixed situation could be predictions based on the experience of the reliability engineers or could be simply the maximum response time given in the maintenance contract for a computer. In either case, MTTR predictions may be subject to some fluctuations. To take an extreme example, if the service engineer has a flat tire while on the way to effect the repair, then the repair time may be many times the predicted MTTR. For some systems no MTTR can be predicted, as they are in situations that make repair impossible or uneconomic. Computers in satellites are a good example. In these cases and all others where no errors in the output can be allowed, fault tolerant approaches must be used in order to extend the MTTF beyond the desired system operational lifetime.

98.7 Mean Time Between Failures

For systems where repair is possible, a figure for the expected time between failures can be defined as

$$\text{MTBF} = \text{MTTF} + \text{MTTR} \tag{98.9}$$

The definitions given for MTTF and MTBF are the most commonly accepted ones. In some texts, MTBF is wrongly used as mean time before failure, confusing it with MTTF. In many real systems, MTTF is very much greater than MTTR, so the values of MTTF and MTBF will be almost identical, in any case.

98.8 Availability

Availability is defined as the probability that the system will be functioning at a given time during its normal working period.

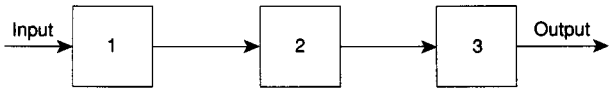


FIGURE 98.2 Series model.

$$\text{Av} = \frac{\text{total working time}}{\text{total time}} \tag{98.10}$$

This can also be written as

$$\text{Av} = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}} \tag{98.11}$$

Some systems are designed for extremely high availability. For example, the computers used by AT&T to control its telephone exchanges are designed for an availability of 0.9999999, which corresponds to an unplanned downtime of 2 min in 40 years. In order to achieve this level of availability, fault tolerant techniques have to be used from the design stage, accompanied by a high level of monitoring and maintenance.

98.9 Calculation of Computer System Reliability

For systems that have not been designed to be fault tolerant it is common to assume that the failure of any component implies the failure of the system. Thus, the system failure rate can be determined by the so-called parts count method. If the system contains m types of component, each with a failure rate λ_m, then the system failure rate λ_s can be defined as

$$\lambda_s = \sum_1^m N_m \cdot \lambda_m \tag{98.12}$$

where N_m is the number of each type of component.

The system reliability will be

$$r_s(t) = \prod_1^m N_m \cdot r_m \tag{98.13}$$

If the system design is such that the failure of an individual component does not necessarily cause system failure, then the calculations of MTTF and $r_s(t)$ become more complicated.

Consider two situations where a computer system is made up of several subsystems. These may be individual components or groups of components, e.g., circuit boards. The first is where failure of an individual subsystem implies system failure. This is known as the series model and is shown in Fig. 98.2. This is the same case as considered previously, and the parts count method, Eqs. (98.12) and (98.13), can be used. The second case is where failure of an individual subsystem does not imply system failure. This is shown in Fig. 98.3. Only the failure of every subsystem means that the system has failed, and the system reliability can be evaluated by the following method. If $r(t)$ is the reliability (or probability of not failing) of each subsystem, then $q(t) = 1 - r(t)$ is the probability of an individual subsystem failing. Hence, the probability of them all failing is

$$q_s(t) = [1 - r(t)]^n \tag{98.14}$$

for n subsystems.

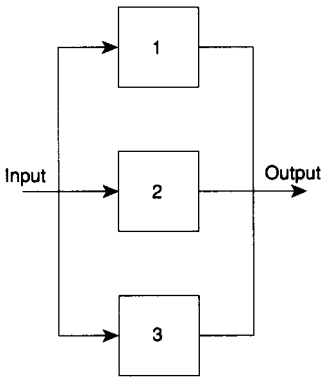


FIGURE 98.3 Parallel model.

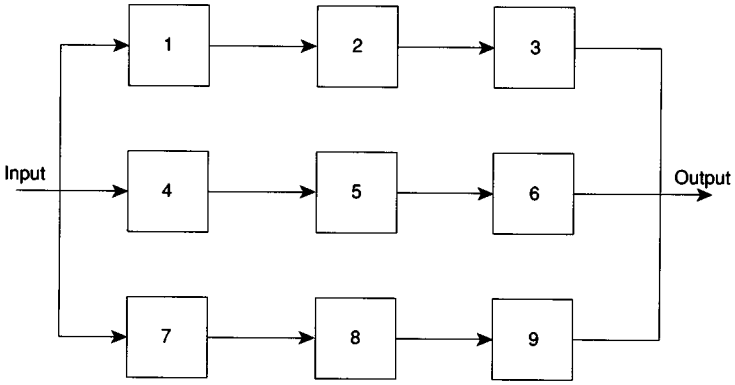


FIGURE 98.4 Parallel series model.

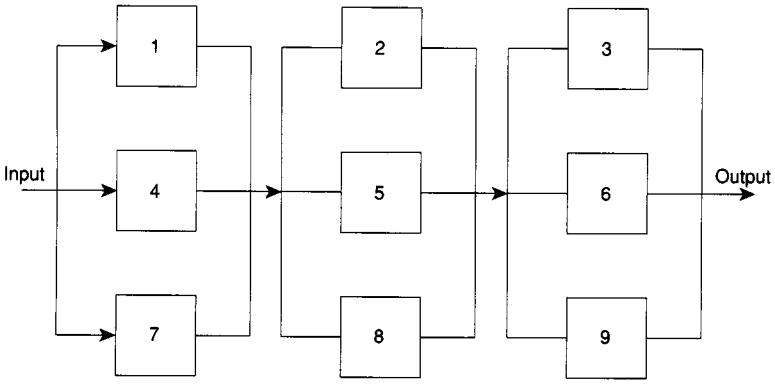


FIGURE 98.5 Series-parallel model.

Hence the system reliability will be:

$$r_s(t) = 1 - [1 - r(t)]^n \tag{98.15}$$

In practice, systems will be made up of differing combinations of parallel and series networks; the simplest examples are shown in Figs. 98.4 and 98.5.

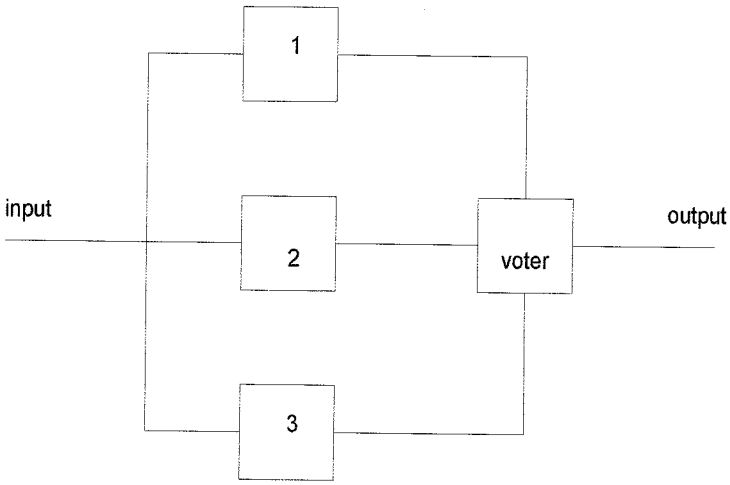


FIGURE 98.6 Triple-modular-redundant system.

Parallel-Series System

Assuming that the reliability of each subsystem is identical, then the overall reliability can be calculated thus. The reliability of one unit is *r;* hence the reliability of the series path is r^n. The probability of failure of each path is then $q = 1 - r^n$. Hence, the probability of failure of all m paths is $(1 - r^n)^m$, and the reliability of the complete system is

$$r_{ps} = 1 - (1 - r^n)^m \tag{98.16}$$

Series-Parallel System

Making similar assumptions, and using a similar method, the reliability can be written as

$$r_{sp} = [1 - (1 - r)^n]^m \tag{98.17}$$

It is straightforward to extend these results to systems with subsystems having different reliabilities and in different combinations. It can be seen that these simple models could be used as the basis for a fault tolerant system, i.e., one that is able to carry on performing its designated function even while some of its parts have failed.

Practical Systems Using Parallel Sub-Systems

A computer system that uses parallel sub-systems to improve reliability must incorporate some kind of arbitrator to determine which output to use at any given time. A common method of arbitration involves adding a voter to a system with N parallel modules, where N is an odd number. For example, if $N = 3$, a single incorrect output can be masked by the two correct outputs outvoting it. Hence, the system output will be correct, even though an error has occurred in one of the sub-systems. This system would be known as Triple-Modular-Redundant (TMR) (Fig. 98.6).

The reliability of a TMR system is the probability that any two out of the three units will be working. This can be expressed as

$$r_{tmr} = r_1 r_2 r_3 + r_1 r_2 (1 - r_3) + r_1 (1 - r_2) r_3 + (1 - r_1) r_2 r_3$$

where r_n $(n = 1, 2, 3)$ is the reliability of each subsystem. If $r_1 = r_2 = r_3 = r$ this reduces to

$$r_{tmr} = 3r^2 - 2r^3$$

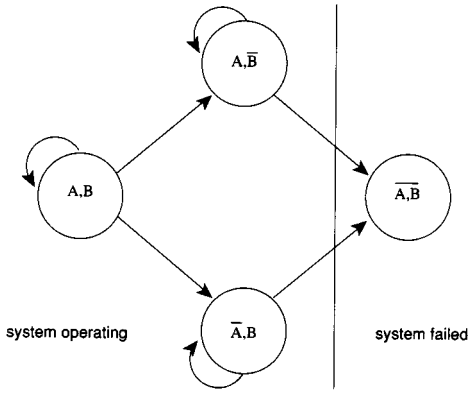


FIGURE 98.7 State diagram for two-unit parallel system.

The reliability of the voter must be included when calculating the overall reliability of such a system. As the voter appears in every path from input to output, it can be included as a series element in a series-parallel model. This leads to

$$r_{tmr} = r_v\left[3r^2 - 2r^3\right] \tag{98.18}$$

where r_v is the reliability of the voter.

More information on methods of using redundancy to improve system reliability can be found in Chapter 93.

98.10 Markov Modeling

Another approach to determining the probability of system failure is to use a Markov model of the system, rather than the combinatorial methods outlined previously. Markov models involve the defining of *system states* and *state transitions*. The mathematics of Markov modeling are well beyond the scope of this brief introduction, but most engineering mathematics textbooks will cover the technique.

To model the reliability of any system it is necessary to define the various fault-free and faulty states that could exist. For example, a system consisting of two identical units (A and B), either of which has to work for the system to work, would have four possible states. They would be (1) A and B working; (2) A working, B failed; (3) B working, A failed; and (4) A and B failed. The system designer must assign to each state a series of probabilities that determine whether it will remain in the same state or change to another after a given time period. This is usually shown in a state diagram, as in Fig. 98.7. This model does not allow for the possibility of repair, but this could easily be added.

98.11 Software Reliability

One of the major components in any computer system is its software. Although software is unlikely to wear out in a physical sense, it is still impossible to prove that anything other than the simplest of programs is totally free from bugs. Hence, any piece of software will follow the first and second parts of the normal bathtub curve (Fig. 98.1). The burn-in phase for hardware corresponds to the early release of a complex program, where bugs are commonly found and have to be fixed. The useful life phase for hardware corresponds to the time when the software can be described as stable, even though bugs may still be found. In this phase, where the failure rate can be characterized as constant (even if it is very low), the hardware performance criteria, such as MTTF and MTTR can be estimated. They must be included in any estimation of the overall availability for the computer system as a whole. Just as with hardware, techniques using redundancy can be used to improve the availability through fault tolerance.

98.12 Reliability Calculations for Real Systems

The most common source of basic reliability data for electronic components and circuits is the military handbook *Reliability Prediction of Electronic Equipment,* published by the U.S. Department of Defense. It has the designation MIL-HDBK-217E in its most recent version. This handbook provides both the basic reliability data and the formulae to modify those data for the application of interest. For example, the formula for predicting the failure rate, λ_p, of a bipolar or MOS microprocessor is given as

$$\lambda_p = \pi_Q(C_1\pi_T\pi_V + C_2\pi_E)\pi_L \text{ failures per } 10^6 \text{ hours}$$

where π_Q is the part quality factor, with several categories, ranging from a full mil-spec part to a commercial part; π_T is the temperature acceleration factor, related to both the technology in use and the actual operating temperature; π_V is the voltage stress derating factor, which is higher for devices operating at higher voltages; π_E is the application environment factor (the handbook gives figures for many categories of environment, ranging from laboratory conditions up to the conditions found in the nose cone of a missile in flight); π_L is the device learning factor, related to how mature the technology is and how long the production of the part has been going on; C_1 is the circuit complexity factor, dependent on the number of transistors on the chip; and C_2 is the package complexity, related to the number of pins and the type of package.

The following figures are given for a 16-bit microprocessor, operating on the ground in a laboratory environment, with a junction temperature of 51°C. The device is assumed to be packaged in a plastic, 64-pin dual in-line package and to have been manufactured using the same technology for several years:

$$\begin{array}{llll} \pi_Q = 20 & \pi_T = 0.89 & \pi_V = 1 & \pi_E = 0.38 \\ \pi_L = 1 & C_1 = 0.06 & C_2 = 0.033 & \end{array}$$

Hence, the failure rate λ_p for this device, operating in the specified environment, is estimated to be 1.32 failures per 10^6 hours. To calculate the predicted failure rate for a system based around this microprocessor would involve similar calculations for all the parts, including the passive components, the PCB, and connectors, and multiplying all the resultant failure rates together. The resulting figure could then be inverted to give a predicted MTTF. This kind of calculation is repetitive, tedious, and therefore prone to errors, so many companies now provide software to perform the calculations. The official Department of Defense program for automating the calculation of reliability figures is called ORACLE. It is regularly updated to include all the changes since MIL-HDBK-217E was released. Versions for VAX/VMS and the IBM PC are available from the Rome Air Defense Center, RBET, Griffiss Air Force Base, NY 13441-5700. Other software to perform the same function is advertised in the publications listed under Further Information.

Defining Terms

Availability: This figure gives a prediction for the proportion of time that a given part or system will be in full working order. It can be calculated from

$$\text{Av} = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}$$

Failure rate: The failure rate, λ, is the (predicted or measured) number of failures per unit time for a specified part or system operating in a given environment. It is usually assumed to be constant during the working life of a component or system.

Mean time to failure: This figure is used to give an expected working lifetime for a given part, in a given environment. It is defined by the equation

$$\text{MTTF} = \int_0^\infty r(t)\,dt$$

If the failure rate λ is constant, then

$$\text{MTTF} = \frac{1}{\lambda}$$

Mean time to repair: The MTTR figure gives a prediction for the amount of time taken to repair a given part or system.

Reliability: Reliability $r(t)$ is the probability that a component or system will function without failure over a specified time period, under stated conditions.

Related Topics

92.2 Local Area Networks • 110.1 Introduction • 110.4 Mean Time to Failure (MTTF) • 110.14 Markov Models • 110.22 Reliability and Economics

References

B. W. Johnson, *Design and Analysis of Fault Tolerant Digital Systems,* Reading, Mass.: Addison-Wesley, 1989.

V. P. Nelson and B. D. Carroll, *Tutorial: Fault Tolerant Computing,* Washington, D.C.: IEEE Computer Society Press, 1987.

D. K. Pradhan, *Fault Tolerant Computing, Theory and Techniques,* vols. I and II, Englewood Cliffs, N. J.: Prentice-Hall, 1986.

Further Information

The quarterly magazine *IEEE Transactions on Reliability* contains much of the latest research on reliability estimation techniques.

The monthly magazine *Microelectronics and Reliability* covers the field of reliability estimation and also includes papers on actual measured reliabilities.

Sometimes, manufacturers make available measured failure rates for their devices.

Hawke, G.L. "The Internet and its Role in the Future"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

99

The Internet and Its Role in the Future

99.1 Introduction
99.2 History
99.3 The Internet Today
Send/Receive Capacity • Login • Password • The World Wide Web (WWW, Web)
99.4 The Future

Gary L. Hawke
University of Kansas

99.1 Introduction

The creation of the computer rivals any invention in our history. The wide use of computers and their resultant affect on all communication is tied directly to their ability to contact and interface with each other. This ability to link computers together is the basis for all global success of computer usage. It is the key to sharing knowledge, data, conversation, and discovery on a global scale.

The Internet is the adolescent offspring developed from the creation of a backbone connecting computers and computer networks. These **Local Area Networks (LANs)** within internal locations can now be linked to other networks worldwide. The Internet, not really a place but a concept, is intended to connect global information resources. The value of the Internet is more than its ability to link computers and networks, its real value is in bringing information to any who need it.

This chapter will attempt to give a broad overview of the Internet, condensing the major points of interest into one short chapter. There are, of course, hundreds of books on the subject. At the end of the chapter there is a list of just a few publications the reader may wish to check for more in-depth information.

The rapid advances in technology and innovations in operating software make any published account of the Internet obsolete before they arrive in the reader's hands. As we look at the Internet as it exists today, we also try to take a glimpse at the future of this creation. There are, however, as many visions of the future as there are dreamers. So, with due respect to those who envision the developments and uses of the Internet in the years to come, here is a basic introduction.

99.2 History

Early in the development of the mainframe computer, the creators realized these machines would have to be able to talk to each other for the technology to be of real meaning. They began by connecting one computer to another through direct cable hook-up, then to printers, and eventually they sought a means of connection outside their own building. The telephone system seemed like a logical choice. It was virtually everywhere. It was flexible, inexpensive, and one could simply dial up the location one wanted to reach. The **modem** was invented to send computer information over telephone lines instead of a direct wire connection. It provided the **modulated** (**Mo**) computer signal from one location and the **demodulated** (**dem**) to the other end of the connection. Protocols were developed that would translate the computer signal to telephone tones. Modems

are limited by the speed at which they can make this translation. Later in the chapter, modem speeds and types will be discussed at greater length.

Computers can speak with each other in various methods over telephone lines. **Circuit switching** is when one computer directly dials up another. As phone lines in and out of locations began to become "busy" at times, the **multiplexing** idea was born. This allows an electronic cue system where computers may line up to access the next available line when one computer is done. Various forms of multiplexing have been developed to hold the data to be transmitted until an outgoing line is available, then delivering it. This eliminates the need for the entire computer waiting for a path. Other tasks can be undertaken while the outgoing data is in the hold position.

This same early system is still the main access to the Internet for businesses and personal use. It allows anyone access to the Internet anywhere in the world that telephone service is available. This now includes cellular applications and even direct satellite transmission for extremely remote locations. The system works very well and is expanding on a local access basis.

As the Cold War continued, the government began the search for a system that would be less vulnerable to damage or sabotage. The **U.S. Advanced Research Projects Agency (ARPA)** began to explore an alternate path for government, business, and research locations to interconnect. This national network would allow different operating systems the ability to talk to each other and to route this path in a multitude of connects. Thus, if one path was damaged, the system would seek another path to its destination. The packet system of transmitting, which divides the information into streams of small packets of information, each with an I.D. number for final assembly, was developed. This new network would allow scientists from around the country to connect to one another to further computer technology. The system would be able to share equipment, transfer data files, and use long distance login capabilities. However, according to one of its early developers, Vint Cerf, vice-president of the Corporation for National Research Initiatives, "...we didn't know that e-mail was important...we weren't even sure what it was at the time" [Dern, 1994]. This new network would be called **ARPAnet**.

Soon researchers at private companies and institutions of higher learning were cranking out new software protocols to allow computers access to this new highway. In 1969, the first three ARPAnet participants were linked between Stanford University in Palo Alto, California, and the University of California at Los Angeles and the University of California at Santa Barbara. The system grew slowly for the first decade, a bit faster over the next 10 years, and exploded in the 1990s as protocols, access, and individual desk top computers increased.

During these years many operating systems, high speed (including fiber optic) transmission lines, access providers, and language protocols were developed. The Defense Advanced Research Projects Agency developed a protocol for network interconnection called **Transmission Control Protocol/Internetworking Protocol (TCP/IP)**. In 1983 this protocol was adopted as the standard for what would later be called *Internet*.

99.3 The Internet Today

Today the Internet is a conglomeration of a huge number of interconnected LANs around the world. These are a series of linked IPs, functioning on their own, but able to instantly connect with any other LAN likewise connected to the Internet. It also covers the thousands of personal computers at home or in the office. The Internet can also connect to a wide range of other networks, including privately operated ones like America On Line, Prodigy, and Compuserve, plus hundreds of other special networks now in use throughout the world. All you need is their IP, email account, or Web site address. As more and more users access the Internet problems of speed, capacity, and reliability begin to crop up. These will be discussed later in the chapter.

The operating system most used by computers on the Internet is **Unix**. This system was developed by AT&T Bell Labs. This system uses characters on a menu or command message to access files. For more detailed information on Unix see [Dern, 1994].

The Internet has four major services: (1) **electronic mail** (email), (2) **telnet**, (3) **file transfer protocol (FTP)**, and (4) **client/server connections**. Discussion on each service follows.

1. Email—Probably the most used process on the Internet, email is a text transfer of messages from one point to another. Email addresses will route a message anywhere in the world in seconds. Several email programs are in use, from free shared systems such as Elm, **Pine**, **Mailx**, and **Eudora**. Recently some

advanced features have been added to mail programs and many have been released at low cost. **Eudora Pro** is an example. Email is a low-cost instant information carrier. It can be used internally or on the Internet and will continue to expand in usage.

2. Telnet—This program allows the user to log onto other computers in distant locations and to operate those computers from their own terminal, running systems or downloading data. This defines how computers speak to each other for login applications. Of course, most computers require a **password** to be able to access their information. The use of telnet to open other computers has spawned a rash of **hacking**, which attempts to forge passwords for unauthorized access. These can be criminal acts and are the very situations seen on occasional newscasts! Telnet is available on a public domain basis and can operate with almost all computer hardware.
3. FTP—This allows you to transfer files from one computer to another or from one remote site to another remote site. The user has the ability to download files from shareware servers or from private files using a password for entry. FTP can download small files and huge databases. It is the primary method for those doing research or fact finding to acquire massive amounts of information in a very short time. As was mentioned previously, the time it takes to download files is directly related to the speed of your modem, or direct connection. This can take a very long time in the case of large files.
4. Client/server—Much of the current use of the Internet is to share information. This can happen when one computer accesses another that operates separately and serves information to the client who has logged in. You request another computer to send files to you, the client. In the early days of the Internet, most of these were text files only. One such text-based browse is **Lynx**, developed at the University of Kansas for exploration of the documents on the Internet and the **World Wide Web (www)** Web, for short. WWW is a system for finding and accessing Internet resources. The Web's most innovative feature is the ability to "point and click" on a text word or graphic to **hypertext link** you to another Web site holding the information you request. This process allows you to move from one computer to the next hop-scotching across the globe to find the information you seek. By using hypertext-based protocol, the world is at your fingertips in an instant. The Lynx program is only text-based, but later developments brought graphics to the Web search vehicles. In order to search the vast reserves of information, **browsers** were developed. An early public domain browser, developed by the National Center for Supercomputing Applications, is **Mosaic**. Mosaic is all you need for Web site access. Later, **Netscape Navigator** was commercially developed with lots of bells and whistles for quick and easy access. There is a cost for Netscape.

Send/Receive Capacity

Computers function using binary numbers. This series of 0 or 1 format exists at the smallest level as a bit. This digital system is transmitted over modems or direct connections in a series of 8-bit packets called bytes. The number of bits per second that can be transmitted or received equates to the speed at which you can send or receive information over the Internet. This **bandwidth** of the transmission line is expressed in bits per second, normally in metric form.

Early computer connects were made at 56 bits per second (56 bps). As the speed of machine and systems improved, rates of transmission have increased to kilobits (kbs), megabits (mps), gigabits (gbs), and so on. Early modems operated at 14.4 kbs (14,000 bps), more modern modems function at 28.8 kbs (28,800 bps). Higher capacity "pipelines" include **Integrated Services Digital Network (ISDN)** lines that can now be installed at home or business that operate at 56 kbs, direct **Ethernet** connections (a technology that can connect computers at a 10 Mbps). Higher capacity lines such as **T-1** (1.54 Mbs) and multiples of those, **T-2**, and **T-3** lines are now being used for extremely high capacity systems. The T-1 and higher lines are very expensive to install and maintain but offer extremely quick Internet work speed.

Login

In order to begin the process of logging in, you will need an account of some kind. This is established through your service provider, which could be your own company or a commercial service such as Compuserve, America

on Line, etc. If you are operating your own system, you'll first need a TCP/IP plus a serial communications protocol such as a **Serial Line Internet Protocol (SLIP)**, **Point to Point Protocol (PPP)**, or the like. If you access one of these from your desktop computer you may start an email program, such as Eudora, without logging in to your multiuser account. Your mail will be downloaded to your desktop, where you can read it, print it, save it, or compose a new message. When you copy files from public acccess sites, your desktop system can bring them directly to it. You can also use the full desktop visual capabilities of your system when attaching to a Web site. If you are using a modem connection to another server/host, you'll need to check the speed. From Internet providers and accounts available you can select:

- full network sites
- dial in IP accounts
- telnet (long distance telephone service)
- public access accounts (terminal servers)
- commercial servers

Email and Web sites are identified by an address. On the Internet, address always refers to an electronic address. The form is **Userid**, the name selected by the user, followed by the @ character, followed by the **computer's name** (all computers on the Internet have a specific name).

Example: ghawke@ukans.edu In this case "ukans" refers to the University of Kansas computer center. No spaces are used in the address. The information after the "@" refers to a **domain**. There can be any number of user names alike but each and every domain will have their own unique name.

After the userid –@– and domain name, there can be one or more sub-domains. In the example above, the sub-domain is "edu" for educational institution. The sub-domains are listed with the most general information about the domain computer being further to the right or end of the address and the most specific name being to the immediate right of the domain name. Domains and their meanings are listed in Table 99.1.

Password

At this point of your login, you'll be asked to create a password. This password is the key to your account and the tighter you guard it, the more secure your system will be. Common passwords are a combination of letter and number characters, usually eight or so in number. Obviously, these can be the subject of unauthorized usage. The more creative the password, the more secure it will be.

The World Wide Web (WWW, Web)

The WWW has more Internet locations than any other resource. It serves as a huge reference book for the world, with the added feature of hypertext links . Hypertext allows one to follow a topic line or reference idea

TABLE 99.1 Organizational Top-Level Domains

Domain	Meaning
art	Cultural and entertainment entities
com	Commercial organization
edu	Educational institution
firm	Businesses or firms
gov	Government
info	Information services
int	International organization
mil	U.S. military
net	Networking organization
nom	Individual or personal nomenclature
org	Non-profit organization
rec	Recreational activity organizations
store	Businesses offering goods to purchase
web	WWW related activity organizations

from one site to another, just by a point and "click" of a mouse. This link can take you to another page of the site you initially addressed or it can automatically link you around the world to another computer with a totally different Web site full of additional information. This is done through a computer program titled **HyperText Markup Language (HTML)**. This program will allow you to create your own Web site, or **home page** as they are sometimes called, or connect your site to any others you wish to link.

With the millions of Web sites out there, the browsers we previously mentioned are invaluable in locating the site you want. You may access these by topic name, individual name, address, or randomly. A list of **searching tools** is provided in the Defining Terms section.

99.4 The Future

Obviously, the future of the Internet is as wide open as imagination itself. The author, therefore, can only propose one person's opinion. Those who have the vision and talent for the future needs of information and services on the Internet will be the millionaires and experts in the years to come. There are a few areas that current trends indicate will be open for expansion.

Advertising and Marketing

It seems clear that what is driving the expanse of the Internet's development is directly linked to advertising and marketing. When *Netscape* and *Yahoo* went public, the company founders became instant millionaires and the value of those and other companies soared. Although theses two companies now charge for their software, these search vehicles will create the majority of their income for advertising on their search pages and by the sale of marketing research they can develop from the chronicling of data received from their visitors. Every time one of these is accessed, computers can archive the visitor's name and address, the sites they visited, the time spent on each page, the links made to other sites and any purchases ordered. Advertisers can focus on the exact customer profile they want in purchasing this information database or to advertise directly on the page. There are huge dollars waiting to be spent on this form of advertising. The user of the Internet may feel they are an independent wanderer through the garden of knowledge, with no forced direction and no guided path. They should know, however, that the most popular search vehicles, most exciting Web site pages, and most creative browsers will be paid for by advertising dollars. Their use of these sites will contribute to databases designed to exploit these visitors.

One of the major reasons the Internet and, more specifically, the World Wide Web will increase in use is its ability for *interactive* use by the public. This takes the form of game playing, research responses, banking, entertainment, and purchases. Recent trade shows have unveiled incredible interactive games designed for all ages that can be played over the Internet from *Scrabble* to *Super Mario*. The Web already contains hundreds of sites that allow you to respond directly to research questions, interest levels and personal inquiry. There are many sites that already involve banking and the ordering of goods and services. Obviously, the popularity of these services is directly related to the public's confidence in their security. Just as with Automated Teller Machines, I believe this aspect of technology will be mastered on the Internet, although any code or security system designed by one can be broken by another.

The entertainment side will grow rapidly. Movies on demand, archival information from existing print media and traditional radio and TV will be available. In December of 1994, *KJHK*, the student FM radio station at the University of Kansas, was the first station to provide a real time, live and continuous programming source to the Internet. The on-air programming of the station can be heard anywhere in the world by attaching to their Web site. There are many radio stations now "broadcasting" over the Internet and several commercial software companies around to facilitate that use. On January 2, 1996, the University of Kansas provided a live television signal to the Internet that could be viewed by computers connected to the Internet with 28.8 modems or higher. These real time audio and video services will be used in all kinds of applications in the future. They not only provide good quality and diverse choice, but can be delivered at very low costs since most long distance telephone charges are not present.

The huge growth in the use of the Internet is even more amazing when one considers that the ownership of personal computers is still rather low when compared with the entire population, although 1995 was the first

year that the purchase of computers outpaced the purchase of television sets! The use of the Internet for email will continue to increase. This will also take the form of video conferencing. The speed and low cost of both of these systems make contact between family and student or between corporations and their clients a very desirable advantage of the Internet.

No discussion about the Internet's future would be complete without a quick look at the legal aspects. We have already seen an attempt by the government to censor content on the Internet. The recent Telecommunications Act called for criminal penalties for certain obscene material provided to the Internet. This was struck down by the Supreme Court, but look for more attempts to regulate this new media. In addition, the concerns about intellectual property and copyright infringement are already being addressed. Many Internet users "use" material on their Web sites without permission from their creators. This applies to music licenses for entertainment, reprints of news stories and books, cartoons, artists creations, and copyrighted research. Be careful in this area. The larger the pockets of the user, the more apt they are to be sued for copyright infringement. As the years go by, you will see many new regulations and judicial decisions in favor of the creators of this material.

Finally, the anonymity for Internet users, the personal soap box it offers, and the ability to gather even the most obscure interest groups together in the privacy of their very own home means a huge use of the Internet for personal expression. This new technology will be as great, if not greater, an influence on world society than television has been. Your predictions on its future are as good as mine.

Defining Terms

The following list of terms is based in large part from the published list compiled by *Academic Computing Services, University of Kansas*, and is reprinted with their permission.

Anonymous FTP: Used to log into public access file sites and download files by logging in the user name "anonymous"

Archie: A search system for locating publicly available files by anonymous FTP.

ARPA: The United States Advanced Research Projects Agency was the original source for the development of a network to inter-connect computers.

ARPAnet: The original inter-connecting network that was the basis for the Internet.

ASCII: American Standard Code for Information Interchange. An industry-wide computer standard for the encoding of numeric characters.

ASCII file: American Standard Code for Information Interchange. A file type where characters are stored as a series of eight binary digits (A=01000001).

Buad rate: The primary signaling rate of a carrier. A 9600 "baud" modem transfers data at 2400 baud, but the signal rate is 4 bits per cycle allowing for a transfer rate of 9600 bps. Baud and bps are used synonymously but they not the same.

BBS: Bulletin board system. An electronic multi-user system that often includes a message database people can login to and leave messages for a particular group.

binary file: A file in which all 8 bits of a byte are used to encode information. Binary is also a file transfer type used for .zip and executable files in ftp.

BITNET: An academically oriented international network using a different protocol than the Internet, although email may be exchanged via gateways. A typical BITNET address might look like *joe@ukanvm.*

bps: Bits per second.

Bridge: A connection using software or hardware to connect two segments of a network not necessarily using the same protocol.

Browser: any program that reads hypertext. *Mosaic, Lynx,* and *Netscape Navigator* are browser clients used to access World Wide Web sites.

Client: A software application that exists to extract some service from a remote server somewhere on the network. (Think of your telephone as the client and the telephone company as the server).

Compression: A utility used on many platforms to make files smaller for transport. On a Mac common compression formats are *.sit* (use *Stuffit* to uncompress) or *.cpx* (use *Compact Pro t*o uncompress). On a PC, *.zip* is common (use *pkunzip* to uncompress).

Domain name server: A distribution database system for translating mnemonic computer addresses (like *kuhub.cc.ukans.edu* into numeric addresses (like *129.237.32.1*) and vice versa.

Domain name: The exclusive name assigned to a site on the Internet.

Email: Electronic mail. The exchange of messages between people on an electronic network.

Ethernet: A computer communications technology designed to connect computer systems together to form local area networks (*LANs*). Ethernet transmits information at 10 million bits per second over different physical media ranging from twisted pair wires to fiber optic cable.

Fetch: An easy to use public domain *ftp* program for the Macintosh.

FTP: File Transfer Protocol. The process of moving a file from one computer to another. The application program that moves files using the file transfer protocol.

Gateway: Device that converts messages from one protocol to another allowing two different networks to communicate.

Gopher: A menu-based system for exploring the Internet.

HTML: HyperText Markup Language. Used to define the various components of a World Wide Web document. HTML tells Web browsers like *Netscape* how to display text.

Hypertext: Documents that contain links to other documents or other areas of the same document. Selecting a link in a hypertext document automatically displays the second location.

IP: Internet Protocol. Allows a packet to transverse multiple networks to find a destination.

LAN: Local area network. A hardware/software combination that allows a group of computers in a limited area to share resources.

Lynx: A cursor-based hypertext browser for exploring the Web. Developed at the University of Kansas by members of the Distributed Computing Group of User Services.

Modem: A piece of equipment that connects a computer to a data transmission line (usually a telephone line).

Mosaic: An Internet navigating tool for exploring the Web.

Multiplexing: As it relates to the modem connection of computer, the ability to process on command while waiting for another to "cue up" and wait for a connection to another site.

Netscape Navigator: An Internet navigating tool for exploring the Web, sometimes called *Netscape*.

NIC: Network Information Center. As close as the Internet gets to a central office.

Packet: A bundle of data. Data on the Internet is broken up into small chunks called packets. When packets arrive at an addressed location, they are reassembled into the original data stream.

PPP: Point to point protocol. A format that allows a computer to use Internet protocols over a serial dial-in connection.

Protocol: A definition of a formal process. For example, a communications protocol allows computers from different manufacturers to talk to each other.

Public domain: A file is said to be in the public domain if it can be downloaded free without restrictions such as shareware fees.

Router: A combined hardware/software system to transfer data between two networks that use the same protocol.

Searching tools: Programs that aide you in finding the many Internet sites you wish to connect to, searchable by name, subject, or category. Some of these are:

- **Archie:** A toll for locating files on publicly accessible sites. The results of an Archie search are the names and directories of files on *anonymous ftp* sites. Access Archie at: *http://www.lerc.nasa.gov/Doc/archieplezhttpd.html*
- **Finger:** Designed to give information about a person with an account on a particular system. From the system prompt enter the command: *talk address*
- **Gopher:** A menu-based system for exploring Internet resources. Accessed files may be on an *anonymous ftp*, library, or database that is accessible only with Gopher. To access Gopher: *http://www.cc.ukans.edu/cwis/ reference/gopher_ resources.html*
- **WAIS:** Access at: *http://www.w3.org/hypertext/DataSources/WAIS/ByHost.html*
- **Whois:** Access from your multiuser account by entering the command: *whois name*

Server: A computer that provides files and other facilities to anyone with proper access and authorization.

Shareware: Software that is distributed for use by the public. If found to be useful, a fee is expected by the developer.

SLIP: Serial Line Internet Protocol. A protocol that allows a computer to use Internet protocols over a serial dial-in connection.

TCP/IP: The Transmission Control Protocol/Internet Protocol. One of the protocols on which the Internet is based.

Telnet: A terminal emulation protocol that allows someone to sign-on to a remote system on the Internet and an application that implements the telnet protocol.

Terminal server: Provides dial-in access to basic services (i.e., telnet) for file transfer and Internet service via PPP of SLIP.

UNIX: A popular operating system developed by AT&T/Bell labs. This is the major system by which Internet servers are programmed. The most common user interfaces on the Internet are character-based menu-choice or command-line interfaces to Unix systems.

vt100: A type of terminal made by Digital Equipment Corp. Many terminal emulation programs provide vt100 emulation.

WAIS: Wide Area Information Services. A system for searching databases across the Internet.

Whois: An application used to access a directory of domain names and addresses using the *Network Information Center* database. To access whois from a multiuser account enter the command: *whois name.* The name may be a registered person, Internet host name, or an organization.

Web (WWW): World Wide Web. A hypertext-based system for finding and accessing Internet resources.

The following are Internet program locaters.

Archie for Macintosh: Via anonymous ftp from *sumex-aim.stanford.edu* in the directory *info-mac/comm.*

Eudora: Commercial version email *eudora-info@qulcomm.com.* No charge version available via anonymous ftp from *sumex-aim.stanford.edu* in the directory *info-mac/comm* or from *ftp.qualcomm.com* in the mac/eudora directory.

PC Eudora: Available via anonymous ftp from *ftp.qualcomm.com.*

Fetch: Via anonymous ftp from *sumex-aim.stanford.edu* in the directory *info-mac* or *dartmouth.edu* in the *pub/mac* directory.

Gopher Book for Widows: *gophbook.zip* is available via anonymous ftp from *sunsite.unc.edu.* Directory *pub/micro/pc-stuff/ms-windows/winsock/apps.*

PC Gopher for DOS: From *sunsite.unc.edu pub/package/gopher/PC_client.*

HGopher for Windows: Available via anonymous ftp from *lister.cc.ic.ac.uk* in the directory *pub/wingopher.*

TurboGopher: Via anonymous ftp from *sumex-aim.stanford.edu* in the directory *info-mac/comm.*

Mosaic: Via anonymous ftp from *ftp.ncsa.uiuc.edu* in the *Mosaic/mosaic-binaries* directory.

NCSA Telnet: Via anonymous ftp from *sumex-sim.stanford.edu* in the directory *info-mac/comm* or from *ftp.nsca.uiuc.edu* in Mac directory.

MacPPP: *pub/pp* directory at *merit.edu.* Combine this with MacTCP to get your Mac directly on the Internet.

PC/TCP Plus for DOS: *info@ftp.com.*

MacTCP: Contact *apda@applelink.apple.com.*

InterSLIP: *pub/sales* directory at *ftp.intercon.com.*

MacSLIP: For information, e-mail *infor@hdepark.com.*

Super-TCP//NFS for Windows: *tcp@frontiertech.com.*

Talk for the Macintosh: Via anonymous ftp from *mac.archive.umich.edu* in the directory *mac/util/comm.*

WAIS-for-Mac: Via anonymous ftp from *ftp/wais.com.* Go to the file *wais-for-mac-1.2-alpha.sea.hqx* from the directory *pub/freeware/mac.*

WinWAIS for Windows: *wnwais21.zip* is available via anonymous ftp from *ridgisd.er.usgs.gov* in the directory *software/wais.*

WWW Browser for Macintosh: In the directory *pub/www/bin/mac* via anonymous ftp from *infor.cern.ch.*

WorldWideWeb Browser: Via anonymous ftp from *info.cern.ch* in the directory *pub.www.bin.next.*

Related Topics

72.3 Local-Area Networks • 92.2 Local Area Networks

References

Hahn, Osborne, and Stout, *The Internet Complete Reference*, New York: McGraw-Hill, 1994.

D. Sachs and H. Stair, *Hands on Internet: A Beginning Guide for PC Users*, Englewood Cliffs, N.J.: Prentice-Hall, 1994.

A. Glossbrenner, *Internet 101: A College Student's Guide*, 2nd ed., New York: McGraw-Hill., 1995.

D. Dern, *The Internet Guide for New Users,* New York: McGraw-Hill, 1994.

B. P. Kohoe, *Zen and the Art of the Internet*, Englewood Cliffs, N.J.: Prentice-Hall.

E. T. L. Hardie and V. Neov, *Internet: Mailing Lists*, Englewood Cliffs, N.J.: Prentice-Hall.

Tummala, R.L. "Section X – Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

This small robotic rover has been sent to Mars inside the Mars Pathfinder lander which was launched from Kennedy Space Center on December 4, 1996. Known as Sojourner, this compact, semi-autonomous six-wheeler will become the first vehicle to traverse the Martian surface.

The rover weighs just 22 lb (10 kg) and when fully deployed is 11 in. (280 mm) high, 25 in. (630 mm) long, and 19 in. (480 mm) wide. It is equipped with three cameras and an alpha proton x-ray spectrometer which must be in contact with rocks or soil to measure elemental composition. A sensor detects alpha particle scattering and proton and x-ray emissions. Sojourner runs on a solar array, sufficient to power the rover for several hours per day, even in the worst dust storms.

The rover control system features operator designation of targets and autonomous control to reach targets and perform tasks. The instructions will be sent to the rover from a controller on earth. The onboard control system is built around an Intel 80C85 processor, which is an 8-bit processor and runs at about 100,000 instructions per second.

Landers and rovers of the future will share the heritage of Mars Pathfinder designs and technologies first tested during this mission. (Photo courtesy of National Aeronautics and Space Administration.)

X

Systems

100 Control Systems *W.L. Brogan, G.K.F. Lee, A.P. Sage, B.C. Kuo, C.L. Phillips, R.D. Harbor, R.G. Jacquot, J.E. McInroy, D.P. Atherton, J.S. Bay, W.T. Baumann, M-Y Chow*
Models • Deynamic Response • Frequency Response Methods: Bode Diagram Approach • Root Locus • Compensation • Digital Control Systems • Nonlinear Control Systems • Optimal Control and Estimation • Neural Control

101 Robotics *T.A. Lasky, T.C. Hsia, R.L. Tummala, N.G. Odrey*
Robot Configuration • Dynamics and Control • Applications

102 Aerospace Systems *C.R. Spitzer, D.A. Martinec, C.T. Leondes, A.H. Rana, W. Check*
Avionics Systems • Communications Satellite Systems: Applications

103 Command, Control and Communications (C^3) *G. Clapp, D. Sworder*
Scope • Background • The Technologies of C^3 • The Dynamics of Encounters • The Role of the Human Decisionmaker in C^3

104 Industrial Systems *G.E. Cook, K. Anderson, R.J. Barnett, A.K. Wallace, R. Spée, M. Sznaier, R.S. Sánchez Peña*
Welding and Bonding • Large Drives • Robust Systems

105 Man-Machine Systems *D. McRuer*
Seven Natures of Man • Machine Control—A Catalog of Behavioral Complexities • Full Attention Compensatory Operations—The Crossover Model

106 Vehicular Systems *L.S. Boehmer*
Design Considerations • Land Transportation Classifications • Propulsion • Microprocessor Controls • Monitoring and Diagnostics

107 Industrial Illuminating Systems *K. Chen*
New Concepts in Designing an Industrial Illuminating System • Factors Affecting Industrial Illumination • System Components • Applications • System Energy Efficiency Considerations

108 Instruments *J.L. Schmalzel*
Physical Variables • Transducers • Instrument Elements • Instrumentation System • Modeling Elements of an Instrumentation System • Summary of Noise Reduction Techniques • Personal Computer-Based Instruments • Modeling PC-Based Instruments • The Effects of Sampling • Other Factors

109 Navigation Systems *M. Kayton*
Coordinate Frames • Categories of Navigation • Dead Reckoning • Radio Navigation • Celestial Navigation • Map Matching Navigation • Navigation Software • Design Trade-Offs

110 Reliability Engineering *R. Ramakumar*
Catastrophic Failure Models • The Bathtub Curve • Mean Time To Failure (MTTF) • Average Failure Rate • *A Posteriori* Failure Probability • Units for Failure Rates • Application of the Binomial Distribution • Application of the Poisson Distribution • The Exponential Distribution • The Weibull Distribution • Combinatorial Aspects • Modeling Maintenance • Markov Models • Binary Model for a Repairable Component • Two Dissimilar Repairable Components • Two Identical Repairable Components • Frequency and Duration Techniques • Applications of Markov Process • Some Useful Approximations • Application Aspects • Reliability and Economics

111 Environmental Effects *K. Blades, B. Allenby*
Industrial Ecology • Design for Environment • Environmental Implications for the Electronics Industry • Emerging Technology • Tools and Strategies for Environmental Design

112 Computer-Aided Control Systems Design *C.M. Rimvall, C.P. Jobling*
A Brief History of CACSD • The State of the Art in CACSD • CACSD Block-Diagram Tools

R. Lal Tummala
Michigan State University

FUNDAMENTAL IDEAS, CONCEPTS, AND TOOLS developed in system theory have contributed significantly to the breakthroughs in aerospace, manufacturing, and medicine, to name a few. In 1990, the National Academy of Engineering identified ten outstanding engineering achievements of the preceding 25 years. These feats included five accomplishments made possible by utilizing modern system theory: Apollo lunar landings, satellites, computer-based manufacturing, computer axial tomography, and the jumbo jet. At present, these ideas are being extended to increase the productivity of the manufacturing sector, improve highway safety, increase fuel efficiency of automobiles, and design and produce environmentally friendly products.

This section discusses conceptual approaches and tools of modern system theory and their applications. The key concepts for the analysis and design of linear and non-linear control systems: modeling, dynamic response, frequency response, root locus, compensation, digital control, describing functions, and phase plane are discussed in Chapter 100. Application of these concepts to a variety of systems is discussed in the following chapters. These systems draw their name from their application, for example, vehicular systems. The topic discussed in Chapter 101 is robotics. A robot is a computer-based mechanical manipulator which can be programmed to perform a variety of tasks. The authors review modeling, control, and application of robots. Chapter 102 describes aerospace systems in avionics and their use in communication satellite systems. The next chapter reviews the command, control, and communication systems used to monitor and control military aerospace systems. Chapter 104 describes two key industrial systems: welding and bonding, and large drives. The authors describe modeling, sensor requirements, control system requirements, and implementation for these systems. Chapter 105 discusses man-machine systems and models used to analyze them. The next two chapters review the key characteristics and electronic controls for vehicular systems, and industrial illumination systems. Chapter 108 describes instruments, which are systems consisting of sensors and electronic circuits, usually for measurement applications. Modern approaches to navigation on the land, sea, or in the air are discussed in Chapter 109. Important topics such as reliability (Chapter 110) and environment (Chapter 111) are included to emphasize their improtance in the design of modern products and processes. With the advent of computer technology, system theory tools are widely available on the computer and the use of this is widespread and thus deserves special attention. Chapter 112 discusses this software.

Brogan, W.L., Lee, G.K.F., Sage, A.P., Kuo, B.C., Phillips, C.L., Harbor, R.D., Jacquot, R.G., McInroy, J.E., Atherton, D.P., Bay, J.S., Baumann, W.T., Chow, M-Y. "Control Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

100

Control Systems

William L. Brogan
University of Nevada, Las Vegas

Gordon K. F. Lee
North Carolina State University

Andrew P. Sage
George Mason University

Benjamin C. Kuo
University of Illinois (Urbana-Champaign)

Charles L. Phillips
Auburn University

Royce D. Harbor
University of West Florida

Raymond G. Jacquot
University of Wyoming

John E. McInroy
University of Wyoming

Derek P. Atherton
University of Sussex

John S. Bay
Virginia Polytechnic Institute and State University

William T. Baumann
Virginia Polytechnic Institute and State University

Mo-Yuen Chow
North Carolina State University

100.1 Models
Classes of Systems to Be Modeled • Two Major Approaches to Modeling • Forms of the Model • Nonuniqueness • Approximation of Continuous Systems by Discrete Models

100.2 Dynamic Response
Computing the Dynamic System Response • Measures of the Dynamic System Response

100.3 Frequency Response Methods: Bode Diagram Approach
Frequency Response Analysis Using the Bode Diagram • Bode Diagram Design-Series Equalizers • Composite Equalizers • Minor-Loop Design

100.4 Root Locus
Root Locus Properties • Root Loci of Digital Control Systems • Design with Root Locus

100.5 Compensation
Control System Specifications • Design • Modern Control Design • Other Modern Design Procedures

100.6 Digital Control Systems
A Simple Example • Single-Loop Linear Control Laws • Proportional Control • PID Control Algorithm • The Closed-Loop System • A Linear Control Example

100.7 Nonlinear Control Systems
The Describing Function Method • The Sinusoidal Describing Function • Evaluation of the Describing Function • Limit Cycles and Stability • Stability and Accuracy • Compensator Design • Closed-Loop Frequency Response • The Phase Plane Method • Piecewise Linear Characteristics • Discussion

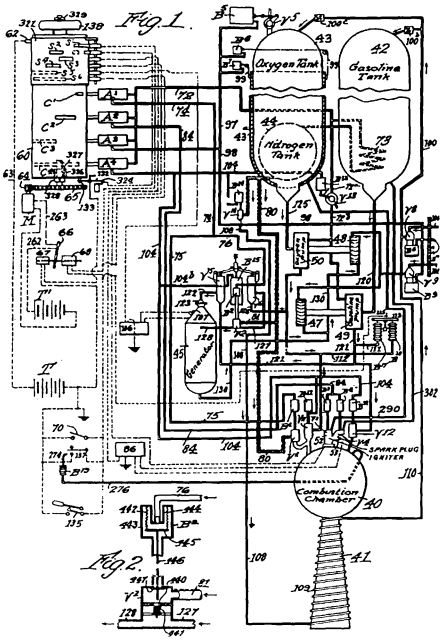
100.8 Optimal Control and Estimation
Linear Quadratic Regulators • Optimal Estimation: The Kalman Filter • Linear-Quadratic-Gaussian (LQG) Control • H_∞ Control • Example • Other Approaches

100.9 Neural Control
Brief Introduction to Artificial Neural Networks • Neural Observer • Neural Control • HVAC Illustration • Conclusion

100.1 Models

William L. Brogan

A naive trial-and-error approach to the design of a control system might consist of constructing a controller, installing it into the system to be controlled, performing tests, and then modifying the controller until satisfactory performance is achieved. This approach could be dangerous and uneconomical, if not impossible. A more rational approach to control system design uses mathematical models. A *model* is a mathematical description of system behavior, as influenced by input variables or initial conditions. The model is a stand-in for the actual system during the control system design stage. It is used to predict performance; to carry out stability, sensitivity, and trade-off



Control Mechanism for Rocket Apparatus

Robert H. Goddard
Patented April 2, 1946
#2,397,657

An excerpt from Robert Goddard's patent application:

This invention relates to rockets and rocket craft which are propelled by combustion apparatus using liquid fuel and a liquid to support combustion, such as liquid oxygen. Such combustion apparatus is disclosed in my prior application Serial No. 327,257 filed April 1, 1940.

It is the general object of my present invention to provide control mechanism by which the necessary operative steps and adjustments for such mechanism will be affected automatically and in predetermined and orderly sequence.

To the attainment of this object, I provide control mechanism which will automatically discontinue flight in a safe and orderly manner.

Dr. Goddard was instrumental in developing rocket propulsion in this country, both solid-fuel rocket engines and later liquid-fuel rocket motors used in missile and spaceflight applications. Goddard died in 1945, before this pivotal patent (filed June 23, 1941) on automatic control of liquid-fuel rockets was granted. He assigned half the rights to the Guggenheim Foundation in New York.

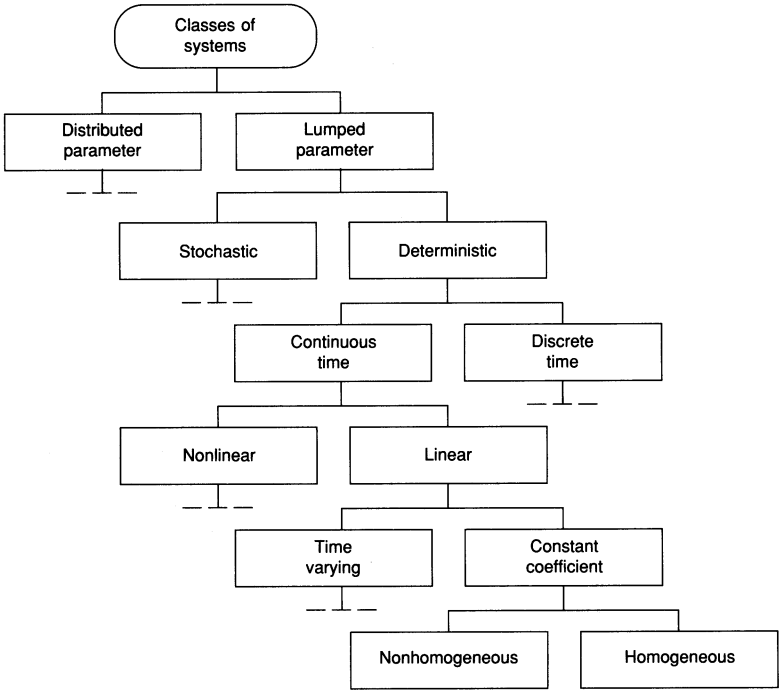


FIGURE 100.1 Major classes of system equations. (*Source:* W.L. Brogan, *Modern Control Theory,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1991, p. 13. With permission.)

studies; and answer various "what-if" questions in a safe and efficient manner. Of course, the validation of the model, and all conclusions derived from it, must ultimately be based upon test results with the physical hardware.

The final form of the mathematical model depends upon the type of physical system, the method used to develop the model, and mathematical manipulations applied to it. These issues are discussed next.

Classes of Systems to Be Modeled

Most control problems are multidisciplinary. The system may consist of electrical, mechanical, thermal, optical, fluidic, or other physical components, as well as economic, biological, or ecological systems. Analogies exist between these various disciplines, based upon the similarity of the equations that describe the phenomena. The discussion of models in this section will be given in mathematical terms and therefore will apply to several disciplines.

Figure 100.1 [Brogan, 1991] shows the classes of systems that might be encountered in control systems modeling. Several branches of this tree diagram are terminated with a dashed line indicating that additional branches have been omitted, similar to those at the same level on other paths.

Distributed parameter systems have variables that are functions of both space and time (such as the voltage along a transmission line or the deflection of a point on an elastic structure). They are described by partial differential equations. These are often approximately modeled as a set of *lumped parameter* systems (described by ordinary differential or difference equations) by using modal expansions, finite element methods, or other approximations [Brogan, 1968]. The lumped parameter continuous-time and discrete-time families are stressed here.

Two Major Approaches to Modeling

In principle, models of a given physical system can be developed by two distinct approaches. Figure 100.2 shows the steps involved in *analytical modeling.* The real-world system is represented by an interconnection of idealized elements. Table 100.1 [Dorf, 1989] shows model elements from several disciplines and their elemental equations. An electrical circuit diagram is a typical result of this physical modeling step (box 3 of Fig. 100.2). Application of the appropriate physical laws (Kirchhoff, Newton, etc.) to the idealized physical model (consisting of point masses, ideal springs, lumped resistors, etc.) leads to a set of mathematical equations. For a circuit these will

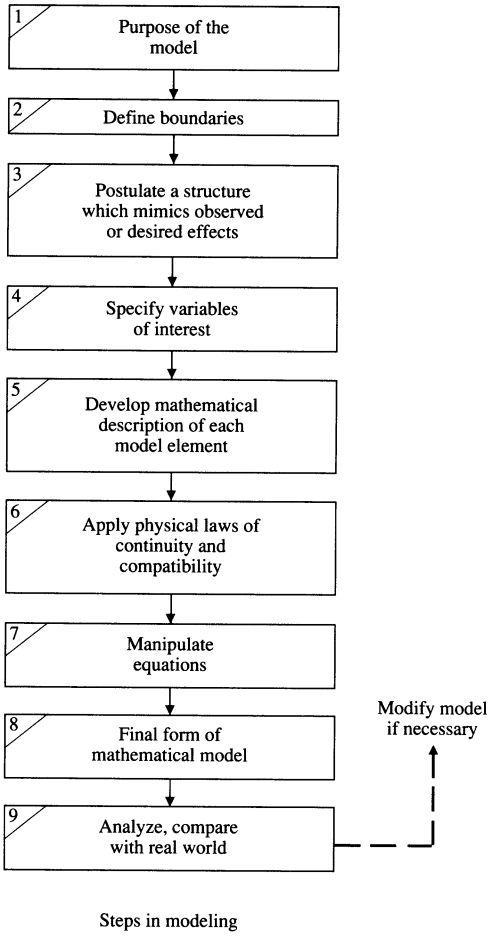


FIGURE 100.2 Modeling considerations. (*Source:* W.L. Brogan, *Modern Control Theory,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1991, p. 5. With permission.)

be mesh or node equations in terms of elemental currents and voltages. Box 6 of Fig. 100.2 suggests a generalization to other disciplines, in terms of continuity and compatibility laws, using through variables (generalization of current that flows through an element) and across variables (generalization of voltage, which has a differential value across an element) [Shearer et al., 1967; Dorf, 1989].

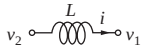
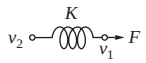
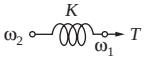
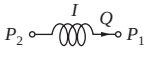
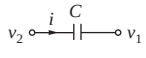
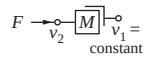
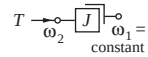
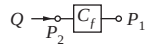
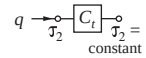
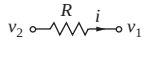
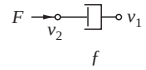
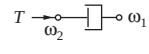
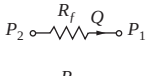
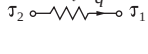
Experimental or *empirical* modeling typically assumes an *a priori* form for the model equations and then uses available measurements to estimate the coefficient values that cause the assumed form to best fit the data. The assumed form could be based upon physical knowledge or it could be just a credible assumption. Time-series models include autoregressive (AR) models, moving average (MA) models, and the combination, called ARMA models. All are difference equations relating the input variables to the output variables at the discrete measurement times, of the form

$$y(k+1) = a_0 y(k) + a_1 y(k-1) + a_2 y(k-2) + \ldots + a_n y(k-n) + b_0 u(k+1) + b_1 u(k) + \ldots + b_p u(k+1-p) + v(k) \quad (100.1)$$

where $v(k)$ is a random noise term. The z-transform transfer function relating u to y is

$$\frac{y(z)}{u(z)} = \frac{b_0 + b_1 z^{-1} + \cdots + b_p z^{-p}}{1 - (a_0 z^{-1} + \cdots + a_{n-1} z^{-n})} = H(z) \quad (100.2)$$

TABLE 100.1 Summary of Describing Differential Equations for Ideal Elements

Type of Element	Physical Element	Describing Equation	Energy E or Power $\mathcal{P}$	Symbol
Inductive storage	Electrical inductance	$v_{21} = L\,\frac{di}{dt}$	$E = \frac{1}{2}\,Li^2$	
	Translational spring	$v_{21} = \frac{1}{K}\,\frac{dF}{dt}$	$E = \frac{1}{2}\,\frac{F^2}{K}$	
	Rotational spring	$\omega_{21} = \frac{1}{K}\,\frac{dT}{dt}$	$E = \frac{1}{2}\,\frac{T^2}{K}$	
	Fluid inertia	$P_{21} = I\,\frac{dQ}{dt}$	$E = \frac{1}{2}\,IQ^2$	
Capacitive storage	Electrical capacitance	$i = C\,\frac{dv_{21}}{dt}$	$E = \frac{1}{2}\,Cv_{21}^2$	
	Translational mass	$F = M\,\frac{dv_2}{dt}$	$E = \frac{1}{2}\,Mv_2^2$	
	Rotational mass	$T = J\,\frac{d\omega_2}{dt}$	$E = \frac{1}{2}\,J\omega_2^2$	
	Fluid capacitance	$Q = C_f\,\frac{dP_{21}}{dt}$	$E = \frac{1}{2}\,C_fP_{21}^2$	
	Thermal capacitance	$q = C_t\,\frac{d\tau_2}{dt}$	$E = C_t\tau_2$	
Energy dissipators	Electrical resistance	$i = \frac{1}{R}\,v_{21}$	$\mathcal{P} = \frac{1}{R}\,v_{21}^2$	
	Translational damper	$F = fv_{21}$	$\mathcal{P} = fv_{21}^2$	
	Rotational damper	$T = f\omega_{21}$	$\mathcal{P} = f\omega_{21}^2$	
	Fluid resistance	$Q = \frac{1}{R_f}\,P_{21}$	$\mathcal{P} = \frac{1}{R_f}\,P_{21}^2$	
	Thermal resistance	$q = \frac{1}{R_t}\,\tau_{21}$	$\mathcal{P} = \frac{1}{R_t}\,\tau_{21}$	

In the MA model all $a_i = 0$. This is alternatively called an all-zero model or a finite impulse response (FIR) model. In the AR model all b_j terms are zero except b_0. This is called an all-pole model or an infinite impulse response (IIR) model. The ARMA model has both poles and zeros and also is an IIR model [Makhoul, 1975].

Adaptive and learning control systems have an experimental modeling aspect. The data fitting is carried out on-line, in real time, as part of the system operation. The modeling described above is normally done off-line [Astrom and Wittenmark, 1989].

Forms of the Model

Regardless of whether a model is developed from knowledge of the physics of the process or from empirical data fitting, it can be further manipulated into several different but equivalent forms. This manipulation is box 7 in Fig. 100.2. The class that is most widely used in control studies is the deterministic lumped-parameter continuous-time constant-coefficient system. A simple example has one input u and one output y. This might be a circuit composed of one ideal source and an interconnection of ideal resistors, capacitors, and inductors.

The equations for this system might consist of a set of mesh or node equations. These could be reduced to a single nth-order linear ordinary differential equation by eliminating extraneous variables.

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_1\frac{dy}{dt} + a_0 y = b_0 u + b_1\frac{du}{dt} + \cdots + b_m\frac{d^m u}{dt^m} \tag{100.3}$$

This nth-order equation can be replaced by an input-output transfer function

$$\frac{Y(s)}{U(s)} = H(s) = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} \tag{100.4}$$

The inverse Laplace transform $\mathcal{L}^{-1}\{H(s)\} = h(t)$ is the system impulse response function. Alternatively, by selecting a set of n internal **state variables,** Eq.(100.3) can be written as a coupled set of first-order differential equations plus an algebraic equation relating the states to the original output y. These equations are called state equations, and one possible choice for this example is, assuming $m = n$,

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} -a_{n-1} & 1 & 0 & 0 & \cdots & 0 \\ -a_{n-2} & 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_1 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} b_{n-1} - a_{n-1}b_n \\ b_{n-2} - a_{n-2}b_n \\ \vdots \\ b_1 - a_1 b_n \\ b_0 - a_0 b_n \end{bmatrix} u(t)$$

and

$$y(t) = [1 \;\; 0 \;\; 0 \;\; \ldots \;\; 0]\mathbf{x}(t) + b_n u(t) \tag{100.5}$$

In matrix notation these are written more succinctly as

$$\dot{\mathbf{x}} = A\mathbf{x} + Bu \quad \text{and} \quad y = C\mathbf{x} + Du \tag{100.6}$$

Any one of these six possible model forms, or others, might constitute the result of box 8 in Fig. 100.2. Discrete-time system models have similar choices of form, including an nth-order difference equation as given in Eq. (100.1) or a z-transform input-output transfer function as given in Eq. (100.2). A set of n first-order difference equations (state equations) analogous to Eq. (100.5) or (100.6) also can be written.

Extensions to systems with r inputs and m outputs lead to a set of m coupled equations similar to Eq. (100.3), one for each output y_i. These higher-order equations can be reduced to n first-order state differential equations and m algebraic output equations as in Eq. (100.5) or (100.6). The **A** matrix is again of dimension $n \times n$, but **B** is now $n \times r$, **C** is $m \times n$, and **D** is $m \times r$. In all previous discussions, the number of state variables, n, is the order of the model. In transfer function form, an $m \times r$ matrix $H(s)$ of transfer functions will describe the input-output behavior

$$Y(s) = H(s)U(s) \tag{100.7}$$

Other transfer function forms are also applicable, including the left and right forms of the matrix fraction description (MFD) of the transfer functions [Kailath, 1980]

$$H(s) = P(s)^{-1}N(s) \quad \text{or} \quad H(s) = N(s)P(s)^{-1} \tag{100.8}$$

Both **P** and **N** are matrices whose elements are polynomials in s. Very similar model forms apply to continuous-time and discrete-time systems, with the major difference being whether Laplace transform or z-transform transfer functions are involved.

When time-variable systems are encountered, the option of using high-order differential or difference equations versus sets of first-order state equations is still open. The system coefficients $a_i(t)$, $b_j(t)$ and/or the matrices $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\mathbf{C}(t)$, and $\mathbf{D}(t)$ will now be time-varying. Transfer function approaches lose most of their utility in time-varying cases and are seldom used. With nonlinear systems all the options relating to the order and number of differential or difference equation still apply.

The form of the nonlinear state equations is

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}, \mathbf{t})$$

$$y = h(\mathbf{x}, \mathbf{u}, \mathbf{t}) \tag{100.9}$$

where the nonlinear vector-valued functions $f(\mathbf{x}, \mathbf{u}, \mathbf{t})$ and $h(\mathbf{x}, \mathbf{u}, \mathbf{t})$ replace the right-hand sides of Eq. (100.6). The transfer function forms are of no value in nonlinear cases.

Stochastic systems [Maybeck, 1979] are modeled in similar forms, except the coefficients of the model and/or the inputs are described in probabilistic terms.

Nonuniqueness

There is not a unique correct model of a given system for several reasons. The selection of idealized elements to represent the system requires judgment based upon the intended purpose. For example, a satellite might be modeled as a point mass in a study of its gross motion through space. A detailed flexible structure model might be required if the goal is to control vibration of a crucial on-board sensor. In empirical modeling, the assumed starting form, Eq. (100.1), can vary.

There is a trade-off between the complexity of the model form and the fidelity with which it will match the data set. For example, a pth-degree polynomial can exactly fit to $p + 1$ data points, but a straight line might be a better model of the underlying physics. Deviations from the line might be caused by extraneous measurement noise. Issues such as these are addressed in Astrom [1980].

The preceding paragraph addresses nonuniqueness in determining an input-output system description. In addition, state models developed from input-output descriptions are not unique. Suppose the transfer function of a single-input, single-output linear system is known exactly. The state variable model of this system is not unique for at least two reasons. An arbitrarily high-order state variable model can be found that will have this same transfer function. There is, however, a unique minimal or irreducible order $n_{\min}$ from among all state models that have the specified transfer function. A state model of this order will have the desirable properties of **controllability** and **observability.** It is interesting to point out that the minimal order may be less than the actual order of the physical system.

The second aspect of the nonuniqueness issue relates not to order, i.e., the *number* of state variables, but to *choice* of internal variables (state variables). Mathematical and physical methods of selecting state variables are available [Brogan, 1991]. An infinite number of choices exist, and each leads to a different set $\{A, B, C, D\}$, called a realization. Some state variable model forms are more convenient for revealing key system properties such as stability, controllability, observability, **stabilizability,** and **detectability.** Common forms include the controllable canonical form, the observable canonical form, the Jordan canonical form, and the Kalman canonical form.

The reverse process is unique in that every valid realization leads to the same model transfer function

$$H(s) = C\{sI - A\}^{-1}B + D \tag{100.10}$$

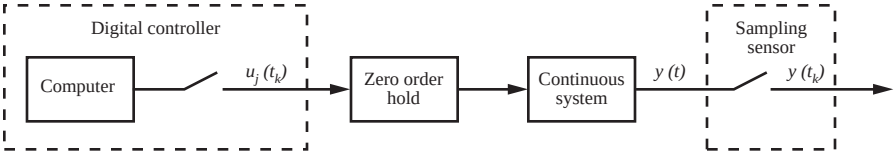


FIGURE 100.3 Digital output provided by modern sensor.

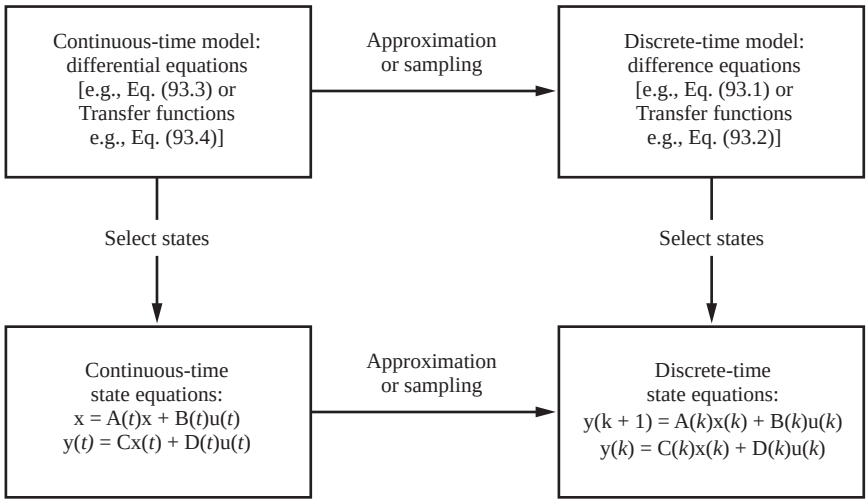


FIGURE 100.4 State variable modeling paradigm.

Approximation of Continuous Systems by Discrete Models

Modern control systems often are implemented digitally, and many modern sensors provide digital output, as shown in Fig. 100.3. In designing or analyzing such systems discrete-time approximate models of continuous-time systems are frequently needed. There are several general ways of proceeding, as shown in Fig. 100.4. Many choices exist for each path on the figure. Alternative choices of states or of approximation methods, such as forward or backward differences, lead to an infinite number of valid models.

Defining Terms

Controllability: A property that in the linear system case depends upon the **A,B** matrix pair which ensures the existence of some control input that will drive any arbitrary initial state to zero in finite time.

Detectability: A system is detectable if all its unstable modes are observable.

Observability: A property that in the linear system case depends upon the **A,C** matrix pair which ensures the ability to determine the initial values of all states by observing the system outputs for some finite time interval.

Stabilizable: A system is stabilizable if all its unstable modes are controllable.

State variables: A set of variables that completely summarize the system's status in the following sense. If all states x_i are known at time t_0, then the values of all states and outputs can be determined uniquely for any time $t_1 > t_0$, provided the inputs are known from t_0 onward. State variables are components in the state vector. State space is a vector space containing the state vectors.

Related Topic

6.1 Definitions and Properties

References

K.J. Astrom, "Maximum likelihood and prediction error methods," *Automatica,* vol. 16, pp. 551–574, 1980.

K.J. Astrom and B. Wittenmark, *Adaptive Control,* Reading, Mass.: Addison-Wesley, 1989.

W.L. Brogan, "Optimal control theory applied to systems described by partial differential equations," in *Advances in Control Systems,* vol. 6, C. T. Leondes (ed.), New York: Academic Press, 1968, chap. 4.

W.L. Brogan, *Modern Control Theory,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1991.

R.C. Dorf, *Modern Control Systems,* 5th ed., Reading, Mass.: Addison-Wesley, 1989.

T. Kailath, *Linear Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE,* vol. 63, no. 4, pp. 561–580, 1975.

P.S. Maybeck, *Stochastic Models, Estimation and Control,* vol. 1, New York: Academic Press, 1979.

J.L. Shearer, A.T. Murphy, and H.H. Richardson, *Introduction to Dynamic Systems,* Reading, Mass.: Addison-Wesley, 1967.

Further Information

The monthly *IEEE Control Systems Magazine* frequently contains application articles involving models of interesting physical systems.

The monthly *IEEE Transactions on Automatic Control* is concerned with theoretical aspects of systems. Models as discussed here are often the starting point for these investigations.

Automatica is the source of many related articles. In particular an extended survey on system identification is given by Astrom and Eykhoff in vol. 7, pp. 123–162, 1971.

Early developments of the state variable approach are given by R. E. Kalman in "Mathematical description of linear dynamical systems," *SIAM J. Control Ser.,* vol. A1, no. 2, pp. 152–192, 1963.

100.2 Dynamic Response

Gordon K. F. Lee

Computing the Dynamic System Response

Consider a linear time-invariant dynamic system represented by a differential equation form

$$\frac{d^n y(t)}{dt^n} + a_{n-1}\frac{d^{n-1} y(t)}{dt^{n-1}} + \cdots + a_1\frac{dy(t)}{dt} + a_0 y(t) = b_m\frac{d^m f(t)}{dt^m} + \cdots + b_1\frac{df(t)}{dt} + b_0 f(t) \tag{100.11}$$

where $y(t)$ and $f(t)$ represent the output and input, respectively, of the system.

Let $p^k(\cdot) \triangleq (d^k/dt^k)(\cdot)$ define the differential operator so that (100.11) becomes

$$(p^n + a_{n-1}p^{n-1} + \ldots + a_1 p + a_0)y(t) = (b_m p^m + \ldots + b_1 p + b_0)f(t) \tag{100.12}$$

The solution to (100.11) is given by

$$y(t) = y_S(t) + y_I(t) \tag{100.13}$$

where $y_S(t)$ is the **zero-input response,** or that part of the response due to the initial conditions (or states) only, and $y_I(t)$ is the **zero-state response,** or that part of the response due to the input $f(t)$ only.

Zero-Input Response: $y_S(t)$

Here $f(t) = 0$, and thus (100.11) becomes

$$(p^n + a_{n-1}p^{n-1} + \ldots + a_1 p + a_0)y(t) = 0 \tag{100.14}$$

That is,

$$D(p)y(t) = 0$$

The roots of $D(p) = 0$ can be categorized as either distinct or multiple. That is, in general,

$$D(p) = \prod_{i=1}^{q} (p - \lambda_i)^{k_i} \prod_{i=1}^{r} (p - \lambda_{q+i})$$

where there are r distinct roots and q sets of multiple roots (each set has multiplicity k_i). Note that $r + \sigma = n$, where $\sigma \triangleq \sum_{i=1}^{q} k_i$. Each distinct root contributes a term to $y_S(t)$ of the form $c_i e^{\lambda_i t}$, where c_i is a constant, while each set of multiple roots contributes a set of terms to $y_S(t)$ of the form $\sum_{j=0}^{k_i-1} c_{i,j} t^j e^{\lambda_i t}$, where $c_{i,j}$ is some constant. Thus, the zero-input response is given by

$$y_S(t) = \sum_{i=1}^{q} \sum_{j=0}^{k_i-1} c_{i,j} t^j e^{\lambda_i t} + \sum_{i=1}^{r} c_{\sigma+i} e^{\lambda_{\sigma+i} t} \tag{100.15}$$

The coefficients $c_{i,j}$ and $c_{\sigma+i}$ are selected to satisfy the initial conditions.

Special Case

If all the roots of $D(p) = 0$ are distinct and the initial conditions for (100.11) are given by

$$\left\{ y(0), \frac{dy(0)}{dt}, \ldots, \frac{d^{n-1}y(0)}{dt^{n-1}} \right\}$$

then the coefficients of (100.15) are given by the solution of

$$\begin{bmatrix} y(0) \\ \dfrac{dy(0)}{dt} \\ \vdots \\ \dfrac{d^{n-1}y(0)}{dt^{n-1}} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{(n-1)} & \lambda_2^{(n-1)} & \cdots & \lambda_n^{(n-1)} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \tag{100.16}$$

Zero-State Response: $y_I(t)$

Here the initial conditions are made identically zero. Observing (100.11), let

$$H(p) = \frac{b_m p^m + \cdots + b_1 p + b_0}{p^n + a_{n-1}p^{n-1} + \cdots + a_1 p + a_0}$$

denote a rational function in the p operator. Consider using partial-fraction expansion on $H(p)$ as

$$H(p) = \sum_{i=1}^{q} \sum_{j=1}^{k_i} \frac{g_{i,j}}{(p - \lambda_i)^j} + \sum_{i=1}^{r} \frac{g_{\sigma+i}}{p - \lambda_{q+i}} \tag{100.17}$$

when the first term corresponds to the sets of multiple roots and the second term corresponds to the distinct roots.

Note the constant residuals are computed as

$$g_{\sigma+i} = [(p - \lambda_{q+i})H(p)]_{p = \lambda q+i}$$

and

$$g_{i,j} = \frac{1}{(k_i - j)!} \frac{d^{(k_i-j)}}{dp^{(k_i-j)}} \left\{(p - \lambda_i)^{k_i} H(p)\right\} \Big|_{p=\lambda_i}$$

Then

$$h(t) = \sum_{i=1}^{q} \sum_{j=1}^{k_i} \frac{g_{i,j}}{(j-1)!} t^{j-1} e^{\lambda_i t} + \sum_{i=1}^{r} g_{\sigma+i} e^{\lambda_{\sigma+i} t} \tag{100.18}$$

is the **impulse response** of the system (100.11). Then the zero-state response is given by

$$y_I(t) = \int_0^t f(\tau) h(t - \tau)\, d\tau \tag{100.19}$$

that is, $y_I(t)$ is the time convolution between input $f(t)$ and impulse response $h(t)$. In some instances, it may be easier to find $y_s(t)$ and $y_I(t)$ using Laplace Transform methods.

Measures of the Dynamic System Response

Several measures may be employed to investigate dynamic response performance. These include:

1. Speed of the response—how quickly does the system reach its final value
2. Accuracy—how close is the final response to the desired response
3. Relative stability—how stable is the system or how close is the system to instability
4. Sensitivity—what happens to the system response if the system parameters change

Objectives 3 and 4 may be analyzed by frequency domain methods (Section 100.3). Time-domain measures classically analyze the dynamic response by partitioning the total response into its steady-state (objective 2) and transient (objective 1) components. The **steady-state response** is that part of the response which remains as time approaches infinity; the **transient response** is that part of the response which vanishes as time approaches infinity.

Measures of the Steady-State Response

In the steady state, the accuracy of the time response is an indication of how well the dynamic response follows a desired time trajectory. Usually a test signal (reference signal) is selected to measure accuracy. Consider Fig. 100.5. In this configuration, the objective is to force $y(t)$ to track a reference signal $r(t)$ as close as possible.

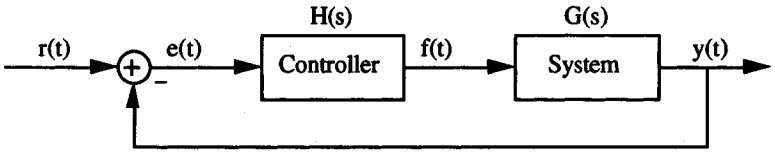


FIGURE 100.5 A tracking controller configuration.

TABLE 100.2 Steady-State Error Constants

$r(t)$ Test Signal	$eSS(t)$	Error Constant
$Ru(t)$: step function	$\frac{R}{1+K_p}$	$K_p = \lim_{s\to 0} G(s)H(s)$
$Rtu(t)$: ramp function	$\frac{R}{K_v}$	$K_v = \lim_{s\to 0} sG(s)H(s)$
$\frac{R}{2}t^2u(t)$: parabolic function	$\frac{R}{K_a}$	$K_a = \lim_{s\to 0} s^2G(s)H(s)$

The **steady-state error** is a measure of the accuracy of the output $y(t)$ in tracking the reference input $r(t)$. Other configurations with different performance measures would result in other definitions of the steady-state error between two signals.

From Fig. 100.5, the error $e(t)$ is

$$e(t) = r(t) - y(t) \tag{100.20}$$

and the steady-state error is

$$e_{SS}(t) = \lim_{t\to\infty} e(t) = \lim_{s\to\infty} sE(s) \tag{100.21}$$

assuming the limits exists, where $E(s)$ is the Laplace transform of $e(t)$, and s is the Laplacian operator. With $G(s)$ the transfer function of the system and $H(s)$ the transfer function of the controller, the transfer function between $y(t)$ and $r(t)$ is found to be

$$T(s) = \frac{G(s)H(s)}{1 + G(s)H(s)} \tag{100.22}$$

with

$$E(s) = \frac{R(s)}{1 + G(s)H(s)} \tag{100.23}$$

Direct application of the steady-state error for various inputs yields Table 100.2. Note $u(t)$ is the unit step function. This table can be extended to an mth-order input in a straightforward manner. Note that for $e_{SS}(t)$ to go to zero with a reference signal $Ct^m u(t)$, the term $G(s)H(s)$ must have at least m poles at the origin (a type m system).

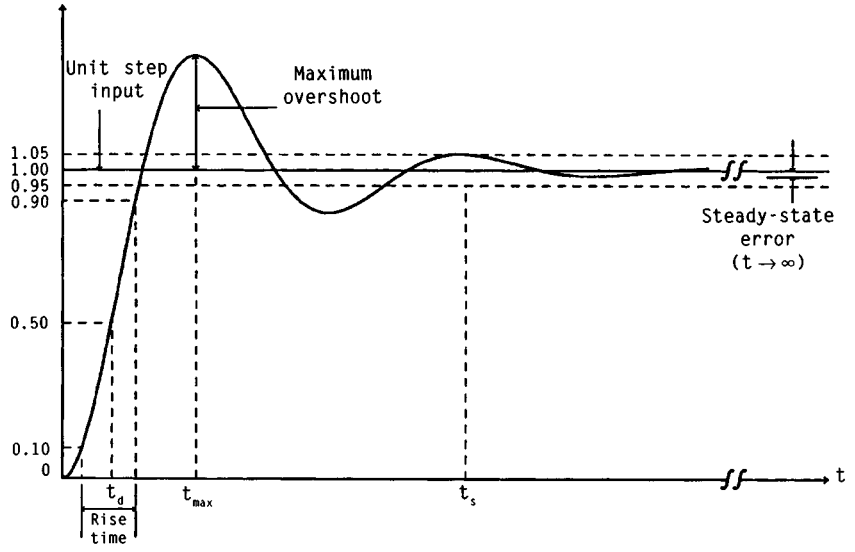


FIGURE 100.6 Step response.

Measures of the Transient Response

In general, analysis of the transient response of a dynamic system to a reference input is difficult. Hence formulating a standard measure of performance becomes complicated. In many cases, the response is dominated by a pair of poles and thus acts like a second-order system.

Consider a reference unit step input to a dynamic system (Fig. 100.6). Critical parameters that measure transient response include:

1. *M:* maximum overshoot
2. % overshoot = $M/A \times 100\%$, where A is the final value of the time response
3. t_d: delay time—the time required to reach 50% of A
4. t_r: rise time—the time required to go from 10% of A to 90% of A
5. t_s: settling time—the time required for the response to reach and stay within 5% of A

To calculate these measures, consider a second-order system

$$T(s) = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} \tag{100.24}$$

where ξ is the damping coefficient and ω_n is the natural frequency of oscillation.

For the range $0 < \xi < 1$, the system response is *underdamped,* resulting in a damped oscillatory output. For a unit step input, the response is given by

$$y(t) = 1 + \frac{e^{-\xi\omega_n t}}{\sqrt{1-\xi^2}} \sin\left(\omega_n\sqrt{1-\xi^2}\,t - \tan^{-1}\frac{\sqrt{1-\xi^2}}{-\xi}\right) \quad (0 < \xi < 1) \tag{100.25}$$

The eigenvalues (poles) of the system [roots of the denominator of $T(s)$] provide some measure of the time constants of the system. For the system under study, the eigenvalues are at

$$-\xi\omega_n \pm j\omega_n\sqrt{1-\xi^2} \qquad \text{where} \qquad j \triangleq \sqrt{-1}$$

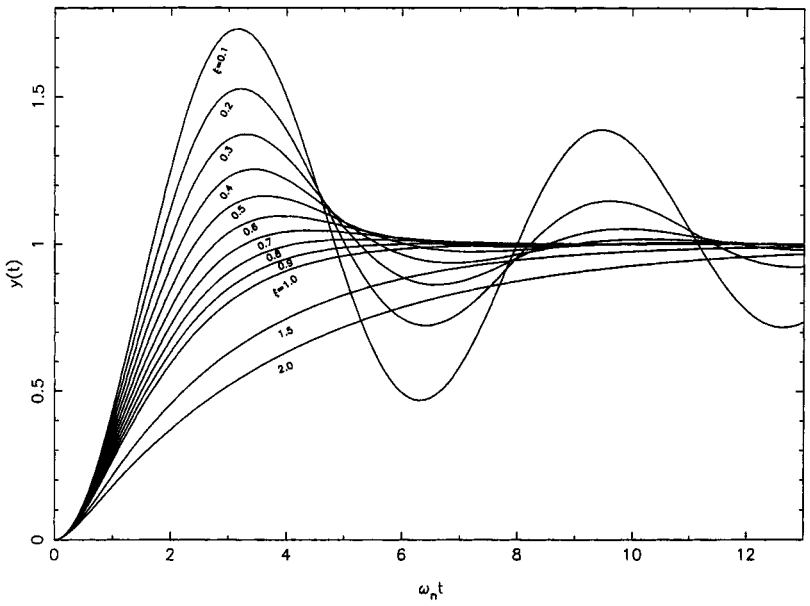


FIGURE 100.7 Effect of the damping coefficient on the dynamic response.

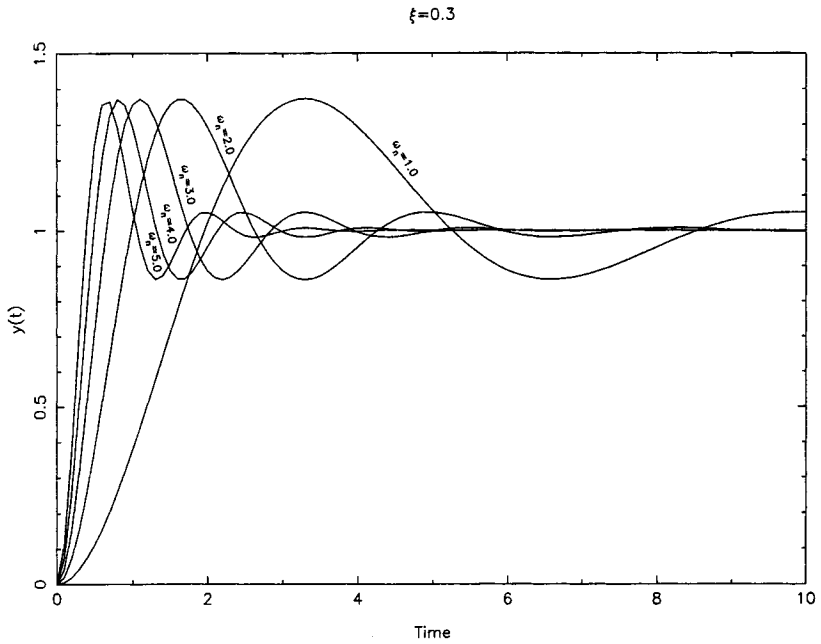


FIGURE 100.8 Effect of the natural frequency of oscillation on the dynamic response.

From the expression of $y(t)$, one sees that the term $\xi\omega_n$ affects the rise time and exponential decay time.

The effects of the damping coefficient on the transient response are seen in Fig. 100.7.

The effects of the natural frequency of oscillation ω_n of the transient response can be seen in Fig. 100.8. As ω_n increases, the frequency of oscillation increases.

For the case when $0 < \xi < 1$, the underdamped case, one can analyze the critical transient response parameters.

To measure the peaks of Fig. 100.6, one finds

$$y_{\text{peak}}(t) = 1 + (-1)^{n-1} \exp \frac{-n\pi\xi}{\sqrt{1-\xi^2}} \qquad n = 0, 1, \ldots \tag{100.26}$$

occurring at

$$t = \frac{n\,\pi}{\omega_n\sqrt{1-\xi^2}} \qquad \begin{matrix} n\text{: odd (overshoot)} \\ n\text{: even (undershoot)} \end{matrix} \tag{100.27}$$

Hence

$$y_{\max} = 1 + \exp \frac{-\pi\xi}{\sqrt{1-\xi^2}} \tag{100.28}$$

occurring at

$$t_{\max} = \frac{\pi}{\omega_n\sqrt{1-\xi^2}} \tag{100.29}$$

With these parameters, one finds

$$t_d \approx \frac{1 + 0.7\xi}{\omega_n}$$

$$t_r \approx \frac{1 + 1.1\xi + 1.4\xi^2}{\omega_n}$$

and

$$t_s \approx \frac{3}{\xi\omega_n}$$

Note that increasing ξ decreases the % overshoot and decreases the settling time but increases t_d and t_r.

When $\xi = 1$, the system has a double pole at $-\omega_n$, resulting in a *critically damped* response. This is the point when the response just changes from oscillatory to exponential in form. For a unit step input, the response is given by

$$y(t) = 1 - e^{-\omega nt}(1 + \omega_n t) \quad (\xi = 1) \tag{100.30}$$

For the range $\xi > 1$, the system is overdamped due to two real system poles. For a unit step input, the response is given by

$$y(t) = 1 + \frac{1}{c_1 - c_2}\left(\frac{1}{c_1}\, e^{c_1\omega_n t} - \frac{1}{c_2}\, e^{c_2\omega_n t}\right) \quad (\xi > 1)$$

$$c_1 = -\xi + \sqrt{\xi^2 - 1} \qquad c_2 = -\xi - \sqrt{\xi^2 - 1} \tag{100.31}$$

Finally, when $\xi = 0$, the response is purely sinusoidal. For a unit step, the response is given by

$$y(t) = 1 - \cos \omega_n t \quad (\xi = 0) \tag{100.32}$$

Defining Terms

Impulse response: The response of a system when the input is an impulse function.
Steady-state error: The difference between the desired reference signal and the actual signal in steady-state, i.e., when time approaches infinity.
Steady-state response: That part of the response which remains as time approaches infinity.
Transient response: That part of the response which vanishes as time approaches infinity.
Zero-input response: That part of the response due to the initial condition only.
Zero-state response: That part of the response due to the input only.

Related Topics

6.1 Definitions and Properties • 7.1 Introduction • 112.2 A Brief History of CACSD

Further Information

J.J. D'Azzo and C.H. Harpis, *Linear Control System Analysis and Design,* New York: McGraw-Hill, 1981.
R.C. Dorf, *Modern Control Systems,* 5th ed., Reading, Mass.: Addison-Wesley, 1989.
M.E. El-Hawary, *Control Systems Engineering,* Reston, Va.: Reston, 1984.
G.H. Hostetter, C. J. Savant, Jr., and R. T. Stefani, *Design of Feedback Control Systems,* Philadelphia: Saunders, 1989.
B.C. Kuo, *Automatic Control Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.
K. Ogata, *Modern Control Engineering,* Englewood Cliffs, N.J.: Prentice-Hall, 1970.
N.K. Sinha, *Control Systems,* New York: Holt, 1986.

100.3 Frequency Response Methods: Bode Diagram Approach

Andrew P. Sage

Our efforts in this section are concerned with analysis and design of linear control systems by frequency response methods. Design generally involves trial-and-error repetition of analysis until a set of design **specifications** has been met. Thus, analysis methods are most useful in the design process, which is one phase of the **systems engineering** life cycle [Sage, 1992]. We will discuss one design method based on **Bode diagrams.** We will discuss the use of both simple **series equalizers** and composite equalizers as well as the use of minor-loop feedback in systems design.

Figure 100.9 presents a flowchart of the frequency response method design process and indicates the key role of analysis in linear systems control design. The flowchart of Fig. 100.9 is applicable to control system design methods in general. There are several iterative loops, generally calling for trial-and-error efforts, that comprise the suggested design process. An experienced designer will often be able, primarily due to successful prior experience, to select a system structure and generic components such that the design specifications can be met with no or perhaps a very few iterations through the iterative loop involving adjustment of equalizer or compensation parameters to best meet specifications.

If the parameter optimization, or parameter refinement such as to lead to maximum phase margin, approach shows the specifications cannot be met, we are then assured that no **equalizer** of the specific form selected will meet specifications. The next design step, if needed, would consist of modification of the equalizer form or

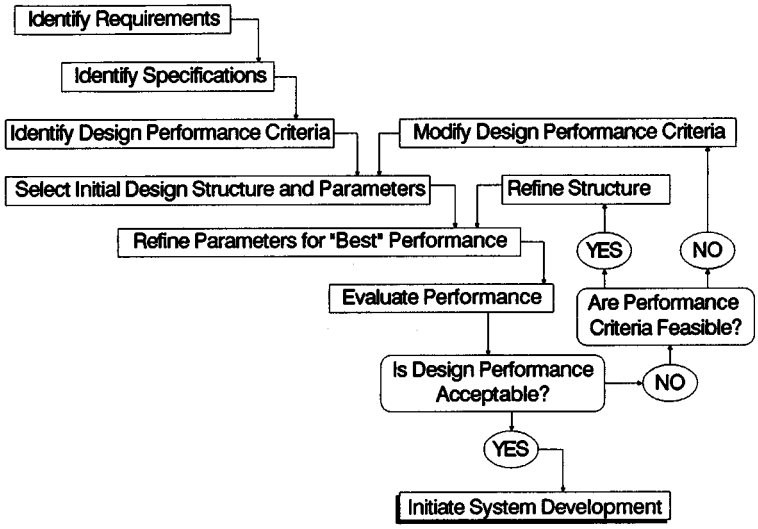


FIGURE 100.9 System design life cycle for frequency-response-based design.

structure and repetition of the analysis process to determine equalizer parameter values to best meet specifications. If specifications still cannot be met, we will usually next modify generic fixed components used in the system. This iterative design and analysis process is again repeated. If no reasonable fixed components can be obtained to meet specifications, then structural changes in the proposed system are next contemplated. If no structure can be found that allows satisfaction of specifications, either the client must be requested to relax the frequency response specifications or the project may be rejected as infeasible using present technology. As we might suspect, economics will play a dominant role in this design process. Changes made due to iteration in the inner loops of Fig. 100.9 normally involve little additional costs, whereas those made due to iterations in the outer loops will often involve major cost changes.

Frequency Response Analysis Using the Bode Diagram

The steady-state response of a stable linear constant-coefficient system has particular significance, as we know from an elementary study of electrical networks and circuits and of dynamics. We consider a stable linear system with input-output transfer function

$$H(s) = \frac{Z(s)}{U(s)}$$

We assume a sinusoidal input $u(t) = \cos \omega t$ so that we have for the Laplace transform of the system output

$$Z(s) = \frac{sH(s)}{s^2 + \omega^2}$$

We expand this ratio of polynomials using the partial-fraction approach and obtain

$$Z(s) = F(s) + \frac{a_1}{s + j\omega} + \frac{a_2}{s - j\omega}$$

In this expression, $F(s)$ contains all the poles of $H(s)$. All of these lie in the left half plane since the system, represented by $H(s)$, is assumed to be stable. The coefficients a_1 and a_2 are easily determined as

$$a_1 = \frac{H(-j\omega)}{2}$$

$$a_2 = \frac{H(j\omega)}{2}$$

We can represent the complex transfer function $H(j\omega)$ in either of two forms,

$$H(j\omega) = B(\omega) + jC(\omega)$$

$$H(-j\omega) = B(\omega) - jC(\omega)$$

The inverse Laplace transform of the system transfer function will result in a transient term due to the inverse transform of $F(s)$, which will decay to zero as time progresses. A steady-state component will remain, and this is, from the inverse transform of the system equation, given by

$$z(t) = a_1 e^{-j\omega t} + a_2 e^{j\omega t}$$

We combine several of these relations and obtain the result

$$z(t) = B(\omega)\left(\frac{e^{j\omega t} + e^{-j\omega t}}{2}\right) - C(\omega)\left(\frac{e^{j\omega t} - e^{-j\omega t}}{2j}\right)$$

This result becomes, using the Euler identity,[1]

$$\begin{aligned} z(t) &= B(\omega)\cos\omega t - C(\omega)\sin\omega t \\ &= [B^2(\omega) + C^2(\omega)]^{1/2}\cos(\omega + \beta) \\ &= |H(j\omega)|\cos(\omega t + \beta) \end{aligned}$$

where $\tan\beta(\omega) = C(\omega)/B(\omega)$.

As we see from this last result, there is a very direct relationship between the transfer function of a linear constant-coefficient system, the time response of a system to any known input, and the sinusoidal steady-state response of the system. We can always determine any of these if we are given any one of them. This is a very important result. This important conclusion justifies a design procedure for linear systems that is based only on sinusoidal steady-state response, as it is possible to determine transient responses, or responses to any given system input, from a knowledge of steady-state sinusoidal responses, at least in theory. In practice, this might be rather difficult computationally without some form of automated assistance.

Bode Diagram Design-Series Equalizers

In this subsection we consider three types of series equalization:

1. Gain adjustment, normally attenuation by a constant at all frequencies
2. Increasing the phase lead, or reducing the phase lag, at the **crossover frequency** by use of a phase **lead network**
3. Attenuation of the gain at middle and high frequencies such that the crossover frequency will be decreased to a lower value where the phase lag is less, by use of a **lag network**

[1]The Euler identity is $e^{j\omega t} = \cos\omega t + j\sin\omega t$.

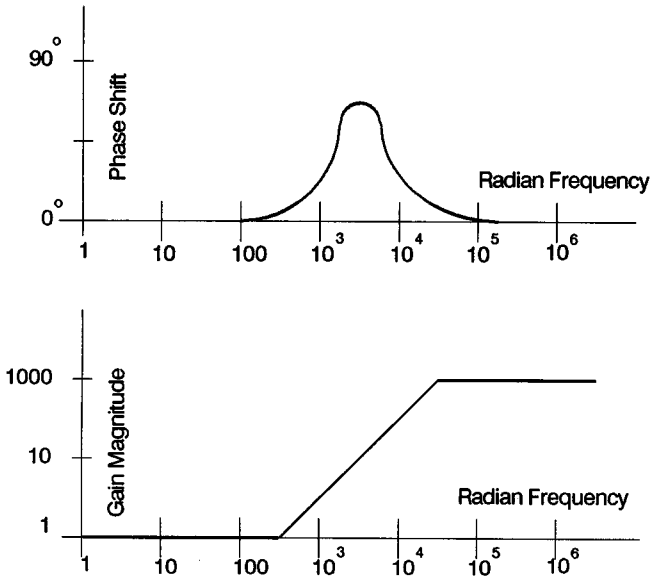


FIGURE 100.10 Phase shift and gain curves for a simple lead network.

In the subsection that follows this, we will first consider use of a composite or **lag-lead network** near crossover to attenuate gain only to reduce the crossover frequency to a value where the phase lag is less. Then we will consider more complex composite equalizers and state some general guidelines for Bode diagram design. Here, we will use Bode diagram frequency domain design techniques to develop a design procedure for each of three elementary types of series equalization.

Gain Reduction

Many linear control systems can be made sufficiently stable merely by reduction of the open-loop system gain to a sufficiently low value. This approach ignores all performance specifications, however, except that of phase margin (PM) and is, therefore, usually not a satisfactory approach. It is a very simple one, however, and serves to illustrate the approach to be taken in more complex cases.

The following steps constitute an appropriate Bode diagram design procedure for compensation by gain adjustment:

1. Determine the required PM and the corresponding phase shift $\beta_c = -\pi + \text{PM}$.
2. Determine the frequency ω_c at which the phase shift is such as to yield the phase shift at crossover required to give the desired PM.
3. Adjust the gain such that the actual crossover frequency occurs at the value computed in step 2.

Phase-Lead Compensation

In compensation using a phase-lead network, we increase the phase lead at the crossover frequency such that we meet a performance specification concerning phase shift. A phase-lead-compensating network transfer function is

$$G_c(s) = \left(1 + \frac{s}{\omega_1}\right) \Big/ \left(1 + \frac{s}{\omega_2}\right) \qquad \omega_1 < \omega_2$$

Figure 100.10 illustrates the gain versus frequency and phase versus frequency curves for a simple lead network with the transfer function of the foregoing equation. The maximum phase lead obtainable from a phase-lead network depends upon the ratio ω_2/ω_1 that is used in designing the network. From the expression for the phase shift of the transfer function for this system, which is given by

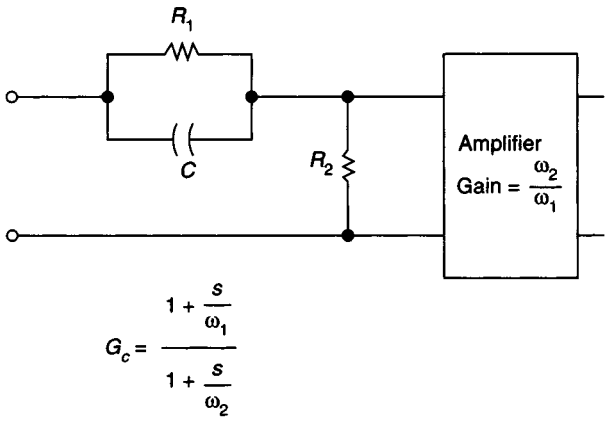


FIGURE 100.11 A simple electrical lead network.

$$\beta = \tan^{-1}\frac{\omega}{\omega_1} - \tan^{-1}\frac{\omega}{\omega_2}$$

we see that the maximum amount of phase lead occurs at the point where the first derivative with respect to frequency is zero, or

$$\left.\frac{d\beta}{d\omega}\right|_{\omega=\omega_m} = 0$$

or at the frequency where

$$\omega_m = (\omega_1\omega_2)^{0.5}$$

This frequency is easily seen to be at the center of the two break frequencies for the lead network on a Bode log asymptotic gain plot. It is interesting to note that this is exactly the same frequency that we would obtain using an arctangent approximation[2] with the assumption that $\omega_1 < \omega < \omega_2$.

There are many ways of realizing a simple phase-lead network. All methods require the use of an active element since the gain of the lead network at high frequencies is greater than 1. A simple electrical network realization is shown in Fig. 100.11.

We now consider a simple design example. Suppose that we have an open-loop system with transfer function

$$G_f(s) = \frac{10^4}{s^2}$$

It turns out that this is often called a type-two system due to the presence of the double integration. This system will have a steady-state error of zero for a constant acceleration input. The crossover frequency, that is to say

[2]The arctangent approximation is $\tan^{-1}(\omega/\alpha) = \omega/\alpha$ for $\omega < \alpha$ and $\tan^{-1}(\omega/\alpha) = \pi/2 - \alpha/\omega$ for $\omega > \alpha$. This approximation is rather easily obtained through use of a Taylor series approximation.

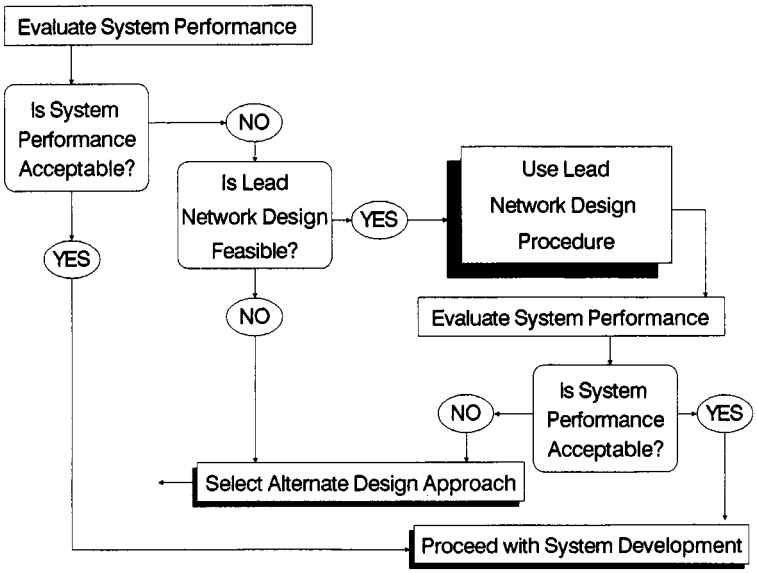


FIGURE 100.12 Life cycle of frequency domain design incorporating lead network compensation.

the frequency where the magnitude of the open-loop gain is 1, is 100 rad/s. The PM for the system without equalization is zero. We will design a simple lead network compensation for this zero PM system. If uncompensated, the closed-loop transfer function will be such that the system is unstable and any disturbance at all will result in a sinusoidally oscillating output.

The asymptotic gain diagram for this example is easily obtained from the open-loop transfer function

$$G_f(s)G_c(s) = \frac{K(1 + s/\omega_1)}{(1 + s/\omega_2)s^2}$$

and we wish to select the break frequencies ω_1 and ω_2 such that the phase shift at crossover is maximum. Further, we want this maximum phase shift to be such that we obtain the specified PM. We use the procedure suggested in Fig. 100.12.

Since the crossover frequency is such that $\omega_1 < \omega_c < \omega_2$, we have for the arctangent approximation to the phase shift in the vicinity of the crossover frequency

$$\beta(\omega) = -\pi + \tan^{-1}\frac{\omega}{\omega_1} - \tan^{-1}\frac{\omega}{\omega_2}$$

$$\approx \frac{-\pi}{2} - \frac{\omega_1}{\omega} - \frac{\omega}{\omega_2}$$

In order to maximize the phase shift at crossover, we set

$$\left.\frac{d\beta}{d\omega}\right|_{\omega=\omega_m} = 0$$

and obtain as a result

$$\omega_m = (\omega_1\ \omega_2)^{0.5}$$

We see that the crossover frequency obtained is halfway between the two break frequencies ω_1 and ω_2 on a logarithmic frequency coordinate. The phase shift at this optimum value of crossover frequency becomes

$$\beta_c = \beta(\omega_c) = \frac{-\pi}{2} - 2\left(\frac{\omega_1}{\omega_2}\right)^{0.5}$$

For a PM of $-3\pi/4$, for example, we have $-3\pi/4 = -\pi/2 - 2(\omega_1/\omega_2)^{0.5}$, and we obtain $\omega_1/\omega_2 = 0.1542$ as the ratio of frequencies. We see that we have need for a lead network with a gain of $\omega_1/\omega_2 = 6.485$. The gain at the crossover frequency is 1, and from the asymptotic gain approximation that is valid for $\omega_1 < \omega < \omega_2$, we have the expressions $|G(j\omega)| = K/\omega\omega_1$ and $|G(j\omega_c)| = 1 = K/\omega_c\omega_1$ which for a known K can be solved for ω_c and ω_1.

Now that we have illustrated the design computation with a very simple example, we are in a position to state some general results. In the direct approach to design for a specified PM we assume a single lead network equalizer such that the open-loop system to transfer function results. This approach to design results in the following steps that are applicable for Bode diagram design to achieve maximum PM within an experientially determined control system structure that comprises a fixed plant and a compensation network with adjustable parameters:

1. We find an equation for the gain at the crossover frequency in terms of the compensated open-loop system break frequency.
2. We find an equation of the phase shift at crossover.
3. We find the relationship between equalizer parameters and crossover frequency such that the phase shift at crossover is the maximum possible and a minimum of additional gain is needed.
4. We determine all parameter specifications to meet the PM specifications.
5. We check to see that all design specifications have been met. If they have not, we iterate the design process.

Figure 100.12 illustrates the steps involved in implementing this frequency domain design approach.

Phase-Lag Compensation

In the phase-lag-compensation frequency domain design approach, we reduce the gain at low frequencies such that crossover, the frequency where the gain magnitude is 1, occurs before the phase lag has had a chance to become intolerably large. A simple single-stage phase-lag-compensating network transfer function is

$$G_c(s) = \frac{1 + s/\omega_2}{1 + s/\omega_1} \qquad \omega_1 < \omega_2$$

Figure 100.13 illustrates the gain and phase versus frequency curves for a simple lag network with this transfer function. The maximum phase lag obtainable from a phase-lag network depends upon the ratio ω_2/ω_1 that is used in designing the network. From the expression for the phase shift of this transfer function,

$$\beta = \tan^{-1}\frac{\omega}{\omega_2} - \tan^{-1}\frac{\omega}{\omega_1}$$

we see that maximum phase lag occurs at that frequency ω_c where $d\beta/d\omega = 0$. We obtain for this value

$$\omega_m = (\omega_1\omega_2)^{0.5}$$

which is at the center of the two break frequencies for the lag network when the frequency response diagram is illustrated on a Bode diagram log-log asymptotic gain plot.

The maximum value of the phase lag obtained at $\omega = \omega_m$ is

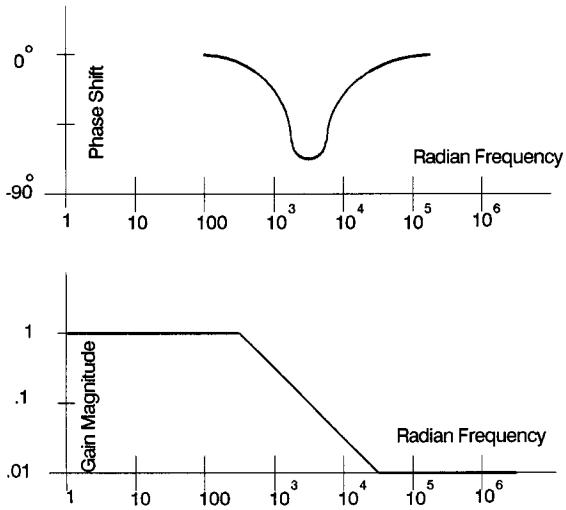


FIGURE 100.13 Phase shift and gain curves for a simple lag network.

$$\beta_m(\omega_m) = \frac{\pi}{2} - 2\tan^{-1}\left(\frac{\omega_2}{\omega_1}\right)^{0.5}$$
$$= \frac{\pi}{2} - 2\tan^{-1}\left(\frac{\omega_1}{\omega_2}\right)^{0.5}$$

which can be approximated in a more usable form, using the arctangent approximation, as

$$\beta_m(\omega_m) \approx \frac{\pi}{2}\sqrt{\frac{\omega_2}{\omega_1}}$$

The attenuation of the lag network at the frequency of minimum phase shift, or maximum phase lag, is obtained from the asymptotic approximation as

$$|G_c(\omega_m)| = \left(\frac{\omega_1}{\omega_2}\right)^{0.5}$$

Figure 100.13 presents a curve of attenuation magnitude obtainable at the frequency of maximum phase lag and the amount of the phase lag for various ratios ω_2/ω_1 for this simple lag network.

There are many ways to physically realize a lag network transfer function. Since the network only attenuates at some frequencies, as it never has a gain greater than 1 at any frequency, it can be realized with passive components only. Figure 100.14 presents an electrical realization of the simple lag network. Figure 100.15 presents a flowchart illustrating the design procedure envisioned here for lag network design. This is conceptually very similar to that for a lead network and makes use of the five-step parameter optimization procedure suggested earlier.

The object of lag network design is to reduce the gain at frequencies lower than the original crossover frequency in order to reduce the open-loop gain to unity before the phase shift becomes so excessive that the system PM is too small. A disadvantage of lag network compensation is that the attenuation introduced reduces

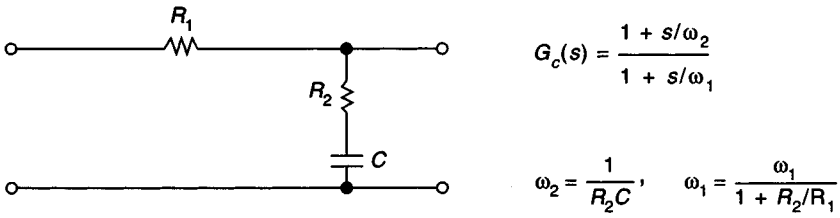


FIGURE 100.14 A simple electrical lag network.

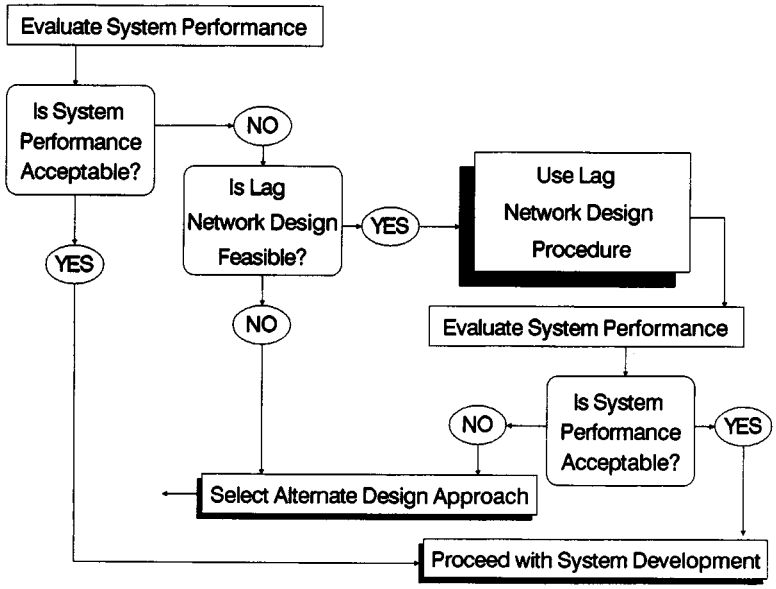


FIGURE 100.15 Life cycle of frequency domain design incorporating lag network compensation.

the crossover frequency and makes the system slower in terms of its transient response. Of course, this would be advantageous if high-frequency noise is present and we wish to reduce its effect. The lag network is an entirely passive device and thus is more economical to instrument than the lead network.

In lead network compensation we actually insert phase lead in the vicinity of the crossover frequency to increase the PM. Thus we realize a specified PM without lowering the medium-frequency system gain. We see that the disadvantages of the lag network are the advantages of the lead network and the advantages of the lag network are the disadvantages of the lead network.

We can attempt to combine the lag network with the lead network into an all-passive structure called a lag-lead network. Generally we obtain better results than we can achieve using either a lead or a lag network. We will consider design using lag-lead networks in our next subsection as well as more complex composite equalization networks.

Composite Equalizers

In the previous subsection we examined the simplest forms of series equalization: gain adjustment, lead network compensation, and lag network compensation. In this subsection we will consider more complex design examples in which composite equalizers will be used for series compensation. The same design principles used earlier in this section will be used here as well.

Lag-Lead Network Design

The prime purpose of a lead network is to add phase lead near the crossover frequency to increase the PM. Accompanied with this phase lead is a gain increase that will increase the crossover frequency. This will sometimes cause difficulties if there is much phase lag in the uncompensated system at high frequencies. There may be situations where use of a phase-lead network to achieve a given PM is not possible due to too many high-frequency poles.

The basic idea behind lag network design is to reduce the gain at "middle" frequencies such as to reduce the crossover frequency to a lower value than for the uncompensated system. If the phase lag is less at this lower frequency, then the PM will be increased by use of the lag network. We have seen that is not possible to use a lag network in situations in which there is not a frequency where an acceptable PM would exist if this frequency were the crossover frequency. Even if use of a lag network is possible, the significantly reduced crossover frequency resulting from its use may make the system so slow and sluggish in response to an input that system performance is unacceptable even though the relative stability of the system is acceptable.

Examination of these characteristics or attributes of lead network and lag network compensation suggests that it might be possible to combine the two approaches to achieve the desirable features of each approach. Thus we will attempt to provide attenuation below the crossover frequency to decrease the phase lag at crossover and phase lead closer to the crossover frequency in order to increase the phase lead of the uncompensated system at the crossover frequency.

The transfer function of the basic lag-lead network is

$$G_c(s) = \frac{(1 + s/\omega_2)(1 + s/\omega_3)}{(1 + s/\omega_1)(1 + s/\omega_4)}$$

where $\omega_4 > \omega_3 > \omega_2 > \omega_1$. Often it is desirable that $\omega_2\omega_3 = \omega_1\omega_4$ such that the high-frequency gain of the equalizer is unity. It is generally not desirable that $\omega_1\omega_4 > \omega_2\omega_3$ as this indicates a high-frequency gain greater than 1, and this will require an active network, or gain, and a passive equalizer. It is a fact that we should always be able to realize a linear minimum phase network using passive components only if the network has a rational transfer function with a gain magnitude that is no greater than 1 at any real frequency.

Figure 100.16 illustrates the gain magnitude and phase shift curves for a single-stage lag-lead network equalizer or compensator transfer function. Figure 100.17 illustrates an electrical network realization of a passive lag-lead network equalizer. Parameter matching can be used to determine the electrical network parameters that yield a specified transfer function. Because the relationships between the break frequencies and the equalizer component values are complex, it may be desirable, particularly in preliminary instrumentation of the control system, to use analog or digital computer programming techniques to construct the equalizer. Traditionally, there has been much analog computer simulation of control systems. The more contemporary approach suggests use of digital computer approaches that require numerical approximation of continuous-time physical systems.

Figure 100.18 presents a flowchart that we may use for lag-lead network design. We see that this flowchart has much in common with the charts and design procedures for lead network and lag network design and that each of these approaches first involves determining or obtaining a set of desired specifications for the control system. Next, the form of a trial compensating network and the number of break frequencies in the network are selected. We must then obtain a number of equations, equal to the number of network break frequencies plus 1. One of these equations shows that the gain magnitude is 1 at the crossover frequency. The second equation will be an equation for the phase shift at crossover. It is generally desirable that there be at least two unspecified compensating network break frequencies such that we may use a third equation, the optimality of the phase shift at crossover equation, in which we set $d\beta/d\omega|_{\omega=\omega c} = 0$. If other equations are needed to represent the design situation, we obtain these from the design specifications themselves.

General Bode Diagram Design

Figure 100.19 presents a flowchart of a general design procedure for Bode diagram design. As we will see in the next subsection, a minor modification of this flowchart can be used to accomplish design using minor-loop feedback

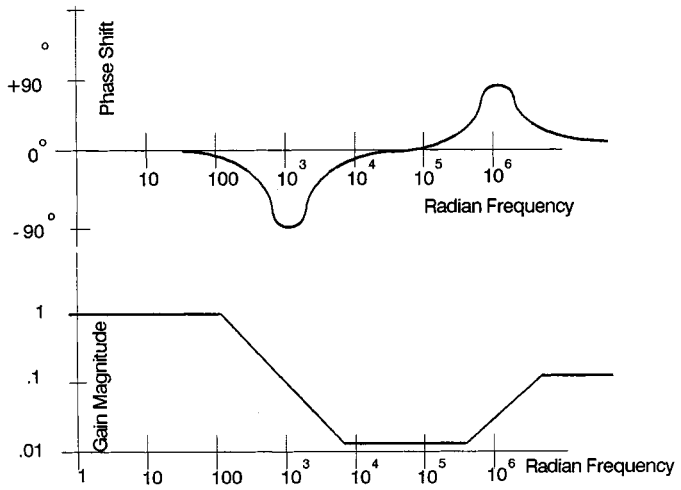
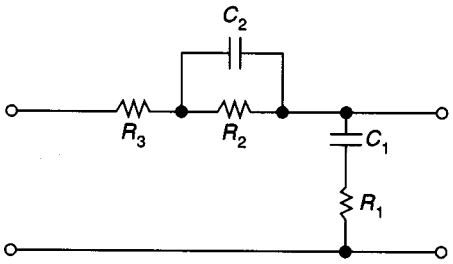


FIGURE 100.16 Phase shift and gain curves for a simple lag-lead network.



$$G_c(s) = \frac{(1 + s/\omega_2)(1 + s/\omega_3)}{(1 + s/\omega_1)(1 + s/\omega_4)}$$

$$= \frac{(1 + R_1C_1s)(1 + R_2C_2s)}{1 + (R_1C_1 + R_2C_1 + R_3C_1 + R_2C_2)s + (R_1R_2C_1C_2 + R_2R_3C_1C_2)s^2}$$

Special case: $R_3 = 0$

$$\omega_2 = \frac{1}{R_1C_1}, \quad \omega_3 = \frac{1}{R_2C_2}, \quad \omega_1\omega_4 = \omega_2\omega_3, \quad \omega_1 + \omega_4 = \omega_2 + \omega_3 + \frac{1}{R_1C_2}$$

FIGURE 100.17 Simple electrical lag-lead network.

or a combination of minor-loop and series equations. These detailed flowcharts for Bode diagram design are, of course, part of the overall design procedure of Fig. 100.9.

Much experience leads to the conclusion that satisfactory linear systems control design using frequency response approaches is such that the crossover frequency occurs on a gain magnitude curve which has a –1 slope at the crossover frequency. In the vicinity of crossover we may approximate any minimum phase transfer function, with crossover on a –1 slope, by

$$G(s) = G_f(s)G_c(s) = \frac{\omega_c\omega_1^{n-1}(1 + s/\omega_1)^{n-1}}{s^n(1 + s/\omega_2)^{m-1}} \qquad \text{for } \omega_1 > \omega_c > \omega_2$$

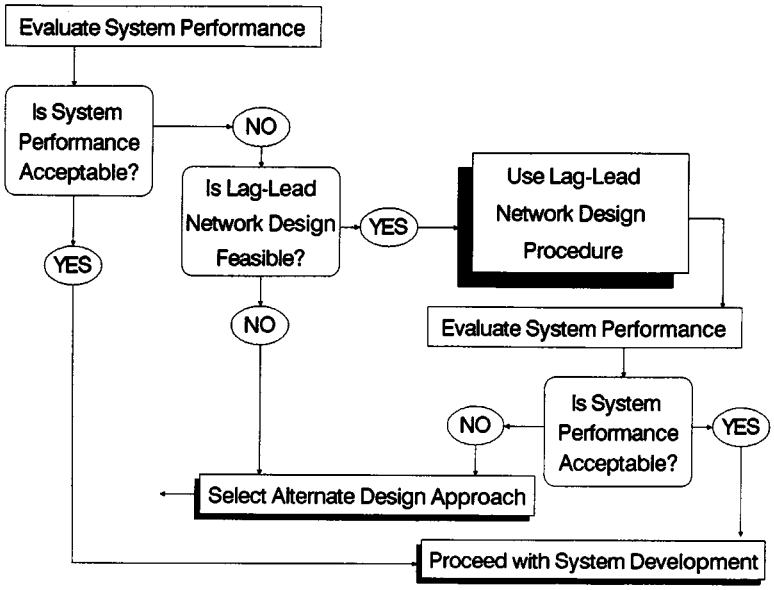


FIGURE 100.18 Life cycle of frequency domain design incorporating lag-lead network compensation.

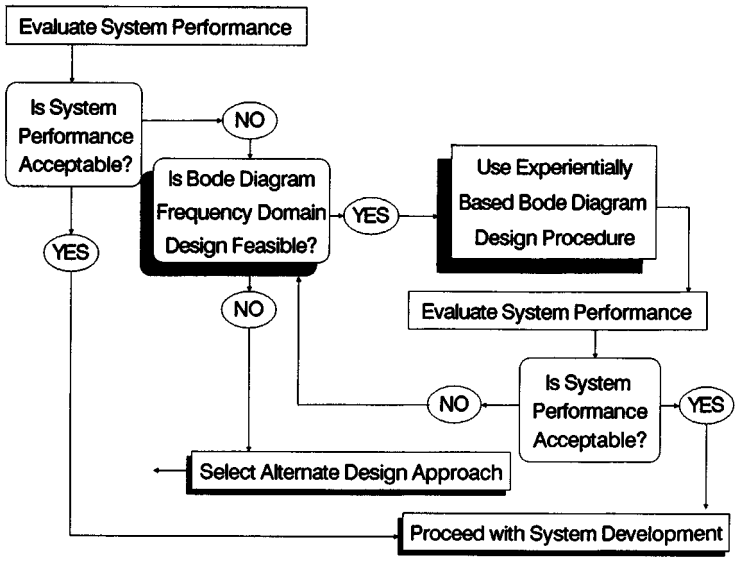


FIGURE 100.19 Life cycle of frequency domain design incorporating general Bode diagram compensation procedure.

Here ω_1 is the break frequency just prior to crossover and ω_2 is the break frequency just after crossover. It is easy to verify that we have $|G(j\omega_c)| = 1$ if $\omega_1 > \omega_c > \omega_2$. Figure 100.20 illustrates this rather general approximation to a compensated system Bode diagram in the vicinity of the crossover frequency. We will conclude this subsection by determining some general design requirements for a system with this transfer function and the associated Bode asymptotic gain magnitude diagram of Fig. 100.20.

There are three unknown frequencies in the foregoing equation. Thus we need three requirements or equations to determine design parameters. We will use the same three requirements used thus far in all our efforts in this section, namely:

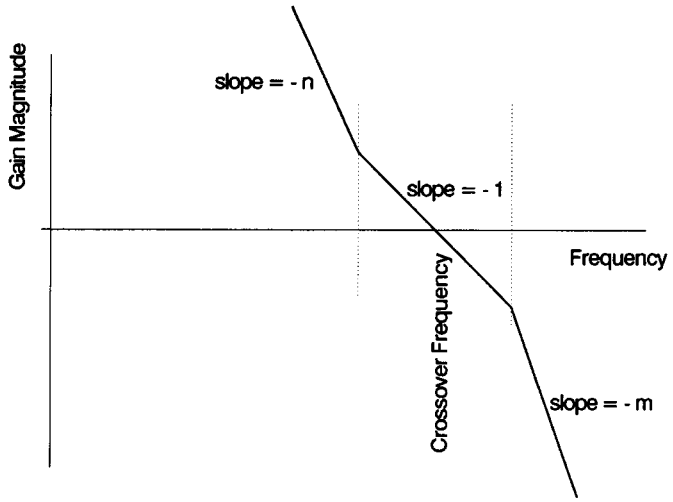


FIGURE 100.20 Illustration of generic gain magnitude in the vicinity of crossover.

1. The gain at crossover is 1.
2. The PM is some specified value.
3. The PM at crossover is the maximum possible for a given ω_2/ω_1 ratio.

We see that the first requirement, that the gain is 1 at the crossover frequency, is satisfied by the foregoing equation if the crossover frequency occurs on the –1 slope portion of the gain curve as assumed in Fig. 100.20. We use the arctangent approximation to obtain the phase shift in the vicinity of crossover as

$$\beta(\omega) = -\frac{n\pi}{2} + (n-1)\left(\frac{\pi}{2} - \frac{\omega_1}{\omega}\right) - (m-1)\frac{\omega}{\omega_2}$$

To satisfy requirement 3 we set

$$\left.\frac{d\beta(\omega)}{d\omega}\right|_{\omega=\omega_c} = 0 = \frac{(n-1)\omega_1}{\omega_c^2} - \frac{m-1}{\omega_2}$$

and obtain

$$\omega_c^2 = \frac{n-1}{m-1}\omega_1\omega_2$$

as the optimum setting for the crossover frequency. Substitution of the "optimum" frequency given by the foregoing into the phase shift equation results in

$$\beta(\omega_c) = \frac{-\pi}{2} - 2\sqrt{(m-1)(n-1)}\sqrt{\frac{\omega_1}{\omega_2}}$$

We desire a specific PM here, and so the equalizer break frequency locations are specified. There is a single parameter here that is unspecified, and an additional equation must be found in any specific application. Alternately, we could simply assume a nominal crossover frequency of unity or simply normalize frequencies ω_1 and ω_2 in terms of the crossover frequency by use of the normalized frequencies $\omega_1 = W_1\omega_c$ and $\omega_2 = W_2\omega_c$.

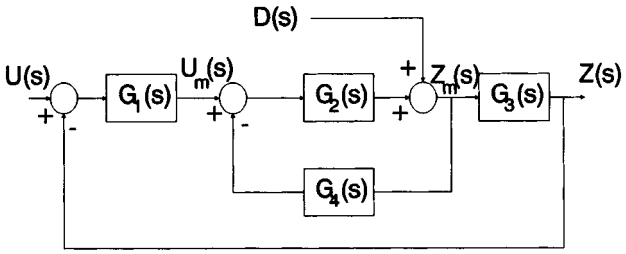


FIGURE 100.21 Feedback control system with a single minor loop and output disturbance.

It is a relatively simple matter to show that for a specified PM expressed in radians, we obtain for the normalized break frequencies

$$W_1 = \frac{\omega_1}{\omega_c} = \frac{\text{PM}}{2(n-1)}$$

$$W_2 = \frac{\omega_2}{\omega_c} = \frac{2(m-1)}{\text{PM}}$$

It is relatively easy to implement this suggested Bode diagram design procedure which is based upon considering only these break frequencies immediately above and below crossover and which approximate all others. Break frequencies far below crossover are approximated by integrations or differentiations, that is, poles or zeros at $s = 0$, and break frequencies far above the crossover frequency are ignored.

Minor-Loop Design

In our efforts thus far in this section we have assumed that compensating networks would be placed in series with the fixed plant and then a unity feedback ratio loop closed around these elements to yield the closed-loop system. In many applications it may be physically convenient, perhaps due to instrumentation considerations, to use one or more minor loops to obtain a desired compensation of a fixed plant transfer function.

For a single-input–single-output linear system there are no theoretical advantages whatever to any minor-loop compensation to series compensation as the same closed-loop transfer function can be realized by all procedures. However, when there are multiple inputs or outputs, then there may be considerable advantages to minor-loop design as contrasted to series compensation design. Multiple inputs often occur when there is a single-signal input and one or more noise or disturbance inputs present and a task of the system is to pass the signal inputs and reject the noise inputs. Also there may be saturation-type nonlinearities present, and we may be concerned not only with the primary system output but also with keeping the output at the saturation point within bounds such that the system remains linear. Thus there are reasons why minor-loop design may be preferable to series equalization.

We have discussed block diagrams elsewhere in this handbook. It is desirable here to review some concepts that will be of value for our discussion of minor-loop design. Figure 100.21 illustrates a relatively general linear control system with a single minor loop. This block diagram could represent many simple control systems. $G_1(s)$ could represent a discriminator and series compensation and $G_2(s)$ could represent an amplifier and that part of a motor transfer function excluding the final integration to convert velocity to position. $G_3(s)$ might then represent an integrator. $G_4(s)$ would then represent a minor-loop compensation transfer function, such as that of a tachometer.

The closed-loop transfer function for this system is given by

$$\frac{Z(s)}{U(s)} = H(s) = \frac{G_1(s)G_2(s)G_3(s)}{1 + G_2(s)G_4(s) + G_1(s)G_2(s)G_3(s)}$$

It is convenient to define several other transfer functions that are based on the block diagram in Fig. 100.21. First there is the minor-loop gain

$$G_m(s) = G_2(s)G_4(s)$$

which is just the loop gain of the minor loop only. The minor loop has the transfer function

$$\frac{Z_m(s)}{U_m(s)} = H_m(s) = \frac{G_2(s)}{1 + G_2(s)G_4(s)} = \frac{G_2(s)}{1 + G_m(s)}$$

There will usually be a range or ranges of frequency for which the minor-loop gain magnitude is much less than 1, and we then have

$$\frac{Z_m(s)}{U_m(s)} = H_m(s) \approx G_2(s) \qquad \left|G_m(\omega)\right| << 1$$

There will also generally be ranges of frequency for which the minor-loop gain magnitude is much greater than 1, and we then have

$$\frac{Z_m(s)}{U_m(s)} = H_m(s) \approx \frac{1}{G_4(s)} \qquad \left|G_m(\omega)\right| >> 1$$

We may use these two relations to considerably simplify our approach to the minor-loop design problem. For illustrative purposes, we will use two major-loop gain functions. First we will consider the major-loop gain with the minor-loop-compensating network removed such that $G_4(s) = 0$. This represents the standard situation we have examined in the last subsection. This uncompensated major-loop transfer function is

$$G_{Mu}(s) = G_1(s)G_2(s)G_3(s)$$

With the minor-loop compensation inserted, the major-loop gain, the input-output transfer function with the unity ratio feedback open, is

$$G_{Mc}(s) = \frac{G_1(s)G_2(s)G_3(s)}{1 + G_m(s)}$$

We may express the complete closed-loop transfer function in the form

$$\frac{Z(s)}{U(s)} = H(s) = \frac{G_{Mc}(s)}{1 + G_{Mc}(s)}$$

A particularly useful relationship may be obtained by combining the last three equations into one equation of the form

$$G_{Mc}(s) = \frac{G_{Mu}(s)}{1 + G_m(s)}$$

We may give this latter equation a particularly simple interpretation. For frequencies where the minor-loop gain $G_m(s)$ is low, the minor-loop–closed major-loop transfer function $G_{Mc}(s)$ is approximately that of the minor-loop–open major-loop transfer function G_{Mu} in that

$$G_{Mc}(s) \approx G_{Mu}(s) \qquad |G_m(\omega)| << 1$$

For frequencies where the minor-loop gain $G_m(s)$ is high, the minor-loop–closed major-loop transfer function is just

$$G_{Mc}(s) \approx \frac{G_{Mu}(s)}{G_m(s)} \qquad |G_m(\omega)| >> 1$$

This has an especially simple interpretation on the logarithmic frequency plots we use for Bode diagrams for we may simply subtract the minor-loop gain $G_m(s)$ from the minor-loop–open major-loop gain $G_{Mu}(s)$ to obtain the compensated system gain as the transfer function $G_{Mc}(s)$.

The last several equations are the key relations for minor-loop design using this frequency response approach. These relations indicate that some forms of series compensation yield a given major-loop transfer function $G_{Mc}(s)$ which will not be appropriate for realization by minor-loop compensation. In particular, a lead network series compensation cannot be realized by means of equivalent minor-loop compensation. The gain of the fixed plant $G_{Mu}(s)$ is raised at high frequencies due to the use of a lead network compensation. Also, we see that $G_{Mc}(s)$ can only be lowered by use of a minor-loop gain $G_m(s)$.

A lag network used for series compensation will result in a reduction in the fixed plant gain $|G_{Mu}(\omega)|$ at all high frequencies. This can only be achieved if the minor-loop transfer gain $G_m(s)$ is constant for high frequencies. In some cases this may be achievable but often will not be. It is possible to realize the equivalent of lag network series equalization by means of a minor-loop equalization for systems where the low- and high-frequency behavior of $G_{Mu}(s)$, or $G_f(s)$, and $G_{Mc}(s)$ are the same and where the gain magnitude of the compensated system $|G_{Mc}(s)|$ is at no frequency any greater than is the gain magnitude of the fixed plant $|G_f(s)|$ or the minor-loop–open major-loop transfer function $|G_{Mu}(s)|$. Thus we see that lag-lead network series equalization is an ideal type of equalization to realize by means of equivalent minor-loop equalization. Figures 100.9 and 100.19 represent flowcharts of a suggested general design procedure for minor-loop compensator design as well as for the series equalization approaches we examined previously.

In our work thus far we have assumed that parameters were constant and known. Such is seldom the case, and we must naturally be concerned with the effects of parameter variations, disturbances, and nonlinearities upon system performance. Suppose, for example, that we design a system with a certain gain assumed as K_1. If the system operates open loop and the gain K_1 is in cascade or series with the other input-output components, then the overall transfer function changes by precisely the same factor as K_1 changes. If we have an amplifier with unity ratio feedback around a gain K_1, the situation is much different. The closed-loop gain would nominally be $K_1/(1 + K_1)$, and a change to $2K_1$ would give a closed-loop gain $2K_1/(1 + 2K_1)$. If K_1 is large, say 10_3, then the new gain is 0.99950025, which is a percentage change of less than 0.05% for a change in gain of 100%.

Another advantage of minor-loop feedback occurs when there are output disturbances such as those due to wind gusts on an antenna. We consider the system illustrated in Fig. 100.21. The response due to $D(s)$ alone is

$$\frac{Z(s)}{D(s)} = \frac{1}{1 + G_2(s)\,G_4(s) + G_1(s)\,G_2(s)}$$

When we use the relation for the minor-loop gain

$$G_m(s) = G_2(s)G_4(s)$$

and the major-loop gain

$$G_{Mc}(s) = \frac{G_1(s)G_2(s)}{1 + G_2(s)G_4(s)}$$

we can rewrite the response due to $D(s)$ as

$$\frac{Z(s)}{D(s)} = \frac{1}{[1 + G_m(s)]G_{Mc}(s)}$$

Over the range of frequency where $|G_{Mc}(j\omega)| >> 1$, such that the corrected loop gain is large, the attenuation of a load disturbance is proportional to the uncorrected loop gain. This is generally larger over a wider frequency range than the corrected loop gain magnitude $|G_{Mc}(j\omega)|$, which is what the attenuation would be if series compensation were used.

Over the range of frequencies where the minor-loop gain is large but where the corrected loop gain is small, that is, where $|G_m(j\omega)| > 1$ and $|G_{Mc}(j\omega)| < 1$, we obtain for the approximate response due to the disturbance

$$\frac{Z(s)}{D(s)} \approx G_m(s)$$

and the output disturbance is therefore seen to be attenuated by the minor-loop gain rather than unattenuated as would be the case if series compensation had been used. This is, of course, highly desirable.

At frequencies where both the minor-loop gain transfer and the major-loop gain are small we have $Z(s)/D(s) \approx 1$, and over this range of frequencies neither minor-loop compensation nor series equalization is useful in reducing the effect of a load disturbance. Thus, we have shown here that there are quite a number of advantages to minor-loop compensation as compared to series equalization. Of course, there are limitations as well.

Summary

In this section, we have examined the subject of linear system compensation by means of the frequency response method of Bode diagrams. Our approach has been entirely in the frequency domain. We have discussed a variety of compensation networks, including:

1. Gain attenuation
2. Lead networks
3. Lag networks
4. Lag-lead networks and composite equalizers
5. Minor-loop feedback

Despite its age, the frequency domain design approach represents a most useful approach for the design of linear control systems. It has been tested and proven in a great many practical design situations.

Defining Terms

Bode diagram: A graph of the gain magnitude and frequency response of a linear circuit or system, generally plotted on log-log coordinates. A major advantage of Bode diagrams is that the gain magnitude plot will look like straight lines or be asymptotic to straight lines. H.W. Bode, a well-known Bell Telephone Laboratories researcher, published *Network Analysis and Feedback Amplifier Design* in 1945. The approach, first described there, has been refined by a number of other workers over the past half-century.

Crossover frequency: The frequency where the magnitude of the open-loop gain is 1.

Equalizer: A network inserted into a system that has a transfer function or frequency response designed to compensate for undesired amplitude, phase, and frequency characteristics of the initial system. Filter and equalizer are generally synonymous terms.

Lag network: In a simple phase-lag network, the phase angle associated with the input-output transfer function is always negative, or lagging. Figures 100.13 and 100.14 illustrate the essential characteristics of a lag network.

Lag-lead network: The phase shift versus frequency curve in a phase lag-lead network is negative, or lagging, for low frequencies and positive, or leading, for high frequencies. The phase angle associated with the input-output transfer function is always positive, or leading. Figures 100.16 and 100.17 illustrate the essential characteristics of a lag-lead network, or composite equalizer.

Lead network: In a simple phase-lead network, the phase angle associated with the input-output transfer function is always positive, or leading. Figures 100.10 and 100.11 illustrate the essential characteristics of a lead network.

Series equalizer: In a single-loop feedback system, a series equalizer is placed in the single loop, generally at a point along the forward path from input to output where the equalizer itself consumes only a small amount of energy. In Fig. 100.21, $G_1(s)$ could represent a series equalizer. $G_1(s)$ could also be a series equalizer if $G_4(s) = 0$.

Specification: A statement of the design or development requirements to be satisfied by a system or product.

Systems engineering: An approach to the overall life cycle evolution of a product or system. Generally, the systems engineering process comprises a number of phases. There are three essential phases in any systems engineering life cycle: formulation of requirements and specifications, design and development of the system or product, and deployment of the system. Each of these three basic phases may be further expanded into a larger number. For example, deployment generally comprises operational test and evaluation, maintenance over an extended operational life of the system, and modification and retrofit (or replacement) to meet new and evolving user needs.

Related Topic

11.1 Introduction

References

J.L. Bower and P.M. Schultheiss, *Introduction to the Design of Servomechanisms*, New York: Wiley, 1958.
A.P. Sage, *Linear Systems Control,* Champaign, Ill.: Matrix Press, 1978.
A.P. Sage, *Systems Engineering,* New York: Wiley, 1992.
M.G. Singh, Ed., *Systems and Control Encyclopedia,* Oxford: Pergamon, 1987.

Further Information

Many of the practical design situations used to test the frequency domain design approach are described in the excellent classic text by Bower and Schultheiss [1958]. A rather detailed discussion of the approach may also be found in Sage [1978] on which this discussion is, in part, based. A great variety of control systems design approaches, including frequency domain design approaches, are discussed in a recent definitive control systems encyclopedia [Singh, 1987], and there are a plethora of new introductory control systems textbooks that discuss it as well. As noted earlier, frequency domain design, in particular, and control systems design, in general, constitute one facet of systems engineering effort, such as described in Sage [1992].

100.4 Root Locus

Benjamin C. Kuo

Root locus represents a trajectory of the roots of an algebraic equation with constant coefficients when a parameter varies. The technique is used extensively for the analysis and design of linear time-invariant control

systems. For linear time-invariant control systems the roots of the characteristic equation determine the stability of the system. For a stable continuous-data system the roots must all lie in the left half of the s plane. For a digital control system to be stable, the roots of the characteristic equation must all lie inside the unit circle $|z| = 1$ in the z plane. Thus, in the s plane, the imaginary axis is the stability boundary, whereas in the z plane the stability boundary is the unit circle. The location of the characteristic equation roots with respect to the stability boundary also determine the relative stability, i.e., the degree of stability, of the system.

For a linear time-invariant system with continuous data, the characteristic equation can be written as

$$F(s) = P(s) + KQ(s) = 0 \tag{100.33}$$

where $P(s)$ is an Nth-order polynomial of s,

$$P(s) = s^N + a_1 s^{N-1} + \ldots + a_{N-1}s + a_N \tag{100.34}$$

and $Q(s)$ is the Mth-order polynomial of s,

$$Q(s) = s^M + b_1 s^{M-1} + \ldots + b_{M-1}s + b_M \tag{100.35}$$

where N and M are positive integers. The real constant K can vary from $-\infty$ to $+\infty$. The coefficients $a_1, a_2, \ldots, a_N, b_1, b_2, \ldots, b_M$ are real. As K is varied from $-\infty$ to $+\infty$, the roots of Eq. (100.33) trace out continuous trajectories in the s plane called the *root loci.*

The above development can be extended to digital control systems by replacing s with z in Eqs. (100.33) through (100.35).

Root Locus Properties

The root locus problem can be formulated from Eq. (100.33) by dividing both sides of the equation by the terms that do not contain the variable parameter K. The result is

$$1 + \frac{KQ(s)}{P(s)} = 0 \tag{100.36}$$

For a closed-loop control system with the loop transfer function $KG(s)H(s)$, where the gain factor K has been factored out, the characteristic equation is known to be the zeros of the rational function

$$1 + KG(s)H(s) = 0 \tag{100.37}$$

Since Eqs. (100.36) and (100.37) have the same form, the general root locus problem can be formulated using Eq. (100.36).

Equation (100.37) is written

$$G(s)H(s) = -\frac{1}{K} \tag{100.38}$$

To satisfy the last equation, the following conditions must be met simultaneously:

$$\text{Condition on magnitude:} \quad |G(s)H(s)| = \frac{1}{|K|} \tag{100.39}$$

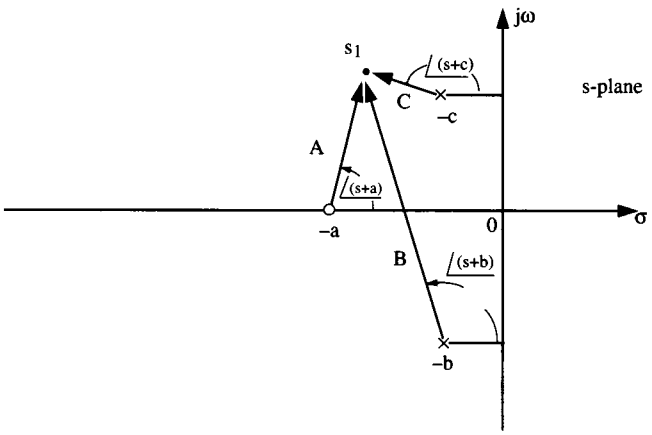


FIGURE 100.22 Graphical interpretation of magnitude and angle conditions of root loci.

where K varies between $-\infty$ and $+\infty$.

$$\text{Conditions on angles: } \angle G(s)H(s) = (2k+1)\pi \quad K \geq 0$$

$$= \text{odd multiples of } \pi \text{ rad} \tag{100.40}$$

$$\angle G(s)H(s) = 2k\pi \quad K \leq 0$$

$$= \text{even multiples of } \pi \text{ rad} \tag{100.41}$$

where $k = 0, \pm 1, \pm 2, \ldots, \pm$ any integer.

In general, the conditions on angles in Eqs. (100.40) and (100.41) are used for the construction of the root loci, whereas the condition on magnitude in Eq. (100.39) is used to find the value of K on the loci once the loci are drawn. Let $KG(s)H(s)$ be of the form

$$KG(s)H(s) = \frac{K(s+a)}{(s+b)(s+c)} \tag{100.42}$$

Applying Eqs. (100.40) and (100.41) to the last equation, the angles conditions are

$$K \geq 0: \quad \angle G(s)H(s) = \angle(s+a) - \angle(s+b) - \angle(s+c)$$

$$= (2k+1)\pi \tag{100.43}$$

$$K \leq 0: \quad \angle G(s)H(s) = \angle(s+a) - \angle(s+b) - \angle(s+c)$$

$$= 2k\pi \tag{100.44}$$

where $k = 0, \pm 1, \pm 2, \ldots$. The graphical interpretation of the last two equations is shown in Fig. 100.22. For the point s_1 to be a point on the root locus, the angles of the phasors drawn from the poles and zeros of $G(s)H(s)$ to s_1 must satisfy Eq. (100.43) or (100.44) depending on the sign of K. Applying the magnitude condition of Eq. (100.39) to (100.42), the magnitude of K is expressed as

$$|K| = \frac{|s+b|\,|s+c|}{|s+a|} = \frac{B \cdot C}{A} \tag{100.45}$$

where A, B, and C are the lengths of the phasors drawn from the poles and zeros of $G(s)H(s)$ to the point s_1.

The following properties of the root loci are useful for sketching the root loci based on the pole-zero configuration of $G(s)H(s)$. Many computer programs, such as the ROOTLOCI in the ACSP software package [Kuo, 1991b], are available for computing and plotting the root loci. The proofs and derivations of these properties can be carried out from Eqs. (100.39), (100.40), and (100.41) [Kuo, 1991a].

Starting Points (K 5 0 Points). The points at which $K = 0$ on the root loci are at the poles of $G(s)H(s)$.

Ending Points (K 56` Points). The points at which $K = \pm\infty$ on the root loci are at the zeros of $G(s)H(s)$. The poles and zeros referred to above include those at s = ∞.

Number of Root Loci. The total number of root loci of Eq. (100.37) equals the higher of the number of poles and zeros of $G(s)H(s)$.

Symmetry of Root Loci. The root loci are symmetrical with respect to the axes of symmetry of the pole-zero configuration of $G(s)H(s)$. In general, the root loci are symmetrical at least to the real axis of the complex s plane.

Asymptotes of the Root Loci. Asymptotes of the root loci refer to the behavior of the root loci at $|s| = \infty$ when the number of poles and zeros of $G(s)H(s)$ is not equal. Let N denote the number of finite poles of $G(s)H(s)$ and M be the number of finite zeros of $G(s)H(s)$. In general, $2|N - M|$ of the loci will approach infinity in the s plane. The properties of the root loci at $|s| = \infty$ are described by the angles and the intersects of the asymptotes. When $N \neq M$, the angles of the asymptotes are given by

$$\phi_k = \begin{cases} \dfrac{(2k + 1)\pi}{|N - M|} & K \geq 0 \quad (100.46) \\ \dfrac{2k\,\pi}{|N - M|} & K \leq 0 \quad (100.47) \end{cases}$$

where $k = 0, 1, 2, \ldots, |N - M| - 1$.

The asymptotes intersect on the real axis at

$$\sigma = \frac{\sum \text{finite poles of } G(s)H(s) - \sum \text{finite zeros of } G(s)H(s)}{N - M} \quad (100.48)$$

Root Loci on the Real Axis. The entire real axis of the s plane is occupied by the root loci. When $K > 0$, root loci are found in sections of the real axis to the right of which the total number of poles and zeros of $G(s)H(s)$ is *odd*. When $K < 0$, root loci are found in sections to the right of which the total number of poles and zeros of $G(s)H(s)$ is *even*.

As a summary of the root locus properties discussed above, the properties of the root loci of the following equation are displayed in Fig. 100.23.

$$s^3 + 2s^2 + 2s + K(s + 3) = 0 \quad (100.49)$$

Dividing both sides of the last equation by the terms that do not contain K we get

$$1 + KG(s)H(s) = 1 + \frac{K(s + 3)}{s(s^2 + 2s + 2)} \quad (100.50)$$

Thus, the poles of $G(s)H(s)$ are at $s = 0$, $s = -1 + j$, and $s = -1 - j$. The zero of $G(s)H(s)$ is at $z = -3$.

As shown in Fig. 100.23, the $K = 0$ points on the root loci are at the poles of $G(s)H(s)$, and the $K = \pm\infty$ points are at the zeros of $G(s)H(s)$. Since $G(s)H(s)$ has two zeros at $s = \infty$, two of the three root loci approach

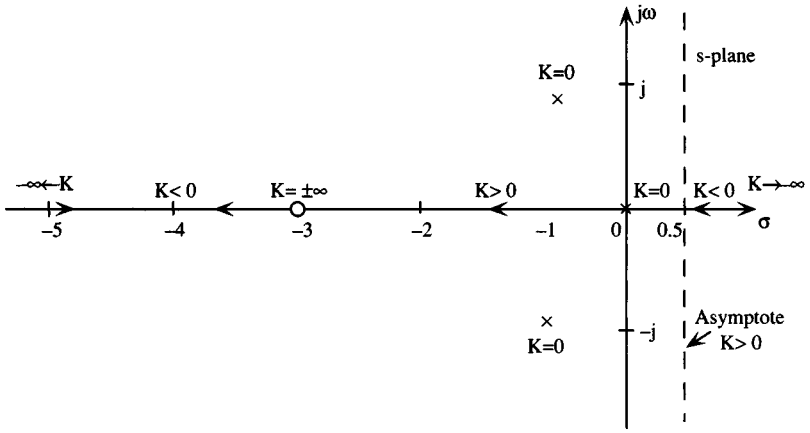


FIGURE 100.23 Some properties of the root loci of $G(s)H(s) = K(s + 3)/[s(s^2 + 2s + 2)]$.

infinity in the s plane. The root loci are symmetrical to the real axis of the s plane, since the pole-zero configuration of $G(s)H(s)$ is symmetrical to the axis. The asymptotes of the two root loci that approach infinity are characterized by Eqs. (100.46) through (100.48). The angles of the asymptotes are:

$$K \geq 0: \quad \phi_k = \frac{(2k + 1)\pi}{3 - 1} \qquad k = 0, 1 \tag{100.51}$$

$$K \leq 0: \quad \phi_k = \frac{2k\pi}{3 - 1} \qquad k = 0, 1 \tag{100.52}$$

Thus, for $K \geq 0$, the angles of the asymptotes are $\phi_0 = 90°$ and $\phi_1 = 270°$. For $K \leq 0$, $\phi_0 = 0°$ and $\phi_1 = 180°$.

The intersect of the asymptotes is at

$$\sigma = \frac{-1 + j - 1 - j - (-3)}{3 - 1} = \frac{1}{2} \tag{100.53}$$

The root loci on the real axis are as indicated in Fig. 100.23.

Angles of Departure and Arrival. The slope of the root locus in the vicinity of a pole of $G(s)H(s)$ is measured at the *angle of departure* and that in the vicinity of a zero of $G(s)H(s)$ is measured at the *angle of arrival.*

The angle of departure (arrival) of a root locus at a pole (zero) of $G(s)H(s)$ is determined by assigning a point s_1 to the root locus that is very close to the pole (zero) and applying the angle conditions of Eqs. (100.40) or (100.41). Figure 100.24 illustrates the calculation of the angles of arrival and departure of the root locus at the pole $s = -1 + j$. We assign a point s_1 that is on the locus for $K > 0$ near the pole and draw phasors from *all* the poles and the zero of $G(s)H(s)$ to this point. The angles made by the phasors with respect to the real axis must satisfy the angle condition in Eq. (100.46). Let the angle of the phasor drawn from $-1 + j$ to s_1 be designated as θ, which is the angle of departure; the angles drawn from the other poles and zero can be approximated by regarding s_1 as being very close to $-1 + j$. Thus, Eq. (100.46) leads to

$$\angle G(s_1)H(s_1) = -\theta - 135° - 90° + 26.6° = -180° \tag{100.54}$$

or $\theta = -18.4°$. For the angle of arrival of the root locus at the pole $s = -1 + j$, we assign a point s_1 on the root loci for $K < 0$ near the pole. Drawing phasors from all the poles and the zero of $G(s)H(s)$ to s_1 and applying the angle condition in Eq. (100.47), we have

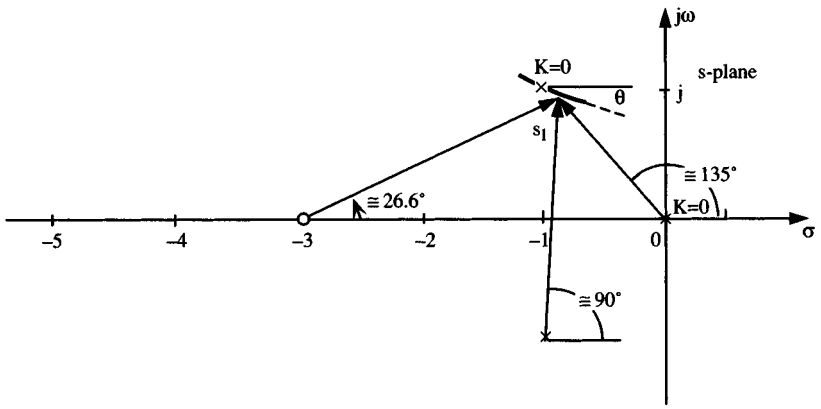


FIGURE 100.24 Angle of arrival and departure calculations.

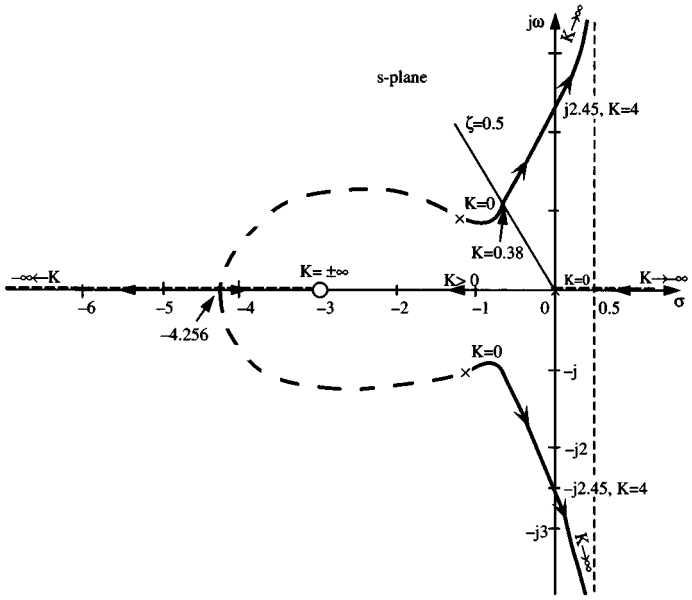


FIGURE 100.25 The complete root loci of $G(s)H(s) = K(s + 3)/[s(s^2 + 2s + 2)]$.

$$\angle G(s_1)H(s_1) = -\theta - 135° - 90° + 26.6° = 0° \tag{100.55}$$

Thus, the angle of arrival of the root locus for $K < 0$ is $\theta = 198.4°$. Similarly, we can show that the angles of arrival and departure of the root locus at $s = -3$ are 180° and 0°, respectively.

Intersection of the Root Loci with the Imaginary Axis. The points where the root loci intersect the imaginary axis (if there is any) in the s plane, and the corresponding values of K, may be determined by means of the Routh-Hurwitz stability criterion [Kuo, 1991a]. The root locus program can also be used on a computer to give the intersects.

The complete root loci in Fig. 100.25 show that the root loci intersect the imaginary axis at $s = \pm j2.45$, and the value of K is 4. The system is stable for $0 \le K < 4$.

Breakaway Points of the Root Loci. Breakaway points on the root loci correspond to multiple-order roots of the equation. At a breakaway point several root loci converge and then break away in different directions. The breakaway point can be real or complex. The latter case must be in complex conjugate pairs.

The breakaway points of the root loci of Eq. (100.37) must satisfy the following condition:

$$\frac{dG(s)H(s)}{ds} = 0 \tag{100.56}$$

On the other hand, not all solutions of Eq. (100.56) are breakaway points. To satisfy as a breakaway point, the point must also lie on the root loci, or satisfy Eq. (100.37). Applying Eq. (100.56) to the function $G(s)H(s)$ given in Eq. (100.50), we have the equation that the breakaway points must satisfy,

$$2s^3 + 11s^2 + 12s + 6 = 0 \tag{100.57}$$

The roots of the last equation are $s = -4.256$, $-0.622 + j0.564$ and $-0.622 - j0.564$. As shown in Fig. 100.25, only the solution $s = -4.256$ is a breakaway point on the root loci.

Root Loci of Digital Control Systems

The root locus analysis presented in the preceding subsections can be applied to digital control systems without modifying the basic principles. For a linear time-invariant digital control system, the transfer functions are expressed in terms of the z-transform rather than the Laplace transform. The relationship between the z-transform variable z and the Laplace transform variable s is

$$z = e^{Ts} \tag{100.58}$$

where T is the sampling period in seconds. Typically, the characteristic equation roots are solutions of the equation

$$1 + KGH(z) = 0 \tag{100.59}$$

where K is the variable gain parameter. The root loci for a digital control system are constructed in the complex z plane. All the properties of the root loci in the s plane apply readily to the z plane, with the exception that the stability boundary is now the unit circle $|z| = 1$. That is, the system is stable if all the characteristic equation roots lie inside the unit circle.

As an illustration, the open-loop transfer function of a digital control system is given as

$$G(z) = \frac{K(z + 0.1)}{z(z - 1)} \tag{100.60}$$

The characteristic equation of the closed-loop system is

$$z(z - 1) + K(z + 0.1) = 0$$

The root loci of the system are shown in Fig. 100.26. Notice that the system is stable for $0 \le K < 2.22$. When $K = 2.22$, one root is at $z = -1$, which is on the stability boundary.

Design with Root Locus

The root locus diagram of the characteristic equation of a closed-loop control system can be used for design purposes. The roots of the characteristic equation can be positioned in the s plane (or the z plane for digital control systems) to realize a certain desired relative stability or damping of the system. It should be kept in mind that the zeros of the closed-loop transfer function also affect the relative stability of the system, although the absolute stability is strictly governed by the characteristic equation roots.

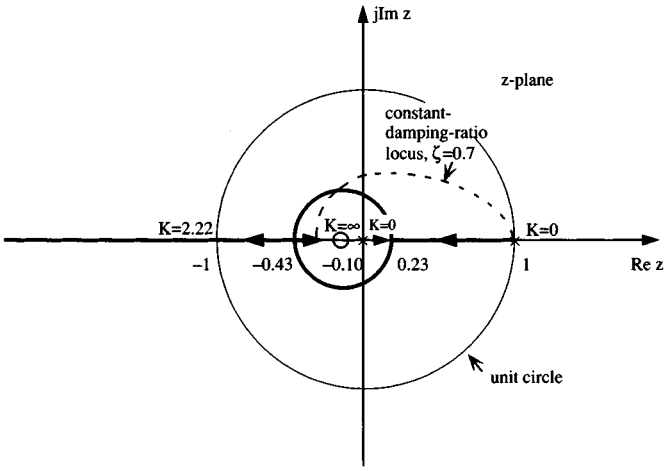


FIGURE 100.26 Root loci in the z plane for a digital control system.

As an illustrative example, the constant-damping ratio line for $\zeta = 0.5$ is shown in Fig. 100.25. The intersect of the $\zeta = 0.5$ line and the root locus corresponds to $K = 0.38$. Let us assume that we want to keep the relative damping at approximately 0.5 but the gain K should be increased tenfold. The following cascade controller is applied to the system [Evans, 1948]:

$$G_c(s) = \frac{1 + 5s}{1 + 50s} \tag{100.61}$$

The open-loop transfer function of the compensated system is now

$$G_c(s)G(s)H(s) = \frac{0.1K(s + 3)(s + 0.2)}{s(s + 0.02)(s^2 + 2s + 2)} \tag{100.62}$$

Figure 100.27 shows the root locus diagram of the compensated system for $K \geq 0$. The shape of the complex root loci is not appreciably affected by the controller, but the value of K that corresponds to a relative damping ratio of 0.5 is now approximately 3.9.

In a similar manner the root loci of digital control systems can be reshaped in the z plane for design. The constant-damping ratio locus in the z plane is shown in Fig. 100.26.

Defining Terms

Angles of departure and arrival: The slope of the root locus in the vicinity of a pole of $G(s)H(s)$ is measured as the angle of departure, and that in the vicinity of a zero of $G(s)H(s)$ is measured as the angle of arrival.

Asymptotes of root loci: The behavior of the root loci at $|s| = \infty$ *when the number of poles and zeros of* $G(s)H(s)$ *is not equal.*

Breakaway points of the root loci: Breakaway points on the root loci correspond to multiple-order roots of the equation.

Root locus: The trajectory of the roots of an algebraic equation with constant coefficient when a parameter varies.

Related Topics

6.1 Definitions and Properties • 12.1 Introduction

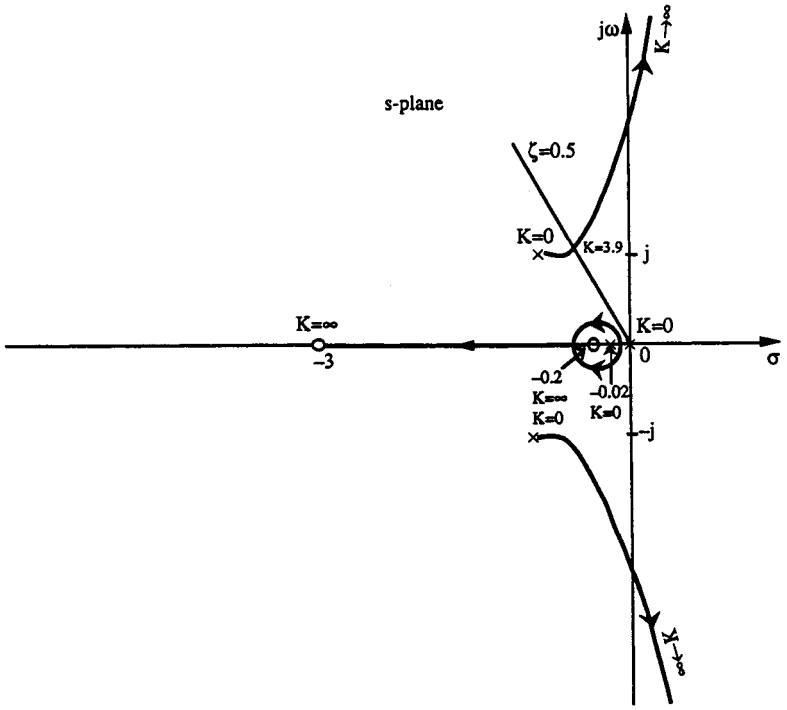


FIGURE 100.27 Root loci of Eq. (100.62).

References and Further Information

R. C. Dorf, *Modern Control Systems,* 5th ed., Reading, Mass.: Addison-Wesley, 1989.
W. R. Evans, "Graphical analysis of control systems," *Trans. AIEE,* vol. 67, pp. 547–551, 1948.
B. C. Kuo, *Automatic Control Systems,* 6th ed., Englewood Cliffs, N.J.: Prentice-Hall, 1991a.
B. C. Kuo, *ACSP Software and Manual,* Englewood Cliffs, N.J.: Prentice-Hall, 1991b.
B. C. Kuo, *Digital Control Systems,* 2nd ed., New York: Holt, 1992a.
B. C. Kuo, *DCSP Software and Manual,* Champaign, Ill.: SRL, Inc., 1992b.

100.5 Compensation

Charles L. Phillips and Royce D. Harbor

Compensation is the process of modifying a closed-loop control system (usually by adding a *compensator* or *controller*) in such a way that the compensated system satisfies a given set of design specifications. This section presents the fundamentals of compensator design; actual techniques are available in the references.

A single-loop control system is shown in Fig. 100.28. This system has the transfer function from input $R(s)$ to output $C(s)$

$$T(s) = \frac{C(s)}{R(s)} = \frac{G_c(s)G_p(s)}{1 + G_c(s)G_p(s)H(s)} \tag{100.63}$$

and the characteristic equation is

$$1 + G_c(s)G_p(s)H(s) = 0 \tag{100.64}$$

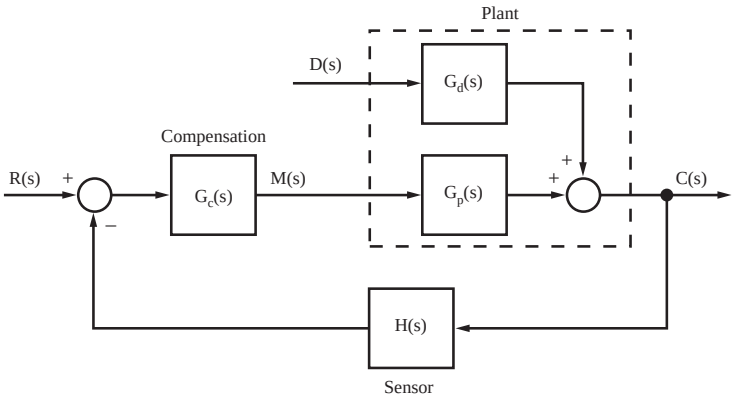


FIGURE 100.28 A closed-loop control system.

where $G_c(s)$ is the *compensator* transfer function, $G_p(s)$ is the *plant* transfer function, and $H(s)$ is the *sensor* transfer function. The transfer function from the disturbance input $D(s)$ to the output is $G_d(s)/[1 + G_c(s)G_p(s)H(s)]$. The function $G_c(s)G_p(s)H(s)$ is called the *open-loop function.*

Control System Specifications

The compensator transfer function $G_c(s)$ is designed to give the closed-loop system certain specified characteristics, which are realized through achieving one or more of the following:

1. Improving the transient response. Increasing the speed of response is generally accomplished by increasing the open-loop gain $G_c(j\omega)G_p(j\omega)H(j\omega)$ at higher frequencies such that the system bandwidth is increased. Reducing overshoot (ringing) in the response generally involves increasing the phase margin ϕ_m of the system, which tends to remove any resonances in the system. The phase margin ϕ_m occurs at the frequency ω_1 and is defined by the relationship

$$\left|G_c(j\omega_1)G_p(j\omega_1)H(j\omega_1)\right| = 1$$

   with the angle of $G_c(j\omega_1)G_p(j\omega_1)H(j\omega_1)$ equal to $(180° + \phi_m)$.
2. Reducing the steady-state errors. Steady-state errors are decreased by increasing the open-loop gain $G_c(j\omega)G_p(j\omega)H(j\omega)$ in the frequency range of the errors. Low-frequency errors are reduced by increasing the low-frequency open-loop gain and by increasing the type number of the system [the number of poles at the origin in the open-loop function $G_c(s)G_p(s)H(s)$].
3. Reducing the sensitivity to plant parameters. Increasing the open-loop gain $G_c(j\omega)G_p(j\omega)\ H(j\omega)$ tends to reduce the variations in the system characteristics due to variations in the parameters of the plant.
4. Rejecting disturbances. Increasing the open-loop gain $G_c(j\omega)G_p(j\omega)H(j\omega)$ tends to reduce the effects of disturbances [$D(s)$ in Fig. 100.28] on the system output, provided that the increase in gain does not appear in the direct path from disturbance inputs to the system output.
5. Increasing the relative stability. Increasing the open-loop gain tends to reduce phase and gain margins, which generally increases the overshoot in the system response. Hence, a trade-off exists between the beneficial effects of increasing the open-loop gain and the resulting detrimental effects of reducing the stability margins.

Design

Design procedures for compensators are categorized as either *classical methods* or *modern methods.* Classical methods discussed are:

- Ph ase-lag frequency response
- Phase-lead frequency response
- Phase-lag root locus
- Phase-lead root locus

Modern methods discussed are:

- Pole placement
- State estimation
- Optimal

Frequency Response Design

Classical design procedures are normally based on the open-loop function of the uncompensated system, $G_p(s)H(s)$. Two compensators are used in classical design; the first is called a *phase-lag compensator,* and the second is called a *phase-lead compensator.*

The general characteristics of phase-lag-compensated systems are as follows:

1. The low-frequency behavior of a system is improved. This improvement appears as reduced errors at low frequencies, improved rejection of low-frequency disturbances, and reduced sensitivity to plant parameters in the low-frequency region.
2. The system bandwidth is reduced, resulting in a slower system time response and better rejection of high-frequency noise in the sensor output signal.

The general characteristics of phase-lead-compensated systems are as follows:

1. The high-frequency behavior of a system is improved. This improvement appears as faster responses to inputs, improved rejection of high-frequency disturbances, and reduced sensitivity to changes in the plant parameters.
2. The system bandwidth is increased, which can increase the response to high-frequency noise in the sensor output signal.

The transfer function of a first-order compensator can be expressed as

$$G_c(s) = \frac{K_c(s/\omega_0 + 1)}{s/\omega_p + 1} \tag{100.65}$$

where $-\omega_0$ is the compensator zero, $-\omega_p$ is its pole, and K_c is its dc gain. If $\omega_p < \omega_0$, the compensator is phase-lag. The Bode diagram of a phase-lag compensator is given in Fig. 100.29 for $K_c = 1$.

It is seen from Fig. 100.29 that the phase-lag compensator reduces the high-frequency gain of the open-loop function relative to the low-frequency gain. This effect allows a higher low-frequency gain, with the advantages listed above. The pole and zero of the compensator must be placed at very low frequencies relative to the compensated-system bandwidth so that the destabilizing effects of the negative phase of the compensator are negligible.

If $\omega_p > \omega_0$ the compensator is phase-lead. The Bode diagram of a phase-lead compensator is given in Fig. 100.30 for $K_c = 1$.

It is seen from Fig. 100.30 that the phase-lead compensator increases the high-frequency gain of the open-loop function relative to its low-frequency gain. Hence, the system has a larger bandwidth, with the advantages listed above. The pole and zero of the compensator are generally difficult to place, since the increased gain of the open-loop function tends to destabilize the system, while the phase lead of the compensator tends to stabilize the system. The pole-zero placement for the phase-lead compensator is much more critical than that of the phase-lag compensator.

A typical Nyquist diagram of an uncompensated system is given in Fig. 100.31. The pole and the zero of a phase-lag compensator are placed in the frequency band labeled *A*. This placement negates the destabilizing

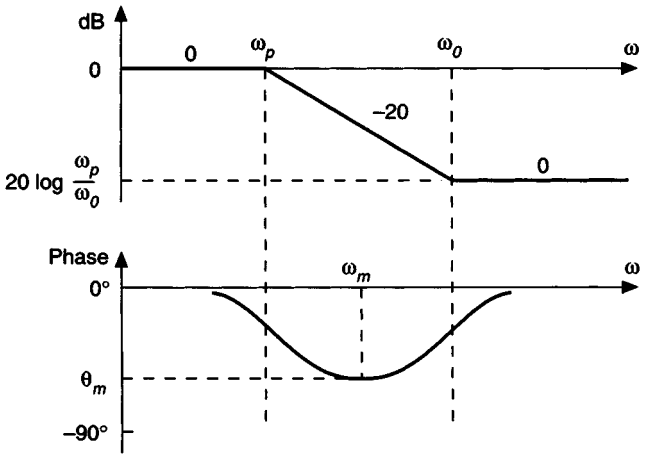


FIGURE 100.29 Bode diagram for a phase-lag compensator.

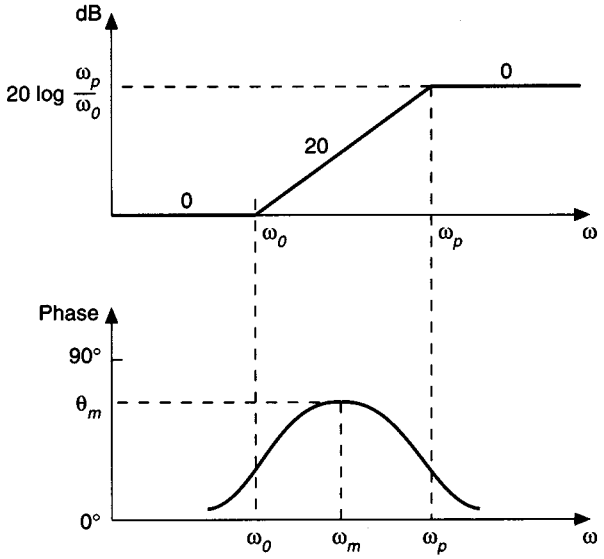


FIGURE 100.30 Bode diagram for a phase-lead compensator.

effect of the negative phase of the compensator. The pole and zero of a phase-lead compensator are placed in the frequency band labeled *B*. This placement utilizes the stabilizing effect of the positive phase of the compensator.

PID Controllers

Proportional-plus-integral-plus-derivative (PID) compensators are probably the most utilized form for compensators. These compensators are essentially equivalent to a phase-lag compensator cascaded with a phase-lead compensator. The transfer function of this compensator is given by

$$G_c(s) = K_P + \frac{K_I}{s} + K_D s \tag{100.66}$$

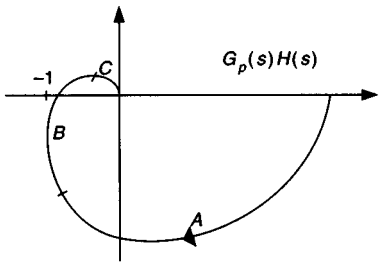


FIGURE 100.31 A typical Nyquist diagram for $G_p(s)H(s)$.

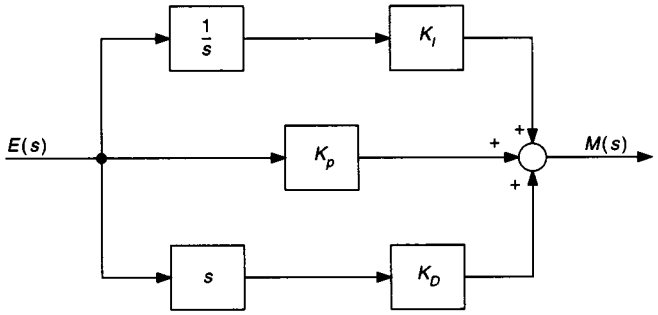


FIGURE 100.32 Block diagram of a PID compensator.

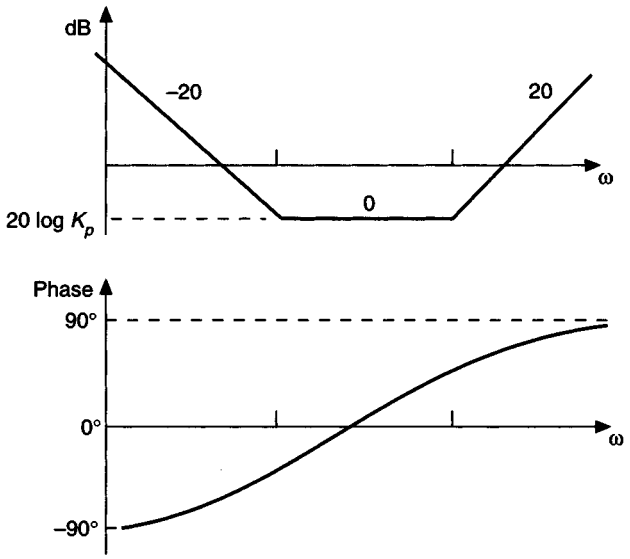


FIGURE 100.33 Bode diagram of a PID compensator.

A block diagram portrayal of the compensator is shown in Fig. 100.32. The integrator in this compensator increases the system type by one, resulting in an improved low-frequency response. The Bode diagram of a PID compensator is given in Fig. 100.33.

With $K_D = 0$, the compensator is phase-lag, with the pole in (100.65) moved to $\omega_p = 0$. As a result the compensator is type one. The zero of the compensator is placed in the low-frequency range to correspond to the zero of the phase-lag compensator discussed above.

With $K_I = 0$, the compensator is phase-lead, with a single zero and the pole moved to infinity. Hence, the gain continues to increase with increasing frequency. If high-frequency noise is a problem, it may be necessary to add one or more poles to the PD or PID compensators. These poles must be placed at high frequencies relative to the phase-margin frequency such that the phase margin (stability characteristics) of the system is not degraded. PD compensators realized using rate sensors minimize noise problems [Phillips and Harbor, 1991].

Root Locus Design

Root locus design procedures generally result in the placement of the two dominant poles of the closed-loop system transfer function. A system has two dominant poles if its behavior approximates that of a second-order system.

The differences in root locus designs and frequency response designs appear only in the interpretation of the control-system specifications. A root locus design that improves the low-frequency characteristics of the system will result in a phase-lag controller; a phase-lead compensator results if the design improves the high-frequency response of the system. If a root locus design is performed, the frequency response characteristics of the system should be investigated. Also, if a frequency response design is performed, the poles of the closed-loop transfer function should be calculated.

Modern Control Design

The classical design procedures above are based on a transfer-function model of a system. Modern design procedures are based on a *state-variable model* of the plant. The plant transfer function is given by

$$\frac{Y(s)}{U(s)} = G_p(s) \tag{100.67}$$

where we use $u(t)$ for the plant input and $y(t)$ for the plant output. If the system model is nth order, the denominator of $G_p(s)$ is an nth-order polynomial.

The state-variable model, or state model, for a single-input–single-output plant is given by

$$\begin{aligned} \frac{d\mathbf{x}(t)}{dt} &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}\mathbf{x}(t) \end{aligned} \tag{100.68}$$

$$y(t) = \mathbf{C}\mathbf{x}(t)$$

where $\mathbf{x}(t)$ is the $n \times 1$ state vector, $u(t)$ is the plant input, $y(t)$ is the plant output, $\mathbf{A}$ is the $n \times n$ *system matrix,* $\mathbf{B}$ is the $n \times 1$ *input matrix*, and $\mathbf{C}$ is the $1 \times n$ *output matrix*. The transfer function of (100.67) is an input-output model; the state model of (100.68) yields the same input-output model and in addition includes an internal model of the system. The state model of (100.68) is readily adaptable to a multiple-input–multiple-output system (*a multivariable system*); for that case, $u(t)$ and $y(t)$ are vectors. We will consider only single-input–single-output systems.

The plant transfer function of (100.67) is related to the state model of (100.68) by

$$G_p(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \tag{100.69}$$

The state model is not unique; many combinations of the matrices **A, B,** and **C** can be found to satisfy (100.69) for a given transfer function $G_p(s)$.

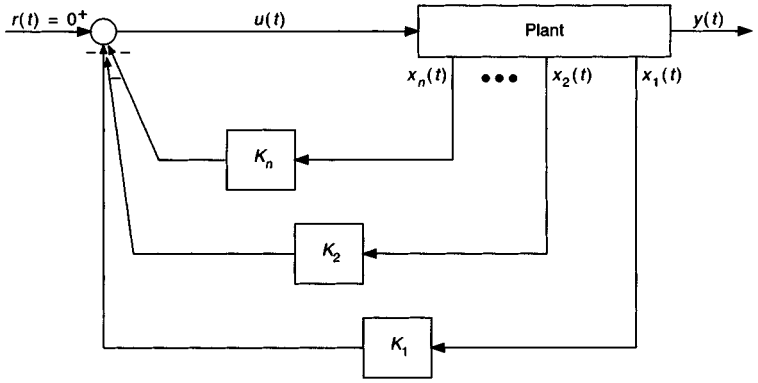


FIGURE 100.34 Implementation of pole-placement design.

Classical compensator design procedures are based on the open-loop function $G_p(s)H(s)$ of Fig. 100.28. It is common to present modern design procedures as being based on only the plant model of (100.68). However, the models of the sensors that measure the signals for feedback must be included in the state model. This problem will become more evident as the modern procedures are presented.

Pole Placement

Probably the simplest modern design procedure is *pole placement.* Recall that root locus design was presented as placing the two dominant poles of the closed-loop transfer function at desired locations. The pole-placement procedure places *all* poles of the closed-loop transfer function, or equivalently, all roots of the closed-loop system characteristic equation, at desirable locations.

The system design specifications are used to generate the desired closed-loop system characteristic equation $\alpha_c(s)$, where

$$\alpha_c(s) = s^n + \alpha_{n-1}s^{n-1} + \ldots + \alpha_1 s + \alpha_0 = 0 \tag{100.70}$$

for an nth-order plant. This characteristic equation is realized by requiring the plant input to be a linear combination of the plant states, that is,

$$u(t) = -K_1x_1(t) - K_2x_2(t) - \ldots - K_nx_n(t) = -\mathbf{K}\mathbf{x}(t) \tag{100.71}$$

where $\mathbf{K}$ is the $1 \times n$ feedback-gain matrix. Hence *all* states must be measured and fed back. This operation is depicted in Fig. 100.34.

The feedback-gain matrix $\mathbf{K}$ is determined from the desired characteristic equation for the closed-loop system of (100.70):

$$\alpha_c(s) = |sI - A + BK| = 0 \tag{100.72}$$

The state feedback gain matrix $\mathbf{K}$ which yields the specified closed-loop characteristic equation $\alpha_c(s)$ is

$$\mathbf{K} = [0\ 0\ \ldots\ 0\ 1][\mathbf{B}\ \mathbf{AB}\ \ldots\ \mathbf{A}^{n-1}\mathbf{B}]^{-1}\alpha_c(\mathbf{A}) \tag{100.73}$$

where $\alpha_c(\mathbf{A})$ is (100.70) with the scalar s replaced with the matrix $\mathbf{A}$. A plant is said to be *controllable* if the inverse matrix in (100.73) exists. Calculation of $\mathbf{K}$ completes the design process. A simple computer algorithm is available for solving (100.73) for $\mathbf{K}$.

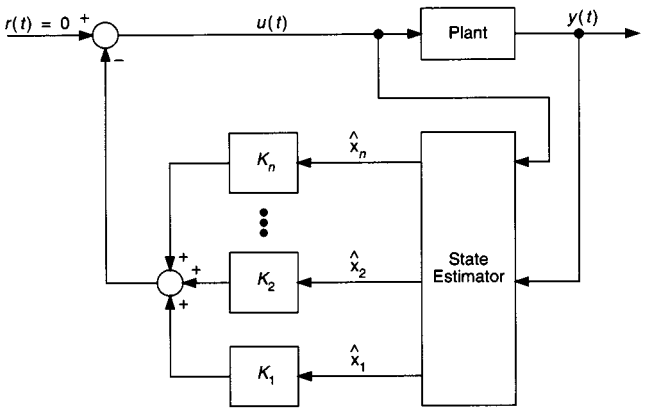


FIGURE 100.35 Implementation of observer-pole-placement design.

State Estimation

In general, modern design procedures require that the state vector $\mathbf{x}(t)$ be fed back, as in (100.71). The measurement of all state variables is difficult to implement for high-order systems. The usual procedure is to estimate the states of the system from the measurement of the output $y(t)$, with the estimated states then fed back.

Let the estimated state vector be $\hat{\mathbf{x}}$. One procedure for estimating the system states is by an *observer,* which is a dynamic system realized by the equations

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = (\mathbf{A} - \mathbf{GC})\hat{\mathbf{x}}(t) + \mathbf{B}u(t) + \mathbf{G}y(t) \tag{100.74}$$

with the feedback equation of (100.71) now realized by

$$u(t) = -\mathbf{K}\hat{\mathbf{x}}(t) \tag{100.75}$$

The matrix $\mathbf{G}$ in (100.74) is calculated by assuming an nth-order characteristic equation for the observer of the form

$$\alpha_e(s) = |sI - A + GC| = 0 \tag{100.76}$$

The estimator gain matrix $\mathbf{G}$ which yields the specified estimator characteristic equation $\alpha_e(s)$ is

$$\mathbf{G} = \alpha_e(\mathbf{A})[\mathbf{C}\ \mathbf{CA}\ \ldots\ \mathbf{CA}^{n-1}]^{-T}[0\ 0\ \ldots\ 0\ 1]^T \tag{100.77}$$

where $[\cdot]^T$ denotes the matrix transpose. A plant is said to be *observable* if the inverse matrix in (100.77) exists. An implementation of the closed-loop system is shown in Fig. 100.35. The observer is usually implemented on a digital computer. The plant and the observer in Fig. 100.35 are both nth-order; hence, the closed-loop system is of order $2n$.

The observer-pole-placement system of Fig. 100.35 is equivalent to the system of Fig. 100.36, which is of the form of closed-loop systems designed by classical procedures. The transfer function of the controller-estimator (equivalent compensator) of Fig. 100.36 is given by

$$\mathbf{G}_{ec}(s) = \mathbf{K}[s\mathbf{I} - \mathbf{A} + \mathbf{GC} + \mathbf{BK}]^{-1}\mathbf{G} \tag{100.78}$$

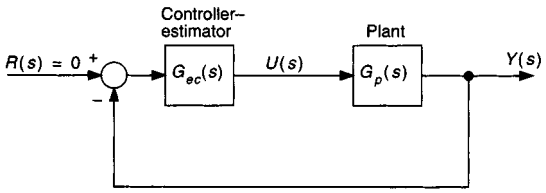


FIGURE 100.36 Equivalent system for pole-placement design.

This compensator is *n*th-order for an *n*th-order plant; hence, the total system is of order 2*n*. The characteristic equation for the compensated system is given by

$$\left|s\mathbf{I} - \mathbf{A} + \mathbf{BK}\right|\left|s\mathbf{I} - \mathbf{A} + \mathbf{GC}\right| = \alpha_c(s)\alpha_e(s) = 0 \tag{100.79}$$

The roots of this equation are the roots of the pole-placement design plus those of the observer design. For this reason, the roots of the characteristic equation for the observer are usually chosen to be faster than those of the pole-placement design.

Linear Quadratic Optimal Control

We define an optimal control system as one for which some mathematical function is minimized. The function to be minimized is called the *cost function*. For steady-state linear quadratic optimal control the cost function is given by

$$V_\infty = \int_t^\infty \left[\mathbf{x}^T(\tau)\mathbf{Q}\mathbf{x}(\tau) + Ru^2(\tau)\right] d\tau \tag{100.80}$$

where Q and R are chosen to satisfy the design criteria. In general, the choices are not straightforward. Minimization of (100.80) requires that the plant input be given by

$$u(t) = -R^{-1}\mathbf{B}^T\mathbf{M}_\infty\mathbf{x}(t) \tag{100.81}$$

where the $n \times n$ matrix $\mathbf{M}_\infty$ is the solution to the *algebraic Riccati equation*

$$\mathbf{M}_\infty\mathbf{A} + \mathbf{A}^T\mathbf{M}_\infty - \mathbf{M}_\infty\mathbf{B}R^{-1}\mathbf{B}^T\mathbf{M}_\infty + \mathbf{Q} = 0 \tag{100.82}$$

The existence of a solution for this equation is involved [Friedland, 1986] and is not presented here. Optimal control systems can be designed for cost functions other than that of (100.80).

Other Modern Design Procedures

Other modern design procedures exist; for example, *self-tuning control systems* continually estimate certain plant parameters and adjust the compensator based on this estimation. These control systems are a type of *adaptive control systems* and usually require that the control algorithms be implemented using a digital computer. These control systems are beyond the scope of this book (see, for example, Astrom and Wittenmark, 1984).

Defining Term

Compensation: The process of physically altering a closed-loop system such that the system has specified characteristics. This alteration is achieved either by changing certain parameters in the system or by adding a physical system to the closed-loop system; in some cases both methods are used.

Related Topics

100.3 Frequency Response Methods: Bode Diagram Approach • 100.4 Root Locus

References

K. J. Astrom and B. Wittenmark, *Computer Controlled Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1984.
W. L. Brogan, *Modern Control Theory,* Englewood Cliffs, N.J.: Prentice-Hall, 1985.
R. C. Dorf, *Modern Control Systems,* 7th ed., Reading, Mass.: Addison-Wesley, 1995.
G. F. Franklin, J. D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems,* Reading, Mass.: Addison-Wesley, 1986.
B. Friedland, *Control System Design,* New York: McGraw-Hill, 1986.
B. C. Kuo, *Automatic Control Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1987.
C. L. Phillips and R. D. Harbor, *Feedback Control Systems,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1991.

100.6 Digital Control Systems

Raymond G. Jacquot and John E. McInroy

The use of the **digital computer** to control physical processes has been a topic of discussion in the technical literature for over four decades, but the actual use of a digital computer for control of industrial processes was reserved only for massive and slowly varying processes such that the high cost and slow computing speed of available computers could be tolerated. The invention of the integrated circuit microprocessor in the early 1970s radically changed all that; now microprocessors are used in control tasks in automobiles and household appliances, applications where high cost is not justifiable.

When the term *digital control* is used, it usually refers to the process of employing a digital computer to control some process that is characterized by continuous-in-time dynamics. The control can be of the open-loop variety where the control strategy output by the digital computer is dictated without regard to the status of the process variables. An alternative technique is to supply the digital computer with digital data about the process variables to be controlled, and thus the control strategy output by the computer depends on the process variables that are to be controlled. This latter strategy is a **feedback control** strategy wherein the computer, the process, and interface hardware form a closed loop of information flow.

Examples of dynamic systems that are controlled in such a closed-loop digital fashion are flight control of civilian and military aircraft, control of process variables in chemical processing plants, and position and force control in industrial robot manipulators. The simplest form of feedback control strategy provides an on-off control to the controlling variables based on measured values of the process variables. This strategy will be illustrated by a simple example in a following subsection.

In the past decade and a half many excellent textbooks on the subject of digital control systems have been written, and most of them are in their second edition. The texts in the References provide in-depth development of the theory by which such systems are analyzed and designed.

A Simple Example

Such a closed-loop or feedback control situation is illustrated in Fig. 100.37, which illustrates the feedback control of the temperature in a simple environmental chamber that is to be kept at a constant temperature somewhat above room temperature.

Heat is provided by turning on a relay that supplies power to a heater coil. The on-off signal to the relay can be supplied by 1 bit of an output port of the microprocessor (typically the port would be 8 bits wide). A second bit of the port can be used to turn a fan on and off to supply cooling air to the chamber. An analog-to-digital (A/D) converter is employed to convert the amplified thermocouple signal to a digital word that is then supplied to the input port of the microprocessor. The program being executed by the microprocessor reads the temperature data supplied to the input port and compares the binary number representing the temperature to a binary

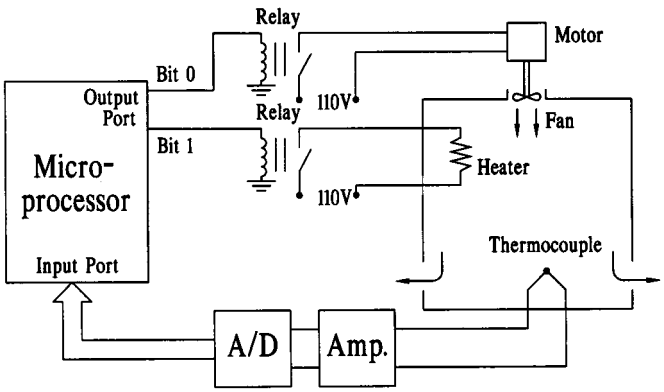


FIGURE 100.37 Microprocessor control of temperature in a simple environmental chamber.

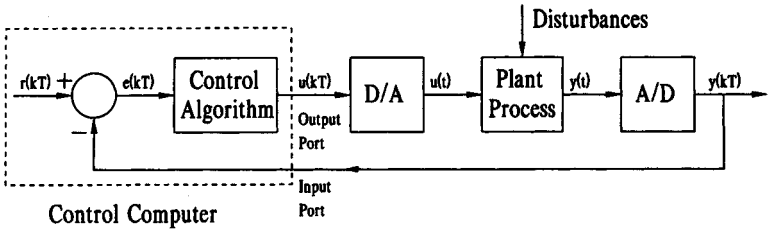


FIGURE 100.38 Closed-loop control of a single process variable.

version of the desired temperature and makes a decision whether or not to turn on the heater or the fan or to do nothing. The program being executed runs in a continuous loop, repeating the operations discussed above.

This simple on-off control strategy is often not the best when extremely precise control of the process variables is required. A more precise control may be obtained if the controlling variable levels can be adjusted to be somewhat larger if the deviation of the process variable from the desired value is larger.

Single-Loop Linear Control Laws

Consider the case where a single variable of the process is to be controlled, as illustrated in Fig. 100.38. The output of the plant $y(t)$ is to be sampled every T seconds by an A/D converter, and this sequence of numbers will be denoted as $y(kT)$, $k = 0, 1, 2, \ldots$. The goal is to make the sequence $y(kT)$ follow some desired known sequence [the reference sequence $r(kT)$]. Consequently, the sequence $y(kT)$ is subtracted from $r(kT)$ to obtain the so-called error sequence $e(kT)$. The control computer then acts on the error sequence, using some control algorithms, to produce the control effort sequence $u(kT)$ that is supplied to the digital-to-analog (D/A) converter which then drives the actuating hardware with a signal proportional to $u(kT)$. The output of the D/A converter is then held constant on the current time interval, and the control computer waits for the next sample of the variable to be controlled, the arrival of which repeats the sequence. The most commonly employed control algorithm or control law is a linear difference equation of the form

$$u(kT) = a_n e(kT) + a_{n-1} e((k-1)T) + \ldots + a_0 e((k-n)T)$$
$$+ b_{n-1} u((k-1)T) + \ldots + b_0 u((k-n)T) \qquad (100.83)$$

The question remains as to how to select the coefficients $a_0, \ldots, a_n$ and $b_0, \ldots, b_{n-1}$ in expression (100.83) to give an acceptable degree of control of the plant.

Proportional Control

This is the simplest possible control algorithm for the digital processor wherein the most current control effort is proportional to the current error or using only the first term of relation (100.83)

$$u(kT) = a_n e(kT) \tag{100.84}$$

This algorithm has the advantage that it is simple to program, while, on the other hand, its disadvantage lies in the fact that it has poor disturbance rejection properties in that if a_n is made large enough for good disturbance rejection, the closed-loop system can be unstable (i.e., have transient responses which increase with time). Since the object is to regulate the system output in a known way, these unbounded responses preclude this regulation.

PID Control Algorithm

A common technique employed for decades in chemical process control loops is that of proportional-plus-integral-plus-derivative (PID) control wherein a continuous-time control law would be given by

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de}{dt} \tag{100.85}$$

This would have to be implemented by an analog filter.

To implement the design in digital form the proportional term can be carried forward as in relation (100.84); however, the integral can be replaced by trapezoidal integration using the error sequence, while the derivative can be replaced with the backward difference resulting in a computer control law of the form [Jacquot, 1995]

$$\begin{aligned} u(kT) = u((k-1)T) &+ \left(K_p + \frac{K_i T}{2} + \frac{K_d}{T} \right) e(kT) \\ &+ \left(\frac{K_i T}{2} - K_p - \frac{2K_d}{T} \right) e((k-1)T) + \frac{K_d}{T} e((k-2)T) \end{aligned} \tag{100.86}$$

where T is the duration of the sampling interval. The selection of the coefficients in this algorithm (K_p, K_d, and K_p) is best accomplished by the Ziegler-Nichols tuning process [Franklin et al., 1990].

The Closed-Loop System

When the plant process is linear or may be linearized about an operating point and the control law is linear as in expressions (100.83), (100.84), or (100.86), then an appropriate representation of the complete closed-loop system is by the so-called z-transform. The z-transform plays the role for linear, constant-coefficient difference equations that the Laplace transform plays for linear, constant-coefficient differential equations. This z-domain representation allows the system designer to investigate system time response, frequency response, and stability in a single analytical framework.

If the plant can be represented by an s-domain transfer function $G(s)$, then the discrete-time (z-domain) transfer function of the plant, the analog-to-digital converter, and the driving digital-to-analog converter is

$$G(z) = \left(\frac{z-1}{z} \right) Z \left\{ L^{-1} \left[\frac{G(s)}{s} \right] \right\} \tag{100.87}$$

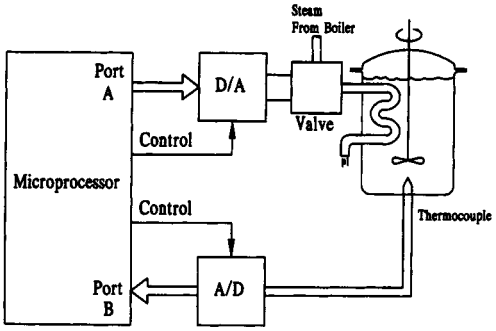


FIGURE 100.39 A computer-controlled thermal mixing tank.

where $Z(\cdot)$ is the z-transform and $L^{-1}(\cdot)$ is the inverse Laplace transform. The transfer function of the control law of (100.83) is

$$D(z) = \frac{U(z)}{E(z)} = \frac{a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0}{z^n - b_{n-1} z^{n-1} - \cdots - b_0} \tag{100.88}$$

For the closed-loop system of Fig. 100.38 the closed-loop z-domain transfer function is

$$M(z) = \frac{Y(z)}{R(z)} = \frac{G(z)D(z)}{1 + G(z)D(z)} \tag{100.89}$$

where $G(z)$ and $D(z)$ are specified above. The characteristic equation of the closed-loop system is

$$1 + G(z)D(z) = 0 \tag{100.90}$$

The dynamics and stability of the system can be assessed by the locations of the zeros of (100.90) (the closed-loop poles) in the complex z plane. For stability the zeros of (100.90) above must be restricted to the unit circle of the complex z plane.

A Linear Control Example

Consider the temperature control of a chemical mixing tank shown in Fig. 100.39. From a transient power balance the differential equation relating the rate of heat added $q(t)$ to the deviation in temperature from the ambient $\theta(t)$ is given as

$$\frac{d\theta}{dt} + \frac{1}{\tau}\theta = \frac{1}{mc} q(t) \tag{100.91}$$

where τ is the time constant of the process and mc is the heat capacity of the tank. The transfer function of the tank is

$$\frac{\Theta(s)}{Q(s)} = G(s) = \frac{1/mc}{s + 1/\tau} \tag{100.92}$$

The heater is driven by a D/A converter, and the temperature measurement is sampled with an A/D converter. The data converters are assumed to operate synchronously, so the discrete-time transfer function of the tank and the two data converters is from expression (100.87):

$$G(z) = \frac{\Theta(z)}{Q(z)} = \frac{\tau}{mc} \frac{1 - e^{-T/\tau}}{z - e^{-T/\tau}} \tag{100.93}$$

If a proportional control law is chosen, the transfer function associated with the control law is the gain $a_n = K$ or

$$D(z) = K \tag{100.94}$$

The closed-loop characteristic equation is from (100.90):

$$1 + \frac{K\tau}{mc} \frac{1 - e^{-T/\tau}}{z - e^{-T/\tau}} = 0 \tag{100.95}$$

If a common denominator is found, the resulting numerator is

$$z - e^{-T/\tau} + \frac{K\tau}{mc}(1 - e^{-T/\tau}) = 0 \tag{100.96}$$

The root of this equation is

$$z = e^{-T/\tau} + \frac{K\tau}{mc}(e^{-T/\tau} - 1) \tag{100.97}$$

If this root location is investigated as the gain parameter K is varied upward from zero, it is seen that the root starts at $z = e^{-T/\tau}$ for $K = 0$ and moves to the left along the real axis as K increases. Initially it is seen that the system becomes faster, but at some point the responses become damped and oscillatory, and as K is further increased the oscillatory tendency becomes less damped, and finally a value of K is reached where the oscillations are sustained at constant amplitude. A further increase in K will yield oscillations that increase with time. Typical unit step responses for $r(k) = 1$ and $T/\tau = 0.2$ are shown in Fig. 100.40.

It is easy to observe this tendency toward oscillation as K increases, but a problem that is clear from Fig. 100.40 is that in the steady state there is a persistent error between the response and the reference $[r(k) = 1]$. Increasing the gain K will make this error smaller at the expense of more oscillations. As a remedy for this steady-state error problem and control of the dynamics, a control law transfer function $D(z)$ will be sought that inserts integrator action into the loop while simultaneously canceling the pole of the plant. This dictates that the controller have a transfer function of the form

$$D(z) = \frac{U(z)}{E(z)} = \frac{K(z - e^{-T/\tau})}{z - 1} \tag{100.98}$$

Typical unit step responses are illustrated in Fig. 100.41 for several values of the gain parameter. The control law that must be programmed in the digital processor is

$$u(kT) = u((k-1)T) + K[e(kT) - e^{-T/\tau}e((k-1)T)] \tag{100.99}$$

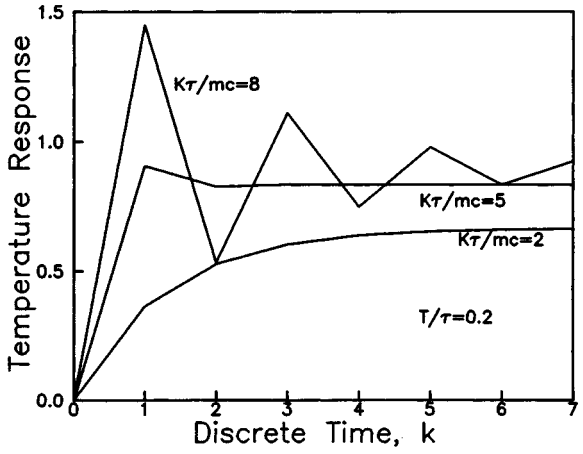


FIGURE 100.40 Step responses of proportionally controlled thermal mixing tank.

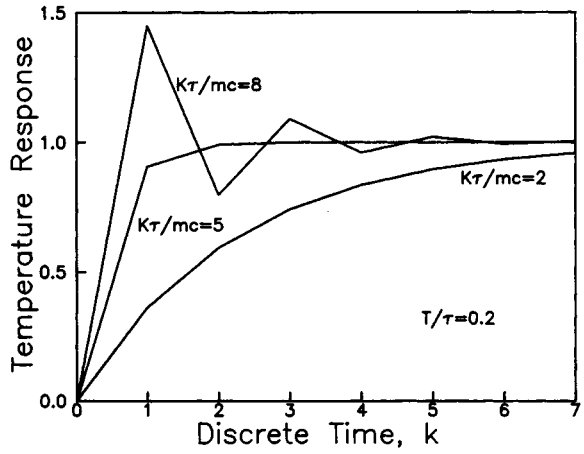


FIGURE 100.41 Step responses of the compensated thermal mixing tank.

The additional effort to program this over that required to program the proportional control law of (100.94) is easily justified since K and $e^{-T/\tau}$ are simply constants.

Defining Terms

Digital computer: A collection of digital devices including an arithmetic logic unit (ALU), read-only memory (ROM), random-access memory (RAM), and control and interface hardware.

Feedback control: The regulation of a response variable of a system in a desired manner using measurements of that variable in the generation of the strategy of manipulation of the controlling variables.

Related Topics

8.1 Introduction • 112.1 Introduction • 112.3 The State of the Art in CACSD

References

K.J. Astrom and B. Wittenmark, *Computer Controlled Systems: Theory and Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1984.

G.F. Franklin, J.D. Powell, and M.L. Workman, *Digital Control of Dynamic Systems,* 2nd ed., Reading, Mass.: Addison-Wesley, 1990.
C.H. Houpis and G.B. Lamont, *Digital Control Systems: Theory, Hardware, Software,* 2nd ed., New York: McGraw-Hill, 1992.
R.G. Jacquot, *Modern Digital Control Systems,* 2nd ed., New York: Marcel Dekker, 1995.
B.C. Kuo, *Digital Control Systems,* 2nd ed., Orlando, Fla.: Saunders, 1992.
C.L. Phillips and H.T. Nagle, *Digital Control System Analysis and Design,* 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1995.
R. J. Vaccaro, *Digital: A State-Space Approach,* New York: McGraw-Hill, 1995.

Further Information

The *IEEE Control Systems Magazine* is a useful information source on control systems in general and digital control in particular. Highly technical articles on the state of the art in digital control may be found in the *IEEE Transactions on Automatic Control, the IEEE Transactions on Control Systems Technology,* and the ASME *Journal of Dynamic Systems, Measurement and Control.*

100.7 Nonlinear Control Systems[3]

Derek P. Atherton

The Describing Function Method

The describing function method, abbreviated as DF, was developed in several countries in the 1940s [Atherton, 1982], to answer the question: "What are the necessary and sufficient conditions for the nonlinear feedback system of Fig. 100.42 to be stable?" The problem still remains unanswered for a system with static nonlinearity, $n(x)$, and linear plant $G(s)$. All of the original investigators found limit cycles in control systems and observed that, in many instances with structures such as Fig. 100.42, the wave form of the oscillation at the input to the nonlinearity was almost sinusoidal. If, for example, the nonlinearity in Fig. 100.42 is an ideal relay, that is has an on-off characteristic, so that an odd symmetrical input wave form will produce a square wave at its output, the output of $G(s)$ will be almost sinusoidal when $G(s)$ is a low pass filter which attenuates the higher harmonics in the square wave much more than the fundamental. It was, therefore, proposed that the nonlinearity should be represented by its gain to a sinusoid and that the conditions for sustaining a sinusoidal limit cycle be evaluated to assess the stability of the feedback loop. Because of the nonlinearity, this gain in response to a sinusoid is a function of the amplitude of the sinusoid and is known as the describing function. Because describing function methods can be used other than for a single sinusoidal input, the technique is referred to as the single sinusoidal DF or sinusoidal DF.

The Sinusoidal Describing Function

For the reasons explained above, if we assume in Fig. 100.42 that $x(t) = a \cos \theta$, where $\theta = \omega t$ and $n(x)$ is a symmetrical odd nonlinearity, then the output $y(t)$ will be given by the Fourier series,

$$y(\theta) = \sum_{n=0}^{\infty} a_n \cos n\theta + b_n \sin n\theta, \tag{100.100}$$

$$\text{where } a_0 = 0, \tag{100.101}$$

[3]The material in this section was previously published by CRC Press in *The Control Handbook,* William S. Levine, Ed., 1996.

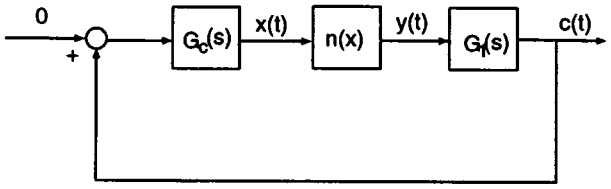


FIGURE 100.42 Block diagram of a nonlinear system.

$$a_1 = (1/\pi)\int_0^{2\pi} y(\theta)\cos\theta d\theta, \tag{100.102}$$

and

$$b_1 = (1/\pi)\int_0^{2\pi} y(\theta)\sin\theta d\theta. \tag{100.103}$$

The fundamental output from the nonlinearity is $a_1 \cos\theta + b_1 \sin\theta$, so that the describing function, DF, defined as the fundamental output divided by the input amplitude, is complex and given by

$$N(a) = (a_1 - jb_1)/a \tag{100.104}$$

which may be written

$$N(a) = N_p(a) + jN_q(a) \tag{100.105}$$

where

$$N_p(a) = a_1/a \text{ and } N_q(a) = -b_1/a\,. \tag{100.106}$$

Alternatively, in polar coordinates,

$$N(a) = M(a)e^{j\Psi(a)} \tag{100.107}$$

where

$$M(a) = \left(a_1^2 + b_1^2\right)^{1/2}/a$$

and

$$\Psi(a) = -\tan^{-1}(b_1/a_1). \tag{100.108}$$

If $n(x)$ is single valued, then $b_1 = 0$ and

$$a_1 = (4/\pi)\int_0^{\pi/2} y(\theta)\cos\theta d\theta \tag{100.109}$$

giving

$$N(a) = a_1/a = (4/a\pi)\int_0^{\pi/2} y(\theta)\cos\theta d\theta \tag{100.110}$$

Although Eqs. (100.102) and (100.103) are an obvious approach to evaluating the fundamental output of a nonlinearity, they are indirect, because one must first determine the output wave form $y(\theta)$ from the known nonlinear characteristic and sinusoidal input wave form. This is avoided if the substitution $\theta = \cos^{-1}(x/a)$ is made. After some simple manipulations,

$$a_1 = (4/a)\int_0^a x n_p(x) p(x) dx \tag{100.111}$$

and

$$b_1 = (4/a\pi)\int_0^a n_q(x) dx. \tag{100.112}$$

The function $p(x)$ is the amplitude probability density function of the input sinusoidal signal given by

$$p(x) = (1/\pi)(a^2 - x^2)^{-1/2}. \tag{100.113}$$

The nonlinear characteristics $n_p(x)$ and $n_q(x)$, called the inphase and quadrature nonlinearities, are defined by

$$n_p(x) = [n_1(x) + n_2(x)]/2 \tag{100.114}$$

and

$$n_q(x) = [n_2(x) - n_1(x)]/2 \tag{100.115}$$

where $n_1(x)$ and $n_2(x)$ are the portions of a double-valued characteristic traversed by the input for $\dot{x} > 0$ and $\dot{x} < 0$, respectively, When the nonlinear characteristic is single-valued, $n_1(x) = n_2(x)$, so $n_p(x) = n(x)$ and $n_q(x) = 0$. Integrating Eq. (100.111) by parts yields

$$a_1 = (4/\pi)n(0^+) + (4/a\pi)\int_0^a n'(x)(a^2 - x^2)^{1/2} dx \tag{100.116}$$

where $n'(x) = dn(x)/dx$ and $n(0^+) = \lim_{\epsilon\to\infty} n(\epsilon)$, a useful alternative expression for evaluating a_1.

An additional advantage of using Eqs. (100.111) and (100.112) is that they yield proofs of some properties of the DF for symmetrical odd nonlinearities. These include the following:

1. For a double-valued nonlinearity, the quadrature component $N_q(a)$ is proportional to the area of the nonlinearity loop, that is,

$$N_q(a) = -(1/a^2\pi)(\text{area of nonlinearity loop}) \tag{100.117}$$

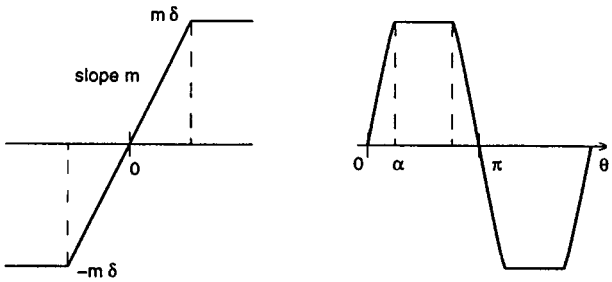


FIGURE 100.43 Saturation nonlinearity.

2. For two single-valued nonlinearities $n_\alpha(x)$ and $n_\beta(x)$, with $n_\alpha(x) < n_\beta(x)$ for all $0 < x < b$, $N_\alpha(a) < N_\beta(a)$ for input amplitudes less than b.
3. For a single-valued nonlinearity with $k_1 x < n(x) < k_2 x$ for all $0 < x < b$, $k_1 < N(a) < k_2$ for input amplitudes less than b. This is the sector property of the DF; a similar result can be obtained for a double-valued nonlinearity [Cook, 1973].

When the nonlinearity is single valued, from the properties of Fourier series, the DF, $N(a)$, may also be defined as:

1. the variable gain, K, having the same sinusoidal input as the nonlinearity, which minimizes the mean squared value of the error between the output from the nonlinearity and that from the variable gain, and
2. the covariance of the input sinusoid and the nonlinearity output divided by the variance of the input.

Evaluation of the Describing Function

To illustrate the evaluation of the DF two simple examples are considered.

Saturation Nonlinearity

To calculate the DF, the input can alternatively be taken as $a \sin \theta$. For an ideal saturation characteristic, the nonlinearity output wave form $y(\theta)$ is as shown in Fig. 100.43. Because of the symmetry of the nonlinearity, the fundamental of the output can be evaluated from the integral over a quarter period so that

$$N(a) = \frac{4}{a\pi} \int_0^{\pi/2} y(\theta) \sin \theta d\theta,$$

which, for $a > \delta$, gives

$$N(a) = \frac{4}{a\pi} \left[\int_0^{\alpha} ma \sin^2 \theta d\theta + \int_{\alpha}^{\pi/2} m\delta \sin \theta d\theta \right]$$

where $a = \sin^{-1} \delta/a$. Evaluation of the integrals gives

$$N(a) = (4m/\pi) \left[\frac{\alpha}{2} - \frac{\sin 2\alpha}{4} + \delta \cos \alpha \right]$$

which, on substituting for δ, give the result

$$N(a) = (m/\pi)(2\alpha + \sin 2\alpha). \tag{100.118}$$

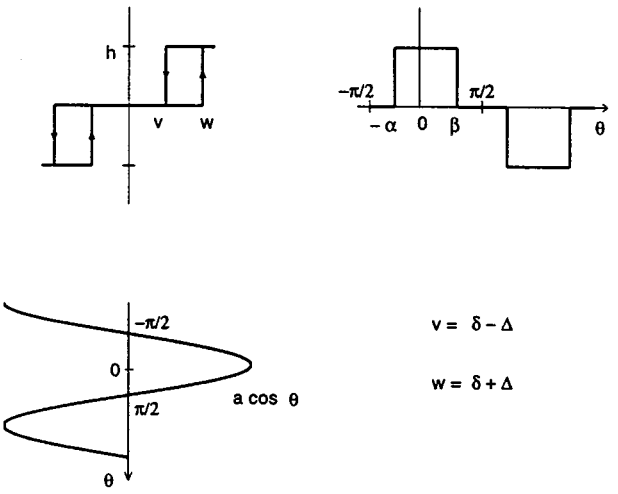


FIGURE 100.44 Relay with dead zone and hysteresis.

Because, for $a < \delta$, the characteristic is linear giving $N(a) = m$, the DF for ideal saturation is $mN_s(\delta/a)$ where

$$N_s(\delta/a) = \begin{cases} 1, & \text{for } a < \delta, \text{ and} \\ (1/\pi)[2\alpha + \sin 2\alpha], & \text{for } a > \delta, \end{cases} \tag{100.119}$$

where $a = \sin^{-1} \delta/a$.

Alternatively, one can evaluate $N(a)$ from Eq. (100.116), yielding

$$N(a) = a_1/a = (4/a^2\,\pi)\int_0^{\delta} m(a^2 - x^2)^{1/2}\,dx.$$

Using the substitution $x = a \sin \theta$,

$$N(a) = (4m/\pi)\int_0^{\alpha} \cos^2 \theta\, d\theta = (m/\pi)(2\alpha + \sin 2\alpha)$$

as before.

Relay with Dead Zone and Hysteresis

The characteristic is shown in Fig. 100.44 together with the corresponding input, assumed equal to $a \cos \theta$, and the corresponding output wave form. Using Eqs. (100.102) and (100.103) over the interval $-\pi/2$ to $\pi/2$ and assuming that the input amplitude a is greater than $\delta + \Delta$,

$$\begin{aligned} a_1 &= \left(2/\pi \int_{-\alpha}^{\beta} h \cos \theta\, d\theta\right) \\ &= (2h/\pi)(\sin \beta + \sin \alpha), \end{aligned}$$

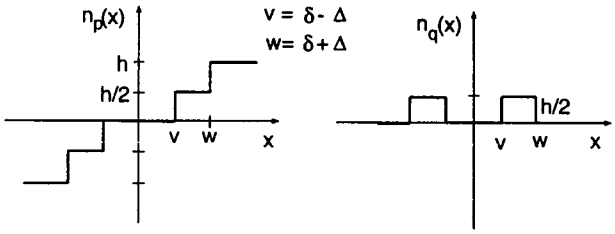


FIGURE 100.45 Function $n_p(x)$ and $n_q(x)$ for the relay of Figure 100.44.

$$\text{where } \alpha = \cos^{-1}\left[(\delta - \Delta)/a\right] \text{ and } \beta = \cos^{-1}\left[(\delta + \Delta)/a\right], \text{ and}$$

$$b_1 = (2/\pi)\int_{-\alpha}^{\beta} h \sin\theta d\theta$$

$$= (-2h/\pi)\left(\frac{(\delta + \Delta)}{a} - \frac{\delta - \Delta}{a}\right) = 4h\Delta/a\pi\,.$$

Thus

$$N(a) = \frac{2h}{a^2\pi}\left\{\left[a^2 - (\delta + \Delta)^2\right]^{1/2} + \left[a^2 - (\delta - \Delta)^2\right]^{1/2}\right\} - \frac{j4h\Delta}{a^2\pi}. \quad (100.120)$$

For the alternative approach, one must first obtain the in-phase and quadrature nonlinearities shown in Fig. 100.45. Using Eqs. (100.111) and (100.112),

$$a_1 = (4/a)\int_{\delta-\Delta}^{\delta+\Delta} x(h/2)p(x)dx + \int_{\delta+\Delta}^{a} xhp(x)dx,$$

$$= \frac{2h}{a\pi}\left\{\left[a^2 - (\delta + \Delta)^2\right]^{1/2} + \left[a^2 - (\delta - \Delta)^2\right]^{1/2}\right\},$$

and

$$b_1 = (4/a\pi)\int_{d-\Delta}^{\delta+\Delta}(h/2\,dx) = 4h\Delta/a\pi$$

$$= (\text{Area of nonlinearity loop})/a\pi$$

as before.

The DF of two nonlinearities in parallel equals the sum of their individual DFs, a result very useful for determining DFs, particularly of linear segmented characteristics with multiple break points. Several procedures [Altherton, 1982] are available for approximating the DF of a given nonlinearity either by numerical integration or by evaluating the DF of an approximating nonlinear characteristic defined, for example, by a quantized characteristic, linear segmented characteristic, or Fourier series. Table 100.3 gives a list of DFs for some commonly used approximations of nonlinear elements. Several of the results are in terms of the DF for an ideal saturation characteristic of unit slope, $N_s(\delta/a)$, defined in Eq. (100.119).

TABLE 100.3 DFs of Single-Valued Nonlinearities

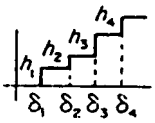
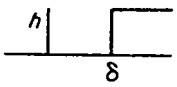
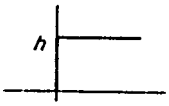
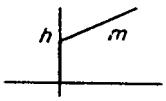
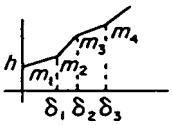
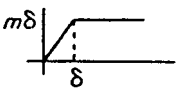
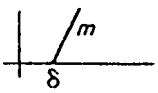
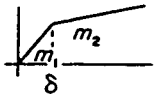
General quantizer 	$a < \delta_1$	$N_p = 0$
	$\delta_{M+1} > a > \delta_M$	$N_p = \left(4/a^2\pi\right)\sum_{m=1}^{M} h_m\left(a^2 - \delta_m^2\right)^{1/2}$
Uniform quantizer $h_1 = h_2 = \cdots h$ $\delta_m = (2m-1)\delta/2$	$a < \delta$	$N_p = 0$
	$(2M+1)\delta > a > (2M-1)\delta$	$N_p = \left(4h/a^2\pi\right)\sum_{m=1}^{M}\left(a^2 - n^2\delta^2\right)^{1/2}$
	$n = (2m-1)/2$	
Relay with dead zone 	$a < \delta$	$N_p = 0$
	$a > \delta$	$N_p = 4h(a^2 - \delta^2)^{1/2}/a^2\pi$
Ideal relay 		$N_p = 4h/a\pi$
Preload 		$N_p = (4h/a\pi) + m$
General piecewise linear 	$a < \delta_1$	$N_p = (4h/a\pi) + m_1$
	$\delta_{M+1} > \alpha > \delta_M$	$N_p = (4h/a\pi) + m_{M+1} + \sum_{i=1}^{M}\left(m_j - m_{j+1}\right)N_s\left(\delta_j/a\right)$
Ideal saturation 		$N_p = mN_s(\delta/a)$
Dead zone 		$N_p = m[1 - N_s(\delta/a)$

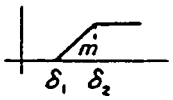
TABLE 100.3 (continued) DFs of Single-Valued Nonlinearities

Gain changing nonlinearity

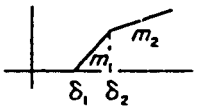


$N_p = (m_1 - m_2)N_s(\delta/a) + m_2$

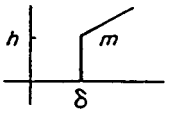
Saturation with dead zone



$N_p = m[N_s(\delta_2/a) - N_s(\delta_1/a)]$

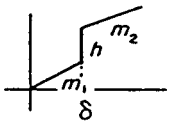


$N_p = -m_1N_s(\delta_1/a) + (m_1 - m_2)N_s(\delta_2/a) + m_2$



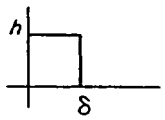
$a < \delta$ $\quad N_p = 0$

$a > \delta$ $\quad N_p = 4h(a^2 - \delta^2)^{1/2}/a_2\pi + m - mN_s(\delta/a)$



$a < \delta$ $\quad N_p = m_1$

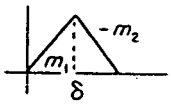
$a > \delta$ $\quad N_p = (m_1 - m_2)N_s(\delta/a) + m_2 + 4h(a^2 - \delta^2)^{1/2}/a^2\pi$



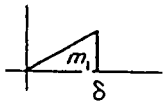
$a < \delta$ $\quad N_p = 4h/a\pi$

$a > \delta$ $\quad N_p = 4h/[a - (a^2 - \delta^2)^{1/2}]/a^2\pi$

Limited field of view



$N_p = (m_1 + m_2)N_s(\delta/a) - m_2N_s[(m_1 + m_2)\delta/m_2a]$



$a < \delta$ $\quad N_p = m_1$

$a > \delta$ $\quad N_p = m_1N_s(\delta/a) - 4m_1\delta(a^2 - \delta^2)^{1/2}/a^2\pi$

$y = x^m$ $\quad m > -2$ Γ is the gamma function

$$N_p = \frac{\Gamma(m+1)a^{m-1}}{2^{m-1}\Gamma[(3+m)/2]\Gamma[(1+m)/2]} = \frac{2}{\sqrt{\pi}}\frac{\Gamma[(m+2)\ 2]a^{m-1}}{\Gamma[(m+3)/2]}$$

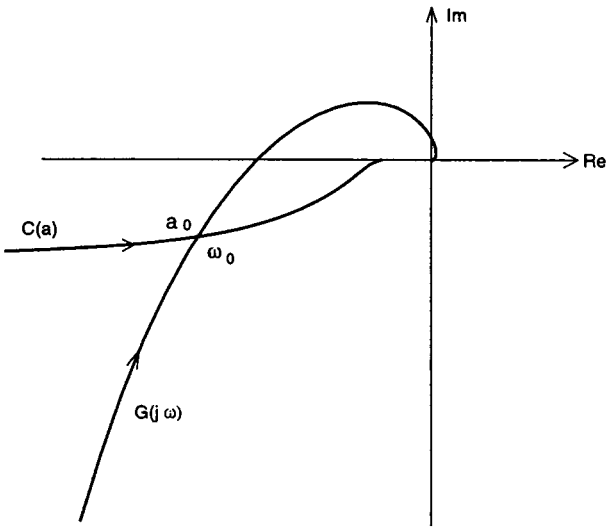


FIGURE 100.46 Nyquist plot showing solution for a limit cycle.

Limit Cycles and Stability

To investigate the possibility of limit cycles in the autonomous closed loop system of Fig. 100.42, the input to the nonlinearity $n(x)$ is assumed to be a sinusoid so that it can be replaced by the amplitude-dependent DF gain $N(a)$. The open loop gain to a sinusoid is thus $N(a)G(j\omega)$ and, therefore, a limit cycle exists if

$$N(a)G(j\omega) = -1 \tag{100.121}$$

where $G(j\omega) = G_c(j\omega)G_1(j\omega)$. As in general, $G(j\omega)$ is a complex function of ω and $N(a)$ is a complex function of a, solving Eq. (100.121) will yield both the frequency ω and amplitude a of a possible limit cycle.

A common procedure to examine solutions of Eq. (100.120) is to use a Nyquist diagram, where the $G(j\omega)$ and $C(a) = -1/N(a)$ loci are plotted as in Fig. 100.46, where they are shown intersecting for $a = a_0$ and $\omega = \omega_0$. The DF method indicates therefore that the system has a limit cycle with the input sinusoid to the nonlinearity, x, equal to $a_0 \sin(\omega_0 t + \phi)$, where ϕ depends on the initial conditions. When the $G(j\omega)$ and $C(a)$ loci do not intersect, the DF method predicts that no limit cycle will exist if the Nyquist stability criterion is satisfied for $G(j\omega)$ with respect to any point on the $C(a)$ locus. Obviously, if the nonlinearity has unit gain for small inputs, the point $(-1, j0)$ will lie on $C(a)$ and may be used as the critical point, analogous to a linear system.

For a stable case, it is possible to use the gain and phase margin to judge the relative stability of the system. However, a gain and phase margin can be found for every amplitude a on the $C(a)$ locus, so it is usually appropriate to use the minimum values of the quantities [Atherton, 1982]. When the nonlinear block includes dynamics so that its response is both amplitude and frequency dependent, that is $N(a, \omega)$, then a limit cycle will exist if

$$G(j\omega) = -1/N(a, \omega) = C(a, \omega). \tag{100.122}$$

To check for possible solutions of this equation, a family of $C(a, \omega)$ loci, usually as functions of a for fixed values of ω, is drawn on the Nyquist diagram.

An additional point of interest is whether when a solution to Eq. (100.120) exists the predicted limit cycle is stable. When there is only one intersection point, the stability of the limit cycle can be found using the Loeb criterion which states that if the Nyquist stability criterion indicates instability (stability) for the point on $C(a)$ with $a < a_0$ and stability (instability) for the point on $C(a)$ with $a > a_0$ the limit cycle is stable (unstable).

When multiple solutions exist, the situation is more complicated and the criterion above is a necessary but not sufficient result for the stability of the limit cycle [Choudhury and Atherton, 1974].

Normally in these cases, the stability of the limit cycle can be ascertained by examining the roots of the characteristic equation

$$1 + N_{i\gamma}(a)G(s) = 0 \tag{100.123}$$

where $N_{iy}(a)$ is known as the incremental describing function (IDF). $N_{iy}(a)$ for a single valued nonlinearity can be evaluated from

$$N_{i\gamma}(a) = \int_{-a}^{a} n'(x)p(x)dx \tag{100.124}$$

where $n'(x)$ and $p(x)$ are as previously defined. $N_{i\gamma}(a)$ is related to $N(a)$ by the equation

$$N_{i\gamma}(a) = N(a) + (a/2)\,dN(a)/da\,. \tag{100.125}$$

Thus, for example, for an ideal relay, making $\delta = \Delta = 0$ in Eq. (100.120) gives $N(a) = 4h/a\pi$, also found directly from Eq. (100.116), and, substituting this value in Eq. (100.125) yields $N_{i\gamma}(a) = 2h/a\pi$. Some examples of feedback system analysis using the DF follow.

Autotuning in Process Control

In 1943 Ziegler and Nichols [1943] suggested a technique for tuning the parameters of a PID controller. Their method was based on testing the plant in a closed loop with the PID controller in the proportional mode. The proportional gain was increased until the loop started to oscillate and then the value of gain and the oscillation frequency were measured. Formulae were given for setting the controller parameters based on the gain named the critical gain, K_c, and the frequency called the critical frequency, ω_c.

Assuming that the plant has a linear transfer function $G_1(s)$, then K_c is its gain margin and ω_c the frequency at which its phase shift is 180°. Performing this test in practice may prove difficult. If the plant has a linear transfer function and the gain is adjusted too quickly, a large amplitude oscillation may start to build up. In 1984 Astrom and Hagglund [1984] suggested replacing the proportional control by a relay element to control the amplitude of the oscillation. Consider therefore the feedback loop of Fig. 100.42 with $n(x)$ an ideal relay, $G_c(s) = 1$, and the plant with a transfer function $G_1(s) = 10/(s + 1)^3$. The $C(a)$ locus, $-1/N(a) = -a\pi/4h$, and the Nyquist locus $G(j\omega)$ in Fig. 100.47 intersect. The values of a and ω at the intersection can be calculated from

$$-a\pi/4h = \frac{10}{(1 + j\omega)^3} \tag{100.126}$$

which can be written

$$\text{Arg}\left(\frac{10}{(1 + j\omega)^3}\right) = 180°, \text{ and} \tag{100.127}$$

$$\frac{a\pi}{4h} = \frac{10}{(1 + \omega^2)^{3/2}}\,. \tag{100.128}$$

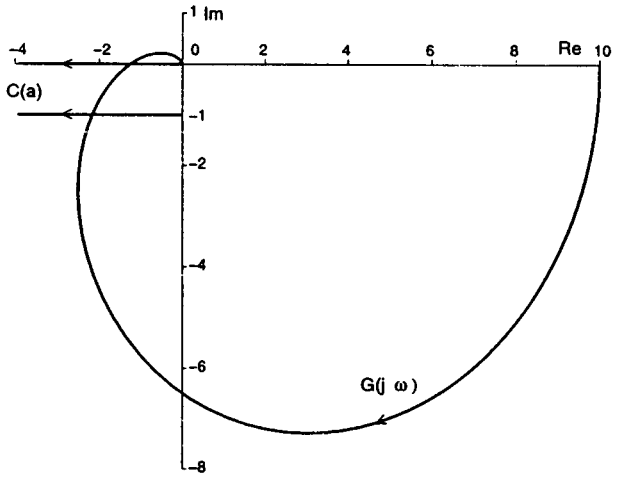


FIGURE 100.47 Nyquist plot $10/(s+1)^3$ and $C(a)$ loci for $\Delta = 0$ and $4h/\pi$.

The solution for ω_c from Eq. (100.127) is $\tan^{-1}\omega_c = 60°$, giving $\omega_c = \sqrt{3}$. Because the DF solution is approximate, the actual measured frequency of oscillation will differ from this value by an amount which will be smaller the closer the oscillation is to a sinusoid. The exact frequency of oscillation in this case will be 1.708 rads/sec in error by a relatively small amount. For a square wave input to the plant at this frequency, the plant output signal will be distorted by a small percentage. The distortion, *d*, is defined by

$$d = \left[\frac{\text{M.S. value of signal} \;-\; \text{M.S. value of fundamental harmonic}}{\text{M.S. value of fundamental harmonic}}\right]^{1/2} \tag{100.129}$$

Solving Eq. (100.128) gives the amplitude of oscillation a as $5h/\pi$. The gain through the relay is $N(a)$ equal to the critical gain K_c. In the practical situation where a is measured, K_c equal to $4h/a\pi$, should be close to the known value of 0.8 for this transfer function.

If the relay has an hysteresis of Δ, then with $\delta = 0$ in Eq. (100.120) gives

$$N(a) = \frac{4h\left(a^2 - \Delta^2\right)^{1/2}}{a^2\pi} - j\,\frac{4h\Delta}{a^2\pi}$$

from which

$$C(a) = \frac{-1}{N(a)} = \frac{-\pi}{4h}\left[\left(a^2 - \Delta^2\right)^{1/2} + j\Delta\right].$$

Thus on the Nyquist plot, $C(a)$ is a line parallel to the real axis at a distance $\pi\Delta/4h$ below it, as shown in Fig. 100.47 for $\Delta = 1$ and $h = \pi/4$ giving $C(a) = -(a^2 - 1)^{1/2} - j$. If the same transfer function is used for the plant, then the limit cycle solution is given by

$$-\left(a^2 - 1\right)^{1/2} - j = \frac{10}{\left(1 + j\omega\right)^3} \tag{100.130}$$

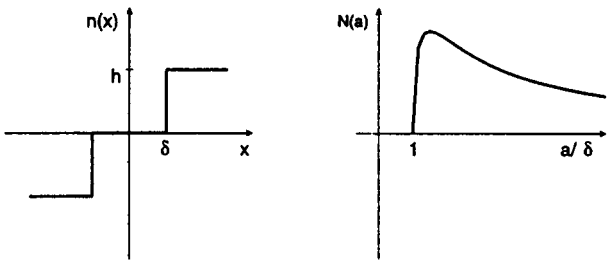


FIGURE 100.48 $N(a)$ for ideal relay with dead zone.

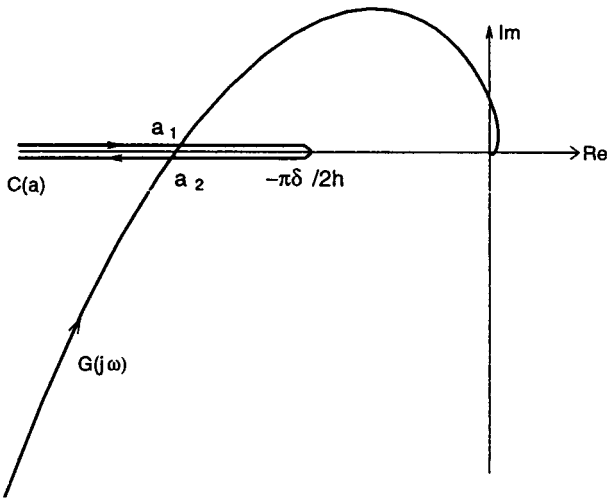


FIGURE 100.49 Two limit cycles: a_1, unstable; a_2, stable.

where $\omega = 1.266$, which compares with an exact solution value of 1.254, and $a = 1.91$. For the oscillation with the ideal relay, Eq. (100.123) with $N_{i\gamma}(a) = 2h/a\pi$ shows that the limit cycle is stable. This agrees with the perturbation approach which also shows that the limit cycle is stable when the relay has hysteresis.

Feedback Loop with a Relay with Dead Zone

For this example the feedback loop of Fig. 100.42 is considered with $n(x)$ a relay with dead zone and $G(s) = 2/s(s+1)^2$. From Equation 19.22 with $\Delta = 0$, the DF for this relay, given by

$$N(a) = 4h\left(a^2 - \delta^2\right)^{1/2} \Big/ a^2\pi \text{ for } a > \delta. \tag{100.131}$$

is real because the nonlinearity is single valued. A graph of $N(a)$ against a is in Fig. 100.48, and shows that $N(a)$ starts at zero, when $a = \delta$, increases to a maximum, with a value of $2h/\pi\delta$ at $a = \delta\sqrt{2}$, and then decreases toward zero for larger inputs. The $C(a)$ locus, shown in Fig. 100.49, lies on the negative real axis starting at $-\infty$ and returning there after reaching a maximum value of $-\pi\delta/2h$. The given transfer function $G(j\omega)$ crosses the negative real axis, as shown in Fig. 100.49, at a frequency of $\tan^{-1}\omega = 45°$, that, is $\omega = 1$ rad/sec and, therefore, cuts the $C(a)$ locus twice. The two possible limit cycle amplitudes at this frequency can be found by solving

$$\frac{a^2\pi}{4h\left(a^2 - \delta^2\right)^{1/2}} = 1$$

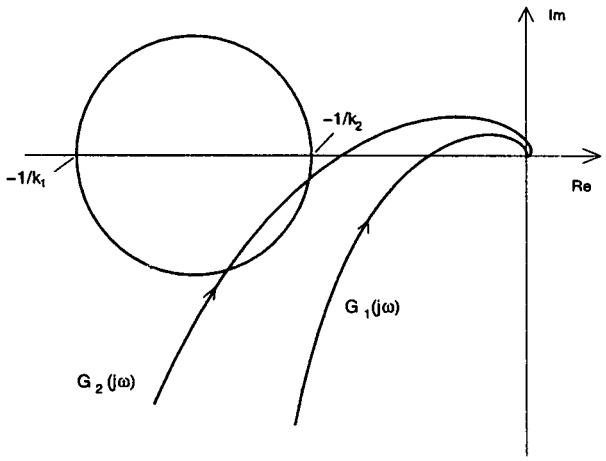


FIGURE 100.50 Circle criterion and stability.

which gives $a = 1.04$ and 3.86 for $\delta = 1$ and $h = \pi$. Using the perturbation method or the IDF criterion, the smallest amplitude limit cycle is unstable and the larger one is stable. If a condition similar to the lower amplitude limit cycle is excited in the system, an oscillation will build up and stabilize at the higher amplitude limit cycle.

Other techniques show that the exact frequencies of the limit cycles for the smaller and larger amplitudes are 0.709 and 0.989, respectively. Although the transfer function is a good low pass filter, the frequency of the smallest amplitude limit cycle is not predicted accurately because the output from the relay, a wave form with narrow pulses, is highly distorted.

If the transfer function of $G(s)$ is $K/s(s + 1)^2$, then no limit cycle will exist in the feedback loop, and it will be stable if

$$\left.\frac{K}{\omega\left(1 + \omega^2\right)}\right|_{\omega=1} < \frac{\pi d}{2h},$$

that is, $K < \pi\delta/h$. If $\delta = 1$ and $h = \pi$, $K < 1$ which may be compared with the exact result for stability of $K < 0.96$.

Stability and Accuracy

Because the DF method is an approximate procedure, it is desirable to judge its accuracy. Predicting that a system will be stable, when in practice it is not, may have unfortunate consequences. Many attempts have been made to solve this problem, but those obtained are difficult to apply or produce too conservative results [Mess and Bergen, 1975].

The problem is illustrated by the system of Fig. 100.42 with a symmetrical odd single-valued nonlinearity confined to a sector between lines of slope k_1 and k_2, that is, $k_1x < n(x) < k_2x$ for $x > 0$. For absolute stability, the circle criterion requires satisfying the Nyquist criterion for the locus $G(j\omega)$ for all points within a circle having its diameter on the negative real axis of the Nyquist diagram between the points $(-1/k_1, 0)$ and $(-1/k_2, 0)$, as shown in Fig. 100.50. On the other hand, because the DF for this nonlinearity lies within the diameter of the circle, the DF method requires satisfying the Nyquist criterion for $G(j\omega)$ for all points on the circle diameter, if the autonomous system is to be stable.

Therefore, for a limit cycle in the system of Fig. 100.42, errors in the DF method relate to its inability to predict a phase shift, which the fundamental harmonic may experience in passing through the nonlinearity, rather than an incorrect magnitude of the gain. When the input to a single-valued nonlinearity is a sinusoid together with some of its harmonics, the fundamental output is not necessarily in phase with the fundamental

input, that is, the fundamental gain has a phase shift. The actual phase shift varies with the harmonic content of the input signal in a complex manner, because the phase shift depends on the amplitudes and phases of the individual input components.

From an engineering viewpoint one can judge the accuracy of DF results by estimating the distortion, *d,* in the input to the nonlinearity. This is straightforward when a limit-cycle solution is given by the DF method; the loop may be considered opened at the nonlinearity input, the sinusoidal signal corresponding to the DF solution can be applied to the nonlinearity, and the harmonic content of the signal fed back to the nonlinearity input can be calculated. Experience indicates that the percentage accuracy of the DF method in predicting the fundamental amplitude and frequency of the limit cycle is less than the percentage distortion in the fedback signal. As mentioned previously, the DF method may incorrectly predict stability. To investigate this problem, the procedure above can be used again, by taking, as the nonlinearity input, a sinusoid with amplitude and frequency corresponding to values of those parameters where the phase margin is small. If the calculated fedback distortion is high, say greater than 2% per degree of phase margin, the DF result should not be relied on.

The limit-cycle amplitude predicted by the DF is an approximation to the fundamental harmonic. The accuracy of this prediction cannot be assessed by using the peak value of the limit cycle to estimate an equivalent sinusoid. It is possible to estimate the limit cycle more accurately by balancing more harmonics, as mentioned earlier. Although this is difficult algebraically other than with loops whose nonlinearity is mathematically simply described, for example a cubic, software is available for this purpose [McNamara and Atherton, 1987]. The procedure involves solving sets of nonlinear algebraic equations but good starting guesses can usually be obtained for the magnitudes and phases of the other harmonic components from the wave form fedback to the nonlinearity, assuming its input is the DF solution.

Compensator Design

Although the design specifications for a control system are often in terms of step-response behavior, frequency domain design methods rely on the premise that the correlation between the frequency and a step response yields a less oscillatory step response if the gain and phase margins are increased. Therefore the design of a suitable linear compensator for the system of Fig. 100.42 using the DF method, is usually done by selecting for example a lead network to provide adequate gain and phase margins for all amplitudes. This approach may be used in example 2 of the previous section where a phase lead network could be added to stabilize the system, say for a gain of 1.5, for which it is unstable without compensation. Other approaches are the use of additional feedback signals or modification of the nonlinearity $n(x)$ directly or indirectly [Atherton, 1982; Gelb and van der Velde, 1968].

When the plant is nonlinear, its frequency response also depends on the input sinusoidal amplitude represented as $G(j\omega, a)$. In recent years several approaches [Nanka-Bruce and Atherton, 1990; Taylor and Strobel, 1984] use the DF method to design a nonlinear compensator for the plant, with the objective of closed-loop performance independent of the input amplitude.

Closed-Loop Frequency Response

When the closed-loop system of Fig. 100.42 has a sinusoidal input $r(t) = R\sin(\omega t + \theta)$, it is possible to evaluate the closed-loop frequency response using the DF. If the feedback loop has no limit cycle when $r(t) = 0$ and, in addition, the sinusoidal input $r(t)$ does not induce a limit cycle, then, provided that $G_c(s)G_1(s)$ gives good filtering, $x(t)$, the nonlinearity input, almost equals the sinusoid $a \sin \omega t$. Balancing the components of frequency ω around the loop,

$$\begin{aligned} &g_c R \sin\left(\omega t + \theta - \phi_c\right) - a g_1 g_c M(a) \\ &\sin\left[\omega t + \phi_1 + \phi_c + \psi(a)\right] = a \sin \omega t \end{aligned} \tag{100.132}$$

where $G_c(j\omega) = g_c e^{j\phi_c}$ and $G_1(j\omega) = g_1 e^{j\phi_1}$. In principle Eq. (100.132), which can be written as two nonlinear algebraic equations, can be solved for the two unknowns a and θ and the fundamental output signal can then be found from

$$c(t) = aM(a)g_1 \sin\left[\omega t + \psi(a) + \phi_1\right] \tag{100.133}$$

to obtain the closed-loop frequency for R and ω.

Various graphical procedures have been proposed for solving the two nonlinear algebraic equations resulting from Eq. (100.132) [Levinson, 1953; Singh, 1965; West and Douce, 1954]. If the system is lightly damped, the nonlinear equations may have more than one solution, indicating that the frequency response of the system has a jump resonance. This phenomenon of a nonlinear system has been studied by many authors, both theoretically and practically [Lamba and Kavanagh, 1971; West et al., 1954].

The Phase Plane Method

The phase plane method was the first method used by control engineers for studying the effects of nonlinearity in feedback systems. The technique which can only be used for systems with second order models was examined and further developed for control engineering purposes for several major reasons,

1. The phase plane approach has been used for several studies of second order nonlinear differential equations arising in fields such as planetary motion, nonlinear mechanics and oscillations in vacuum tube circuits.
2. Many of the control systems of interest, such as servomechanisms, could be approximated by second order nonlinear differential equations.
3. The phase plane was particularly appropriate for dealing with nonlinearities with linear segmented characteristics which were good approximations for the nonlinear phenomena encountered in control systems.

The next section considers the basic aspects of the phase plane approach but later concentration is focused on control engineering applications where the nonlinear effects are approximated by linear segmented nonlinearities.

Background

Early analytical work [Andronov et al., 1966], on second order models assumed the equations

$$\begin{aligned} \dot{x}_1 &= P(x_1, x_2) \\ \dot{x}_2 &= Q(x_1, x_2) \end{aligned} \tag{100.134}$$

for two first-order nonlinear differential equations. Equilibrium, or singular points, occur when

$$\dot{x}_1 = \dot{x}_2 = 0$$

and the slope of any solution curve, or trajectory, in the $x_1 - x_2$ state plane is

$$\frac{dx_2}{dx_1} = \frac{\dot{x}_2}{\dot{x}_1} = \frac{Q(x_1, x_2)}{P(x_1, x_2)} \tag{100.135}$$

A second order nonlinear differential equation representing a control system can be written

$$\ddot{x} + f(x, \dot{x}) = 0 \tag{100.136}$$

If this is rearranged as two first-order equations, choosing the phase variables as the state variables, that is $x_1 = x$, $x_2 = \dot{x}$, then Eq. (100.136) can be written as

$$\dot{x}_1 = \dot{x}_2 \qquad \dot{x}_2 = -f(x_1, x_2) \tag{100.137}$$

which is a special case of Eq. (100.135). A variety of procedures has been proposed for sketching state [phase] plane trajectories for Eqs. (100.135) and (100.137). A complete plot showing trajectory motions throughout the entire state (phase) plane is known as a state (phase) portrait. Knowledge of these methods, despite the improvements in computation since they were originally proposed, can be particularly helpful for obtaining an appreciation of the system behavior. When simulation studies are undertaken, phase plane graphs are easily obtained and they are often more helpful for understanding the system behavior than displays of the variables x_1 and x_2 against time.

Many investigations using the phase plane technique were concerned with the possibility of limit cycles in the nonlinear differential equations When a limit cycle exists, this results in a closed trajectory in the phase plane. Typical of such investigations was the work of Van der Pol, who considered the equation

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0 \tag{100.138}$$

where μ is a positive constant. The phase plane form of this equation can be written as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -f(x_1, x_2) = \mu(1 - x_1^2)x_2 - x_1 \end{aligned} \tag{100.139}$$

The slope of a trajectory in the phase plane is

$$\frac{dx_2}{dx_1} = \frac{\dot{x}_2}{\dot{x}_1} = \frac{\mu(1 - x_1^2)x_2 - x_1}{x_2} \tag{100.140}$$

and this is only singular (that is, at an equilibrium point), when the right hand side of Eq. (100.140) is 0/0, that is $x_1 = x_2 = 0$.

The form of this singular point which is obtained from linearization of the equation at the origin depends upon μ, being an unstable focus for $\mu < 2$ and an unstable node for $\mu > 2$. All phase plane trajectories have a slope of r when they intersect the curve

$$rx_2 = \mu(1 - x_1^2)x_2 - x_1 \tag{100.141}$$

One way of sketching phase plane behavior is to draw a set of curves given for various values of r by Eq. (100.141) and marking the trajectory slope r on the curves. This procedure is known as the method of isoclines and has been used to obtain the limit cycles shown in Fig. 100.51 for the Van der Pol equation with $\mu = 0.2$ and 4.

Piecewise Linear Characteristics

When the nonlinear elements occurring in a second order model can be approximated by linear segmented characteristics then the phase plane approach is usually easy to use because the nonlinearities divide the phase

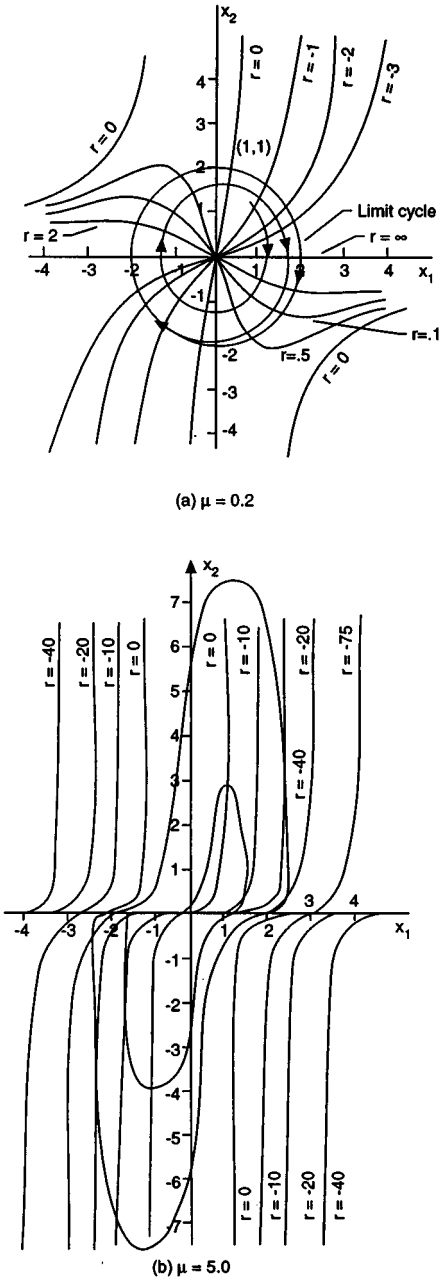


FIGURE 100.51 Phase portraits of the Van der Pol equation for different values, of μ.

plane into various regions within which the motion may be described by different linear second-order equations [Atherton, 1982]. The procedure is illustrated by the simple relay system in Fig. 100.52.

The block diagram represents an "ideal" relay position control system with velocity feedback. The plant is a double integrator, ignoring viscous (linear) friction, hysteresis in the relay, or backlash in the gearing. If the system output is denoted by x_1 and its derivative by x_2, then the relay switches when $-x_1 - x_2 = \pm 1$; the equations of the dotted lines are marked switching lines on Fig. 100.53.

Because the relay output provides constant values of ±2 and 0 to the double integrator plant, if we denote the constant value by *h*, then the state equations for the motion are

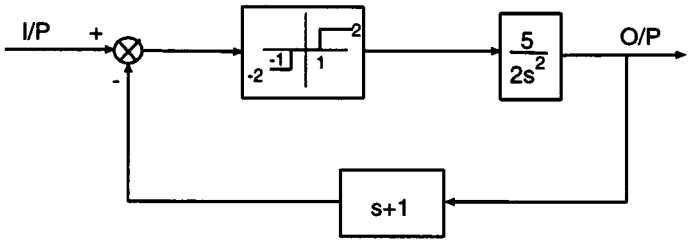


FIGURE 100.52 Relay system.

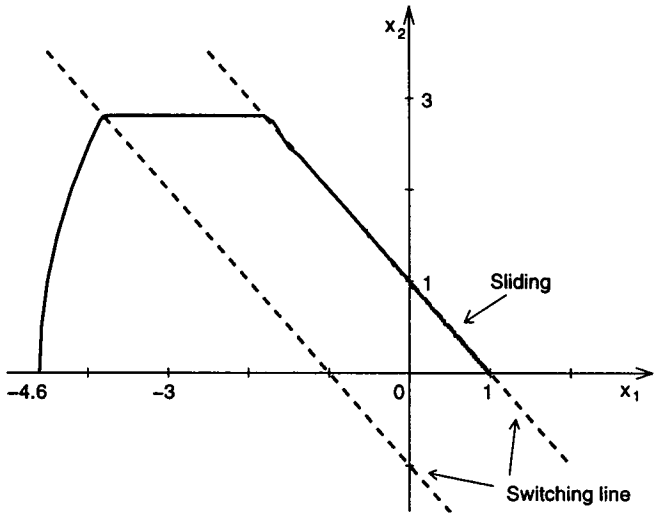


FIGURE 100.53 Phase plane for relay system.

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= h \end{aligned} \tag{100.142}$$

which can be solved to give the phase plane equation

$$x_2^2 - x_{20}^2 = 2h\left(x_1 - x_{10}\right) \tag{100.143}$$

which is a parabola for h finite and the straight line $x_2 = x_{20}$ for $h = 0$, where x_{20} and x_{10} are the initial values of x_2 and x_1. Similarly, more complex equations can be derived for other second-order transfer functions. Using Eq. (100.143) with the appropriate values of h for the three regions in the phase plane, the step response for an input of 4.6 units can be obtained as shown in Fig. 100.53.

In the step response, when the trajectory meets the switching line $x_1 + x_2 = -1$ for the second time, trajectory motions at both sides of the line are directed towards it, resulting in a sliding motion down the switching line. Completing the phase portrait by drawing responses from other initial conditions shows that the autonomous system is stable and also that all responses will finally slide down a switching line to equilibrium at $x_1 = \pm 1$.

An advantage of the phase plane method is that it can be used for systems with more than one nonlinearity and for those situations where parameters change as functions of the phase variables. For example, Fig. 100.54 shows the block diagram of an approximate model of a servomechanism with nonlinear effects due to torque saturation and Coulomb friction.

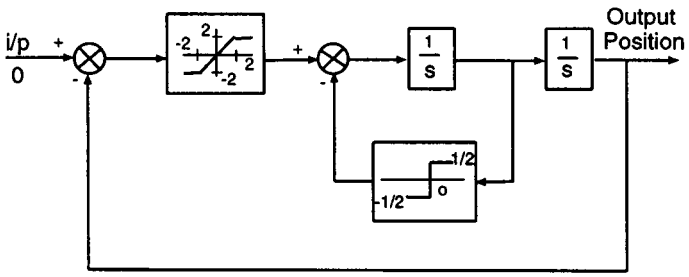


FIGURE 100.54 Block diagram of servomechanism.

The differential equation of motion in phase variable form is

$$\dot{x}_2 = f_s(-x_1) - (1/2)\,\mathrm{sgn}\,x_2 \tag{100.144}$$

where f_s denotes the saturation nonlinearity and sgn the signum function, which is +1 for $x_2 > 0$ and –1 for $x_2 < 0$. There are six linear differential equations describing the motion in different regions of the phase plane. For x_2 positive, Eq. (100.144) can be written

$$\dot{x}_1 + f_s(x_1) + 1/2 = 0$$

so that for

(a) x_2+ve, $x_1 < -2$, we have $\dot{x}_1 = x_2$, $\dot{x}_2 = 3/2$, a parabola in the phase plane.
(b) x_2+ve$|x_1| < 2$, we have $\dot{x}_1 = x_2$, $\dot{x}_2 + x_1 + 1/2 = 0$.
(c) x_2+ve, $x_1 > 2$, we have $\dot{x}_1 = x_2$, $\dot{x}_2 = -5/2$, a parabola in the phase plane. Similarly for x_2 negative,
(d) x_2–ve, $x_1 - 2$, we have $\dot{x}_1 = x_2$, $\dot{x}_2 = -5/2$, a parabola in the phase plane.
(e) x_2–ve, $|x_2| < 2$, we have $\dot{x}_1 = x_2$, $\dot{x}_2 + x_1 - 1/2 = 0$, a circle in the phase plane.
(f) x_2–ve,$x_1 > 2$, we have $\dot{x}_1 = x_2$, $\dot{x}_2 = -3/2$, a parabola in the phase plane.

Because all the phase plane trajectories are described by simple mathematical expressions, it is straightforward to calculate specific phase plane trajectories.

Discussion

The phase plane approach is useful for understanding the effects of nonlinearity in second order systems, particularly if it may be approximated by a linear segmented characteristic. Solutions for the trajectories with other nonlinear characteristics may not be possible analytically so that approximate sketching techniques were used in early work on nonlinear control. These approaches are described in many books, for example, [Blaquiere, 1966; Cosgriff, 1958; Cunningham, 1958; Gibson, 1963; Graham and McRuer, 1961; Hayashi, 1964, Thaler and Pastel, 1962; West, 1960]. Although the trajectories are now easily obtained with modern simulation techniques, knowledge of the topological aspects of the phase plane are still useful for interpreting the responses in different regions of the phase plane and appreciating the system behavior.

Related Topics

5.2 Limiters • 12.1 Introduction • 12.3 Lyapunov Stability Theory

References

A.A. Andronov, A.A. Vitt, and S.E. Khaikin, *Theory of Oscillators,* Reading, Mass.: Addison-Wesley, 1966. (First edition published in Russia in 1937.)

K.J. Astrom, and T. Haggland, *Automatic tuning of single regulators,* Budapest: Proc IFAC Congress, Vol. 4, 267–272, 1984.
D.P. Atherton, *Nonlinear Control Engineering, Describing Function Analysis and Design,* London: Van Nostrand Reinhold, 1975.
D.P. Atherton, *Non Linear Control Engineering,* Student Ed., New York: Van Nostrand Reinhold, 1982.
A. Blaquiere, *Nonlinear Systems Analysis,* New York: Academic Press, 1966.
S.K. Choudhury, and D.P. Atherton, "Limit cycles in high order nonlinear systems," *Proc. Inst. Electr. Eng.,* 121, 717–724, 1974.
P.A. Cook, "Describing function for a sector nonlinearity," *Proc. Inst. Electr. Eng.,* 120, 143–144, 1973.
R. Cosgriff, *Nonlinear Control Systems,* New York: McGraw-Hill, 1958.
W.J. Cunningham, *Introduction to Nonlinear Analysis,* New York: McGraw-Hill, 1958.
A. Gelb and W.E. van der Velde, *Multiple Input Describing Functions and Nonlinear Systems Design,* New York: McGraw-Hill, 1968.
J.E. Gibson, *Nonlinear Automatic Control,* New York: McGraw-Hill, 1963.
D. Graham and D. McRuer, *Analysis of Nonlinear Control Systems,* New York: John Wiley & Sons, 1961.
C. Hayashi, *Nonlinear Oscillations in Physical Systems,* New York, McGraw-Hill, 1964.
S.S. Lamba and R.J. Kavanagh, "The phenomenon of isolated jump resonance and its application," *Proc. Inst. Electr. Eng.,* 118, 1047–1050, 1971.
E. Levinson, "Some saturation phenomena in servomechanims with emphasis on the techometer stabilised system," *Trans, Am. Inst. Electr. Eng.,* Part 2, 72, 1–9, 1953.
O.P. McNamara, and D.P. Atherton, "Limit cycle prediction in free structured nonlinear systems," *IFAC Congress,* Munich, 8, 23–28, July 1987.
A.I. Mees and A.R. Bergen, "Describing function revisited," *IEEE Trans. Autom. Control,* 20, 473–478, 1975.
O. Nanka-Bruce and D.P. Atherton, "Design of nonlinear controllers for nonlinear plants," *IFAC Congress,* Tallinn, 6, 75–80, 1990.
T.P. Singh, "Graphical method for finding the closed loop frequency response of nonlinear feedback control systems," *Proc. Inst. Electr. Eng.,* 112, 2167–2170, 1965.
J.H. Taylor and K.L. Strobel, "Applications of a nonlinear controller design approach based on the quasilinear system models," *Prof ACC,* San Diego, 817–824, 1984.
G.J. Thaler and M.P. Pastel, *Analysis and Design of Nonlinear Feedback Control Systems,* New York: McGraw-Hill, 1962.
J.C. West, *Analytical Techniques of Nonlinear Control Systems,* London: E.U.P., 1960.
J.C. West and J.L. Douce, "The frequency response of a certain class of nonlinear feedback systems," *Br. J. Appl. Phys.,* 5, 201–210, 1954.
J.C. West, B.W. Jayawant, and D.P. Rea, "Transition characteristics of the jump phenomenon in nonlinear resonant circuits," *Proc. Inst. Electr. Eng.,* 114, 381–392, 1967.
J.G. Ziegler and N.B. Nichols, "Optimal setting for automatic controllers," *Trans. ASME,* 65, 433–444, 1943.

Further Information

Many control engineering text books contain material on nonlinear systems where the describing function is discussed. The coverage, however, is usually restricted to the basic sinusoidal DF for determining limit cycles in feedback systems. The basic DF method, which is one of quasilinearisation, can be extended to cover other signals, such as random signals, and also to cover multiple input signals to nonlinearities and feedback system analysis. The two books with the most comprehensive coverage of this are Gelb and Van der Velde [1968] and Atherton [1975]. More specialized books on nonlinear feedback systems usually cover the phase plane method and the DF, together with other topics such as absolute stability, exact linearization, etc.

100.8 Optimal Control and Estimation

John S. Bay and William T. Baumann

Consider the closed-loop feedback control of linear time-invariant, multi-input/multi-output (MIMO) state-space systems of the form:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{100.145}$$

In this form, the vector $x \in \Re^n$ represents the internal state, $u \in \Re^p$ represents the input, and $y \in \Re^q$ represents the measured outputs. It is well known that if the system in Eq. (100.145) is *stabilizable* (i.e., its unstable part is ***controllable***), then it can be asymptotically stabilized with static state feedback. If it is ***detectable*** (i.e., its unstable part is ***observable***), then a state estimator can be found, whose state variables asymptotically approach the true state variables in Eq. (100.145). However, merely determining the state feedback gain or observer gain leaves considerable design freedom for satisfying criteria other than stabilization and asymptotic observation. In this chapter section, we will provide results of some basic techniques of optimal control and estimation, which provide a mechanism to find *optimal* feedback and **observer (estimator)** gains according to selected optimality criteria.

Linear Quadratic Regulators

The linear quadratic regulator (LQR) problem is to find an *optimal* control input $u^*(t)$ that minimizes the performance criterion

$$J(x,u) = \tfrac{1}{2} x^{\mathrm{T}}(t_f) S x(t_f) + \tfrac{1}{2} \int_{t_o}^{t_f} \left[x(t) Q x(t) + u^{\mathrm{T}}(t) R u(t) \right] dt \tag{100.146}$$

where S and Q are symmetric, ***positive-semidefinite*** weighting matrices; and R is a symmetric, *positive-definite* weighting matrix. In this criterion, the term $\frac{1}{2}x^{\mathrm{T}}(t_f)Sx(t_f)$ represents a penalty for the state at the final time t_f being different from zero. The term inside the integral, $x^{\mathrm{T}}(t)Qx(t)$, represents a penalty on the transient response of the state vector. The term $u^{\mathrm{T}}(t)Ru(t)$ represents a penalty on the size of the control input $u(t)$. We allow S and Q to be positive-semidefinite because we can generally tolerate unbounded state variables, provided they are not observed at the output. However, by forcing R to be positive-definite, we can guarantee that the process of minimizing Eq. (100.146) gives a bounded input. Minimization of this control energy is one of the primary reasons for using optimal control.

The optimal control $u^*(t)$ can be found via a number of techniques, including dynamic programming [Bay, 1999] and variational techniques [Kirk, 1970]. The result of any of these methods is that the optimal control $u^*(t)$ is a linear function of the state vector (linear state feedback) of the form:

$$u^*(t) = -R^{-1}B^{\mathrm{T}}P(t)x(t) \tag{100.147}$$

where the $n \times n$ matrix function $P(t)$ satisfies the following equation:

$$\dot{P} = PBR^{-1}B^{\mathrm{T}}P - Q - PA - A^{\mathrm{T}}P \tag{100.148}$$

Equation (100.148) is known as the differential matrix Riccati equation, and it is solved in backward time, with the end-time condition $P(t_f) = S$.

It may be noted that a reasonable optimization criterion may have no finite final time t_f. Instead, it may be desired that the controller be continually active, implying $t_f \to \infty$ and eliminating the possibility of the final state term in Eq. (100.146). In this case, the optimization criterion is more properly written as

$$J(x,u) = \tfrac{1}{2}\int_{t_o}^{\infty}\left[x^{\mathrm{T}}(t)Qx(t) + u^{\mathrm{T}}(t)Ru(t)\right]dt \tag{100.149}$$

Fortunately, this criterion simplifies the optimal control solution to the steady-state solution of the finite-time problem. That is, for the criterion of Eq. (100.149), the optimal control is

$$u^{*}(t) = -R^{-1}B^{\mathrm{T}}Px(t) \tag{100.150}$$

where in this case P is the matrix solution to the following algebraic Riccati equation:

$$0 = PBR^{-1}B^{\mathrm{T}}P - Q - PA - A^{\mathrm{T}}P \tag{100.151}$$

Such a steady-state optimal solution exists whenever the system (A, B) is stabilizable. Furthermore, this constant P is the unique positive-definite solution of Eq. (100.151) if and only if the pair (A, T) is detectable, where T is defined as the square-root of Q, $Q = T^{\mathrm{T}}T$. **Stabilizability** of (A, B) ensures convergence of the cost criterion integral Eq. (100.149), and detectability of (A, T) guarantees that no unstable part of $x(t)$ escapes integration as part of the integrand of Eq. (100.149).

We should point out also that for the corresponding discrete-time system:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k) \end{aligned} \tag{100.152}$$

minimization of the cost criterion:

$$J = \tfrac{1}{2}\sum_{k=k_0}^{\infty}\left[x^{\mathrm{T}}(k)Qx(k) + u^{\mathrm{T}}(k)Ru(k)\right] \tag{100.153}$$

over all inputs $u(k)$ results in the optimal control

$$u^{*}(k) = -\left[R + B^{\mathrm{T}}SB\right]^{-1}B^{\mathrm{T}}SAx(k) \tag{100.154}$$

where S is the solution to the corresponding discrete-time algebraic Riccati equation:

$$S = A^{\mathrm{T}}SA - A^{\mathrm{T}}SB\left[R + B^{\mathrm{T}}SB\right]^{-1}B^{\mathrm{T}}SA + Q \tag{100.155}$$

Note that in both the continuous- and the discrete-time infinite-horizon cases, the optimal control is actually static state feedback of the form $u = Kx$.

Optimal Estimation: The Kalman Filter

It was noted above that the optimal controller for the linear quadratic cost criterion takes the form of full-state feedback. However, it is often the case that the full state is not physically available for feedback. Rather, it is usually the *output* that is measurable, so that we prefer a technique that uses $y(t)$ (and possibly $u(t)$) to construct the control signal instead of the state $x(t)$.

The simple solution to this problem is to design an ***observer (or estimator)***, which produces an *estimated* state vector $\hat{x}(t)$. If this estimate asymptotically approaches the true state $x(t)$, then we can simply combine the observer and the state feedback to produce a feedback control $u(t) = K\hat{x}(t)$. That we can simply substitute the observed value $\hat{x}(t)$ for $x(t)$ in the feedback function is a fortunate property called the separation principle, which ensures that our controller calculations and observer calculations do not interfere with one another.

Just as we have improved static state feedback by introducing the optimal LQR above, we can take the principles of observer design and extend them with some guarantees of optimality. Such an optimal estimator is the Kalman filter.

The Kalman filter is derived assuming the system model:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Gv(t) \\ y(t) &= Cx(t) + w(t) \end{aligned} \tag{100.156}$$

where $v(t)$ and $w(t)$ are two white, Gaussian, zero-mean, mutually uncorrelated noise signals with $E[w(t)v^{\mathrm{T}}(\tau)] = 0$ and

$$E\left[v(t)\right] = 0, \quad E\left[v(t)v^{\mathrm{T}}(\tau)\right] = V\delta(t-\tau) \tag{100.157}$$

and

$$E\left[w(t)\right] = 0, \quad E\left[w(t)w^{\mathrm{T}}(\tau)\right] = W\delta(t-\tau) \tag{100.158}$$

where $\delta(t)$ is the Dirac delta. Noise $v(t)$ is called the *plant noise*, and $w(t)$ is the *measurement noise*, often representing sensor noise. (By assuming these signals are first passed through auxiliary filters with specified dynamics, these noises can be made to resemble harmonic or narrow-band disturbances.)

The goal is now to design a system that produces an estimated state $\hat{x}(t)$ for Eq. (100.156) while rejecting the influence of the signals $v(t)$ and $w(t)$. To do this, we need the further assumptions that the system's initial state is guessed to be $x(t_0) = x_0$ and that this guess is uncorrelated with the plant and measurement noise:

$$E\left[x_0 v^{\mathrm{T}}(\tau)\right] = 0 \qquad E\left[x_0 w^{\mathrm{T}}(\tau)\right] = 0 \tag{100.159}$$

The covariance of the initial guess is defined as

$$E\left\{\left[x_0 - E(x_0)\right]\left[x_0 - E(x_0)\right]^{\mathrm{T}}\right\} \triangleq P_0 \tag{100.160}$$

The Kalman filter is the system that performs this estimation, rejecting the influence of the noise. The filter itself is given by the equation:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(t)\left[y(t) - C\hat{x}(t)\right] \tag{100.161}$$

which can be seen to resemble the standard full-order observer [Bay, 1999]. However, rather than choosing an observer gain L in Eq. (100.161) to simply stabilize the error dynamics of the observer, the *Kalman gain* $L(t)$ is

$$L(t) = P(t)C^{\mathrm{T}}W^{-1} \tag{100.162}$$

where $P(t)$ is the solution to the following differential Riccati equation:

$$\dot{P}(t) = AP(t) + P(t)A^{\mathrm{T}} - P(t)C^{\mathrm{T}}W^{-1}CP(t) + GVG^{\mathrm{T}} \tag{100.163}$$

whose initial condition is $P(t_0) = P_0$.

For the discrete-time system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + Gv(k) \\ y(k) &= Cx(k) + w(k) \end{aligned} \tag{100.164}$$

with assumptions analogous to Eq. (100.157) through (100.158) for plant noise $v(k)$ and measurement noise $w(t)$, and with $E[(x_0 - E(x_0)(x_0 - E(x_0))^{\mathrm{T}}] \triangleq S_0$, the Kalman filter is given by the two-stage estimator

$$\bar{x}(k+1) = A\hat{x}(k) + Bu(k) \tag{100.165}$$

$$\hat{x}(k+1) = \bar{x}(k+1) + L(k+1)\left[y(k+1) - C\bar{x}(k+1)\right] \tag{100.166}$$

where the Kalman gain $L(k+1)$ is computed from the equation

$$L(k+1) = \left[AS(k)A^{\mathrm{T}} + GVG^{\mathrm{T}}\right]C^{\mathrm{T}}\left\{C\left[AS(k)A^{\mathrm{T}} + GVG^{\mathrm{T}}\right]C^{T} + W\right\}^{-1} \tag{100.167}$$

In Eq. (100.167), the term $S(k)$ is determined from the Riccati equation:

$$\begin{aligned} S(k+1) = &\left[I - L(k+1)C\right]\left[AS(k)A^{\mathrm{T}} + GVG^{\mathrm{T}}\right]\left[I - L(k+1)C\right]^{\mathrm{T}} \\ &+ L(k+1)WL^{\mathrm{T}}(k+1) \end{aligned} \tag{100.168}$$

with initial condition $S(k_0) = S_0$. (Note that these equations can be combined and rearranged to produce a number of alternate formulations.)

Equation (100.165) is often referred to as the *time update* equation. It represents the estimate of the state vector that results from knowledge of the system dynamics. Equation (100.166) is sometimes called the *measurement* update equation because it revises the time update with a so-called *innovations* term $L(y - \bar{y})$ that adjusts this time update according to the error between the output expected from the time update, $\bar{y}(k+1)$, and the actual, measured output, $y(k+1)$.

The matrices $P(t)$ in the continuous-time filter, and $S(k)$ in the discrete-time filter are the error covariance matrices for the state estimate. That is,

$$S(k) \triangleq E\left[e(k)e^{\mathrm{T}}(k)\right] \quad \text{and} \quad P(t) \triangleq E\left[e(t)e^{\mathrm{T}}(t)\right] \tag{100.169}$$

where $e(k) \triangleq x(k) - \hat{x}(k)$ and $e(t) \triangleq x(t) - \hat{x}(t)$. Thus, the size of these matrices is an indicator of the error variance in various components in the estimates, and it can be seen in Eq. (100.161) and (100.167) that as these covariances decrease, the estimators rely less and less on the innovations term and more on the system dynamics for an accurate estimate. Early in the estimation process, the situation is usually reversed, with the innovations having the larger effect.

Linear-Quadratic-Gaussian (LQG) Control

It can be shown that the Kalman filter is the *optimal* estimator for the state of system Eq. (100.156) or (100.164) in the sense that it minimizes the squared error due to the noise input terms. Therefore, it becomes a likely candidate for combination with the LQR of the previous section. Together, the combination of LQR and Kalman filter is known as the LQG (linear-quadratic-Gaussian) controller, and is a useful controller in many situations. This controller is optimal in the sense that it minimizes the expected root-mean-square value of the optimization criterion

$$\lim_{T\to\infty} E\left\{\frac{1}{T}\int_0^T x^{\mathrm{T}}(t)Qx(t) + u^{\mathrm{T}}(t)Ru(t)dt\right\}^{1/2} \tag{100.170}$$

when the noise inputs are unit variance white noise. In the frequency domain, this is equivalent to minimizing the H_2-norm

$$\|G\|_2 = \left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} trace\left(G^*(j\omega)G(j\omega)\right)d\omega\right\}^{1/2} \tag{100.171}$$

of the transfer function G from the white noise inputs $\begin{bmatrix} v \\ w \end{bmatrix}$ to the output $z = \begin{bmatrix} Tx \\ R^{1/2}u \end{bmatrix}$ where $Q = T^{\mathrm{T}}T$, as before. We should point out that the so-called H_2 controller is equivalent to the LQG controller but provides a unified framework for control and observation that explains the striking similarity between the LQ controller equations and the Kalman filter equations (for example, in the Riccati equations of (100.148) and (100.63)). See Zhou and Doyle [1998] for further information.

H_∞ Control

The standard H_∞ control problem considers a system of the form

$$\begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u \\ z &= C_1 x + D_{12} u \\ y &= C_2 x + D_{21} w \end{aligned} \tag{100.172}$$

where w is a deterministic disturbance input vector, y is the measured output vector, and z is a vector of variables to be controlled. The objective of H_∞ control is to minimize the H_∞ norm

$$\|G\|_\infty = \sup_\omega \bar{\sigma}\left[G(j\omega)\right] \tag{100.173}$$

(where $\bar{\sigma}$ denotes the maximum ***singular value*** of a matrix, and sup denotes "supremum," or least upper bound) of the transfer function G from the disturbance input w to the output z. In the time domain, the square of this

norm corresponds to the maximum possible energy magnification between the input and the output, where energy is defined as the integral of the square of a signal; for example, $\int_0^\infty w^T(t)w(t)dt$.

One of the major reasons for the development of H_∞ control is that many performance and robustness criteria for MIMO systems involve the maximum **singular value** of certain system transfer functions. To optimize the performance or robustness of these systems requires the minimization of the maximum singular value of these transfer functions, which is exactly the objective of H_∞ control.

From a disturbance rejection point of view, the H_∞ controller can be used to minimize the root-mean-square value of the controlled variable z due to the worst-case unit-energy disturbance w. This is to be contrasted with LQG (or H_2) controller, which minimizes the average response to a unit-variance random disturbance.

In practice, it is common to solve the suboptimal H_∞ problem where it is desired to find an output feedback controller such that $\|G\|_\infty < \gamma$, where γ is specified by the designer. For large values of γ, there will always be a solution to the problem. In fact, as γ approaches infinity in the equations below, the central H_∞ controller will approach the LQG controller. By decreasing the value of γ until just before a solution to the problem ceases to exist, the designer can get as close to the optimal H_∞ controller as desired. To ensure that a solution for some value of γ exists, the following standard assumptions on the system are made [Green and Limebeer, 1995]:

1. The pair (A, B_2) is stabilizable and the pair (A, C_2) is detectable
2. $D_{12}^T D_{12} = I$ and $D_{21}D_{21}^T = \mathrm{I}$
3. Rank $\begin{bmatrix} A - j\omega I & B_2 \\ C_1 & D_{12} \end{bmatrix} = n + m$ for all real ω
4. Rank $\begin{bmatrix} A - j\omega I & B_1 \\ C_2 & D_{21} \end{bmatrix} = n + q$ for all real ω

where n is the dimension of x, m is the dimension of u, and q is the dimension of y.

Under these assumptions, it can be shown that there exists a stabilizing, measurement feedback solution to the suboptimal H_∞ control problem if and only if the following three conditions are met.

1. The algebraic Riccati equation $X_\infty \tilde{A} + \tilde{A}^T X_\infty + \tilde{C}^T\tilde{C} - X_\infty (B_2B_2^T - \gamma^{-2} B_1B_1^T)X_\infty = 0$ has a positive semi-definite solution such that $\tilde{A} - (B_2B_2^T - \gamma^{-2} B_1B_1^T)X_\infty$ is stable, where $\tilde{A} = A - B_2D_{12}^T C_1$ and $\tilde{C}^T\tilde{C} = C_1^T(I - D_{12}D_{12}^T)C_1$.
2. The algebraic Riccati equation $\overline{A}Y_\infty + Y_\infty\overline{A}^T + \overline{B}\overline{B}^T - Y_\infty (C_2^TC_2 - \gamma^{-2}C_1^TC_1)Y_\infty = 0$ has a positive semi-definite solution such that $\overline{A} - Y_\infty (C_2^TC_2 - \gamma^{-2}C_1^TC_1)$ is stable, where $\overline{A} = A - B_1D_{12}^T C_2$ and $\overline{B}\overline{B}^T = B_1(I - D_{21}^T D_{12}) B_1^T$.
3. $\rho(X_\infty Y_\infty) < \gamma^2$, where $\rho(\cdot)$ denotes the maximum of the absolute values of the matrix's eigenvalues.

The so-called central controller that solves the suboptimal H_∞ problem can be written in a form that closely resembles the state-estimate feedback form of the LQG controller:

$$\begin{aligned} \dot{\hat{x}} &= A\hat{x} + B_1\hat{w}^* + B_2u + \left[B_1D_{21}^T + Z_\infty C_{2z}^T\right]\left(y - C_2\hat{x} - D_{21}\hat{w}^*\right) \\ u &= -F_\infty\hat{x} \\ \hat{w}^* &= \gamma^{-2}B_1^T X_\infty\hat{x} \end{aligned} \tag{100.174}$$

where $C_{2z} = C_2 + \gamma^{-2}D_{21}B_1^T X_\infty$, $F_\infty = D_{12}^T C_1 + B_2^T X_\infty$, and $Z_\infty = Y_\infty(I - \gamma^{-2}X_\infty Y_\infty)^{-1}$. The dynamic part of the above compensator can be interpreted as an estimate of the state assuming that the worst-case disturbance w^* is present. The control signal is a linear feedback of the estimated state, just as in the LQG case. Although the controller formulas above look more complicated than in the LQG case, this is largely due to the fact that the controlled variable z has a more general form in the H_∞ problem statement above. It should be noted, however, that unlike the LQG case, the solution of the Riccati equations is coupled in the H_∞ case due to condition 3 above.

Example

Consider the following state-space system, which represents a plant with two lightly damped modes at approximately $\omega_1 \approx 1$ and $\omega_2 \approx 3.2$:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & -.1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -10 & -.1 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ .2 & 0 \end{bmatrix} d(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix} u(t) \tag{100.175}$$

$$y(t) = \begin{bmatrix} 1 & 0 & .5 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 & .01 \end{bmatrix} d(t)$$

Here, the term $d(t) = [d_1(t) \quad d_2(t)]^T$ is a vector whose first term represents the plant disturbance, and whose second term represents the measurement disturbance (deterministic). To pose the LQG control problem, we can propose minimizing the cost function

$$\int_0^\infty \left(x^T T^T T x + r u^2 \right) dt = \int_0^\infty z^T z \; dt \tag{100.176}$$

where

$$z \triangleq \begin{bmatrix} T \\ 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ r \end{bmatrix} u \tag{100.177}$$

and $T = [.5 \quad 0 \quad -1 \quad 0]$. However, Eq. (100.175) and (100.177) are also in the form of Eq. (100.172), facilitating an H_∞ design that minimizes the H_∞ norm of the transfer function from the disturbance d to the controlled variable z. We can compare this design to an H_2 controller design that minimizes the H_2 norm (Eq. (100.176)). The two controllers will therefore minimize different norms of the same transfer function.

The results of this comparison are shown in the curves of Fig. 100.55. The distinguishing feature of this comparison is the flattening effect of the H_∞ controller. Although this is a plot of a frequency response magnitude and not the maximum **singular value,** it is apparent that theH_∞ controller is reducing the peak response, while the H_2 controller is reducing the average response and providing faster roll-off in the frequency domain.

Other Approaches

Although the LQG and H_∞ design methodologies are probably the most commonly used techniques for linear MIMO controller design, there are many other optimization-based techniques available. A well-developed theory exists for L_1 control, which minimizes the maximum magnification of the peak of the input signal using a linear programming algorithm [Dahleh and Diaz-Bobillo, 1995]. This approach departs from traditional controller design methodologies that have attempted to arrive at a set of formulas for the optimal controller. But with the advent of powerful computers, it makes sense to consider a problem solved if it can be reduced to a problem that can be efficiently solved using a computer algorithm. This is also the premise underlying the design methodology of *linear matrix inequalities*, in which the control design problem is reduced to a convex programming problem [Boyd et al., 1994].

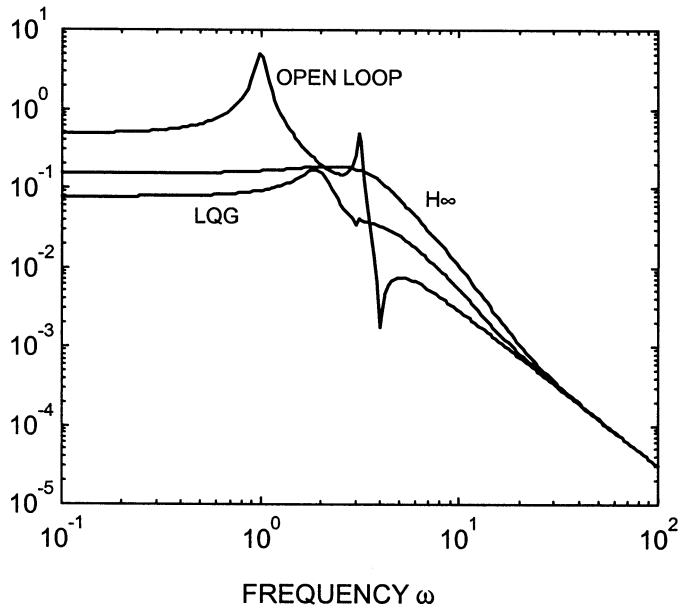


FIGURE 100.55 Frequency response of the closed-loop transfer function from d_1 to Tx, comparing the H_2 (LQG) and H_∞ control designs, using $r = 0.1$.

Defining Terms

Controllability: A linear system is said to be controllable if a control input exists that will drive a system with an arbitrary initial condition to a desired final state in a finite time.

Stabilizability: A linear system is said to be stabilizable if its unstable part is controllable.

Observability: A linear system is said to be observable if its state vector can be reconstructed from finite-length observations of its input and output.

Detectability: A linear system is said to be detectable if its unstable part is observable.

Positive-(semi)definite: A positive- (semi)definite matrix is a symmetric matrix A such that for any nonzero vector x, the quadratic form $x^T Ax$ is positive (non-negative).

Observer (or **estimator**)**:** A linear system whose state output approximates the state vector of a different system, rejecting noise and disturbances in the process.

Singular value: Singular values are non-negative real numbers that measure the magnification effect of an operator in the different basis directions of the operator's space.

References

B. D. O. Anderson and J. B. Moore, *Optimal Filtering,* Prentice-Hall, 1979.

J. S. Bay, *Fundamentals of Linear State Space Systems,* WCB/McGraw-Hill, 1999.

S. Boyd et al., *Linear Matrix Inequalities in System and Control Theory,* Society for Industrial and Applied Mathematics, 1994.

A. E. Bryson, Jr. and Y. C. Ho, *Applied Optimal Control,* Hemisphere Publishing, 1975.

M. Dahleh and I. J. Diaz-Bobillo, *Control of Uncertain Systems: A Linear Programming Approach,* Prentice-Hall, 1995.

J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, State-space solutions to standard H_2 and H_∞ control problems, *IEEE Trans. on Automatic Control,* AC-34, 831–847, 1988.

B. A. Francis, *A Course in H_∞ Control Theory,* Lecture Notes in Control and Information Sciences, Vol. 88, Springer-Verlag, 1987.
M. Green and D. J. N. Limebeer, *Linear Robust Control,* Prentice-Hall, 1995.
R. E. Kalman, A new approach to linear filtering and prediction problems, *Transactions of the ASME, Journal of Basic Engineering,* 82, 35–45, 1960.
D. E. Kirk, *Optimal Control Theory,* Prentice-Hall, 1970.
K. Zhou and J. C. Doyle, *Essentials of Robust Control,* Prentice-Hall, 1998.

Further Information

Optimal control and estimation is an actively growing field of control systems research. Classic texts in the area include [Kirk, 1970] and [Bryson and Ho, 1975] for optimal control systems, and [Anderson and Moore, 1979] for Kalman filtering, with the original source being [Kalman, 1960]. Further information on H_2 and H_∞ theory can be found in [Doyle, et al., 1989], [Zhou and Doyle, 1998], [Green and Limebeer, 1995], and [Francis, 1987].

100.9 Neural Control

Mo-Yuen Chow

Artificial intelligence had strong ties with automatic control during its early development stages several decades ago. Typical examples of these ties are the development of cybernetics, robotics, and early learning systems. Recent efforts to incorporate aspects of artificial intelligence into the design and operation of automatic control systems have focused on using techniques such as artificial neural networks, fuzzy logic, and expert systems. The application of one or more of these techniques in the design of control systems has come to be known as *intelligent control* [Antsaklis et al., 1994], a term questioned by some for its implications. Whether or not such systems should be classified as intelligent, they represent significant contributions to the field of automatic control, as evidenced by the rapidly growing wealth of literature devoted to the successful application of such systems to complex control problems [Chow and Menozzi, 1994a; 1994b; Chow and Teeter, 1997; Chow and Yee, 1991; Hunt et al., 1992; Miller et al., 1990; Nguyen and Widrow, 1990; Psaltis et al., 1988; Werbos, 1990].

The nonlinear functional mapping properties of Artificial Neural Networks (ANNs) are central to their use in system identification and control [Narendra and Parthasarathy, 1990]. Although a number of key theoretical problems remain, results pertaining to the approximation capabilities of neural networks demonstrate that they have great promise in the modeling and control of nonlinear systems. The artificial neural network technology has become increasingly popular as a tool for performing tasks such as automatic control, system identification, pattern recognition, and time series prediction. Most of the *conventional methods*, such as PI control, are based on mathematical and statistical procedures for the modeling of the system and the estimation of the optimal controller parameters. In practice, the plant to be controlled is often highly nonlinear and a mathematical model may be difficult to derive. In such cases, conventional techniques may prove to be suboptimal and may lack robustness in the face of modeling error, because they are only as accurate as the model that was used to design them. With the advancement of technology, however, sophisticated control using artificial neural network techniques has been developed and used successfully to improve the control of systems that cannot be easily handled by conventional control, thus giving rise to terminology such as *neural control* and *intelligent control.*

Usually, a human operator is responsible for adjusting the controller's parameters in order to use his/her own idea of good performance. Indirectly, the operator is performing a minimization of a cost function based on his/her knowledge. More generally, when confronted with a problem situation, humans execute a mapping between a set of events and the set of corresponding appropriate actions. The appropriateness of these actions is due to some basic acquired knowledge, or even instinct, that guides the initial stages of the mapping. Then, through experience and a set of implicit guidelines, the human learns to perform a better mapping. This is an ongoing process throughout a person's life. Similarly, an ANN, if given initial guidance, can learn to improve its performance through a set of guidelines, (e.g., minimize a cost function). In fact, a properly structured ANN can learn any arbitrarily complicated mapping [Cybenko, 1989; Rumelhart and McClelland, 1986; Werbos, 1974].

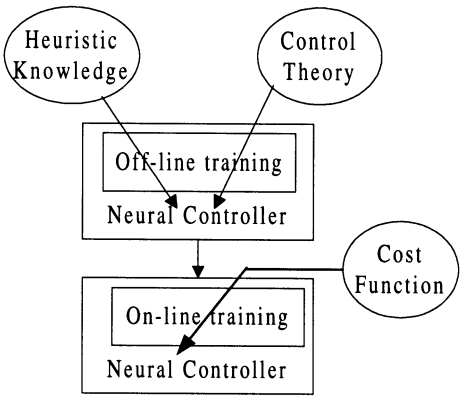


FIGURE 100.56 A two-step training process for a neural controller.

Successful adaptation of online neural controllers in many cases requires careful and substantial design effort. One of the common practices is to pre-train a neural controller offline, based on some simplified design methods, in the same way that a PI controller is designed based on *approximated* system models that can provide reasonable control performance, before putting the neural controller for online fine-tuning [Chow and Menozzi, 1993; Chow and Teeter, 1995; Teeter et al., 1996]. This approach, as shown in Fig. 100.56, can speed up the neural controller online adaptation process and increase the closed-loop online adaptation stability because the initial weights of the neural controller are much closer to the optimal final weights (if they exist) after the pre-training process. By learning online, the ANN controller can adapt to changing operating environments.

This chapter section briefly describes the feedforward net paradigm to facilitate the discussion of the *neural observer* and *neural controller* concepts in later sections. An example of using neural control for an HVAC system will then be provided to demonstrate its effectiveness for solving real-world problems.

Brief Introduction to Artificial Neural Networks

Increasing interest in studying the mechanisms and structure of the brain has led to the development of new computational models for solving problems such as pattern recognition, fast information processing, and adaptation. In the 1940s, McCulloch and Pitts studied the potential and capabilities of the interconnection of components based on a model of a biological neuron. Since then, many different models and architectures have been developed and analyzed for a variety of applications [Zurada, 1992]. One of the most common neuron models is shown in Fig. 100.57.

The inputs x to the neuron model are scaled by connection weights w and summed. An additional input b, often referred to as a *bias*, is added to this sum and the result becomes the input to a function $f(\cdot)$, called the *activation function*, which computes the output of the neuron. The bias can be considered as a connection weight for a constant input of +1. The terms *neuron model* and *neuron* are used interchangeably in this chapter section.

The individual neurons are not very powerful in terms of computation or representation, but the interconnection of neurons to form an *artificial neural network* (ANN) can provide a means of encoding complex relationships between input and output variables. Of the many ANN architectures that have been proposed, the *multilayer feedforward artificial neural network* (MFANN) shown in Fig. 100.58 is one of the most popular.

Bias terms have been omitted in Fig. 100.58 for simplicity. The layers between the input and output layers of an MFANN are usually referred to as *hidden layers* because their inputs and outputs are not measurable at the inputs or outputs of the network. It has been shown that an MFANN with a single hidden layer can approximate any continuous function to an arbitrary degree of accuracy [Cybenko, 1989; Werbos, 1974]. The process of adjusting ANN connection weights in an effort to obtain a desired input/output mapping is usually referred to as *training*. Training represents an optimization problem for which the solution is a set of weights that minimizes some measure of approximation error. The choices of activation functions, number of neurons, error measures, and optimization methods can significantly affect training results.

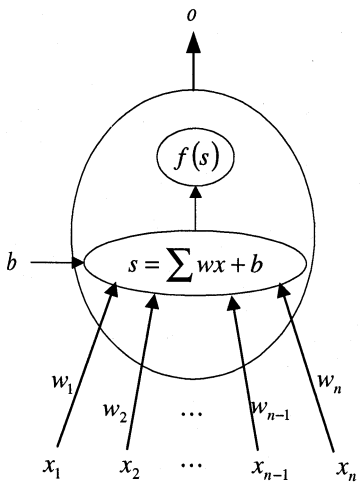


FIGURE 100.57 Computational model of a biological neuron.

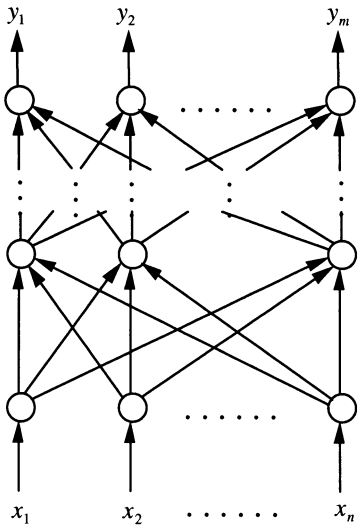


FIGURE 100.58 A multilayer feedforward ANN.

When the desired mapping is described by a set of input/output data, weights are usually modified after the presentation of each input/output pattern. This method is referred to as *pattern update* and represents an approximation of true gradient descent that is generally valid when a sufficiently small stepsize is used. *Batch update*, for which weight changes are accumulated over one sweep of the set of training patterns before being applied, is sometimes used in an effort to more closely mimic true gradient descent. Variants of back-propagation and other training methods can be found in [Zurada, 1992].

Neural Observer

The nonlinear functional mapping properties of *neural networks* are central to their use in identification and control [Chow and Teeter, 1995; Hunt et al., 1992; Poggio and Girosi, 1990; Teeter and Chow, 1998; Teeter et al. 1994]. Although a number of key theoretical problems remain, results pertaining to the approximation capabilities of neural networks demonstrate that they have great promise in the modeling of nonlinear systems. An important question in system identification is whether a system under study can be adequately represented within a given model structure [Hunt et al., 1992]. In the absence of such concrete theoretical results for neural networks, it is usually assumed that the system under consideration belongs to the class of systems that the chosen network is able to represent. Two system identification techniques are now introduced: *forward modeling* and *inverse modeling*.

The procedure of training a neural network to represent the forward dynamics of a system is often referred to as the *forward system identification* approach [Hunt et al., 1992]. A schematic diagram of this process is shown in Fig. 100.59.

The neural network is placed in parallel with the system, and the error, $\boldsymbol{e}$, between the system outputs, $\boldsymbol{y}$, and network outputs, $\hat{\boldsymbol{y}}$, is used to train the network. This represents a classical *supervised learning* problem for which the teacher (i.e., the system) provides target values (i.e., system outputs) directly in the output coordinate system of the learner (i.e., the network model) [Jordan and Rumelhart, 1991].

In an *inverse system identification* approach, a network is trained in an effort to model the inverse of the plant mapping [Hunt et al., 1992]. One of the simplest approaches, known as *direct inverse system identification*, is shown schematically in Fig. 100.60.

A synthetic training signal, $\boldsymbol{s}$, is introduced to the system, and the system output, $\boldsymbol{y}$, is used as the input to the network. The network output is compared to the training signal and this error is used to train the network.

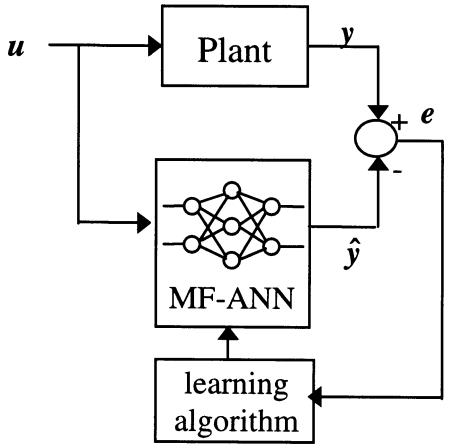


FIGURE 100.59 Forward system identification approach.

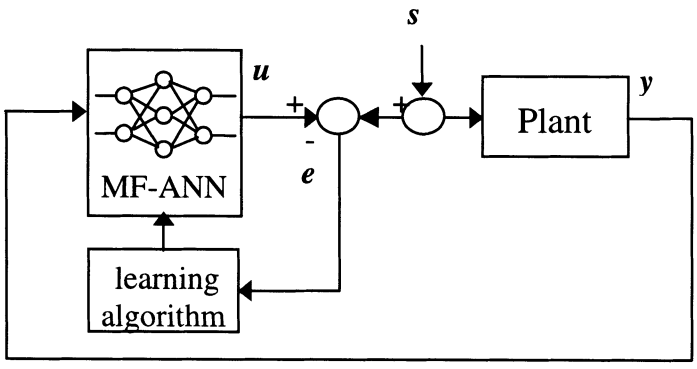


FIGURE 100.60 Direct inverse system identification approach.

The inverse modeling structure shown in Fig. 100.60 tends to force the network to represent the inverse of the plant, but there are potential drawbacks to this approach. The training signal must be chosen to sample over a wide range of system inputs, and the actual operational inputs may be difficult to define *a priori* [Jordan and Rumelhart, 1991]. This point is strongly related to the concept of persistent excitation discussed in adaptive control literature. A second drawback is that an incorrect inverse model can be obtained if the nonlinear system mapping is not one-to-one. An approach called *specialized inverse modeling* has been proposed in an effort to overcome these problems. The details of this approach can be found in [Psaltis et al., 1988]. The neural network identification models can be used in the adaptive control of unknown nonlinear plants.

Neural Control

A method of *direct adaptive control* is depicted in Fig. 100.61.

Methods for directly adjusting control parameters based on the output error e_c are generally not available. This is because the unknown nonlinear plant in Fig. 100.61 lies between the controller and the output error. Until such methods are developed, adaptive control of nonlinear plants must be performed using *indirect* methods [Narendra and Parthasarathy, 1990]. Figure 100.62 depicts a general method of indirect adaptive control using artificial neural networks.

Tapped delay lines (TDL) provide delayed inputs and outputs of the plant to the neural controller and neural observer. Error e_i is used to adapt the neural observer, while the parameters of the neural observer along with

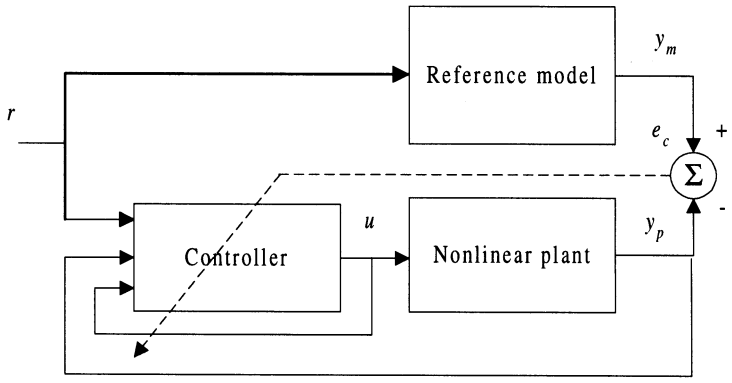


FIGURE 100.61 Direct adaptive control.

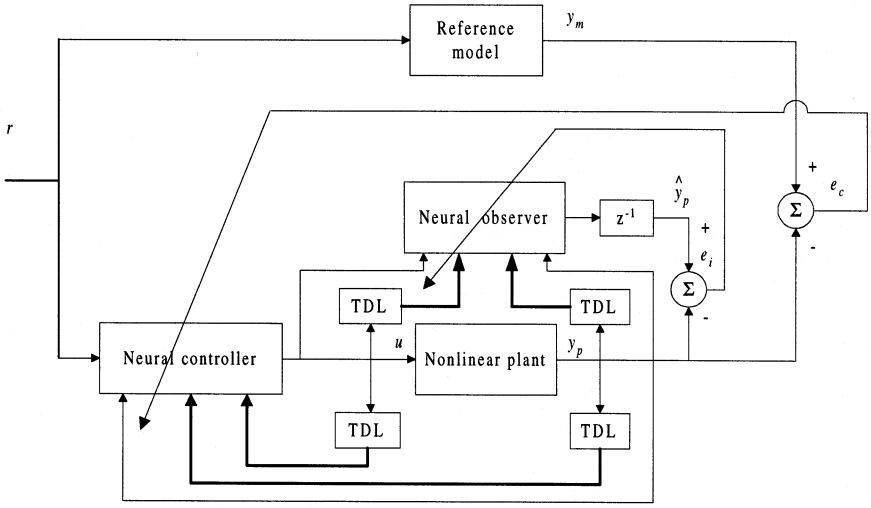


FIGURE 100.62 A method of indirect adaptive control using neural networks.

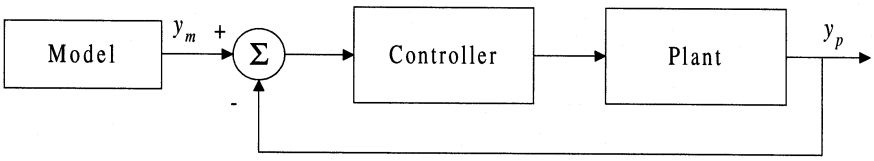


FIGURE 100.63 Explicit model-following.

error e_c are used to adapt the neural controller. The model reference approach depicted in Fig. 100.62 is commonly referred to as *implicit model-following* in adaptive control literature [Narendra and Annaswamy, 1989]. This type of model-following is performed when the dynamics of the closed-loop system are forced to asymptotically match the dynamics of the reference model. The method of *explicit model-following* is depicted in Fig. 100.63. In this case, the reference model acts as a prefilter, and the value of the system output is forced to asymptotically track the value of the reference model output [Narendra and Annaswamy, 1989].

A neural control and identification scheme can be described as the following: at each time step, the plant states are measured and a neural network controller computes the plant inputs. Controller parameters are then

adjusted between samples so that the system approaches optimal performance with respect to a given cost index. The general form of the cost index used to adapt the neural controller is:

$$J_k = \sum_{i=k+1}^{N} L\left(\mathbf{x}\left(i\right), \mathbf{u}\left(i\right)\right) \tag{100.178}$$

where L is the cost function as a function of system state $\mathbf{x}$ and control $\mathbf{u}$. Thus, at each time step, the goal is to minimize J_k subject to system dynamics and control constraints, with k denoting the current time step and N the prediction horizon.

The back-propagation training algorithm can be used to adapt neural networks for the identification and control of nonlinear plants [Teeter and Chow, 1998; Werbos, 1990]. For system identification, a network can be trained offline using plant input/output data obtained from simulation of a mathematical model or from observation of the physical system. When the network is used for adaptive identification, training can be performed using *batch update* with a window of sampled data, or with the *pattern update* method in which training patterns consist of past inputs and outputs measured at each sample time.

In order to adaptively train a neural *controller* using gradient descent, the partial derivatives of a cost index, J_k, with respect to the network weights, $\mathbf{w}$, must be obtained [Chow and Yee, 1991]. Let J_k have the form $J_k = L(\mathbf{y}(k+1),\mathbf{u}(k+1))+L(\mathbf{y}(k+2),\mathbf{u}(k+2)) + \ldots + L(\mathbf{y}(k+n),\mathbf{u}(k+n))$ where k is the current sample. For simplicity of notation, let $L(\mathbf{y}(k+i),\mathbf{u}(k+i))$ be denoted by $L(k+i)$. Application of the chain rule yields

$$\frac{\partial L\left(k+i\right)}{\partial \mathbf{w}} = \frac{\partial L\left(k+i\right)}{\partial \mathbf{u}\left(k\right)}\frac{\partial \mathbf{u}\left(k\right)}{\partial \mathbf{w}} = \left[\frac{\partial L\left(k+i\right)}{\partial \mathbf{y}\left(k+i\right)}\frac{\partial \mathbf{y}\left(k+i\right)}{\partial \mathbf{u}\left(k\right)} + \frac{\partial L\left(k+i\right)}{\partial \mathbf{u}\left(k+i\right)}\frac{\partial \mathbf{u}\left(k+i\right)}{\partial \mathbf{u}\left(k\right)}\right]\frac{\partial \mathbf{u}\left(k\right)}{\partial \mathbf{w}} \tag{100.179}$$

The $\partial\mathbf{u}(k)/\partial\mathbf{w}$ term can be calculated using the backpropagation approach since the controller is a neural network. The $\partial\mathbf{y}(k+i)/\partial\mathbf{u}(k)$ and $\partial\mathbf{u}(k+i)/\partial\mathbf{u}(k)$ terms are obtained using the neural controller and identifiers. First, future inputs and outputs of the plant are predicted. The partial derivatives are then obtained by recursively computing the input/output sensitivities of the plant and controller through i samples. This approach is often referred to as *back-propagation through time* [Chow and Yee, 1991; Werbos, 1990].

The training algorithm resembles methods used by Nguyen and Widrow [1990] and others for training a neural controller to achieve an end goal. In this case, however, the output *trajectory* is of interest and the training is performed in realtime (i.e., output values must be repeatedly predicted rather than observed over several trials). A flowchart of the control methodology is shown in Fig. 100.64.

After controller outputs are computed, the weights of the controller are adjusted N times before the next sample time. The value of N can be selected based on time constraints or convergence properties of the neural controller and observers. If N is large, the neural observers are inaccurate; and if a large prediction horizon is used, the adaptation of controller parameters may cause performance to deteriorate.

HVAC Illustration

In order to demonstrate the ability of the neural identification and control schemes to handle disturbances and changes in nonlinear plant dynamics, a Heat, Ventilation, and Air-Conditioning (HVAC) system where a thermal load is added and the actual output vs. the commanded output of each actuator is modified. The neural observers are adapted at each time step in one simulation, while adaptation of the observers is stopped at time step $k =$ 200 in another simulation. Both simulations are performed for 1000 time steps. The reference trajectory is plotted in Fig. 100.65, along with the output of the system that uses nonadaptive identifiers after time step $k =$ 200. Tracking errors for both simulations are plotted in Fig. 100.66, where T_3 is the temperature to be controlled and r is the reference signal to be tracked.

The performance of the system with nonadaptive observers deteriorates due to the disturbance and the changes in plant dynamics. In this case, adapting the neural observers enables them to more accurately predict

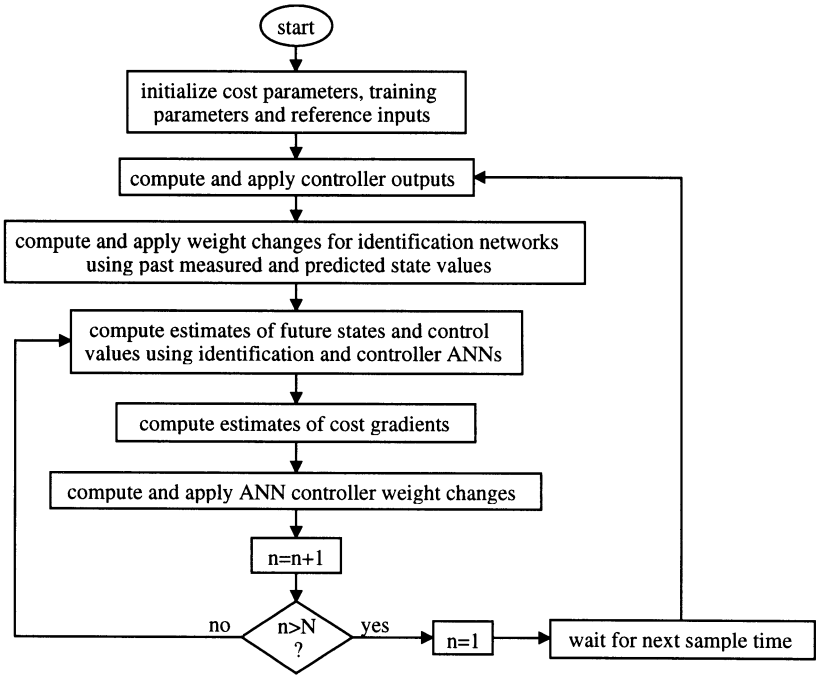


FIGURE 100.64 Flowchart of the neural control training methodology.

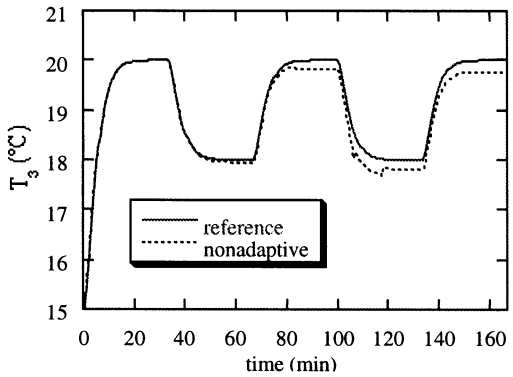


FIGURE 100.65 Reference and output trajectories using nonadaptive neural identifiers.

future states and inputs, and results in better tracking. It is important to select appropriate adaptation stepsizes for the observers because only one training pattern, consisting of the most recent set of plant inputs and states, is available for training. In order to compare the neural identification and control methodology with another potential control methodology, a PI-type controller has been designed for the HVAC system. Typical responses of the system with PI-type controller are shown in Fig. 100.67.

The simple PI-type controller satisfies the conservative performance specifications for the cases tested, but does not always use its resources efficiently. For example, if the outside air temperature of the HVAC system is close to the steady-state reference temperature, it may be more efficient to increase the room volumetric flow rate for a period of time in order to reduce the amount of heating or cooling performed by the heat exchanger. The neural and PI-type control schemes are tested using initial conditions $T_3(0) = 15°C$, a constant outside temperature of 22°C, and a steady-state reference value of 20°C. The tracking errors and heat exchanger outputs for both methods are shown in Figs. 100.68 and 100.69, respectively.

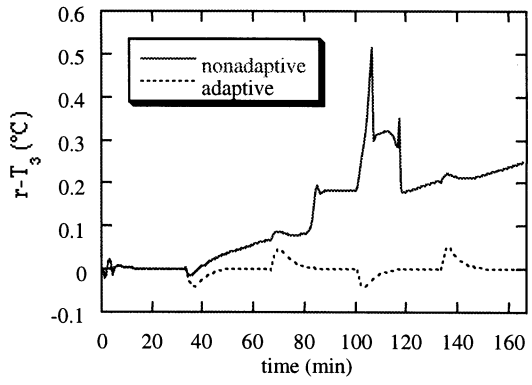


FIGURE 100.66 Tracking errors for the system using adaptive and nonadaptive neural identifiers.

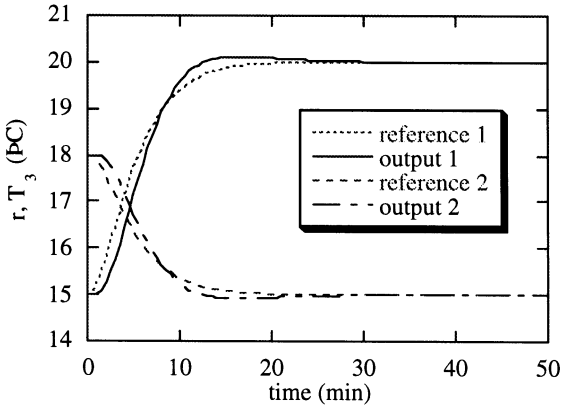


FIGURE 100.67 Typical responses of the system with PI-type controller.

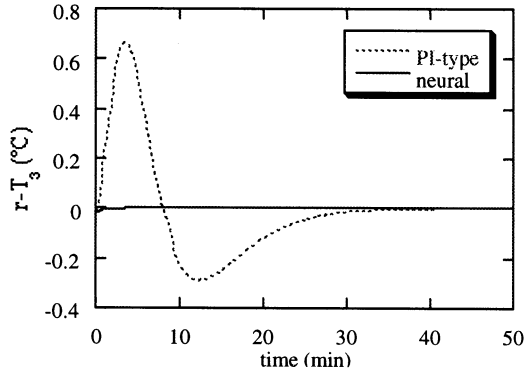


FIGURE 100.68 Tracking errors of the PI-type and neural control systems.

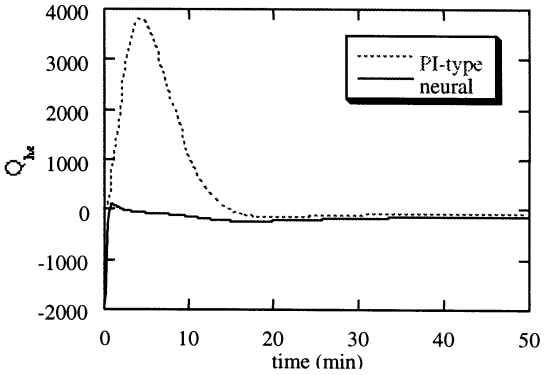


FIGURE 100.69 Heat exchanger outputs for the PI-type and neural control systems.

Conclusion

The use of neural networks for system identification and control provides a means of adapting a controller online in an effort to minimize a given cost index. The cost index includes typical measures associated with system performance and can be modified without significantly increasing the computational complexity of the adaptation process. For nonlinear systems, the identification networks demonstrate the capacity to learn changes in the plant dynamics. The performance of the neural control and identification methodology compares favorably with many types of conventional approaches.

References

Antsaklis, P. J., Albus, S., Lemmon, M. D., Mystel, A., Passino, K. M., Saridis, G. N., and Werbos, P. (1994). Defining intelligent control, *IEEE Control Systems,* 4, 5, 58–66.

Chow, M.-Y. and Menozzi, A. (1993). Design Methodology of an Intelligent Controller Using Artificial Neural Networks, *IECON'93,* Maui, Hawaii.

Chow, M.-Y. and Menozzi, A. (1994). A Self-Organized CMAC Controller for Robot Arm Movements, *IEEE International Conference on Industrial Technology,* Guangzhou, China.

Chow, M.-Y. and Menozzi, A. (1994). Using a Cerebellar Model for FES Control of the Upper Limb, *16th Annual International IEEE Engineering in Medicine and Biology Society Conference,* Baltimore, MD.

Chow, M.-Y. and Teeter, J. (1995). A knowledge-based approach for improved neural network control of a servomotor system with nonlinear friction characteristics, *Mechatronics,* 5(8), 949–962.
Chow, M.-Y. and Teeter, J. (1997). Reduced-Order Functional Link Neural Network for HVAC Thermal System Identification and Modeling, *1997 International Conference on Neural Networks,* Houston, TX.
Chow, M.-Y. and Yee, S. O. (1991). An adaptive backpropagation through time training algorithm for a neural controller, *1991 IEEE International Symposium on Intelligent Control,* Arlington, VA, 170–175.
Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems,* 2, 303–314.
Hunt, K. J., Sbarbaro, D., Zbikowski, R., and Gawthrop, P. J. (1992). Neural networks for control systems — a survey, *Automatica,* 28(6), 1083–1112.
Jordan, M. I. and Rumelhart, D. E. (1991). Forward models: supervised learning with a distal teacher. *Occasional Paper No. 40,* Center for Cognitive Science, MIT.
Miller, III, W. Thomas, Sutton, Richard S., Werbos, Paul J. (1990). *Neural Networks for Control,* The MIT Press, Cambridge, MA.
Narendra, K. S. and Annaswamy, A. M. (1989). *Stable Adaptive Systems,* Prentice-Hall, Englewood Cliffs, NJ.
Narendra, K. S. and Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks, *IEEE Transactions on Neural Networks,* 1(1), 4–27.
Nguyen, D. and Widrow, B. (1990). The truck backer-upper: an example of self-learning in neural networks, *Neural Networks for Control,* The MIT Press, Cambridge, MA, 287–299.
Poggio, T. and Girosi, F. (1990). Networks for approximation and learning, *Proceedings of the IEEE,* 78(9), 1481–1497.
Psaltis, D., Sideris, A., and Yamamura, A. A. (1988). A multilayered neural network controller, *IEEE Control Systems Magazine,* 17–21.
Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* The MIT Press, Cambridge, MA.
Teeter, J. and Chow, M.-Y. (1998). Application of functional link neural network to HVAC thermal dynamic system identification, *IEEE Transactions on Industrial Electronics,* 45(1), 170–176.
Teeter, J., Chow, M.-Y., and Brickley, J. J. Jr. Use of a Fuzzy Gain Tuner for Improved Control of a DC Motor System with Nonlinearities, *IEEE International Conference on Industrial Technology,* Guangzhou, China.
Teeter, J. T., Chow, M.-Y., and Brickley, J. J. Jr.(1996). A novel fuzzy friction compensation approach to improve the performance of a dc motor control system, *IEEE Transactions on Industrial Electronics,* 43(1), 113–120.
Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in Behavioral Science,* Harvard University Press, Cambridge, MA.
Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it, *Proceedings of the IEEE,* 78(10), 1550–1560.
Zurada, J. M. (1992). *Introduction to Artificial Neural Systems,* West Publishing Company, St. Paul, MN.

Lasky, T.A., Hsia, T.C., Tummala, R.L., Odrey, N.G. "Robotics"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

101

Robotics

Ty A. Lasky
University of California, Davis

Tien C. Hsia
University of California, Davis

R. Lal Tummala
Michigan State University

Nicholas G. Odrey
Lehigh University

101.1 Robot Configuration
Cartesian Configuration • Cylindrical Configuration • Spherical Configuration • Articulated Configuration • SCARA Configuration • Gantry Configuration • Additional Information

101.2 Dynamics and Control
Independent Joint Control of the Robot • Dynamic Models • Computed Torque Methods • Adaptive Control • Resolved Motion Control • Compliant Motion • Flexible Manipulators

101.3 Applications
Justification • Implementation Strategies • Applications in Manufacturing • Emerging Issues

101.1 Robot Configuration

Ty A. Lasky and Tien C. Hsia

Configuration is a fundamental classification for industrial robots. Configuration refers to the geometry of the robot manipulator, i.e., the manner in which the links of the manipulator are connected at each joint. The Robotic Industries Association (RIA) defines a robot as *a manipulator designed to move material, parts, tools, or specialized devices, through variable programmed motions for the performance of a variety of tasks*. With this definition, attention here is focused on industrial manipulator arms, typically mounted on a fixed pedestal base. Mobile robots and hard automation [e.g., Computer Numerical Control (CNC) machines] are excluded. The emphasis here is on serial-chain manipulator arms, which consist of a serial chain of linkages, where each link is connected to exactly two other links, with the exception of the first and last, which are connected to only one other link. Additionally, the first three links, called the major linkages, are focused on, with only a brief mention of the last three links, or wrist joints, also called the minor linkages.

Robot configuration is an important consideration in the selection of a manipulator. Configuration refers to the way the manipulator links are connected at each joint. Each link will be connected to the subsequent link by either a linear (sliding or prismatic) joint, which can be abbreviated with a P, or a revolute (or rotary) joint, abbreviated with an R. Using this notation, a robot with three revolute joints would be abbreviated as RRR, while one with a rotary joint followed by two linear (prismatic) joints would be denoted RPP. Each configuration type is well suited to certain types of tasks and ill suited to others. Some configurations are more versatile than others. In addition to the geometrical considerations, robot configuration affects the structural stiffness of the robot, which may be an important consideration. Also, configuration impacts the complexity of the forward and inverse **kinematics**, which are the mappings between the robot actuator (joint) space, and the Cartesian position and orientation of the robot end-effector, or tool.

There are six major robot configurations commonly used in industry. Details for each configuration are presented in subsequent subsections. The simplest configuration is the Cartesian robot, which consists of three orthogonal, linear joints (PPP), so that the robot moves in the x, y, and z directions in the joint space. The

cylindrical configuration consists of one revolute and two linear joints (RPP), so that the robot joints correspond to a cylindrical coordinate system. The spherical configuration consists of two revolute joints and one linear joint (RRP), so that the robot moves in a spherical, or polar, coordinate system. The articulated (arm-and-elbow) configuration consists of three revolute joints (RRR), giving the robot a somewhat human-like range of motion. The SCARA (Selectively Compliant Assembly Robot Arm) configuration consists of two revolute joints and one linear joint (RRP), arranged in a different fashion than the spherical configuration. It may also be equipped with a revolute joint on the final sliding link. The gantry configuration is essentially a Cartesian configuration, with the robot mounted on an overhead track system. One can also mount other robot configurations on an overhead gantry system to give the robot an extended workspace, as well as free up valuable factory floor space. The percentage usage of the first five configuration types is listed in Table 101.1. This table does not include gantry robots, which are assumed to be included in the Cartesian category. Additionally, this information is from 1988, and does not accurately represent current usage.

TABLE 101.1 Robot Arm Geometry Usage

Arm Geometry	Percent of Use
Cartesian	18
Cylindrical	15
Spherical	10
Articulated	42
SCARA	15

Source: V. D. Hunt, *Robotics Sourcebook,* New York: Elsevier, 1988. With permission.

In general, robots with a rotary base have a speed advantage. However, they have more variation in resolution and dynamics compared to Cartesian robots. This can lead to inferior performance if a fixed controller is used over the robot's entire workspace.

Cartesian Configuration

The Cartesian configuration consists of three orthogonal, linear axes, abbreviated as PPP, as shown in Fig. 101.1. Thus, the joint space of the robot corresponds directly with the standard right-handed Cartesian *xyz*-coordinate system, yielding the simplest possible kinematic equations. The work envelope of the Cartesian robot is shown in Fig. 101.2. The work envelope encloses all the points that can be reached by the robot arm or the mounting point for the end-effector or tool. The area reachable by an end effector or tool is not considered part of the work envelope. All interaction with other machines, parts, or processes must take place within this volume [Critchlow, 1985]. Here, the workspace of a robot is assumed to be equivalent to the work envelope.

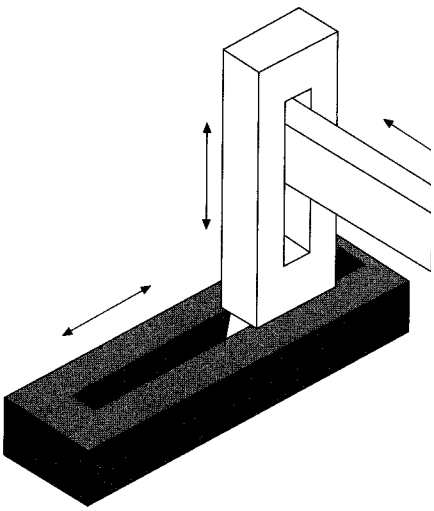


FIGURE 101.1 The Cartesian configuration. (*Source:* T. Owen, *Assembly with Robots*, Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.)

There are several advantages to this configuration. As noted above, the robot is kinematically simple, since motion on each Cartesian axis corresponds to motion of a single actuator. This eases the programming of linear motions. In particular, it is easy to do a straight vertical motion, the most common motion in assembly tasks. The Cartesian geometry also yields a constant arm resolution throughout the workspace; i.e., for any configuration, the resolution for each axis corresponds directly to the resolution for that joint. The simple geometry of the Cartesian robot leads to correspondingly simple manipulator dynamics. The disadvantages of this configuration include inability to reach objects on the floor or points invisible from the base of the robot, and slow speed of operation in the horizontal plane compared to robots with a rotary base. Additionally, the Cartesian configuration requires a large operating volume for a relatively small workspace.

Cartesian robots are used for several applications. As noted above, they are well suited for assembly operations, as they easily perform vertical straight-line insertions. Because of the ease of straight-line motions, they are also well suited to machine loading and unloading. They are also used in clean room tasks.

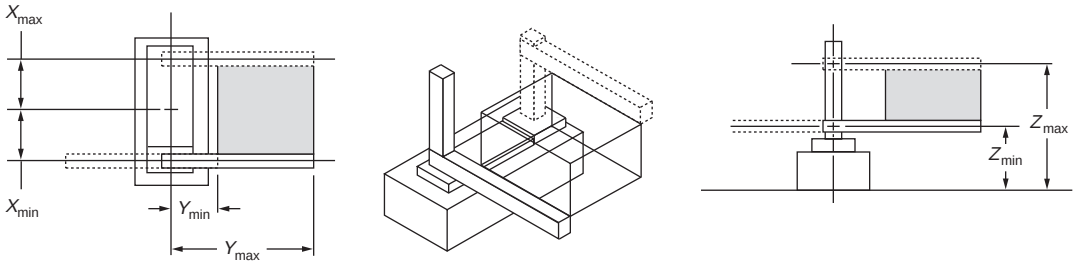


FIGURE 101.2 Cartesian robot work envelope.

Cylindrical Configuration

The cylindrical configuration consists of one vertical revolute joint and two orthogonal linear joints (RPP), as shown in Fig. 101.3. The resulting work envelope of the robot is a cylindrical annulus, as shown in Fig. 101.4. This configuration corresponds with the cylindrical coordinate system.

As with the Cartesian robot, the cylindrical robot is well suited for straight-line vertical and horizontal motions, so it is useful for assembly and machine loading operations. It is capable of higher speeds in the horizontal plane due to the rotary base. However, general horizontal straight-line motion is more complex and correspondingly more difficult to coordinate. Additionally, the end-point resolution of the cylindrical robot is not constant but depends on the extension of the horizontal linkage. A cylindrical robot cannot reach around obstacles. Additionally, if a monomast construction is used on the horizontal linkage, then there can be clearance problems behind the robot.

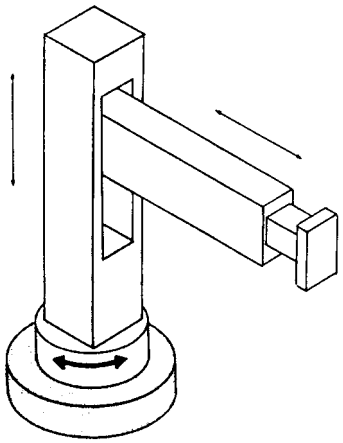


FIGURE 101.3 The cylindrical configuration. (*Source:* T. Owen, *Assembly with Robots*, Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.)

Spherical Configuration

The spherical (or polar) configuration consists of two revolute joints and one linear joint (RRP), as shown in Fig. 101.5. This results in a set of joint coordinates that matches with the spherical coordinate system. A typical work envelope for a spherical robot is shown in Fig. 101.6.

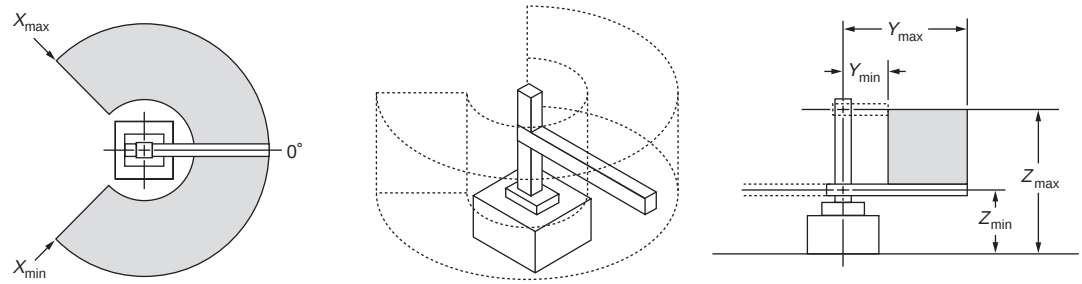


FIGURE 101.4 Cylindrical robot work envelope.

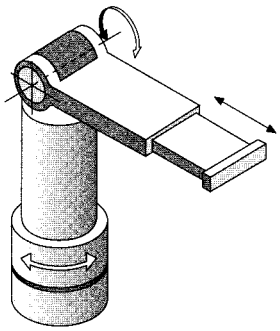


FIGURE 101.5 The spherical configuration. (*Source:* T. Owen, *Assembly with Robots*, Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.)

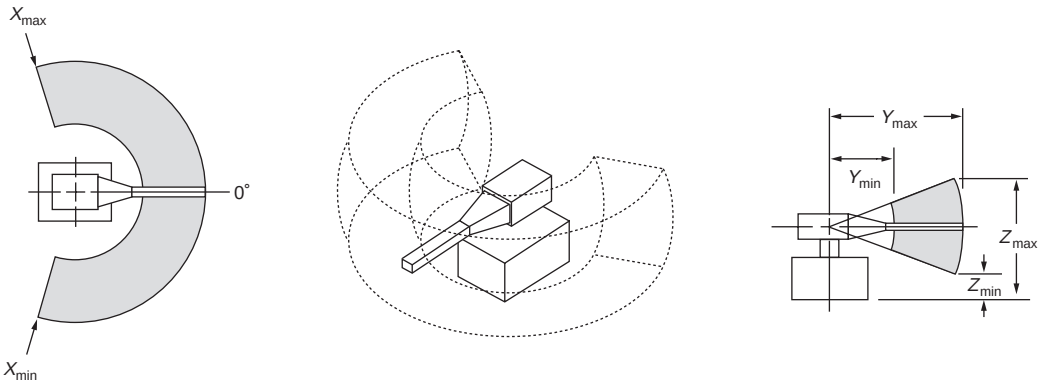


FIGURE 101.6 Spherical robot work envelope.

Spherical robots are typically heavy-duty robots. They have the advantages of high speed due to the rotary base, and a large work volume, but are more kinematically complex than either Cartesian or cylindrical robots. Generally, they are used for heavy-duty tasks in, for example, automobile manufacturing. They do not have the dexterity to reach around obstacles in the workspace. Spherical robots also do not have fixed resolution throughout the workspace.

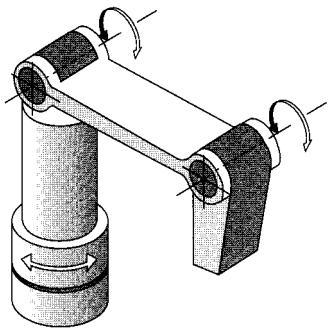


FIGURE 101.7 The articulated configuration. (*Source:* T. Owen, *Assembly with Robots*, Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.)

Articulated Configuration

The articulated (or anthropomorphic, jointed, arm-and-elbow) configuration consists of three revolute joints (RRR), as shown in Fig. 101.7. The resulting joint coordinates do not directly match any standard coordinate system. A slice of a typical work envelope for an articulated robot is shown in Fig. 101.8.

The articulated robot is currently the most commonly used in research. It has several advantages over other configurations. It is closest to duplicating the motions of a human assembler, so there should be less need to redesign an existing workstation to utilize an articulated robot. It has a very large, dexterous work envelope; i.e., it can reach most points in its work envelope from a variety of orientations. Thus, it can more easily reach around or over obstacles in the workspace or into parts or machines. Because all the joints are revolute, high

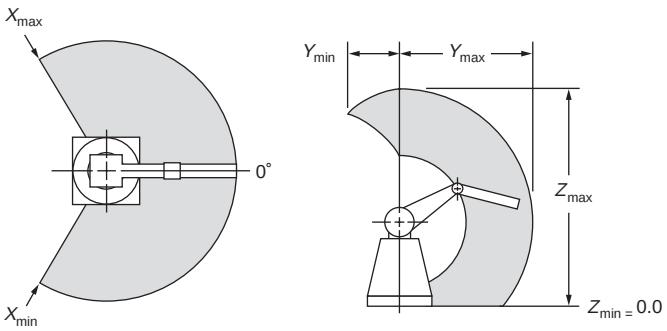


FIGURE 101.8 Articulated robot work envelope.

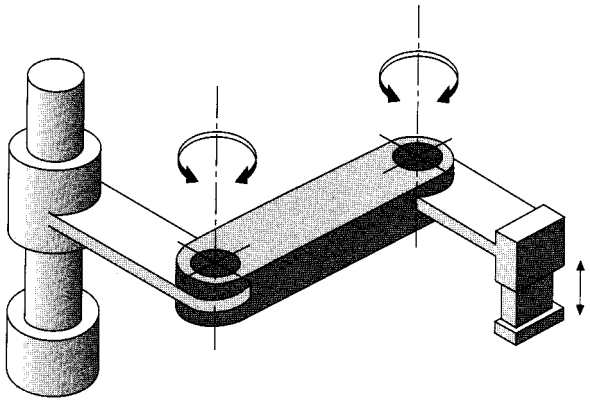


FIGURE 101.9 The SCARA configuration. (*Source:* T. Owen, *Assembly with Robots*, Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.)

speeds are possible. The articulated arm is good for tasks involving multiple insertions, complex motions, and varied tool orientations. The versatility of this configuration makes it applicable to a variety of tasks, so the user has fewer limitations on the use of the robot. However, the same features that give this robot its advantages lead to certain disadvantages. The geometry is complex, and the resulting kinematic equations are quite intricate. Straight-line motion is difficult to coordinate. Control is generally more difficult than for other geometries, with associated increase in cost. Here again, arm resolution is not fixed throughout the workspace. Additionally, the dynamics of an articulated arm vary widely throughout the workspace, so that performance will vary over the workspace for a fixed controller. In spite of these disadvantages, the articulated arm has been applied to a wide variety of research and industrial tasks, including spray painting, clean room tasks, machine loading, and parts-finishing tasks.

SCARA Configuration

The SCARA (Selectively Compliant Assembly Robot Arm) configuration consists of two revolute joints and a linear joint (RRP), as shown in Fig. 101.9. This configuration is significantly different from the spherical configuration, since the axes for all joints are always vertical. In addition to the first three **degrees of freedom** (DOF), the SCARA robot will often include an additional rotation about the last vertical link to aid in orientation of parts. The work envelope of the SCARA robot is illustrated in Fig. 101.10. The SCARA configuration is the newest of the configurations discussed here, and was developed by Professor Hiroshi Makino of Yamanashi University, Japan.

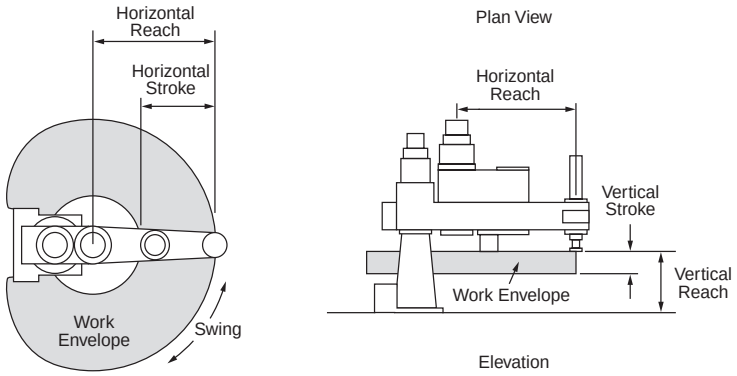


FIGURE 101.10 SCARA robot work envelope.

This configuration has many advantages and is quite popular in industry. The configuration was designed specifically for assembly tasks [Truman, 1990], so has distinct advantages when applied in this area. Because of the vertical orientation of the joints, gravity does not affect the dynamics of the first two joints. In fact, for these joints, the actuators can be shut off and the arm will not fall, even without the application of brakes. As the name SCARA implies, this allows compliance in the horizontal directions to be selectively varied; therefore, the robot can comply to horizontal forces. Horizontal compliance is important for vertical assembly operations. Because of the vertical linear joint, straight-line vertical motions are simple. Also, SCARA robots typically have high positional repeatability. The revolute joints allow high-speed motion. On the negative side, the resolution of the arm is not constant throughout the workspace, and the kinematic equations are relatively complex. In addition, the vertical motion of the SCARA configuration is typically quite limited. While the SCARA robot can reach around objects, it cannot reach over them in the same manner as an articulated arm.

Gantry Configuration

The gantry configuration is geometrically equivalent to the Cartesian configuration, but is suspended from an overhead crane and typically can be moved over a large workspace. It consists of three linear joints (PPP), and is illustrated in Fig. 101.11. In terms of work envelope, it will have a rectangular volume that sweeps out most of the inner area of the gantry system, with a height limited by the length of the vertical mast, and the headroom above the gantry system. One consideration in the selection of a gantry robot is the type of vertical linkage employed in the z axis. A monomast design is more rigid, yielding tighter tolerances for repeatability and accuracy, but requires significant headroom above the gantry to have a large range of z axis motion. On the other hand, a telescoping linkage will require significantly less headroom but is less rigid, with corresponding reduction in repeatability and accuracy. Other robot configurations can be mounted on gantry systems, thus gaining many of the advantages of this geometry.

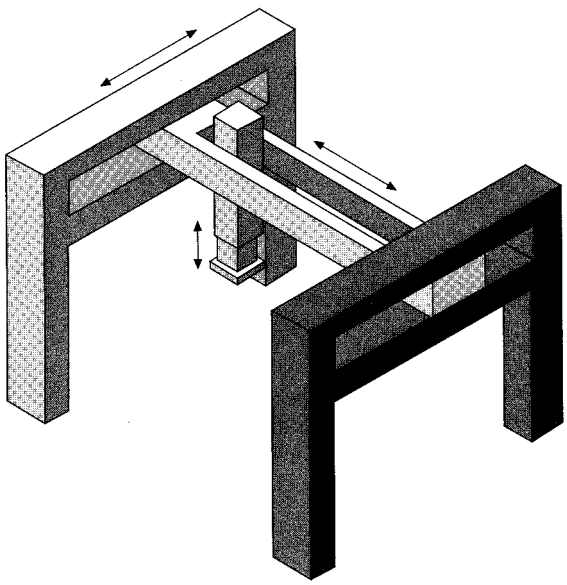


FIGURE 101.11 The gantry configuration. (*Source:* T. Owen, *Assembly with Robots*, Englewood Cliffs, N.J.: Prentice-Hall, 1985. With permission.)

Gantry robots have many advantageous properties. They are geometrically simple, like the Cartesian robot, with the corresponding kinematic and dynamic simplicity. For the same reasons, the gantry robot has a constant arm resolution throughout the workspace. The gantry robot has better dynamics than the pedestal-mounted Cartesian robot, as its links are not cantilevered. One major advantage over revolute-base robots is that its dynamics vary much less over the workspace. This leads to less vibration and more even performance than typical pedestal-mounted robots in full extension. Gantry robots are much stiffer than other robot configurations, although they are still much less stiff than Numerical Control (NC) machines. The gantry robot can straddle a workstation, or several workstations for a large system, so that one gantry robot can perform the work of several pedestal-mounted robots. As with the Cartesian robot, the gantry robot's simple geometry is similar to that of an NC machine, so technicians will be more familiar with the system and require less training time. Also, there is no need for special path or trajectory computations. A gantry robot can be programmed directly from a Computer-Aided Design (CAD) system with the appropriate interface, and straight-line motions are particularly simple to program. Large gantry robots have a very high payload capacity. Small, table-top systems can achieve linear speeds of up to 40 in./s (1.025 m/s), with a payload capacity of 5.0 lb (2.26 kg), making them suitable for assembly operations. However, most gantry robot systems are not as precise as other configurations, such as the SCARA configuration. Additionally, it is sometimes more difficult to apply a gantry robot to an existing workstation, as the workpieces must be brought into the gantry's work envelope, which may be harder to do than for a pedestal-mounted manipulator.

Gantry robots can be applied in many areas. They are used in the nuclear power industry to load and unload reactor fuel rods. Gantry robots are also applied to materials-handling tasks, such as parts transfer, machine loading, palletizing, materials transport, and some assembly applications. In addition, gantry robots are used for process applications such as welding, painting, drilling, routing, cutting, milling, inspection, and nondestructive testing.

The gantry robot configuration is the fastest-growing segment of the robotics industry. While gantry robots accounted for less than 5% of the units shipped in 1985, they are projected to account for about 30% of the robots by the end of the 1990s. One reason for this is summed up in Long [1990], which contains much more information on gantry robots in general:

> Currently gantry robot cells are being set up which allow manufacturers to place a sheet of material in the gantry's work envelope and begin automatic cutting, trimming, drilling, milling, assembly and finishing operations which completely manufacture a part or subassembly using quick-change tools and programmed subroutines.

Additional Information

The above six configurations are the main types currently used in industry. However, there are other configurations used in either research or specialized applications. Some of these configurations have found limited application in industry and may become more prevalent in the future.

All the above configurations are serial-chain manipulators. An alternative to this common approach is the parallel configuration, known as the Stewart platform [Waldron, 1990]. This manipulator consists of two platforms connected by three prismatic linkages. This arrangement yields the full six DOF motion (three position, three orientation) that can be achieved with a six-axis serial configuration but has a comparably very high stiffness. It is used as a motion simulator for pilot training and virtual reality applications. The negative aspects of this configuration are its relatively restricted motion capability and geometric complexity.

The above configurations are restricted to a single manipulator arm. There are tasks that are either difficult or impossible to perform with a single arm. With this realization, there has been significant interest in the use of multiple arms to perform coordinated tasks [Bonitz and Hsia, 1996]. Possible applications include carrying loads that exceed the capacity of a single robot, and assembling objects without special fixturing. Multiple arms are particularly useful in zero-gravity environments. While there are significant advantages to the use of multiple robots, the complexity, in terms of kinematics, dynamics, and control, is quite high. However, the use of multiple robots is opening new areas of application for robots.

Typical industrial robots have six or fewer DOF. With six DOF, the robot can, within its work envelope, reach arbitrary positions and orientations. At the edge of the work envelope, a six-DOF robot can attain only one orientation. To increase the geometric dexterity of the manipulator, it is useful to consider robots with more than six DOF, i.e., redundant robots. These robots are highly dexterous and can use the extra DOF in many ways: avoidance obstacle, joint torque minimization, kinematic **singularity** (points where the manipulator cannot move in certain directions) avoidance, bracing strategies where part of the arm is braced against a structure, which raises the lowest structural resonant frequency of the arm, etc. While the **redundant manipulator** configuration has many desirable properties, the geometric complexity has limited their application in industry.

For any of the six standard robot configurations, the orientation capability of the major linkages is severely limited. Thus, it is critical to provide additional joints, known as the minor linkages, to provide the capability of varied orientations for a given position. Most robots include a three-DOF revolute joint wrist that is connected to the last link of the major linkages. The three revolute axes will be orthogonal and will usually intersect in a common point, known as the wrist center point. Then, the kinematic equations of the manipulator can be partitioned into locating the Cartesian position of the wrist center point and then determining the orientation of a Cartesian frame fixed to the wrist axes.

Conclusions

Each of the six standard configurations has specific advantages and disadvantages. When choosing a manipulator for a task, the properties of the manipulator geometry are one of the most important considerations. If the manipulator will be used for a wide variety of tasks, one may need to trade off performance for any given task for the flexibility that will allow the manipulator to work for the various tasks. In such a case, a more flexible geometry should be considered. The future of robotics will be interesting. With the steady increase in computational capabilities, the more complex geometries, including redundant and multiple robots, are beginning to see increased applications in industry.

Defining Terms

Degrees of freedom: The number of degrees of freedom (DOF) of a manipulator is the number of independent position variables that must be specified in order to locate all parts of the manipulator. For a typical industrial manipulator, the number of joints equals the number of DOF.

Kinematics: The kinematics of the manipulator refers to the geometric properties of the manipulator. *Forward kinematics* is the computation of the Cartesian position and orientation of the robot end-effector given the set of joint coordinates. *Inverse kinematics* is the computation of the joint coordinates given the Cartesian position and orientation of the end-effector. The inverse kinematic computation may not be possible in closed form, may have no solution, or may have multiple solutions.

Redundant manipulator: A redundant manipulator contains more than six DOF.

Singularity: A *singularity* is a location in the workspace of the manipulator at which the robot loses one or more DOF in Cartesian space, i.e., there is some direction (or directions) in Cartesian space along which it is impossible to move the robot end-effector no matter which robot joints are moved.

Related Topic

101.2 Dynamics and Control

References

R.G. Bonitz and T.C. Hsia, "Internal force-based impedance control for cooperating manipulators," *IEEE Transactions on Robotics and Automation*, Feb. 1996.

J.J. Craig, *Introduction to Robotics: Mechanics and Control*, Reading, Mass.: Addison-Wesley, 1986.

A. J. Critchlow, *Introduction to Robotics*, New York: Macmillan, 1985.

E. Long, "Gantry robots," in *Concise International Encyclopedia of Robotics*, R. C. Dorf, Ed., New York: Wiley-Interscience, 1990.
R. Truman, "Component assembly onto printed circuit boards," in *Concise International Encyclopedia of Robotics*, R. C. Dorf, Ed., New York: Wiley-Interscience, 1990.
K. J. Waldron, "Arm design," in *Concise International Encyclopedia of Robotics*, R. C. Dorf, Ed., New York: Wiley-Interscience, 1990.

Further Information

The journal *IEEE Transactions on Robotics and Automation* is a valuable source for a wide variety of robotics research topics, occasionally including new robot configurations. Additionally, IEEE's *Control Systems Magazine* occasionally publishes an issue devoted to robotic systems. The home page for the IEEE Robotics and Automation Society can be found at "http://www.acim.usi.edu/RAS/".

Another journal that often has robotics-related articles is the ASME *Journal of Dynamic Systems, Measurement and Control.*

An additional source of robotics information is *The Proceedings of the IEEE International Conference on Robotics and Automation.* This conference is held annually.

Useful sources on the World Wide Web include the *Robotics Internet Resources* page, located at "http://piglet.cs.umass.edu:4321/robotics.html", and *Robotics Resources,* located at "http://www.eg.bucknell.edu/~robotics/rirc.html". Consult your system administrator for information on this web access.

101.2 Dynamics and Control

R. Lal Tummala

The primary purpose of the robot control system is to issue commands to joint actuators to faithfully execute a planned trajectory in the **tool space**. This may involve position control when the manipulator is following a trajectory through free space or a combination of position and force control if the manipulator is to react continuously to contact forces at the tool or end-effector.

Control systems can operate either in open loop or closed loop. In open-loop systems, the output has no effect on the input. On the other hand, closed-loop systems continuously sense the output and make appropriate adjustments to the input in order to keep the output at the desired level.

The majority of the current industrial robots use the **independent joint control** method and close the loop around the joints of the robot. The desired joint positions corresponding to a tool trajectory are either taught by using a *teach box* or generated by solving an inverse kinematics problem. The independent joint control method, however, is effective only at low speeds. As the speeds increase, the coupling effects between the joints increase and warrant the inclusion of these effects in the controller development. Advanced controller development and implementation based on full dynamics is one of the active areas of current research. New advances in sensor technology, faster computers, advanced robots such as direct drive arms and industrial competition provide new opportunities and motivation for accelerating the development and implementation of advanced controllers for robots in the near future.

Independent Joint Control of the Robot

The independent joint control method assumes that a single joint of a robot is moving while all the other joints are fixed. A typical joint position control system is shown in Fig. 101.12, where the actuator used is a dc servomotor [Luh, 1983]. In general, any one or a combination of electric motors and hydraulic or pneumatic pistons can be used to move the joint through the desired positions. These motors may be connected directly to the joint or indirectly through gears, chains, cables, or lead screws. The desired joint positions that are inputs to the position loops are obtained from the trajectory planner. The actual position of the joint is obtained by using a position sensor, such as a potentiometer or an optical encoder. An amplifier is used for increasing the system gain, denoted by K_a. The velocity feedback K_v is used to reinforce the effect of back emf for controlling

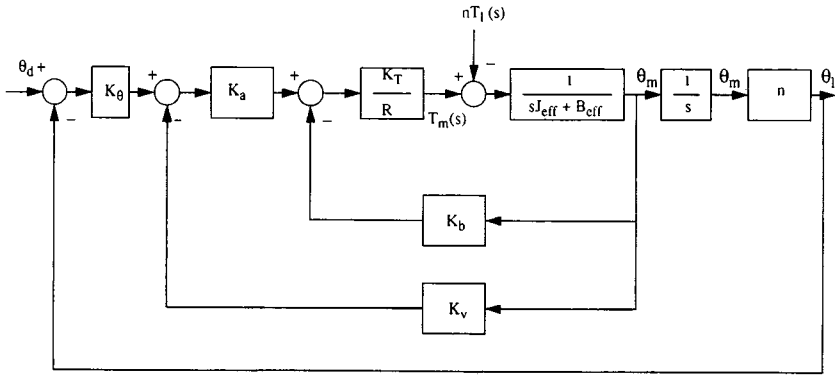


FIGURE 101.12 Closed-loop control of a robot joint. (*Source:* Adapted from J. Y. S. Luh, "Conventional controller design for industrial robots: A tutorial," *IEEE Trans. Systems, Man, Cybernetics,* vol. SMC-13, no. 3, June 1983. © 1983 IEEE.)

the damping of the system. This can be done either using a tachometer or computing the difference in angular displacements of the actuator shaft over a fixed time interval.

The design of the control system involves fixing the values of K_a and K_v to achieve the desired response. Consider the closed-loop transfer function of the system shown in Fig. 101.12 (assuming $nT_l = 0$),

$$\frac{\theta_l(s)}{\theta_d(s)} = \frac{nK_aK_TK_\theta}{s^2RJ_{\text{eff}} + s(RB_{\text{eff}} + K_TK_b + K_aK_TK_v) + nK_TK_aK_\theta} \tag{101.1}$$

where K_a = gain of the amplifier, K_T = torque constant of the motor, K_b = back emf constant, K_θ = position sensor constant (volts/rad), R = resistance of the motor winding (ohms), and n = gear ratio. θ_L = link position (rad) and θ_m = angular displacement at the actuator side (rad).

The effective inertia, J_{eff}, and damping, B_{eff}, are defined as

$$J_{\text{eff}} = J_m + n^2J_L \tag{101.2}$$

and

$$B_{\text{eff}} = B_m + n^2B_L \tag{101.3}$$

where J_m = total inertia on the motor side, B_m = damping coefficient at the motor side, J_L = inertia of the robot link, and B_L = damping coefficient at the load side.

This is a second-order system and stable for all values of K_a and K_v. The values of K_a and K_v are selected to achieve a desired transient response by fixing the damping ratio and the natural frequency of the system and are described below.

The characteristic equation for the above system is

$$s^2 + s\frac{RB_{\text{eff}} + K_TK_b + K_aK_TK_v}{RJ_{\text{eff}}} + \frac{nK_TK_aK_\theta}{RJ_{\text{eff}}} = 0 \tag{101.4}$$

This can be conveniently written as

$$s^2 + 2\zeta\omega_n s + \omega_n^2 = 0 \tag{101.5}$$

where the natural frequency ω_n and the damping ratio ζ of the system are given as

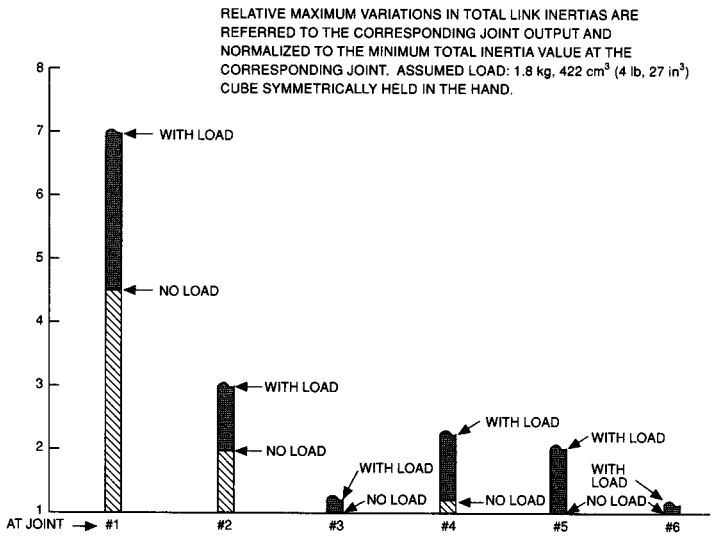


FIGURE 101.13 Variations of link inertias for JPL-Stanford manipulator. (*Source:* A.K. Bejczy, Jet Propulsion Lab, Pasadena, Calif., American Automatic Control Conference Tutorial Workshop, Washington, D.C., June 18, 1982.)

$$\omega_n = \sqrt{\frac{nK_aK_TK_\theta}{RJ_{\text{eff}}}} > 0 \tag{101.6}$$

$$\zeta = \frac{RB_{\text{eff}} + K_TK_b + K_aK_TK_v}{2\sqrt{nK_aK_TRJ_{\text{eff}}K_\theta}} \tag{101.7}$$

These systems are designed to operate with critical damping ($\zeta = 1$) because an underdamped system ($\zeta < 1$) has fast response but results in an overshoot, whereas an overdamped system ($\zeta > 1$) is too slow. However, this is not always possible, because the damping ratio given by Eq. (101.7) depends on B_{eff} and J_{eff} which vary during the actual operation of the manipulator. B_{eff} changes with age or repeated use of the manipulator. J_{eff} varies with the payload. For example, the variation of J_{eff} for the Stanford manipulator under various loading conditions is shown in Fig. 101.13. J_{eff} also varies with the configuration of the manipulator during the actual operation. So a compromise solution will be to design the controller such that $\zeta \geq 1$ throughout the intended operation.

The undamped natural frequency ω_n is selected to be no more than half the resonance frequency of the robot to avoid any structural damage to the robot [Paul, 1981]. These resonances are possible due to the flexibilities associated with the links of the robot and the shafts within the drive system, to name a few. These are called *unmodeled* resonances because they are not explicitly included in the model. In our case, if K_{eff} and J_{eff} are the effective stiffness and the inertias of the joint, respectively, then the resonance frequency ω_r is given by

$$\omega_r = \sqrt{\frac{K_{\text{eff}}}{J_{\text{eff}}}} \tag{101.8}$$

Since K_{eff} is difficult to estimate but constant for a given joint, we can experimentally determine the resonance frequencies for a known inertia and use this information for fixing the gain. Suppose ω is the resonance frequency for a given value of effective inertia J, then

$$\omega = \sqrt{\frac{K_{\text{eff}}}{J}} \tag{101.9}$$

To minimize the effects of unmodeled resonances, we use

$$\omega_n \leq \frac{\omega_r}{2} = \frac{\omega}{2}\sqrt{\frac{J}{J_{\text{eff}}}} \tag{101.10}$$

The selection of K_a and K_v depends on selecting ζ and ω_n. Using Eqs. (101.6) and (101.10), we can find an upper bound on K_a given by

$$K_a \leq \frac{J\omega^2 R}{4nK_T K_\theta} \tag{101.11}$$

The upper bound on K_v is obtained by setting $\zeta \geq 1$. Using Eq. (101.7),

$$RB_{\text{eff}} + K_T K_b + K_a K_T K_v \geq 2\sqrt{nK_a K_T R J_{\text{eff}} K_\theta} \tag{101.12}$$

Substituting the upper bound for K_a from Eq. (101.11), we get

$$K_v \geq \left(\omega R\sqrt{JJ_{\text{eff}}} - RB_{\text{eff}} - K_T K_b\right)\frac{4nK_\theta}{J\omega^2 R} \tag{101.13}$$

The steady-state errors to step position commands for the system shown in Fig. 101.12 are zero. However, in the presence of disturbances such as external load torques or **gravitational torques**, the system will have steady-state errors. For example, if T_L is the load torque as shown in Fig. 101.12, the available torque for the joint motion is given by

$$(J_{\text{eff}}s^2 + B_{\text{eff}}s)\,\theta_m(s) = T_m(s) - nT_L(s) \tag{101.14}$$

Using the superposition property, we get

$$\theta_L(s) = F_1(s)\theta_d(s) + F_2(s)T_L(s) \tag{101.15}$$

where

$$\begin{aligned} F_1(s) &= \frac{nK_a K_T K_\theta}{\Omega(s)} \\ F_2(s) &= -\frac{n^2 R}{\Omega(s)} \end{aligned} \tag{101.16}$$

$$\Omega(s) = RJ_{\text{eff}}s^2 + (RB_{\text{eff}} + K_b K_T + K_v K_a K_T)s + nK_a K_T K_\theta$$

Now if $T_L(s) = C_L/s$ and $\theta_d(s) = C_\theta/s$, then the steady-state error is

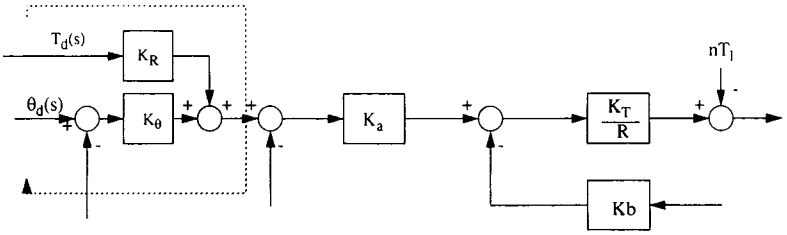


FIGURE 101.14 Feedback compensation method for disturbances. (*Source:* Adapted from J. Y. S. Luh, "Conventional controller design for industrial robots: A tutorial," *IEEE Trans. Systems, Man, Cybernetics,* vol. SMC-13, no. 3, June 1983. ©1983 IEEE.)

$$e_{ss} = \frac{nC_L R}{K_a K_T K_\theta} \tag{101.17}$$

Since the value of K_a has an upper bound, this error cannot be made arbitrarily small. A possible way to reduce this error is to add a feedforward term, as shown in Fig. 101.14 [Luh, 1983]. The feedforward signal $T_d(s)$ is chosen such that the steady-state error is zero. In this case,

$$T_d(s) = \frac{R}{K_T K_a K_R} n\hat{T}_L(s) \tag{101.18}$$

Similar considerations apply for other disturbances such as frictional torques and gravitational torques. Notice from Eq. (101.18) that the feedforward signal is a function of the estimated torque. The burden of determining these torques should not be underestimated. The other factor that was not mentioned earlier is the centrifugal torque, a nonlinear function of velocity. In the case of positioning applications, the velocity tends to zero as $t \to \infty$. However, if the robot is required to follow a conveyor with constant speed, then the input is a velocity. In this case, the centrifugal contribution will affect the steady-state velocity error. A feedforward compensation can be used in this case as well. Another method of compensating for the steady-state errors caused by gravitational and load torque disturbances is by adding an integral feedback (PID control), which of course increases the order of the system. The system is no longer stable for all values of the gains and thus adds another constraint in the selection of K_a and K_v.

So far we have considered the control of one joint of the robot while the other joints are fixed. Implementation of this control by successively positioning each joint while the other joints are fixed slows the robot operation and can also result in awkward hand motions, which is undesirable especially when the robot is supposed to follow a continuous path. Simultaneous fast motion of the joints, on the other hand, requires the inclusion of dynamic interactions between the joints. The controllers designed without considering these dynamic interactions tend to make the arm move slower and can potentially cause overshoots, oscillations, and path errors. To estimate the dynamic effects, one needs to obtain the equations of motion (dynamic models) of the robot. These equations are, in general, complex and take the form of coupled nonlinear differential equations.

Dynamic Models

Two of the most popular methods used to obtain dynamic models of the robot are the *Newton-Euler method* and the *Lagrange-Euler method*. The equations obtained using Lagrangian formulation are more suitable for the application of modern control theory than the recursive equations obtained using the Newton-Euler method. In the Lagrangian formulation, the dynamic models are obtained using kinetic and potential energies associated with the rigid bodies in motion. The derivation is systematic and conceptually simple. This method yields closed-form dynamic equations that explicitly express joint variables in terms of joint torques. To arrive

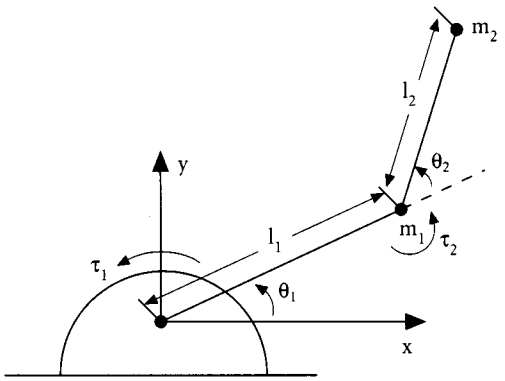


FIGURE 101.15 Two-degree-of-freedom planar manipulator.

at these equations, one starts with a set of generalized coordinates q_i, $i = 1, 2, 3, \ldots, n$, that completely locate the dynamic system and finds the total kinetic energy K and potential energy P of the system [Paul, 1981]. Then the equations of motion are given by

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = T_i \qquad \text{for } i = 1, 2, 3 \ldots n \tag{101.19}$$

where T_i is the generalized force and $L(q,q\cdot) = K - P$ is the Lagrangian. A simple example is given next to illustrate these ideas.

Example. Consider a planar arm with two degrees of freedom, shown in Fig. 101.15. For simplicity, we assume that the masses m_1 and m_2 of the links are represented by point masses at the end of the links. The link lengths are l_1 and l_2, respectively. The variables θ_1 and θ_2 are the joint angles. We know that the kinetic energy of a mass m moving at a linear velocity v is given by $1/2\ mv^2$ and the potential energy associated with a mass m located at a height h in a gravitational field is given by mgh, where g is the gravitational constant.

The kinetic energy K_1 for mass m_1 is found by observing that

$$\begin{aligned} x_1 &= l_1 \cos\theta_1 \\ y_1 &= l_1 \sin\theta_1 \\ v_1^2 &= \dot{x}_1^2 + \dot{y}_1^2 \\ K_1 &= \tfrac{1}{2} m_1 l_1^2 \dot{\theta}_1^2 \end{aligned} \tag{101.20}$$

Similarly, the kinetic energy K_2 for mass m_2 is given by

$$\begin{aligned} K_2 &= \tfrac{1}{2} m_2 v_2^2 \\ v_2^2 &= \dot{x}_2^2 + \dot{y}_2^2 \end{aligned} \tag{101.21}$$

From Fig. 101.15, we have

$$\begin{aligned} x_2 &= l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2) \\ y_2 &= l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2) \\ v_2^2 &= l_1^2\dot{\theta}_1^2 + l_2^2(\dot{\theta}_1 + \dot{\theta}_2)^2 + 2l_1 l_2 \cos\theta_2(\dot{\theta}_1^2 + \dot{\theta}_1\dot{\theta}_2) \end{aligned} \tag{101.22}$$

The potential energies for the masses m_i, $i = 1, 2$, are given by

$$\begin{aligned} P_1 &= m_1 g\, l_1 \sin\theta_1 \\ P_2 &= m_2 g[l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2)] \end{aligned} \tag{101.23}$$

The next step is to form the Lagrangian,

$$L = \sum_{i=1}^{2} K_i - P_i$$

The dynamic model of the robot is obtained by using Eq. (101.19),

$$\tau_1 = \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\theta}_1}\right] - \frac{\partial L}{\partial \theta_1} \tag{101.24}$$

$$\tau_2 = \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\theta}_2}\right] - \frac{\partial L}{\partial \theta_2} \tag{101.25}$$

where τ_i, $i = 1, 2$, are the joint torques.

The equations for a general n-degrees-of-freedom robot can be derived by following the same procedure and are compactly written in the generalized coordinates q as

$$D(q)\ddot{q} + H(q, \dot{q}) + V\dot{q} + G(q) = \tau \tag{101.26}$$

where $D(q)$ is the $n \times n$ inertia matrix, $H(\cdot)$ is an $n \times 1$ vector describing the centripetal and Coriolis terms, V is the coefficient of friction and $G(q)$ is an $n \times 1$ vector describing gravitational torques. For the above example, $q_1 = \theta_1$ and $q_2 = \theta_2$. Thus,

$$D(\theta) = \begin{bmatrix} l_2^2 m_2 + 2 l_1 l_2 m_2 \cos\theta_2 + l_1^2 (m_1 + m_2) & l_2^2 m_2 + l_1 l_2 m_2 \cos\theta_2 \\ l_2^2 m_2 + l_1 l_2 m_2 \cos\theta_2 & l_2^2 m_2 \end{bmatrix} \tag{101.27}$$

$$H(\theta, \dot{\theta}) = \begin{bmatrix} -m_2 l_1 l_2 \sin\theta_2 \dot{\theta}_2^2 - 2 m_2 l_1 l_2 \sin\theta_2 (\dot{\theta}_1 \dot{\theta}_2) \\ (m_2 l_1 l_2 \sin\theta_2)\dot{\theta}_1^2 \end{bmatrix} \tag{101.28}$$

$$G(\theta) = \begin{bmatrix} m_2 l_2^2 g \cos(\theta_1 + \theta_2) + (m_1 + m_2) l_1 g \cos\theta_1 \\ m_2 l_2 g \cos(\theta_1 + \theta_2) \end{bmatrix} \tag{101.29}$$

where

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad \dot{\theta} = \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} \quad \ddot{\theta} = \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} \tag{101.30}$$

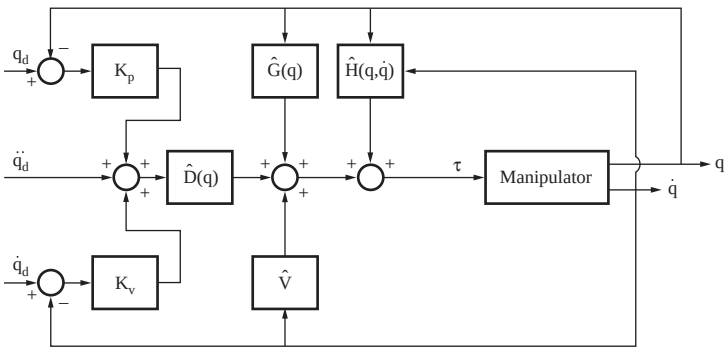


FIGURE 101.16 Computed torque method.

Notice that the inertia matrix $D(\theta)$ is a function of only the position θ. In general, the inertia matrix is symmetric and positive definite and thus invertible. The diagonal elements of this matrix represent the effective inertias at the respective joints, while the off-diagonal elements represent the coupling inertias. For example, the term $m_2l^2_2$ represents the effective inertia at the joint 2, and the term $l^2_2m_2 + l_1l_2m_2 \cos\theta_2$ represents the coupling inertia between joints 1 and 2, i.e., the effect of acceleration of joint 1 on joint 2.

The terms in the matrix $H(\cdot)$ contain all the terms associated with the **centripetal** and **Coriolis forces**. The terms that depend upon the square of the joint velocity are *centripetal forces*. The terms that contain the product of joint velocities are *Coriolis forces*. In our example, the term $-(m_2l_1l_2\sin\theta_2)\,\dot{\theta}^2_2$ represents the centripetal force acting at joint 1 due to the velocity at joint 2. Similarly, the term $(m_2l_1l_2\sin\theta_2)\,\dot{\theta}^2_1$ represents the centripetal force acting at joint 2 due to the velocity at joint 1. The term $-(2m_2l_1l_2\sin\theta_2)\,\dot{\theta}_1\dot{\theta}_2$ is the Coriolis force acting at joint 2 due to the velocities at joints 1 and 2.

The term $G(\theta)$ contains all the terms involving gravitational constant g. Note that these terms depend only on the position of the arm in the gravitational field. If the arm is operating in the gravity-free environment, then these terms become zero. The term $V\cdot\theta$ reflects the frictional forces present in the robot system. In our example, these terms are assumed to be zero. However, in practical robots a substantial amount of friction stiction can be present that if not considered will overestimate the torque available for accelerating the joints.

The above example illustrates that the existence of significant coupling between the joints, if ignored, can cause positioning and tracking errors when the joints are moving simultaneously. However, all these coupling terms become small at low speeds. In this case, independent joint control with appropriate compensations as discussed earlier may be quite adequate. As the operational speeds increase, one needs to take into consideration the full dynamics in the development of control algorithms.

Computed Torque Methods

Several control algorithms that incorporate dynamics have been developed. Many of these are variations of the computed torque method, which is similar to the feedback linearization method used for the control of nonlinear systems [Spong and Vidyasagar, 1989]. In the computed torque method shown in Fig. 101.16, the required input forces or torques are computed as follows:

$$\tau = \hat{D}(q)\left[\ddot{q}_d + K_v(\dot{q}_d - \dot{q}) + K_p(q_d - q)\right] + \hat{H}(q, \dot{q}) + \hat{V}\dot{q} + \hat{G}(q) \qquad (101.31)$$

where K_v and K_p are diagonal matrices with diagonal elements representing velocity and position gains. If this torque is chosen as the input in Eq. (101.26), and assuming that the model is accurate, i.e., $\hat{D}(q) = D(q)$, $\hat{H}(q,\dot{q}) = H(q,\dot{q})$ etc., we get

$$D(q)\left[\ddot{q}_d - \ddot{q} + K_v(\dot{q}_d - \dot{q}) + K_p(q_d - q)\right] = 0 \qquad (101.32)$$

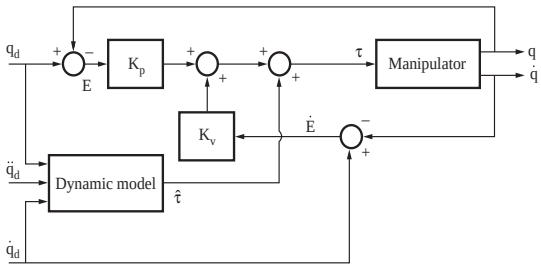


FIGURE 101.17 Dynamic model outside the feedback loop.

Since the inertia matrix, $D(q)$ is nonsingular, we get

$$\ddot{E} + K_v\dot{E} + K_pE = 0 \tag{101.33}$$

which represents a set of decoupled equations where the error $E = q_d - q$. If we select the values of K_v and K_p such that the characteristic roots of Eq. (101.33) have negative real parts, then E approaches zero asymptotically. Effectiveness of this algorithm depends heavily on two factors: (1) the accuracy of the model and (2) the ability to compute the coefficient matrices of the equations of motion in real time.

If the model is not an exact representation of the system, Eq. (101.33) becomes

$$\ddot{E} + K_v\dot{E} + K_pE = R(\ddot{q}, \dot{q}, q) \tag{101.34}$$

where R is the mismatch between the model and the actual dynamics of the robot. This is given by

$$R(\ddot{q}, \dot{q}, q) = \hat{D}^{-1}(q)[(D(q) - \hat{D}(q))\ddot{q} + (H(q, \dot{q}) - \hat{H}(q, \dot{q})) + G(q) - \hat{G}(q)] \tag{101.35}$$

Observe that if the model is an exact match, Eq. (101.34) leads to Eq. (101.33) and the convergence of q to q_d can be guaranteed.

Even if the model is accurate, the ability to compute the dynamics at sample rate (60 to100 Hz is typical) is still a problem. It is estimated that the Stanford manipulator requires 2000 floating-point additions and 1500 multiplications to compute all joint torques. A way to overcome this problem is to use the control scheme where the model is outside the feedback loop shown in Fig. 101.17, [Craig, 1989]. In this case, the desired torques are calculated *a priori* using the model given in Eq. (101.26) as follows:

$$\hat{\tau} = \hat{D}(q_d)\ddot{q}_d + \hat{H}(q_d, \dot{q}_d) + \hat{V}\dot{q}_d + \hat{G}(q_d) \tag{101.36}$$

Then from Fig. 101.17, we get

$$D\ddot{q} + H(q, \dot{q}) + V\dot{q} + G(q) = \hat{\tau} + K_v(\dot{q}_d - \dot{q}) + K_p(q_d - q) \tag{101.37}$$

If the mismatch between the model and the robot is small, then we get

$$D(\ddot{q}_d - \ddot{q}) + K_v(\dot{q}_d - \dot{q}) + K_p(q_d - q) = 0 \tag{101.38}$$

Since the inertia matrix is nonsingular, we can rewrite the above equation as

$$\ddot{E} + D^{-1}K_v\dot{E} + D^{-1}K_pE = 0 \tag{101.39}$$

where $E = q_d - q$ and can be made to go to zero asymptotically by selecting the gains K_v and K_p appropriately. This method has a definite advantage over the earlier method because the model need not be evaluated in real time. However, it does not provide complete decoupling because the inertia matrix is not diagonal. Furthermore, since the gains are continuously modified by the inertia matrix, the response is a function of the configuration and the payload. A way to circumvent this problem is to continuously modify the gains K_v and K_p. This obviously suggests an adaptive control approach.

Adaptive Control

In an attempt to reduce the errors caused by the mismatch of the model with the real system, several adaptive control schemes have been investigated. Model reference adaptive control (MRAC) is one such approach. Dubowsky and DesForges [1979] were the first to use this method for manipulator control. This method is illustrated in Fig. 101.18. They have chosen a linear second-order system with desired ζ and ω_n as a reference model for each joint. Their scheme works as long as the manipulator changes configuration slowly relative to the adaptation rate. Since then several researchers have extended the concepts well developed for linear systems to manipulator control. Two aspects that are central to all these methods are identification of the plant or its parameters and use of this new information to update the control law. An extensive review of recent work in this area is given by Craig [1988] and Hsia [1986]. In spite of many approaches suggested for this problem, no attempt has been made to implement these methods by the robot industry.

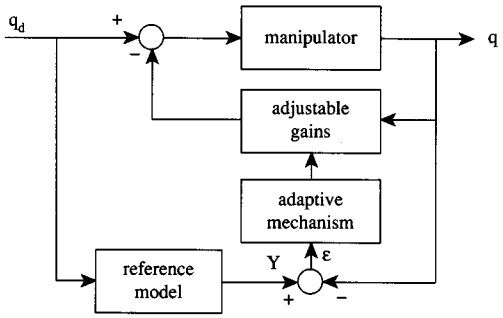


FIGURE 101.18 Model reference adaptive control.

Resolved Motion Control

So far we have discussed the methods to achieve desired joint motion. In practice, the desired motion is specified in terms of hand motions. Resolved motion control methods such as *resolved motion rate control (RMRC)* and *resolved motion acceleration control* have been suggested [papers by Whitney and Luh et al. in Brady et al., 1982]. In these methods, the joint motions are coordinated to achieve coordinated hand motion along any world coordinate axis. Given the relationship between the position and orientation of the hand, $x(t)$, and the joint coordinates $q(t)$ as

$$x(t) = f(q(t)) \tag{101.40}$$

then RMRC transforms the linear and angular velocity of the hand (end effector) to joint velocities using the relationship

$$\dot{q} = J^{-1}(q(t))\dot{x} \tag{101.41}$$

where $J(q(t))$ is the **Jacobian** matrix. Using the above equation, the combination of the joint rates for a given hand motion can be obtained. However, special consideration must be given when the inverse of the Jacobian matrix does not exist. This occurs when the dimension of $x(t)$ and $q(t)$ are not the same (robots with redundant

degrees of freedom) or when a nonredundant robot loses one or more degrees of freedom in its workspace (singular configurations).

Resolved motion acceleration control extends the concepts of RMRC by including desired acceleration of the hand as well. Differentiating Eq. (101.41) twice with respect to time, we get

$$\ddot{x} = J(q)\ddot{q} + \dot{J}(q, \dot{q})\dot{q}(t) \tag{101.42}$$

where $\ddot{x}$ is the acceleration of the hand and $\ddot{q}$ is the joint acceleration. To reduce the position and orientation errors of the hand to zero,

$$\ddot{x}(t) = \ddot{x}_d(t) + K_v[\dot{x}_d(t) - \dot{x}(t)] + K_p[x_d(t) - x(t)] \tag{101.43}$$

By selecting the gains K_p and K_v we can force the error $e(t) = x_d(t) - x(t)$ to zero as before. The desired joint acceleration can be obtained by from Eqs. (101.41) and (101.42), and is given as follows:

$$\ddot{q}(t) = J^{-1}(q)\begin{bmatrix} \ddot{x}_d(t) + K_v\left(\dot{x}_d(t) - \dot{x}(t)\right) \\ + K_p\left(x_d(t) - x(t) - \dot{J}(q, \dot{q})\dot{q}(t)\right) \end{bmatrix} \tag{101.44}$$

Since the inverse of the Jacobian is involved, this method suffers from drawbacks similar to the RMRC method.

Compliant Motion

The position control methods described above are not sufficient when the robot has to react continuously to contact forces at the end effector. Consider, for example, a simple operation of sliding a block of wood on a table along a desired path. Pure position control will not work because any small errors orthogonal to the table may result in the block either losing contact with the surface of the table or forcing the block through the table, which can either damage the table or the end effector. To perform this task, we need to control the position in the plane of the table and control force normal to the table. This is called *compliant motion control* and is required whenever the robot is in contact with its "environment." To perform the above task, for example, a coordinate system called a *compliance frame* or *constraint frame* is defined such that at each instant and along each axis the task can be expressed as a pure position control or pure force control. Suppose we associate a coordinate system with the z-axis normal to the table surface. Then to perform this task, we need to control the position along the x and y directions and force control in the z direction to maintain continuous contact with the table surface. In this case, the position along the z direction is not controlled because one cannot control both position and force in the same direction, just as we cannot control both voltage and current across a resistor. Hence, this framework will provide a natural separation between the axes that need to be position controlled and the axes that need to be force-controlled. This is the idea behind the hybrid position/force control developed by Raibert and Craig [1981]. Another approach for complaint motion control is the impedance control [Hogan, 1985]. Impedance control does not control the end-point position or force directly, rather a desired dynamic relationship between position and force (mechanical impedance). For a good comparison between these two approaches, the reader is referred to Asada and Slotine [1986].

In general, compliance motion control is very important whenever the robot is required to make contact with its "environment." This is true for many assembly tasks. Apart from the active methods of control discussed above, passive methods can be used to introduce the desired compliance. One such passive scheme is the use of a *remote center compliance* (RCC) device developed at Draper Laboratories. The RCC is a purely mechanical device consisting of a spring with six degrees of freedom that is inserted between the wrist and the end effector. By adjusting the stiffness of the springs, various levels of compliance can be obtained. However, passive methods suffer from lack of programmability achieved through active methods. Active control methods, however, require

sensing of contact forces and torques at the end effector. *Joint torque sensors, wrist sensors, fingertip tactile sensors,* and *force pedestals* can be used for this purpose.

Joint torque sensors, as the name implies, are placed at the joints of the manipulator. If **F** represent the vector of forces at the end effector, then the corresponding vector of joint torques is obtained by using $\tau = [J(q)]^T \mathbf{F}$, where $J(q)$ is the Jacobian and q are the generalized joint coordinates. Joint torque sensing has some drawbacks. First, to obtain the endpoint forces **F**, the Jacobian which changes with the configuration has to be inverted in real time. Second, the sensors at the joints not only measure the forces and torques applied at the hand but also those applied at the other points of the manipulator. *Wrist sensors* are better at reducing this uncertainty because they are placed close to the end effector and below the last powered joint of the manipulator. Several wrist sensors are available commercially with necessary electronics to obtain force/torque measurements at high speeds suitable for real-time force control. Another method for providing information about the gripping forces is by mounting *tactile sensors* at the fingertips. However, these may not be suitable in situations where high gripping forces are required. *Force pedestals* are employed when a common platform is used for many tasks. In this case, the platform is instrumented to measure interacting forces and torques.

Flexible Manipulators

The discussion so far assumed that the links of the robot are rigid. These are designed intentionally to minimize the vibrations. Most of the present-day industrial robots fall into this category. These robots, however, cannot handle objects heavier than about 5% their weight. In contrast, lightweight flexible arms consume less energy, achieve faster speeds, and can potentially perform precision assembly tasks. However, it is not possible to move these arms quickly without the onset of structural vibrations due to inadequate structural damping. Efforts have been underway to increase the damping without substantial increase in weight by using composite materials or actively controlling the vibrations, or both.

Defining Terms

Centripetal forces: Forces that are present during the robot motion. They depend upon the square of the joint velocities of the robot and tend to reduce the power available from the actuators.

Compliant motion: Motion of the manipulator (robot) when it is in contact with its "environment," such as writing on a chalkboard or assembling parts.

Coriolis forces: Forces/torques that depend upon the product of joint velocities.

Gravitational torques: Torques that depend upon the position of the robot in the gravitational field.

Independent joint control: A method where each joint is controlled as a single input/single output system. The coupling effects due to motion of other joints are either ignored or treated as disturbances.

Inverse kinematics: A model that maps end-effector positions and orientations to joint variables.

Jacobian of the manipulator: A matrix that maps the joint velocities into end effector velocities.

Tool space: Space of a 6×1 vector representing the positions and orientations of the tool or end effector of the robot.

Related Topics

100.2 Dynamic Response • 101.1 Robot Configuration

References

H. Asada and J. J. E. Slotine, *Robot Analysis and Control*, New York: John Wiley & Sons, 1986.

M. Brady, J.M. Hollerbach, T.L. Johnson, T. Lozano-Perez, and M.T. Mason, *Robot Motion: Planning and Control*, Cambridge, Mass.: The MIT Press, 1982.

J. J. Craig, *Adaptive Control of Mechanical Manipulators*, Reading, Mass.: Addison-Wesley, 1988.

J. J. Craig, *Introduction to Robotics*, Reading, Mass.: Addison-Wesley, 1989.

S. Dubowsky and D.T. DesForges, "The application of model-referenced adaptive control of robotic manipulators," *ASME J. Dyn. Syst. Meas. Control*, 1979.

K. Fu, R. Gonzalez, and C.S.G. Lee, *Robotics: Control, Sensing, Vision, and Intelligence*, New York: McGraw-Hill, 1987.
T.C. Hsia, "Adaptive control of robot manipulators—a review," IEEE Conference on Robotics and Automation, San Francisco, 1986.
N. Hogan, "Impedance control: An approach to Manipulation, Part I, II, and III" *ASME J. Dyn. Sys. Meas. Control*, vol. 107, Mar. 1985.
A. Koivo, *Control of Robotic Manipulators*, New York: John Wiley & Sons, 1989.
J.Y.S. Luh, "Conventional controller design for industrial robots—a tutorial," *IEEE Trans. Syst., Man and Cybern.*, vol. SMC-13, no. 3, June 1983.
R.P. Paul, *Robot Manipulators: Mathematics, Programming and Control*, Cambridge, Mass.: The MIT Press, 1981.
M. Raibert and J. Craig, "Hybrid position/force control of manipulators," *ASME J. Dyn. Syst. Meas. Control*, June 1981.
M. W. Spong and M. Vidyasagar, *Robot Dynamics and Control*, New York: Wiley, 1989.

Further Information

More information about this subject can be obtained by referring to many of the textbooks available on this subject. These are given in the References. Readers who are interested in current research may refer to several journals published by the Institute of Electrical and Electronics Engineers. In particular, *IEEE Transactions on Robotics and Automation*, *IEEE Transactions on Automatic Control*, and *IEEE Transactions on Systems, Man and Cybernetics* along with the conference proceedings published by the respective societies are useful in this regard.

101.3 Applications

Nicholas G. Odrey

An important utilization of robotics has traditionally been in manufacturing operations. By their very design and reprogrammable features, robots have enhanced the capabilities for flexibility in automation. Robot applications initially focused on replacing repetitive, boring, and hazardous manual tasks. Such initial applications required minimal control, programming, or sensory capability and have evolved to applications that use enhanced controller designs and sophisticated sensory capability. The first recorded commercial application of an industrial robot was at the Ford Motor Company in 1961 that used a Unimate robot to unload a die-casting machine. Since then, robots have been used in various manufacturing processes, fabrication, and assembly operations. Current issues relate to the degree of integration with the total manufacturing system and to the degree of autonomy and/or complexity one wishes to implement for a robotic system. In potential applications, it is necessary to determine the degree of sophistication that one wishes to implement coupled with a detailed economic analysis. The focus in this section is to present a practical implementation strategy for robots within a manufacturing environment, to review particular applications, and to discuss issues relevant to enhancing robot applications on the manufacturing shop floor. Such issues include sensors and their integration within an intelligent control system, the development of grippers for enhanced dexterity, and integration topics within a flexible **cellular manufacturing** system.

Justification

Reprogrammable automated devices such as robots provide the flexible automation capability for modern production systems. To evaluate a potential robotic application within a manufacturing environment, both technical and economic issues must be addressed. Typical technical issues include the choice of the number of **degrees of freedom** to perform a task, the level of controller and programming complexity, end effector and sensor choices, and degree of integration within the overall production system. Economic issues have typically been addressed from a traditional point of view, but it is important to note that other criteria should also be evaluated before a final decision is made to implement a robotic system. Such criteria may be both quantitative and qualitative.

Traditional economic approaches analyze investments and costs to compare alternative projects. Three methods are commonly used: (1) payback period method, (2) equivalent uniform annual costs (EUAC) method, and (3) return on investment (ROI). The payback method balances initial investment cost against net annual cash flow during the life of the project to determine the time required to recoup the investment. Many corporations today require relatively short (1- to 3-year) payback periods to justify an investment. In the current environment with the drive toward shortened product life cycles, it is not unusual to see payback period requirements of no greater than 1 year. The payback technique does not consider the time value of money and should be considered only as a first part attempt at justification.

The EUAC and ROI methods consider the time value of money (continuous or discrete compounding) and convert all investments, cash flows, salvage values, and any other revenues and costs into their equivalent uniform annual cash flow over the anticipated life of the project. In the EUAC method, the interest rate is known and set at a minimal acceptable rate of return, whereas the ROI method has the objective to determine the interest rate earned on the investment. Details to such techniques are presented in various engineering economy texts such as those by White et al. [1977] and Thuesen and Fabrycky [1989].

Various more sophisticated approaches have been taken to justify robotic and automated system implementation. Estimates of indirect factors such as taxes, capital gain or losses, variability consideration, and associated expected value analysis along with **decision tree analysis** and Markovian decision analysis [Michel, 1986] are but a few methods to justify such systems. Other factors to be recognized in robotic justification are that robots are reusable from one project to the next and there is a difference in production rates for a robotic implementation over a manual process. A changeover from a manual method to a robotic implementation would have the potential to affect revenues for any project. Many companies have also developed standard investment analysis forms for an economic evaluation of a proposed robot project. These forms are helpful in displaying costs and savings for a project. Groover et al. [1986] presents one such proposed form and gives several references to examples of forms specifically designed for projects devoted to robotics and related automation areas.

The aforementioned techniques are important in performing an economic justification for a proposed robotic installation. Still, in general, there are other issues that should be included in the overall analysis. These issues are of particular importance if one is considering installing a more comprehensive system such as a flexible manufacturing system that may include many robots and automated systems. As noted by Proth and Hillion [1990], these issues give rise to criteria that are both quantitative and qualitative. Quantitative criteria include not only reduced throughput time and work-in-process inventory but also criteria related to increased productivity coupled with fewer resources. Another measurable criterion is the reduction in management and monitoring staff as a result of smaller quantities and automatic monitoring by sensors. Quality improvement can also be measured both quantitatively and qualitatively. Qualitative benefits from quality improvement can include increased customer satisfaction, increased competitiveness, simplified production management, and other factors. It should be noted that any benefits and cost reductions for installation of an automated system are difficult to evaluate and reflect a long-term commitment of the corporation.

Strategic factors should be incorporated in the overall economic justification process, but they are difficult to access and incorporate due to their inherent complexity. Verk [1990] proposes a general framework that attempts to integrate both qualitative and quantitative factors in an economic justification process. The approach taken is being tested at Cincinnati Milacron and the Mazak Corporation.

Implementation Strategies

A logical approach is a prerequisite to robotic implementation within a manufacturing firm. The following steps have been proposed by Groover et al. [1986] to implement a robotic system:

1. Initial familiarization with the technology.
2. Plant survey to identify potential applications.
3. Selection of an application(s).
4. Selection of a robot(s) for the application(s).
5. Detailed economic analysis and capital authorization.
6. Plan and engineer the installation.
7. Installation.

It should be noted that a particular company may have nuances that could modify the above steps. Also of note is that the underlying issue is systems integration and any robotic application should consider total system impact as well as include the equipment, controllers, sensors, software, and other necessary hardware to have a fully functional and integrated system. Another good source of information on robot implementation is the text by Asfahl [1992].

Critical factors for the introduction of robotics technology within a corporation are management support and production personnel acceptance of the technology. Companies such as General Electric have developed checklists to determine the degree of workforce acceptance. Given that the above two factors are met, a plant survey is conducted to determine suitability for automation or robotic implementation. Two general categories of robot applications may be distinguished: (1) a project for a new plant, or (2) placing a robot project in an existing facility. We focus here on the latter category.

General considerations for a robot installation include hazardous, repetitive, or uncomfortable working conditions, difficult handling jobs, or multishift operations. High- and medium-volume production typically has many examples of repetitive operations. It can prove useful to investigate injury (particularly muscular) reports with medical personnel and ergonomics experts to identify potential manual operations that may be alleviated with the aid of robotics or automation. Multishift operations associated with high demand for a product are likely candidates for robot applications. As compared to manual work that typically has a high variable labor cost, a robot substitution would have a high fixed cost which can be distributed over the number of shifts plus a low variable cost. The overall effect of a robot application would then be to reduce the total operating cost.

Once potential robot applications are identified, one typically must determine which application is the best to pursue. Economic and technical criteria must both be considered. Usually, a simple application that is easy to integrate into the overall system is a good initial choice. A fundamental rule is to implement any straightforward application to minimize the risk of failure. The General Electric Company has been successful in choosing robot applications by considering the following technical criteria:

- Operation is simple and repetitive.
- Cycle time for the operation is greater than five seconds.
- Parts can be delivered with the proper POSE (position and orientation).
- Part weight is suitable (typical upper weight limit is 1100 lb).
- No inspection is required for the operation.
- One to two workers can be replaced in a 24-hour period.
- Setups and changeovers are infrequent.

A choice of a robot for a selected application can be a very difficult decision. Vendor information, expert opinion, and various sources such as the *Robotics Product Database* [Flora, 1989] can aid in the selection. Selection needs to consider the appropriate combination of parameters suitable for the application. These parameters or technical features include the degrees of freedom, the type of drive and control system, sensory capability, programming features, accuracy and precision requirements, and load capacity of the selected robot. Various point or weighing schemes can be applied to rate different robot models.

The planning and engineering of a robot installation must address many issues, including the operational methods to be employed, workcell design and its control, the choice or design of end effectors and other fixturing and tooling requirements, and sensory and programming requirements. In addition, one needs to focus on safety considerations for the workcell as well as overall systems integration. Computer-aided design (CAD) is very helpful to study potential **machine interference** and various layout problems as well as estimating various performance parameters. Various commercial CAD software packages exist to analyze such problems. One such example is McAuto's PLACE System. The study at this stage should consider the basic purpose and function of the planned workcell. Consideration needs to be given to analyzing the cycle time that is basic to determining the production rate. An approach developed by Nof and Lechtman [1982], called Robot Time and Motion (RTM), is useful for analyzing the cycle time of robots.

Applications in Manufacturing

Robots have proven to be beneficial in many industrial and nonindustrial environments. Here, we focus on applications within a traditional manufacturing (shop floor) setting and, in particular, on applications which fall into the following three broad categories:

1. Material handling and machine loading/unloading
2. Processing
3. Assembly and inspection

The discussion that follows is not all-inclusive but rather is intended to present (1) an overview of such applications and (2) a few of the more current topics which are impacting the shop floor, particularly as related to flexible manufacturing systems. In the latter case, such issues include developments in **sensor integration**, mobility, sensory interactive grippers/hands, and issues pertaining to intelligent machines and robots. An important reference for many if not all robotic topics is the *International Encylopedia of Robotics* edited by Dorf [1988].

Material Handling and Machine Loading/Unloading

Applications in this category pertain to the grasping and movement of a workpart or item from one location to another. General considerations for such applications pertain to the gripper design, distances moved, robot weight capacity, the POSE, and robot-dependent issues pertaining to the configuration, degrees of freedom, accuracy and precision, the controller, and programming features. POSE information is particularly important if there are no sensors (e.g. vision) to provide such information prior to pick-up. Specialized grippers have been designed for various applications in all three of the listed categories [Engelberger, 1980]. Quick-change wrists enabling the robot to change grippers (or tools in processing applications) during the production cycle have also become more common since their introduction [Vranich, 1984], as have multiple grippers mounted turret-like at the end of a robotic arm. Various factors need to be considered in the selection and design of grippers. One such checklist of factors can be found in Groover et al. [1986]. It should be noted that certain applications may require a high degree of accuracy and precision whereas others do not. Higher requirements result in more sophisticated drive mechanisms and controllers with associated increased costs.

Material handling applications are typically unsophisticated with minimal control requirements. Two- to four-degrees-of-freedom robots may be sufficient in many tasks. More sophisticated operations such as palletizing may require up to six degrees of freedom with stricter control requirements and more programming features. Various criteria that have proved to contribute to the success of material handling and machine load/unload applications can be found in Groover et al. [1986]. In addition, excellent examples and case studies on robotic loading/unloading are given in the text by Asfahl [1992].

Processing

Robotic processing applications are considered here to be those applications in which a robot actually performs work on a part and requires that the end effector is a tool. Examples include spot welding electrodes, arc welding, and spray-painting nozzles. The most common robotic applications in manufacturing processes are listed in Table 101.2 [Odrey, 1992a]. Many more processing applications are possible.

TABLE 101.2 Most Common Robotic Applications in Manufacturing Processes

Spot welding	Grinding
Continuous arc welding	Deburring
Spray coating	Polishing
Drilling	Wire brushing
Routing	Riveting
Waterjet cutting	Laser machining

Spot welding and arc welding represent two major applications of industrial robots. It has been noted that industrial robot usage in welding tasks may be as high as 40% [Ross, 1984]. Spot welding robots have found wide use in automotive assembly lines and have been found to improve weld quality and provide more consistent welds and better repeatability of weld locations. Continuous arc welding is a more difficult application than spot welding. Welding of dissimilar materials, variations in weld joints, dimensional variations from part to part, irregular edges, and gap variations are some of the difficulties encountered in the continuous arc welding processes. Typical arc welding processes include gas metal arc welding (GMAW), shielding metal arc welding (SMAW), i.e., the commonly known "stick" welding, and submerged arc welding (SAW). The most heavily employed

robotic welding process is GMAW in which a current is passed through a consumable electrode and into a base metal, and a shielding gas (typically CO_2, argon, or helium) minimizes contamination during melting and solidification.

In welding, a worker can compensate automatically by varying welding parameters such as travel speed, deposition rate by current adjustment, weave patterns, and multiple welds where required. Duplicating human welding ability and skill requires that industrial robots have sensor capability and complex programming capability. A wide variety of sensors for robotic arc welding are commercially available and are designed to track the welding seam and provide feedback information for the purpose of guiding the welding path.

Two basic categories of sensors exist to provide feedback information: noncontact sensors and contact sensors. Noncontact sensors include arc-sensing systems and machine vision systems. The former, also referred to as a *through-the-arc* system, uses feedback measurements via the arc itself. Specifically, measurements for feedback may be the current (constant-voltage welding) or the voltage (constant-current welding) obtained by programming the robot to perform a weave pattern. The motion results in measurements that are interpreted as vertical and cross-seam position. Adaptive positioning is possible by regulating the arc length (constant-current systems) as irregularities in gaps or edge variations are encountered.

Vision systems track the weld seam, and any deviations from the programmed seam path are detected and fed back to the controller for automatic tracking. Single-pass systems detect variations and make corrections in one welding pass. Double-pass systems first do a high-speed scan of the joint to record in memory deviations from the programmed seam path, with actual welding corrections occurring on the second "arc-on" pass. Single-pass systems give the advantages of reduced cycle time and of being able to compensate for thermal distortions during the welding operation. One recent example of a microcomputer-based single-pass system using a welding torch and laser-ranging sensor on a six-axis robot is given by Nayak and Ray [1990]. Their system, dubbed ARTIST for adaptive, real-time, intelligent, seam tracker, has a two-level integrated control system in which the high level contains rule-based heuristics and model-based reasoning to arrive at real-time decisions, whereas the low level enables tracking of a three-dimensional welding seam.

It should be noted that arc welding, like many manufacturing processes, is not well enough understood physically that one can formulate an exact mathematical model to describe the process. Attempts to optimize welding schedules for any arc welding process have led to expert systems for such processes [Tonkay and Knott, 1989]. Other examples of such work can be found in publications of the *Welding Journal* [e.g., Lucas, 1987; Fellers, 1987].

A robotic arc welding cell provides several advantages over manual welding operations. These advantages include higher productivity as measured by "arc-on" time, elimination of worker fatigue, decreased idle time, and improved safety. It is also important to correct upstream production operations to reduce variations. This is best accomplished during the design and installation phase of a robotic welding cell. During this phase, issues to consider include delivery of materials to the cell, fixtures and welding positioners, methods required for the processes, and any production and inventory control problems related to the efficient utilization and operation of the cell.

Other processing applications for robot use include spray coating and various machining or cutting operations. Spray coating is a major application in the automotive industry where robots have proven suitable in overcoming various hazards such as fumes, mist, nozzle noise, fire, and possible carcinogenic ingredients. The advantages of robotic spray coating are lower energy consumption, improved consistency of finish, and reduced paint quantities used. To install a robotic painting application, one needs to consider certain manual requirements. These include continuous-path control to emulate the motion of a human operator, a hydraulic drive system to minimize electrical spark hazards, and manual lead-through programming with multiple program storage capability [Groover, et al., 1986]. Newer schemes have considered geometric modeling, painting mechanics, and robot dynamics to output an optimal trajectory based on CAD data describing the objects [Suh et al., 1991]. The objective of such work is to plan an optimal robot trajectory that gives uniform coating thickness and minimizes coating time.

Machining operations utilizing robots typically employ end effectors that are powered spindles attached to the robot wrist. A tool is attached to the spindle to perform the processing operation. Examples of tools would be wire brushes or a grinding wheel. It should be noted that such applications are inherently flexible and have

the disadvantage that such operations would be less accurate than a regular machine tool. Finishing operations, such as deburring, have provided excellent opportunities for robotic application. Force control systems have proven particularly useful in regulating the contact force between the tool and the edge of the work to be deburred. One such example for robotic deburring is given by Stepien et al. [1987]. In general, force-torque sensors mounted at the robot wrist have proven extremely useful in many applications in processing and assembly operations. The Lord Corporation and JR3 are two manufacturers of such commercial sensors.

Assembly Applications

Automated assembly has become a major application for robotics. Assembly applications consider two basic categories: parts mating and parts joining. Parts mating refers to peg-in-hole or hole-on-peg operations, whereas joining operations are concerned not only with mating but also a fastening procedure for the parts. Typical fastening procedures could include powered screwdrivers with self-tapping screws, glues, or similar adhesives.

In parts-mating applications, remote center compliance (RCC) devices have proven to be an excellent solution. In general, compliance is necessary for avoiding or minimizing impact forces, for correcting positioning error, and for allowing relaxation of part tolerances. In choosing an RCC device, the following parameters need to be determined prior to an application:

- Remote center distance (center of compliance). This is the point about which the active forces are at a minimum. The distance is chosen by considering the length of the part and the gripper.
- Axial force capacity. Maximum designed axial force to function properly.
- Compressive stiffness. Should be high enough to withstand any press fitting requirements.
- Lateral stiffness. Refers to force required to deflect RCC perpendicular to direction of insertion.
- Angular stiffness. Relates to forces that rotate the part about the compliant center (also called the cocking stiffness).
- Torsional stiffness. Relates to moments required to rotate a part about the axis of insertion.

Other parameters also include the maximum allowable lateral and angular errors as determined by the size of the part and by its design. These errors must be large enough to compensate for errors due to parts, robots, and fixturing. Passive and instrumented (IRCC) devices have been developed for assembly applications. One such device that combines a passive compliance with active control is described by Xu and Paul [1992]. In addition, the SCARA (Selective Compliance Articulated Robot for Assembly) class of robots is stiff vertically but relatively compliant laterally.

Many opportunities exist for flexible assembly systems. Many of the issues for such systems have been addressed by Soni [1991]. The reader is also referred to the *Design for Robotic Assembly Handbook* [Boothroyd and Dewhurst, 1985] for quantitative methods to evaluate a product's ease of assembly by robots. Carter [1990] presents a method for determining robot assembly task time as derived from tests and industrial experience. Carter also addresses the relationship between product design and robotic assembly cycle time. Some of the current trends in automated assembly include coordinating multiple robots to increase the flexibility and reliability of an assembly cell [Coupes et al., 1989; Zheng and Sias, 1986], interaction with CAD databases to automatically generate assembly plans [Wolter, 1989; Nnaji, 1989] and the application of sensors to automatic assembly systems [Cook, 1991]. Meijer and Jonker [1991] consider an architecture for an intelligent assembly cell and its subsequent implementation. An article by Jarneteg [1990] considers the strategies necessary for developing adaptive assembly systems.

Inspection

Inspection involves checking of parts, products, and assemblies as a verification of conformation to the specification of the engineering design. With the emphasis on product quality, there is a growing emphasis for 100% inspection. Machine vision systems, robot-manipulated active sensing for inspection, and automatic test equipment are being integrated into total inspection systems. Robot application of vision systems include part location, part identification, and bin picking. Machine vision systems for inspection typically perform tasks which include dimensional accuracy checks, flaw detection, and correctness and completeness of an assembled product. Current vision inspection systems are predominantly two-dimensional systems capable of extracting

feature information, analyzing such information, and comparing to known patterns previously trained into the system. As documented by Nurre and Hall [1989], various techniques for three-dimensional measurements have also been developed by many researchers. Primary factors to be considered in the design or application of a vision system include the resolution and field of view of the camera, the type of camera, lighting requirements, and the required throughput of the vision system.

Machine vision application can be considered to have three levels of difficulty, namely, that (1) the object can be controlled in both appearance and position, (2) it can be controlled in either appearance or position, or (3) neither can be controlled. The ability to control both position and appearance requires advanced, potentially three-dimensional vision capabilities. The objective in an industrial setting is to lower the level of difficulty involved. It should be noted that inspection is but one category of robotic applications of machine vision. Two other broad categories are identification and visual servoing and navigation. In the latter case, the purpose of the vision system is to direct the motion of the robot based on visual input. The reader is directed to Groover et al. [1986] for further details.

Emerging Issues

Robotics, by definition, is a highly multidisciplinary field. Applications are broad, and even those applications focused on the manufacturing shop floor are too numerous to cover in full here. The reader is referred to the various journals published by the IEEE and other societies and publishers, a few of which have been listed in the references. Still, it is worthwhile to note a few issues relevant to manufacturing shop-floor applications that could have an impact over the next decade. These issues include gripper development, mobility, and intelligent robots. The objective of this work is the overall integration of a flexible manufacturing system.

In a manufacturing process or assembly operation, the actions required of a gripper will vary with the task. Much work has been done in developing multifigured hands such as the Utah-MIT hand, the Salisbury hand, and others with an increasing interest of adding tactile sensory input for dexterous manipulation [Allen et al., 1989]. As noted by Allen and his colleagues, robotic systems need to process multiple source data and be easily programmable for grasping and manipulation tasks. One study focused on capturing a machinist's skill in working with parts and tools and codifying this knowledge in a grip taxonomy has been done by Cutkosky and Wright [1986]. Their study suggests some general principles for the design, construction, and control of hands in a manufacturing (particularly machining) environment. The reader is also referred to the work of Feddema and Ahmad [1986] for the development of an algorithm for a static robot grasp for automated assembly and the work of Cutkosky [1991] on robotic grasping and manipulation. This latter work considers dynamic contact and the application of dynamic tactile sensors in manipulation tasks. An application to identify and locate circuit board fixtures within a robotic workcell that integrates a vision system with a tactile probe is given by DeMeter and Deisenroth [1987].

Automated guided vehicles (AGVs) currently dominate the movement of parts through a flexible manufacturing system (FMS). AGVs typically restrict the path to predetermined routes and subsequently decrease the "flexibility" of the system. Work is being done on mobile robots to address this issue. Research by Arkin and Murphy [1990] focuses on intelligent mobility within a manufacturing environment. The reader is also referred to the research of Wiens and Black [1992] who address a mobile robot system within a manufacturing cell as a means to increase the flexibility, capability, and capacity of a robot-based manufacturing cell.

The issues involved with intelligent robots have been surveyed by Nitzan [1985], where he notes that future proliferation of robotic applications will depend strongly on machine (robotic) intelligence. Such applications will lead to a greater diversity of applications and will not be just manufacturing oriented. The reader is also referred to work on intelligent machines by Weisbin [1986]. It should be noted that particular interest has been directed toward integration of multiple sensors as a means to enhance robot intelligence [Luo and Lin, 1989; Pin et al., 1991]. The text by Klafter et al. [1989] categorizes the major sensory needs for robotic tasks and gives valuable insights to current and future robotics applications. **Intelligent control** systems, particularly hierarchical control systems, are being developed by many organizations and research institutes [Odrey, 1992b]. Such systems are expected to have an impact both at the shop-floor level and the management levels of production facilities well into the next century.

Robotic Tools

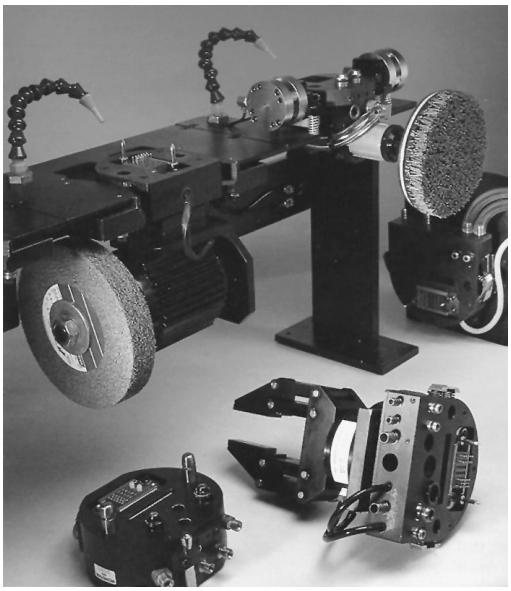
Robotics and Automation Corporation, Minneapolis, Minnesota, manufactures equipment for robotic systems, in particular a variety of tools known as "end effectors", devices attached to the end of a robot arm for picking up, grasping, manipulating, and transferring objects.

Among the company's newer products is the Automatic Robotics Tool-change System (ARTS), a system designed to meet the growing demand for multiple task work cells for welding and plasma spray functions that require grinding and finishing; deburring, deflashing, routing, hole drilling or parts replacement; and multiple tool disk operations.

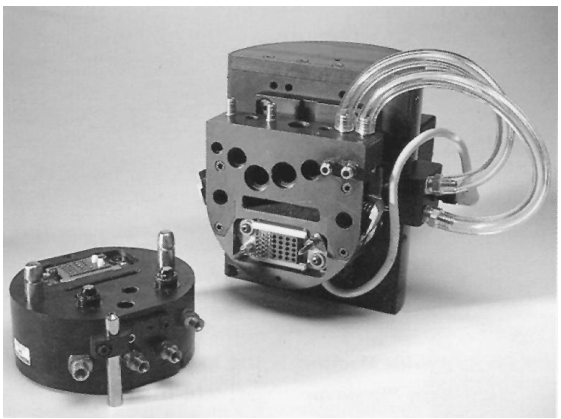
The ARTS systems were designed to work with the company's CFD (Constant/controlled Force Device) product line, a series of end effectors and bench mounted devices for controlling the constant pressure of abrasive tools used to deburr, grind, polish, and finish products fabricated by welding, casting, molding, forging, or machining.

Robotics and Automation Corporation's CFD line includes three end-of-arm devices and two bench-mounted devices. They do not require that the robot apply and control the force, only that it move along a normal programmed path over the work piece; the CFD applies and maintains the required processing pressure.

When the surface to be finished is very rough and course, several different grades of finishing media may be needed, as well as different speeds and power as the surface finish is transformed. To accommodate this multistep process within a single work cell, and with a single robot, Robotics and Automation Corporation developed the automated tool-change system.



The tool rack of the Automatic Robotics Tool-Change System includes a two-finger gripper; a grinder, a coated abrasive brush, and a welding torch. (Photo courtesy of National Aeronautics and Space Administration.)



The quick disconnect system allows changing tools with hydraulic, pneumatic, or electric power. (Photo courtesy of National Aeronautics and Space Administration.)

The ARTS-I is being used in industrial applications with six tool positions ranging from coarse sanding disks and abrasive wheels to cloth polishing wheels with motors of various horsepower. The ARTS-II allows a robot to exchange a welding torch for a CFD end effector to finish a welded assembly with a welding robot; using a second tool-changer (ARTS-I) enables finishing the surface conditioning process. (Courtesy of National Aeronautics and Space Administration.)

Defining Terms

Cellular manufacturing: Grouping of parts by design and/or processing similarities such that the group (family) is manufactured on a subset of machines which constitute a cell necessary for the group's production.

Decision tree analysis: Decomposing a problem into alternatives represented by branches where nodes (branch intersections) represent a decision point or chance event having probabilistic outcome. Analysis consists of calculating expected values associated with the chain of events leading to the various outcomes.

Degrees of freedom: The total number of individual motions typically associated with a machine tool or robot.

Intelligent control: A sensory-interactive control structure incorporating cognitive characteristics that can include artificial intelligence techniques and contain knowledge-based constructs to emulate learning behavior with an overall capacity for performance and/or parameter adaptation.

Machine interference: The idle time experienced by any one machine in a multiple-machine system that is being serviced by an operator (or robot) and is typically measured as a percentage of the total idle time of all the machines in the system to the operator (or robot) cycle time.

Sensor fusion: Combining of multiple sources of sensory information into one representational format.

Sensor integration: The synergistic use of multiple sources of sensory information to assist in the accomplishment of a task.

Related Topic

112.1 Introduction

References

P.K. Allen, P. Michelman, and K.S. Roberts, "An integrated system for dextrous manipulation," IEEE International Conference on Robotics and Automation, 1989, pp. 612–616.

R.C. Arkin and R.R. Murphy, "Autonomous navigation in a manufacturing environment," *IEEE Trans. Robotics Autom.,* vol. 6, no. 4, pp. 445–454, 1990.

C.R. Asfahl, *Robots and Manufacturing Automation*, New York: Wiley, 1992.

G. Boothroyd and P. Dewhurst, "Design for Robotic Assembly," Department of Industrial and Manufacturing Engineering, University of Rhode Island, Kingston, 1985.

P.W. Carter, "Estimating cycle time in design for robotic assembly," *J. Manu. Syst.,* vol. 9, no. 1, pp. 1–12, 1990.

J.W. Cook, "Applying sensors to automatic assembly systems," *IEEE Trans. Ind. Appl.,* vol. 27, no. 2, pp. 282–285, 1991.

D. Coupes, A. Delchambre, and P. Gaspart, "The supervision and management of a two robots flexible assembly cell," Proceedings of IEEE Conference on Robotics and Automation, 1989, pp. 540–550.

M.R. Cutkosky, "Robotic grasping and manipulation," *Proceedings of NSF Design and Manufacturing Systems Conference*, Dearborn, Mich.: Society of Manufacturing Engineers, 1991, pp. 423–430.

M.R. Cutkosky and P.K. Wright, "Modeling manufacturing grips and correlations with the design of robotic hands," IEEE International Conference on Robotics and Automation, San Francisco, Calif., April 7–10, 1986, pp. 1533–1539.

E.C. DeMeter and M.P. Deisenroth, "The integration of visual and tactile sensing for the definition of regions within a robot workcell," Robots 11/17th ISIR, Chicago, Il., April 26–30, 1987, pp. 10-51 to 10-61.

R.C. Dorf, Ed., *International Encyclopedia of Robotics*, vols. 1–3, New York: Wiley, 1988.

J.F. Engelberger, "Robotics in practice," AMA COM: A Division of American Management Associations, 1980.

J.T. Feddema and S. Ahmad, "Determining a static robot grasp for automated assembly," IEEE International Conference on Robotics and Automation, San Francisco, Calif., April 7–10, 1986, pp. 918–924.

K.G. Fellers, "A PC approach to welding variables," *Weld. J.,* vol. 66, pp. 31–40, 1987.

P.C. Flora, Ed., *Robotics Product Database*, 6th ed., Orlando, Fla.: TecSpec, 1989.

M.P. Groover, M. Weiss, R.N. Nagel, and N.G. Odrey, *Industrial Robotics: Technology, Programming, and Applications,* New York: McGraw-Hill, 1986.

B.G. Jarneteg, "FAS control strategies for adaptive assembly systems," 21st CIRP International Seminar on Manufacturing Systems, Stockholm, Sweden, 1990.

R.D. Klafter, T.A. Chmielewski, and M. Negin, *Robotic Engineering: An Integrated Approach*, Englewood Cliffs, N.J.: Prentice-Hall, 1989.

W. Lucas, "Microcomputer systems, software and expert systems for welding engineering," *Weld. J.,* vol. 66, pp. 19–30, 1987.

R.C. Luo and M.-H. Lin, "Intelligent robot multi-sensor data fusion for flexible manufacturing systems," Proceedings of NSF 15th Conference on Production Research and Technology, University of California-Berkeley, Jan. 9–13, 1989, pp. 73–85.

B.R. Meijer and P.P. Jonker, "The architecture and philosophy of the DIAC (Delft Intelligent Assembly Cell)," IEEE Conference on Robotics and Automation, Sacramento, Calif., 1991, pp. 2218–2223.

M. Michel, "Justification models for flexible manufacturing," *Robots' 10 Conference Proceedings*, Dearborn, Mich.: Society of Manufacturing Engineers, 1986, pp. 2-55 to 2-81.

N. Nayak and A. Ray, "An integrated system for intelligent seam tracking in robotic welding: part 1—conceptual and analytical development; part 2—design and implementation," IEEE International Conference on Robotics and Automation, 1990.

D. Nitzan, "Development of intelligent robots: achievements and issues," *IEEE J. Robotics Autom.* vol. RA-1, no. 1, pp. 3–13, 1985.

B.O. Nnaji, "RALPH: An automatic robot assembly language programmer: an overview," Proceedings of Robots 13 Conference, Gaithersburg, Md., May 7–11, 1989, pp. 16-41 to 16-63.

S.Y. Nof and H. Lechtman, "The RTM method of analyzing robot work," *Ind. Eng.,* April 1982, pp. 38–48.

J.H. Nurre and E.L. Hall, "Three dimensional vision for automated inspection," Proceedings of Robots 13 Conference, Gaithersburg, Md., May 7–11, 1989, pp. 16-1 to 16-11.

N.G. Odrey, "Robotics and automation," *Maynard's Industrial Engineering Handbook*, 4th ed., W.K. Hodson, Ed., New York: McGraw-Hill, 1992a.

N.G. Odrey, "Control systems," *1992 McGraw-Hill Yearbook of Science and Technology,* New York: McGraw-Hill, 1992b, pp. 87–90.

F.G. Pin et al., "Robotic learning from distributed sensory sources," *IEEE Trans. Syst. Man and Cybern.,* vol. 21, no. 5, pp. 1216–1223, 1991.

J.M. Proth and H.P. Hillion, *Mathematical Tools in Production Management,* New York: Plenum Press, 1990.

B. Ross, "Machines that can see: here comes a new generation," *Bus. Week.,* January 1984, p. 118.

A.H. Soni, "Flexible assembly systems: Opportunities and challenges," Proceedings of the 1991 NSF Design and Manufacturing Systems Conference, University of Texas at Austin, Jan. 9–11, 1991, pp. 367–373.

T.M. Stepien, L.M. Sweet, M.C. Good, and M. Tomizuka, "Control of tool/workpiece contact force with application of robotic deburring," *IEEE J. Robotics Autom.*, vol. RA-3, no. 1, pp. 7–18, 1987.

S.-H. Suh, I.-K. Woo, and S.-K. Noh, "Automatic trajectory planning system (ATPS) for spray painting robots," *J. Manu. Syst.*, vol. 10, no. 5, pp. 396–406, 1991.

G.J. Thuesen and W.J. Fabrycky, *Engineering Economy,* 7th ed., Englewood Cliffs, N.J.: Prentice-Hall, 1989.

G.L. Tonkay and K. Knott, "Intelligent process specification for robotic arc welding," Proceedings of World Conference on Robotics Research: The Next Five Years and Beyond. Robotics International of the Society of Manufacturing Engineers, Gaithersburg, Md., May 11–17, 1989.

S. Verk, "Strategic optimization cycle as a competitive tool for economic justification of advanced manufacturing systems," *J. Manu. Syst.,* vol. 9, no. 3, pp. 194–205, 1990.

J.M. Vranich, "Quick change system for robots," SME Paper MS84-418, Conference on Robotics Research—The Next Five Years and Beyond, Lehigh University, Bethlehem, 1984.

C.R. Weisbin, "CESAR research in intelligent machines," SME Paper MS586-772, Robotics Research Conference, Scottsdale, Ariz., Aug. 18-21, 1986.

J.A. White, M.H. Agee, and K.E. Case, *Principles of Engineering Economic Analysis*, New York: John Wiley & Sons, 1977.

G.J. Wiens and J.T. Black, "Design for mobility within a manufacturing cell," Proceedings of the NSF Design and Manufacturing Systems Conference, Georgia Institute of Technology, Jan. 8–10, 1992, pp. 1147–1150.

J.D. Wolter, "On the automatic generation of assembly plans," Proceedings of the 1989 IEEE Conference on Robotics and Automation, 1989, pp. 62–68.
Y. Xu and R.P. Paul, "Robotic instrumented compliant wrist," *ASME J. Eng. for Ind.*, vol. 114, pp. 120–123, 1992.
Y.F. Zheng and F.R. Sias, Jr., "Two robot arms in assembly," IEEE Conference on Robotics and Automation, San Francisco, Calif., April 7–10, 1986, pp. 1230–1235.

Further Information

Various journals publish on topics pertaining to robots. Sources include the bimonthly *IEEE Journal of Robotics and Automation,* the quarterly journal *Robotics and Computer-Integrated Manufacturing* (published by Pergamon Press), *Robotics* (published by Cambridge University Press since 1983), and the *Journal of Robotic Systems* (published by Wiley).

IEEE has sponsored since 1984 the annual "International Conference on Robotics and Automation." IEEE conference proceedings and journals are available from the IEEE Service Center, Piscataway, N.J.

The Society of Manufacturing Engineers (SME) is another source for robot publications that are concerned with both research issues and applications. Robots 1 through 13 (1989) conference proceedings are available as well as the Robot Research conference proceedings (three to date) of Robotics International (RI) of SME. A directory of robot research laboratories is also available. Contact SME, Dearborn, Mich.

The three-volume *International Encyclopedia of Robotics: Applications and Automation* (R.C. Dorf, ed.), published by Wiley (1988), brings together the various interrelated fields constituting robotics and provides a comprehensive reference.

Spitzer, C.R., Martinec, D.A., Leondes, C.T., Rana, A.H., Check, W. "Aerospace Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

102

Aerospace Systems

Cary R. Spitzer
AvioniCon Inc.

Daniel A. Martinec
Aeronautical Radio, Inc.

Cornelius T. Leondes
University of California, San Diego

Abdul Hamid Rana
GE LogistiCom

William Check
GE Spacenet

102.1 Avionics Systems
A Modern Example System • Data Buses • Displays • Power • Software in Avionics • CNS/ATM • Navigation Equipment • Emphasis on Communications • Impact of "Free Flight" • Avionics in the Cabin • Avionics Standards

102.2 Communications Satellite Systems: Applications
Satellite Launch • Spacecraft and Systems • Earth Stations • VSAT Communication System • Video • Audio • Second-Generation Systems

102.1 Avionics Systems

Cary R. Spitzer, Daniel A. Martinec, and Cornelius T. Leondes

Avionics (aviation electronics) systems perform many functions: (1) for both military and civil aircraft, avionics are used for flight controls, guidance, navigation, communications, and surveillance; and (2) for military aircraft, avionics also may be used for electronic warfare, reconnaissance, fire control, and weapons guidance and control. These functions are achieved by the application of the principles presented in other chapters of this handbook, e.g., signal processing, electromagnetic, communications, etc. The reader is directed to these chapters for additional information on these topics. This section focuses on the system concepts and issues unique to avionics that provide the traditional functions listed in (1) above.

Development of an avionics system follows the traditional systems engineering flow from definition and analysis of the requirements and constraints at increasing level of detail, through detailed design, construction, **validation,** installation, and maintenance. Like some of the other aerospace electronic systems, avionics operate in real time and perform mission- and life-critical functions. These two aspects combine to make avionics system design and **verification** especially challenging.

Although avionics systems perform many functions, there are three elements common to most systems: data buses, displays, and power. Data buses are the signal interfaces that lead to the high degree of integration found today in many modern avionics systems. Displays are the primary form of crew interface with the aircraft and, in an indirect sense, through the display of synoptic information also aid in the integration of systems. Power, of course, is the life blood of all electronics.

The generic processes in a typical avionics system are signal detection and preprocessing, signal fusion, computation, control/display information generation and transmission, and feedback of the response to the control/display information. (Of course, not every system will perform all of these functions.)

A Modern Example System

The B-777 Airplane Information Management System (AIMS) is the first civil transport aircraft application of the integrated, modular avionics concept, similar to that being used in the U.S. Air Force F-22. Figure 102.1 shows the AIMS cabinet with eight modules installed and three spaces for additional modules to be added as the AIMS functions are expanded. Figure 102.2 shows the AIMS architecture.

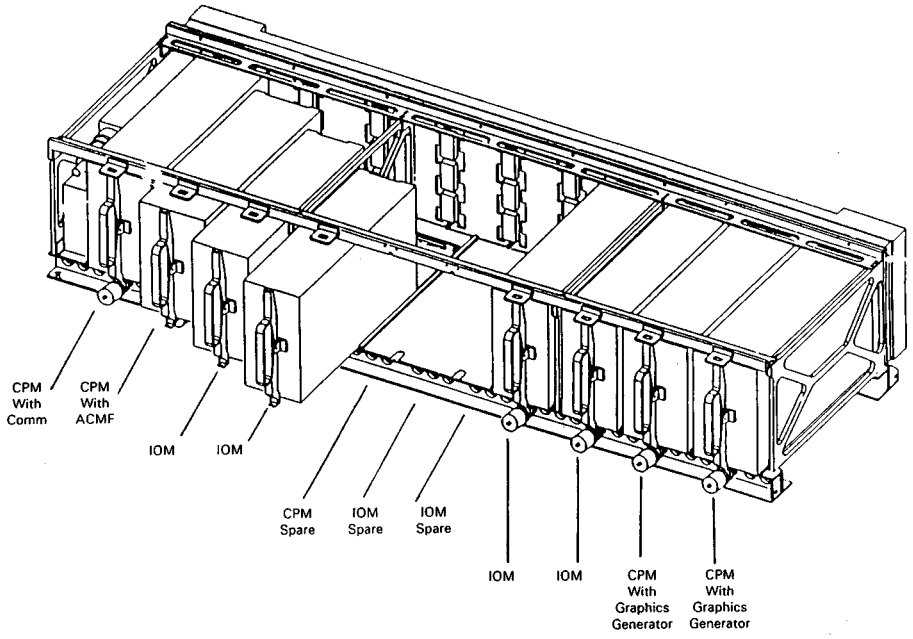


FIGURE 102.1 Cabinet assembly outline and installation (typical installation). (Courtesy of Honeywell, Inc.)

AIMS functions performed in both cabinets include flight management, electronic flight instrument system (EFIS) and engine indicating and crew alerting system (EICAS) displays management, central maintenance, airplane condition monitoring, communications management, data conversion and gateway (ARINC 429 and ARINC 629), and engine data interface. AIMS does not control the engines nor flight controls, nor operate any internal or external voice or data link communications hardware but does select the data link path as part of the communications management function. Subsequent generations of AIMS may include some of these latter functions.

In each cabinet the line replaceable modules (LRMs) are interconnected by dual ARINC 659 backplane data buses. The cabinets are connected to the quadraplex (not shown) or triplex redundant ARINC 629 system and fly-by-wire data buses and are also connected via the system buses to the three multifunction control display units (MCDU) used by flight crew and maintenance personnel to interact with AIMS. The cabinets transmit merged and processed data over quadruple redundant custom designed 100 Mhz buses to the EFIS and EICAS displays.

In the AIMS the high degree of function integration requires levels of system availability and integrity not found in traditional distributed, federated architectures. These extraordinary levels of availability and integrity are achieved by the extensive use of **fault-tolerant** hardware and software maintenance diagnostics and promise to reduce the chronic problem of unconfirmed removals and low mean time between unscheduled removals (MTBUR).

Figure 102.3 is a top-level view of the U.S. Air Force F-22 Advanced Tactical Fighter avionics. Like many other aircraft, the F-22 architecture is hybrid, part federated and part integrated. The left side of the figure is the highly integrated portion, dominated by the two Common Integrated Processors (CIPs) that process, fuse, and distribute signals received from the various sensors on the far left. The keys to this portion of the architecture are the Processor Interconnect (PI) buses within the CIPs and the High Speed Data Buses (HSDBs). (There are provisions for a third CIP as the F-22 avionics grow in capability.) The right side of the figure shows the federated systems including the Inertial Reference, Stores Management, Integrated Flight and Propulsion Control, and Vehicle Management systems and the interface of the latter two to the Integrated Vehicle System Control. The keys to this portion of the architecture are the triple or quadruple redundant AS 15531 (formerly MIL-STD-1553) command/response two-way data buses.

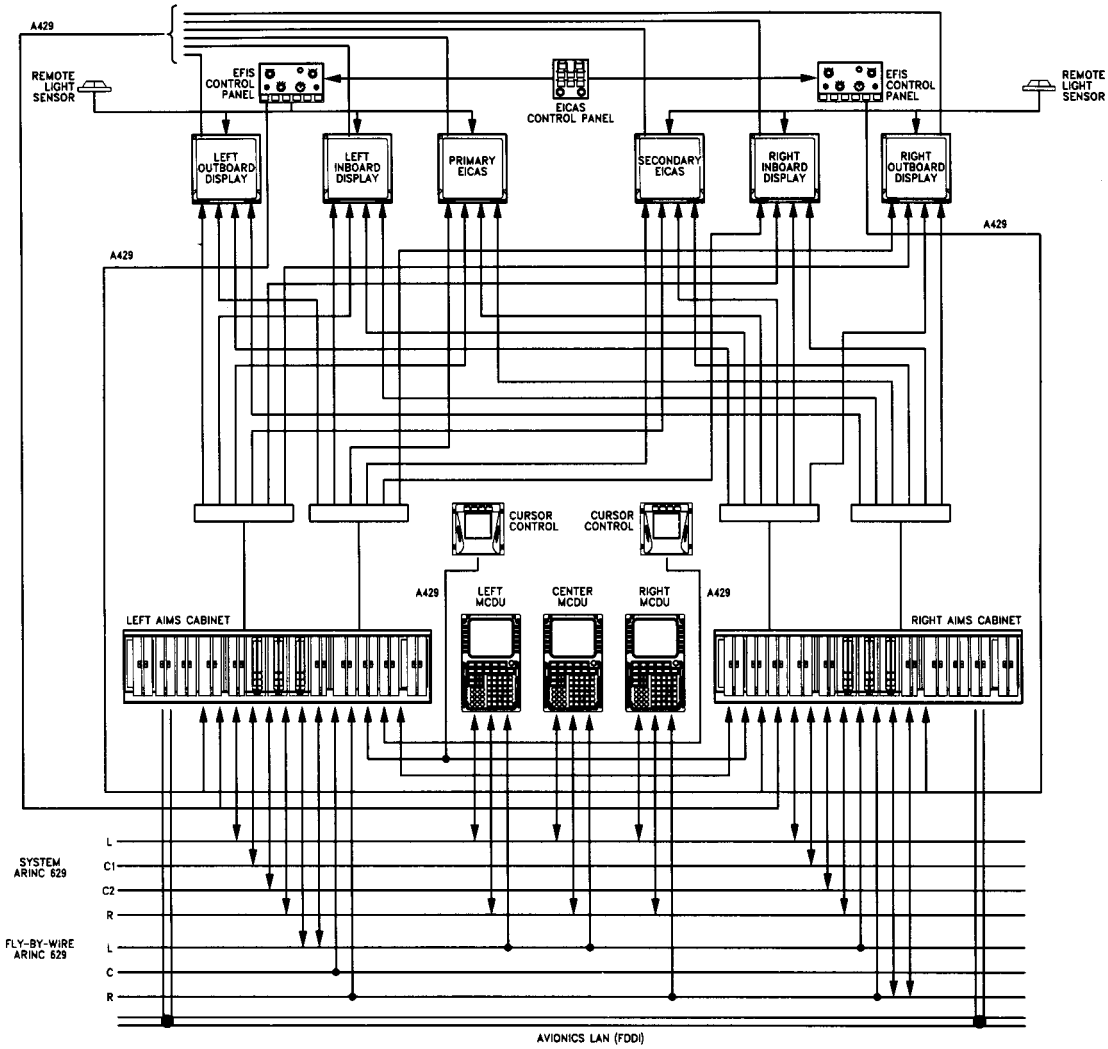


FIGURE 102.2 Architecture for AIMSbaseline configuration. (Courtesy of Honeywell, Inc.)

Data Buses

As noted earlier, data buses are the key to the emerging integrated avionics architectures. Table 102.1 summarizes the major features of the most commonly used system buses. MIL-STD-1553 and ARINC 429 were the first data buses to be used for general aircraft data communications. These are used today widely in military and civil avionics, respectively, and have demonstrated the significant potential of data buses. The others listed in the table build on their success.

Displays

All modern avionics systems use electronic displays, either CRTs or flat-panel LCDs that offer exceptional flexibility in display format and significantly higher reliability than electromechanical displays. Because of the very bright ambient sunlight at flight altitudes the principal challenge for an electronic display is adequate brightness. CRTs achieve the required brightness through the use of a shadow mask design coupled with narrow bandpass optical filters. Flat-panel LCDs also use narrow bandpass optical filters and a bright backlight to achieve the necessary brightness.

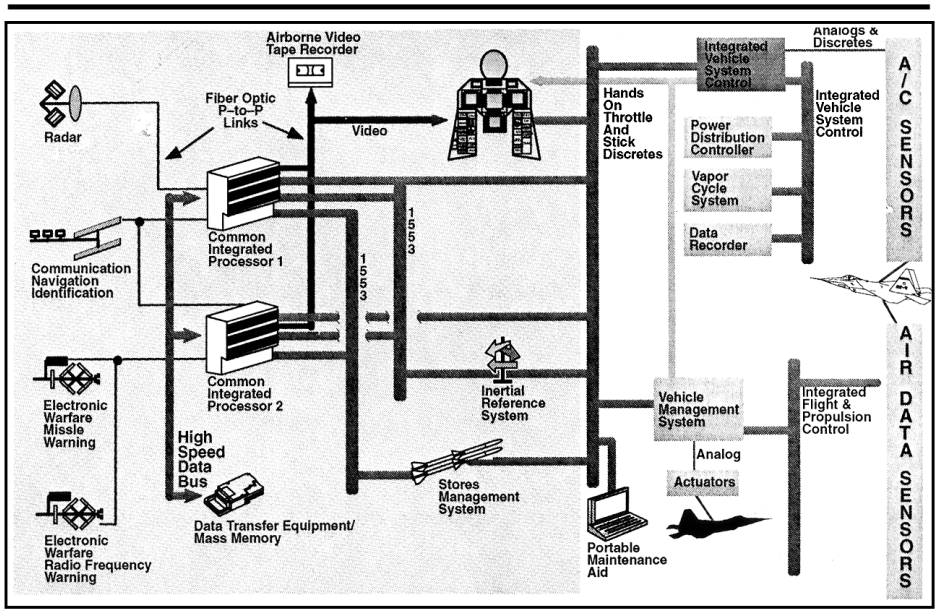


FIGURE 102.3 F-22 EMD Architecture.

TABLE 102.1 Characteristics of Common Avionics Buses

Bus Name	Word Length	Bit Rate	Transmission Media
MIL-STD-1553	20 bits	1 Mb/s	Wire
DOD-STD-1773	20 bits	TBS	Fiber optic
High-speed data bus	32 bits	50 Mb/s	Wire or fiber optic
ARINC 429	32 bits	14.5/100 kb/s	Wire
ARINC 629	20 bits	2 Mb/s	Wire or fiber optic
ARINC 659	32 bits	100 MB/s	Wire

Because of the intrinsic flexibility of electronic displays, a major issue is the design of display formats. Care must be taken not to place too much information in the display and to ensure that the information is comprehendible in high workload (aircraft emergency or combat) situations.

Power

Aircraft power is generally of two types: 28 vdc, and 115 vac, 400 Hz. Some 270 vdc is also used on military aircraft. Aircraft power is of poor quality when compared to power for most other electronics. Under normal conditions, there can be transients of up to 100% of the supply voltage and power interruptions of up to 1 second. This poor quality places severe design requirements on the avionics power supply, especially where the avionics are performing a full-time, flight-critical function. Back-up power sources include ram air turbines and batteries, although batteries require very rigorous maintenance practices to guarantee long-term reliable performance.

Software in Avionics

Most avionics currently being delivered are microprocessor controlled and are software intensive. The "power" achieved from software programs hosted on a sophisticated processor results in very complex avionics with many functions and a wide variety of options. The combination of sophistication and flexibility has resulted

Weather Information Systems

Weather is a critical factor in aircraft operations. It is the largest single contributor to flight delays and a major cause of aircraft accidents.

A study conducted for NASA by Ohio State University reported that the principal difficulties in making proper flight decisions are the timeliness and clarity of weather data dissemination.

To advance the technology of in-flight weather reporting, Langley Research Center developed in the early 1990's a cockpit weather information system known as CWIN (Cockpit Weather Information). The system draws on several commercial data sensors to create radar maps of storms and lightning, together with reports of surface observations.

Shown above is a CWIN display in the simulation cockpit of Langley's Transport Systems Research Vehicle, a modified jetliner used to test advanced technologies. The CWIN display is the lower right screen among the four center panel screens. By pushing a button, the pilot may select from a menu of several displays, such as a ceiling and visibility map, radar storm map, or lightning strike map. (Courtesy of National Aeronautics and Space Administration.)

in lengthy procedures for validation and certification. The **brickwalling** of software modules in a system during the initial development process to ensure isolation between critical and noncritical modules has been helpful in easing the certification process.

There are no standard software programs or standard software certification procedures. RTCA has prepared Document DO-178 to provide guidance (as opposed to strict rules) regarding development and certification of avionics civil software. The techniques for developing, categorizing, and documenting avionics civil software in DO-178 are widely used.

For military avionics software, the principal document is DOD-STD-498. This standard defines a set of activities and documentation suitable for the development of both weapon systems and automated information systems. Many software languages have been used in the past in avionics applications; however, today there is a strong trend for both military and avionics civil software to use Ada wherever reasonably possible.

The evolving definition of a standards for Applications Exchange (APEX) software promises to provide a common software platform whereby the specialized requirements of varying hardware (processor) requirements are minimized. APEX software is a hardware interface that provides a common link with the functional software within an avionics system. The ultimate benefit is the development of software independent of the hardware platform and the ability to reuse software in systems with advanced hardware while maintaining most, if not all, of the original software design.

CNS/ATM

The last decade of this century has seen much attention focused on Communication/Navigation/Surveillance for Air Traffic Management (CNS/ATM), a satellite-based concept developed by the Future Air Navigation System (FANS) Committees of the International Civil Aviation Organization (ICAO), a special agency of the United Nations. Many studies have predicted enormous economic rewards of CNS/ATM for both aircraft operators and air traffic services providers.

The new CNS/ATM system should provide for:

- Global communications, navigation, and surveillance coverage at all altitudes and embrace remote, offshore, and oceanic areas.
- Digital data exchange between air-ground systems (voice backup).
- Navigation/approach service for runways and other landing areas which need not be equipped with precision landing aids.

Navigation Equipment

A large portion of the avionics on an aircraft are dedicated to navigation. The following types of navigation and related sensors are commonly found on aircraft:

- Flight control computer (FCC)
- Flight management computer (FMC)
- Inertial navigation system (INS)
- Attitude heading and reference system (AHRS)
- Air data computer (ADC)
- Low range radio altimeter (LRRA)
- Radar
- Distance Measuring Equipment (DME)
- Instrument Landing System (ILS)
- Microwave Landing System (MLS)
- VHF OmniRange (VOR) Receiver
- Global Navigation Satellite System (GNSS)

Emphasis on Communications

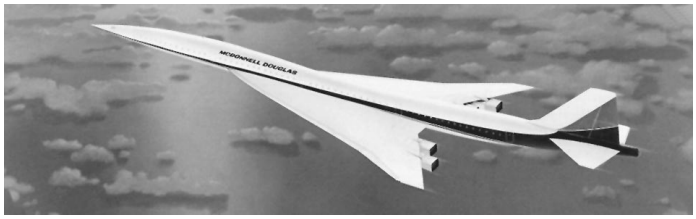
An ever-increasing portion of avionics is dedicated to communications. Much of the increase comes in the form of digital communications for either data transfer or digitized voice. Military aircraft typically use digital communications for security. Civil aircraft use digital communications to transfer data for improved efficiency of operations and RF spectrum utilization. Both types of aircraft are focusing more on enhanced communications to fulfill the requirements for better operational capability.

Various types of communications equipment are used on aircraft. The following list tabulates typical communications equipment:

- VHF transceiver (118–136 MHz)
- UHF transceiver (225–328 MHz/335–400 MHz for military)
- HF transceiver (2.8–24 MHz)
- Satellite (1530–1559/1626.5–1660.5 MHz, various frequencies for military)
- Aircraft Communications Addressing and Reporting (ACARS)
- Joint Tactical Information Distribution System (JTIDS)

In the military environment the need for communicating aircraft status and for aircraft reception of crucial information regarding mission objectives are primary drivers behind improved avionics. In the civil environment

High Speed Research



This McDonnell Douglas conceptual design for a Mach 2.4 supersonic transport is sized to carry about 300 passengers over a distance of 5,000 nautical miles. A NASA/industry high speed civil transport research effort is a first step toward determining whether such a plane can be economically viable and environmentally acceptable. (Photo Courtesy of National Aeronautics and Space Administration.)

Aircraft manufacturers of several nations are developing technology for the next plateau of international aviation: the long-range, environmentally acceptable, second generation supersonic passenger transport, which could be flying by 2010.

NASA's High Speed Research (HSR) program is intended to demonstrate the technical feasibility of a high speed civil transport (HSCT) vehicle. The program is being conducted as a national team effort with shared government/industry funding and responsibilities.

The team has established a baseline design concept that serves as a common configuration for investigations. A full-scale craft of this design would have a maximum cruise speed of Mach 2.4, only marginally faster than the Anglo-French Concorde supersonic transport. However, the HSCT would have double the capacity of the Concorde, and it would operate at an affordable ticket price.

Phase I of the HSR program, which began in 1990, focused on environmental challenges: engine emission effects on the atmosphere, airport noise, and sonic boom. Phase II, initiated in 1994, focuses on the technology advances needed for economic viability, principally weight reductions in every aspect of the baseline configuration. In materials, the HSR team is developing, analyzing, and verifying the technology for trimming the baseline airframe by 30 to 40%. In aerodynamics, a major goal is to minimize air drag to enable a substantial increase in range. Phase II also includes computational and wind tunnel analyses of the baseline HSCT and alternative designs. Additional research involves ground and flight simulations aimed at development of advanced control systems, flight deck instrumentation, and displays. (Courtesy of National Aeronautics and Space Administration.)

(particularly commercial transport), the desire for improved passenger services, more efficient aircraft routing and operation, safe operations, and reduced time for aircraft maintenance are the primary drivers for improving the communications capacity of the avionics.

The requirements for digital communications for civil aircraft have grown so significantly that the industry as a whole embarked on a virtually total upgrade of the communications system elements. The goal is to achieve a high level of flexibility in processing varying types of information as well as attaining compatibility between a wide variety of communication devices. The approach bases both ground system and avionics design on the ISO Open System Interconnect (OSI) model. This seven-layer model separates the various factors of communications into clearly definable elements of physical media, protocols, addressing, and information identification.

The implementation of the OSI model requires a much higher level of complexity in the avionics as compared to avionics designed for simple dedicated point-to-point communications. The avionics interface to the physical

medium will generally possess a higher bandwidth. The bandwidth is required to accommodate the overhead of the additional information on the communications link for the purpose of system management. The higher bandwidths pose a special problem for aircraft designers due to weight and electromagnetic interference (EMI) considerations. Additional avionics are required to perform the buffering and distribution of the information received by the aircraft. Generally a single unit, commonly identified as the communications management unit (CMU), will perform this function.

The CMU can receive information via RF transceivers operating in conjunction with terrestrial, airborne, or space-based transceivers. The capability also exists for transceiver pairs employing direct wire connections or very short-range optical links to the aircraft. The CMU also provides the routing function between the avionics, when applicable. Large on-board databases, such as an electronic library, may be accessed and provide information to other avionics via the CMU.

The increasing demand on data communication system capacity and flexibility is dictating the development of a system without the numerous limitations of current systems. Current communication systems require rather rigid protocols, message formatting, and addressing. The need for a more flexible and capable system has led to the initial work to develop an Aeronautical Telecommunications Network (ATN). The characteristics envisaged for the ATN are the initiation, transport, and application of virtually any type of digital message in an apparently seamless method between virtually any two end systems. The ATN is expected to be a continually evolving system.

Impact of "Free Flight"

"Free Flight" is a term describing an airspace navigation system in which the "normal" air traffic controls are replaced by the regular transmission of position information from the airplane to the ground. The ground system, by projecting the aircraft position and time, can determine if the intended tracks of two aircraft would result in a cohabitation of the same point in the airspace. This is commonly called "conflict probe". If a potential conflict occurs, then a message is transmitted to one or more of the aircraft involved to make a change to course and/or speed.

"Free Flight" dictates special requirements for the avionics suite. A highly accurate navigation system with high integrity is required. The communications and surveillance functions must exhibit an extremely high level of availability.

GNSS Avionics performing the position determination functions will require augmentation to achieve the necessary accuracy. The augmentation will be provided by a data communications system and will be in the form of positional information correction. A data communications system will also be required to provide the frequent broadcast of position information to the ground. A modified Mode S transponder squitter is expected to provide that function.

The free flight concept will require the equipage of virtually all aircraft operating within the designated free-flight airspace with a commensurate level of avionics capability. The early stages of the concept development uncovered the need to upgrade virtually all aircraft with enhanced CNS/ATM avionics. The air transport industry resolved this problem on older airplanes by developing improved and new avionics for retrofit applications. The new avionics design addresses the issues of increased accuracy of position and enhancement of navigation management in the form of the GNSS Navigation and Landing Unit (GNLU) housed in a single unit and designed to be a physical and functional replacement for the ILS and/or MLS receivers. A built-in navigator provides enhanced navigation functionality for the airplane. The GNSS can provide ILS lookalike signals and perform landing guidance functions equivalent to Category I.

Avionics in the Cabin

Historically, the majority of avionics have been located in the electronics bay and the cockpit of commercial air transport airplanes. Cabin electronics had generally been limited to the cabin interphone and public address system, the sound and central video system, and the lighting control system. More recently the cabin has been updated with passenger telephones using both terrestrial and satellite systems. The terrestrial telephone system operates in the 900-MHz band in the United States and will operate near 1.6 GHz in Europe. The satellite

system, when completely operational, will also operate near 1.6 GHz. Additional services available to the passengers are the ability to send facsimiles (FAXes) and to view virtually real-time in-flight position reporting via connection of the video system with the flight system. Private displays at each seat will allow personal viewing of various forms of entertainment including movies, games, casual reading, news programming, etc.

Avionics Standards

Standards play an important role in avionics. Military avionics are controlled by the various standards (MIL-STDs, DOD-STDs, etc.) for packaging, environmental performance, operating characteristics, electrical and data interfaces, and other design-related parameters. General aviation avionics are governed by fewer and less stringent standards. Technical Standard Orders (TSOs) released by the Federal Aviation Administration (FAA) are used as guidelines to ensure airworthiness of the avionics. TSOs are derived from and, in most cases, reference RTCA documents characterized as Minimum Operational Performance Standards and Minimum Avionics System Performance Standards. EUROCAE is the European counterpart of RTCA.

The commercial air transport industry adheres to multiple standards at various levels. The International Civil Aviation Organization (ICAO) is commissioned by the United Nations to govern aviation systems including but not limited to Data Communications Systems, On-Board Recorders, Instrument Landing Systems, Microwave Landing Systems, VHF OmniRange Systems, and Distance Measuring Equipment. The ICAO Standards and Recommended Practices (SARPS) control system performance, availability requirements, frequency utilization, etc. at the international level. The SARPS in general maintain alignment between the national avionics standards such as those published by EUROCAE and RTCA.

The commercial air transport industry also uses voluntary standards created by the Airlines Electronic Engineering Committee and published by Aeronautical Radio Inc. (ARINC). The ARINC "characteristics" define form, fit, and function of airline avionics.

Defining Terms

ACARS: A digital communications link using the VHF spectrum for two-way transmission of data between an aircraft and ground. It is used primarily in civil aviation applications.

Brickwalling: Generally used in software design in critical applications to ensure that changes in one area of software will not impact other areas of software or alter their desired function.

Distance measuring equipment: The combination of a receiver and a transponder for determining aircraft distance from a remote transmitter. The calculated distance is based on the time required for the return of an interrogating pulse set initiated by the aircraft transponder.

Fault tolerance: The built-in capability of a system to provide continued correct execution in the presence of a limited number of hardware or software faults.

JTIDS: Joint Tactical Information Distribution System using spread spectrum techniques for secure digital communication. It is used for military applications.

Validation: The process of evaluating a product at the end of the development process to ensure compliance with requirements.

Verification: (1) The process of determining whether the products of a given phase of the software development cycle fulfill the requirements established during the previous phase. (2) Formal proof of program correctness. (3) The act of reviewing, inspecting, testing, checking, auditing, or otherwise establishing and documenting whether items, processes, services, or documents conform to specified requirements (IEEE).

Related Topic

78.1 Introduction

References

Airlines Electronic Engineering Committee Archives, Aeronautical Radio Inc.
FANS Manual, International Air Transport Association, Montreal, Version 1.1, May 1995.

Federal Radionavigation Plan, DOT-VNTSC-RSPA-90-3/DOD4650.4, Departments of Transportation and Defense, 1990.
M.J. Morgan, "Integrated modular avionics for next generation commercial airplanes," *IEEE/AES Systems Magazine,* pp. 9–12, August 1991.
C.R. Spitzer, *Digital Avionics Systems,* 2nd ed., New York: McGraw-Hill, 1992.

Further Information

K. Feher, *Digital Communications,* Englewood Cliffs, N.J.: Prentice Hall, 1981.
J.L. Farrell, *Integrated Aircraft Navigation,* New York: Academic Press, 1976.
L.E. Tannas, Jr., *Flat Panel Displays and CRTs,* New York: Van Nostrand Reinhold, 1985.
M. Kayton and W.R. Fried, *Avionics Navigation Systems,* New York: John Wiley and Sons, 1969.

102.2 Communications Satellite Systems: Applications

Abdul Hamid Rana and William Check

The history of satellites began in 1957 when the Soviet Union launched Sputnik I, the world's first satellite. In the 1960s the commercial sector became actively involved in satellite communications with the launch of Telstar I by the Bell System followed by the use of a **geosynchronous orbit.** With this type of an orbit, an object 22,753 miles above the earth will orbit the earth once every 24 hours above the equator, and from the earth's surface appear to be stationary. The first geostationary orbit was achieved by NASA using a SYNCOM in 1963. The Communications Satellite Act was signed by the United States Congress in 1962 and created the Communications Satellite Corporation (COMSAT). This was followed by the formation of INTELSAT, an organization that is composed of over 120 countries and provides global satellite communication services. In the 1970s, multiple companies in the private sector in the United States began to operate their own domestic satellite systems. Today there are numerous companies providing this service in the United States: e.g., GE Americom, Hughes, Loral, COMSAT, and American Mobile Satellite Corporation. Other nations such as Canada, Australia, Indonesia, Japan, etc. have their own satellite systems. Several international and regional satellite systems have also been formed. Examples of these are INTELSAT, EUTELSAT, Intersputnik, ARABSAT, AsiaSat, etc. [Pritchard and Sciulli, 1986].

The satellite-based communications systems have significantly evolved over a three-decade period. In the 1960s, satellite communications for commercial use became a viable alternative because of the demand for reliable communications (telephony and voice). In the 1970s, technical innovations made larger, more powerful and more versatile satellites possible. Advanced modulation and multiple-access schemes resulted in smaller, less expensive **earth stations** and better service offerings that were lower cost and higher quality. In the 1980s **very small aperture terminals (VSATs)** emerged and the Ku-band frequency spectrum became widely used. In the 1990's satellites support data, voice, and video communications applications. The VSAT industry has given an overall boost to the entire satellite communication industry.

As new satellites are launched, they will have long-term applications which have expanded opportunities. These include private long-haul networks for internal communications, cable TV, pay TV, business voice and data, satellite news gathering, direct broadcast to the home, integrated VSATs, private international satellite service, high-definition TV, mobile service, personal communications, and ISDN. Disaster recovery planning increasingly includes satellites in order to overcome the coverage limitations of existing terrestrial networks. With the allocation of frequencies for personal communications, the promise of global communications and the reality of a personal phone will soon push satellite communications to a new age.

This section describes satellite communications from the application point-of-view. Since VSATs initiated the growth in satellite communication, a significant portion of the section is devoted to this topic. After a review of the satellites' launch and their characteristics, VSAT networks are discussed in detail. Video/audio applications are described next, along with the equipment necessary for these applications. The section is concluded with a summary of next-generation trends.

Satellite Launch

Launching a communications satellite into orbit is a complex and expensive process. This first stage in a satellite's airborne life may cost several million dollars. The cost for launching is primarily a function of the satellite's weight and size. Traditional geosynchronous communications satellites tend to be large and more costly to launch, although the more compact digital communications payloads and longer satellite life will reduce life cycle costs. Low earth orbit communications satellites tend to be smaller and more economical to launch, but will have shorter in orbit life.

A shortage of launch vehicles influenced the economics of the launch industry following the 1986 U.S. Space Shuttle *Challenger* disaster. The shortage has now given way to other launch alternatives. The dominant player in the satellite launch business is the French company Arianespace. Major U.S. players in the satellite launch business are Lockheed Martin, McDonnell Douglas, and Orbital Sciences Corporation. China and Russia have also begun providing launch services.

The launch of a satellite payload into the geosynchronous orbit involves many complex steps. Using the launch vehicle, the payload is first placed in a parking orbit. This is a nearly circular orbit which places the satellite approximately 300 km above the earth's surface. After reaching this orbit, the next step is to fire a motor known as the payload assist module (PAM) to place the payload in a transfer orbit. The PAM motor is discarded afterwards. The transfer orbit is an elliptical orbit whose perigee matches the parking orbit and whose apogee matches the geostationary orbit. Perigee is defined as the point in the orbit closest to the earth, while apogee is the point in the orbit furthest from the earth. The payload itself consists of the satellite with an apogee kick motor (AKM). Once in a transfer orbit, the AKM is fired at the point when the satellite has reached apogee. This firing will place the satellite in a nearly circular orbit. Final positioning of the satellite in geosynchronous orbit can then take place [Pritchard and Sciulli, 1986].

Spacecraft and Systems

A satellite spacecraft employs several major subsystems. These are propulsion, electrical, tracking, telemetry command and control, and the communications subsystem. Figure 102.4 is a diagram of a typical commercial satellite. The propulsion subsystem consists of thrusters oriented in north-south and east-west directions and is used to maintain the spacecraft in the proper orbit and orientation. An electrical subsystem is used to generate electricity in the spacecraft by means of solar cells. Backup batteries are used during periods of equinoxes. The solar cells are also used to charge the batteries. The tracking, telemetry, and command subsystem is used to receive commands from the controlling ground station, as well as to allow the ground station to monitor on-board systems.

The spacecraft requires some form of stabilization to prevent it from tumbling in space. There are two types of stabilization techniques: spin stabilization and three-axis stabilization. Spin stabilization uses an outside cylinder to spin, creating the effect of a gyroscope providing spacecraft stabilization. An internal platform is decoupled from the cylinder, whose orientation is fixed towards the earth. Three-axis stabilization uses internal gyros which sense movement of the spacecraft. Any movement in the axes is detected and can be compensated by firing thruster jets.

The communications subsystem consists of receiver and transmitter sections. The receiver system consists of wideband redundant units. The transmitter subsystem consists of separate amplifiers (transponders) for each channel utilized. Satellite systems make use of orthogonal polarized signals in order to transmit two signals simultaneously on the same frequency, a technique known as "frequency reuse." Two different polarization methods for signals are used: horizontal and vertical linear polarization, or clockwise and counterclockwise circular polarization.

Figure 102.5 shows a simplified block diagram of a typical satellite. A matrix-type switching arrangement is provided on the input and output of the transmitter subsystem for switching to backup transponders. This satellite is three-axis stabilized and operates at Ku-band. There are 16 operational transponders with a bandwidth of 54 MHz each. The employment of frequency reuse provides nearly 1000 MHz of usable bandwidth. Fourteen of the 16 operational transponders use 20-W traveling wave tube amplifiers (TWTA) to provide ground-commandable east or west regional coverage, for 48-state (CONUS) coverage. The remaining two transponders provide 50-state coverage using 27-W TWTAs. For the 50-state channels, one spare 27-W TWTA provides

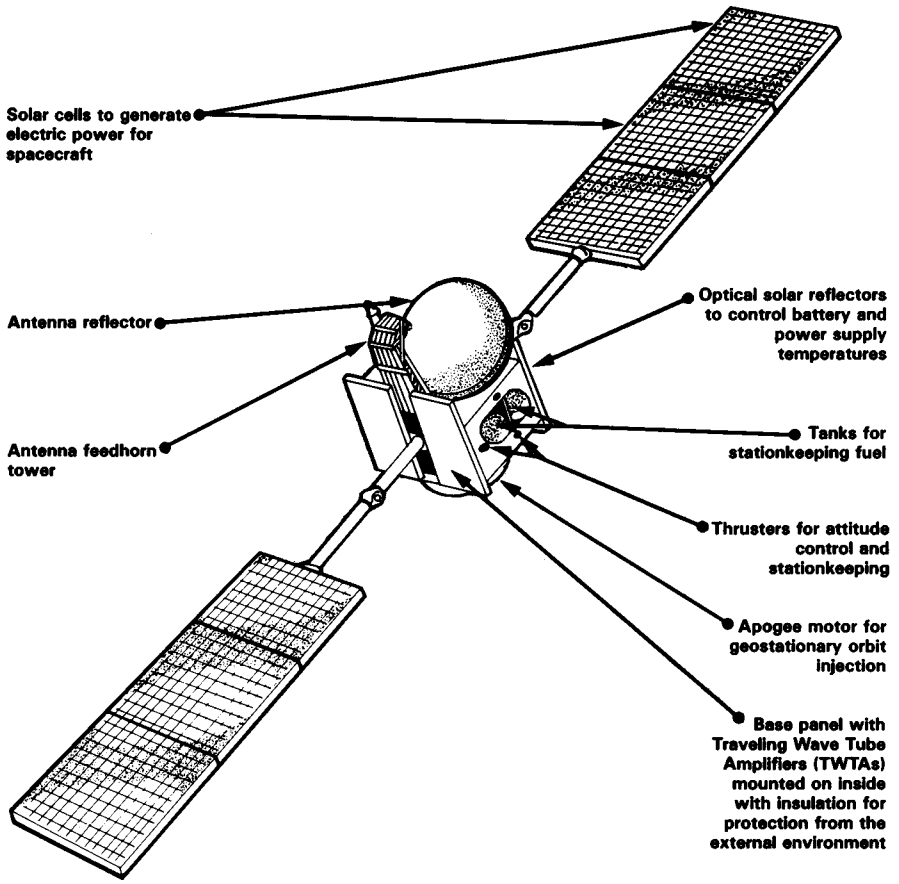


FIGURE 102.4 Simplified block diagram of a communications satellite.

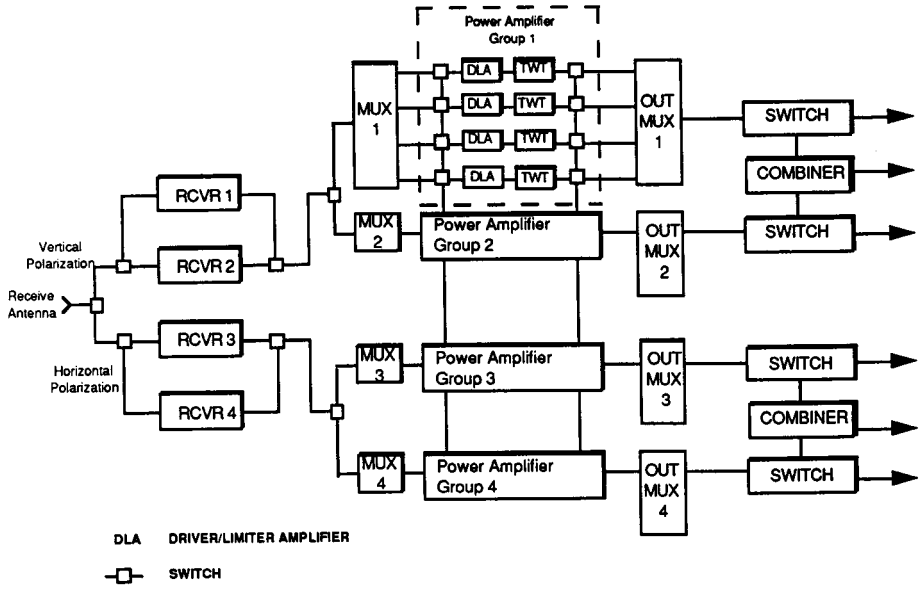


FIGURE 102.5 Simplified block diagram of GSTAR satellite.

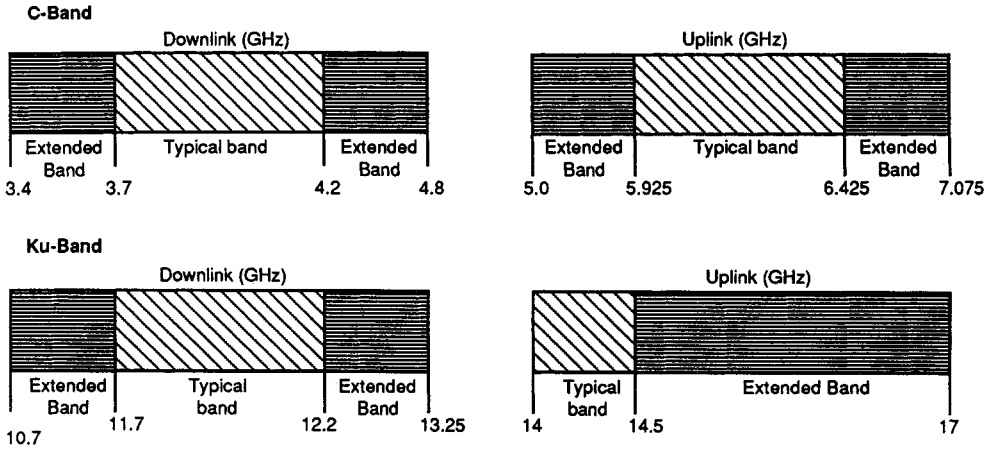


FIGURE 102.6 Ku- and C-band frequency allocation chart.

protection for the two operating TWTAs (3-for-2 redundancy). For the remaining transponder channels, 5 spare 20-W TWTAs provide protection for 14 operating TWTAs (19-for-14 redundancy). Redundant communications receivers are provided on a 4-for-2 basis.

The power radiated from a satellite is described as its effective radiated isotopic power (EIRP) and is the radiated power of the satellite in decibels referenced to one watt of power. The units are in dBW. The strength of the signal received on the ground is a function of the spacecraft location and that of the ground station and will vary depending upon location. A map of the signal strength contours is called the satellite's "footprint."

Geosynchronous Satellites

There are over 500 Ku-band and C-band satellites in geosynchronous orbit. These satellites are typically spaced anywhere between 1 to 3 degrees apart. Older satellites no longer in active service may be spaced less than one degree in an inclined orbit.

The frequency plan for C- and Ku-band satellite services is shown in Fig. 102.6. The typical transmit frequency band used for fixed satellite services in the Ku-band is 14.0–14.5 GHz. Receive frequency is 11.7–12.2 GHz. Some satellites also use the extended band. For C-band satellites, the typical operating transmit frequency is 5.925–6.425 GHz and the receive frequency is 3.7–4.2 GHz. The operating band was extended at WARC '79 to 7.075 GHz to be assigned to individual countries for domestic satellite systems. Ka-band satellites have downlinks in the frequency range 17–23 GHz and uplinks in the range 27–31 GHz. Some European and Japanese satellites operate in this range [Long, 1991].

The satellite performance data indicate a wide range of variation in the specifications among the various satellites. Most satellites have a design lifetime of 10 years. The newer GEO satellites tend to have an extended life of 12–15 years. Most domestic U.S. satellites have 24 transponders. The older generation of Asian satellites have very few transponders per satellite. Some planned satellites will have a large number of transponders. Nominal transponder bandwidths include 36, 54, and 72 MHz.

Satellite power is increasing in the newer generation of satellites. Lower-power satellites have an EIRP in the 20–35 dBW range. There are a significantly large number of medium-power satellites in the 35–45 dBW range. Newer high-power satellites tend to have power in the 50–60 dBW range. Direct broadcast satellites are planned for transponder power in the 60–120 W range. The power generally varies with polarization, frequency, and beam. Table 102.2 is a profile of typical satellite performance characteristics.

Mobile Satellite Systems

Mobile satellite systems encompass communications on land, in the air, or over the oceans ideally allowing a person to communicate with anyone anywhere [Long, 1991]. The Inmarsat system is a mobile communications system providing global coverage through a variety of communication paths. In the United States, the FCC has authorized American Mobile Satellite Corporation (AMSC) to provide domestic mobile satellite services. AMSC makes use of geostationary satellites to provide a domestic offering similar to the international offering of Inmarsat.

TABLE 102.2 Typical Satellite Performance

Satellite Operator	System Name	Configuration	EIRP in dBW at Edge	Comments
GE Americom	GSTAR series	Ku-band	38–48	Domestic coverage
	Spacenet series	C- and Ku-band	C-band: 34–36 Ku-band: 39	
Hughes Comm	Galaxy Series	C- and Ku-band	C-band: 34–38 Ku-band: 45–49.5	Domestic coverage
Intelsat	Intelsat VA (IBS)	C- and Ku-band	C-band: 20–26 Ku-band: 38–41	International service, worldwide
	Intelsat VI	C- and Ku-band	C-band: 20–26 Ku-band: 38–41	
	Intelsat VII	C- and Ku-band	C-band: 26–36 Ku-band: 41–46	
Eutelsat	Eutelsat I series	Ku-band	35–43.5	Covers all of Europe
	Eutelsat II	Ku-band	42–47	
NASDA-NTT (Japan)	Sakura 2	C- and Ka-band	C-band: 30 `Ka-band: 37	CS-4a, CS-4b in the Sakura series is scheduled for launch during 1992–94

Over the past two decades, there has been active work in the area of low earth orbit (LEO) satellite systems. In general, LEOs are designed to provide a full range of communication services, both voice and data. Proposed systems are designed to complement existing cellular communications technology. Several companies have proposed LEO systems and have made application to the FCC for a "Pioneer's Preference" license. This license allows the use of new and innovative technology. Motorola's Iridium system is potentially the largest, using 66 satellites to provide coverage over the entire globe.

Because of the low altitude of the orbit, LEO systems use multiple satellites to provide coverage over a regional area or over the entire globe. Satellites operating at a low orbit are less costly due to the reduced launch costs and reduced weight. However, a low orbit requires the use of multiple satellites since the low altitude of the system provides smaller beam coverage. Since these satellites are not geostationary, ground stations must track an LEO satellite as it passes overhead.

Due to potential growth of mobile satellite communications, several systems are proposed to be in operation in the 1990s. Examples of these systems are the Iridium, Globalstar, ICO, Orbcomm, Starsys, Odyssey, and Teledisc.

Direct Broadcast Satellites

The direct broadcast satellites (DBS) concept is to transmit programming directly to homes using a small receive-only antenna via high-powered satellites. Through the use of a high-powered satellite, a small receive-only satellite antenna may be used for home reception, with the ultimate goal to offer antennas less than one foot in diameter. High-powered DBS satellites use high-powered transponders, i.e., 60–120 W. To prevent interference into the small receive antennas at these high power levels, the DBS satellites will be spaced further apart in geosynchronous orbit.

The first efforts in DBS began in the early 1980s when COMSAT built several DBS satellites, but did not launch them. Internationally, many countries currently have DBS services. Several European countries have high-powered DBS satellites; many others use medium-powered satellites. The DBS industry in the United States is being revitalized by advances in digital video compression technology and the announcement of new players such as Hughes, Primestar, Echostar, etc. to offer DBS services. Hughes Communications and United States Satellite Broadcasting (USSB) system using a high-powered DBS satellite is in operation. As an alternative to the launch of a high-powered satellite, medium-powered DBS systems make use of existing satellites in orbit. However, larger home antennas are required, approximately 2 feet or greater in diameter. A medium-powered DBS in the U.S. is Primestar. Digital video compression techniques using the MPEG-2 standard are used to allow multiple video channels in a transponder. DIRECTV® service, launched in the summer of 1994 by Hughes Electronics, is an example of the direct satellite system.

Earth Stations

Earth stations are the interface point for communications to and from the satellite [Ha, 1986]. An earth station can be divided into two subsystems, the transmit chain and the receive chain. A common element between the transmit and receive chain is the antenna. Because of the large signal attenuation at RF frequencies, the earth station antenna must have high signal gain and be highly directional to focus the power to and from the satellite. A parabolic-shaped reflector antenna is used by earth stations since it can provide these characteristics.

The transmit chain consists of several major components: baseband equipment, modulators, frequency upconverters, high-power amplifiers (HPA), and combiner circuitry used to switch the output of the HPAs to the antenna. The receive chain uses a low-noise amplifier to receive the satellite signals, frequency downconverters, demodulator, and baseband equipment.

In the transit chain the signals are modulated, combined, and frequency-shifted with an upconverter to the desired satellite transmit frequency. After upconversion, the signals are amplified by HPAs. In a large earth station, there may be many HPAs which feed to a single antenna. These signals must be switched and combined appropriately. At microwave frequencies, waveguide combiners are used to route the output of the HPAs to the antenna.

In the receive chain, the counterpart to the HPA is the low-noise amplifier (LNA), which is used to amplify the signals received from the antenna. This amplifier must be designed for maximum gain with a very small noise contribution. The noise generated in this unit contributes significantly to the overall performance of the receive side of the earth station. Gallium arsenide (GaAs) FETs are commonly used in the amplifier section of the LNA because of their low-noise characteristics. The LNA feeds the signal to the frequency downconverter, which converts it to IF frequency suitable for demodulator.

A hub monitoring and control (M&C) system provides the monitoring and control of the RF equipment and baseband equipment. Redundant RF equipment is common at a hub, and the M&C system is used to monitor the components and provide automatic switchover in the event of equipment failure. Switchover between equipment can occur either by operator initiation or automatically by the M&C upon sensing an equipment failure.

Technical characteristics of *large earth stations* have been established for use with the INTELSAT system. INTELSAT categorizes two types of earth stations: multipurpose and special purpose. A multipurpose earth station can be used with any service, while a special-purpose earth station is restricted. Multipurpose standard A, B, and C earth stations have antenna diameters from 11 to 33 meters. Special-purpose standard D, E, and F earth stations have antenna diameters between 3.5 to 11 meters.

In addition to fixed earth stations, "portable" earth stations, called *transportables,* have been manufactured which can be taken to locations originating the programming. These transportables are usually mounted on a truck or trailer and include all the components necessary for an earth station. In the case of the transportable, the antenna size is selected to be as small as 4 meters in diameter. A transportable earth station is designed to be upgraded with "building blocks" to handle heavy, medium, and thin route traffic. Transportable earth stations are designed to meet the requirements for various applications such as temporary business communications, temporary carrier service, backup during the retrofit of an existing earth station, and disaster recovery.

Another type of earth station is the *flyaway.* This is a small remote satellite terminal which can be packed into suitcases for shipment on an airline for delivery anywhere in the world. These systems consist of a small antenna, RF unit, and baseband equipment to provide a complete satellite communications station. An example is an L-band version which provides audio communications via the Inmarsat system. Fitting into a small suitcase, it contains a telephone handset, RF electronics, and antenna that can be assembled to provide audio communications anywhere in the world. Mobile satellite terminals are even smaller, suitable to be carried as handheld or briefcase units.

VSAT Communication System

Advances in technology have revolutionized the satellite communications industry by deployment of very small aperture terminal (VSAT) networks for data, voice, and video communication. Since the mid-1980s, VSAT networks have become widely used in the oil, lodging, financial, auto, retail, and manufacturing industries. By the 1990s, VSATs were operating in C and Ku-bands. Also by the mid-1990s, over 70% of the VSAT market

TABLE 102.3 Typical VSAT Systems Features

Feature	Interactive	Point-to-point	Broadcast
Topology	Star	Point-to-point, mesh	Point-to-multipoint
Communication	Between hub and VSATs, VSAT to VSAT through hub	VSAT to VSAT	Hub to VSATs
Frequency	Ku-, C-band	Ku-, C-band	Ku-, C-band
Hub antenna	3–11 m	—	3–11m
VSAT antenna	0.9–2.4 m	1.8, 2.4 m	0.5–2.4 m
Hub to remote access	TDM, SCPC, spread spectrum	SCPC	SCPC, spread spectrum, FM[2]
Remote to hub access	ALOHA, reservation stream, CDMA	SCPC	—
Outbound data rate (Kbps)	56–512	9.6–2048	9.6–2048
Inbound data rate (Kbps)	9.6–256	9.6–2048	—
Modulation	BPSK, QPSK, DPSK	BPSK, QPSK	BPSK, QPSK, FM[2]
FEC	Rate 1/2, convolutional or block	Rate 1/2, convolutional	Rate 1/2, convolutional or block
Protocols	SDLC, Bisync, Async, X.25, TCP/IP Burroughs and others	Clear channel	Clear channel, synchronous, HDLC format

was accounted for by the retail, automotive, and financial industries. VSATs are making private networks a viable alternative for many companies, for applications such as point-of-sale, reservation systems, remote monitoring and control, branch office administration, financial transactions, etc. A VSAT is a small earth station suitable for installation at a customer's premises. A VSAT typically consists of antenna less than 2.4 m, an outdoor unit to receive and transmit signals, and an indoor unit containing the satellite and terrestrial interface units [Rana et al., 1990].

VSAT networks fall into three general categories: broadcast networks, point-to-point networks, and interactive networks. In a broadcast network, a centralized hub station broadcasts data, audio, and/or video to a group of receive-only VSATs. Low-cost receive-only VSATs can receive news, weather services, and financial information. Music distribution and video broadcast via broadcast networks is widely used. Point-to-point networks provide direct communication between two locations without the requirement of a large hub for data, voice, and image transmission. Variations of these networks include point-to-multipoint dedicated circuits or demand-assigned mesh topologies. Interactive networks are used for two-way communications services between a central hub station and a large number of VSATs in a star topology. Table 102.3 is a summary of the salient features of VSAT networks. VSATs are available for both C- and Ku-band frequency. Most VSAT systems use BPSK modulation with Rate 1/2 FEC. For interactive networks, the inbound channel is shared on contention basis to conserve space segment. More advanced systems use concatenated codes to improve performance. Recently, hybrid VSATs have been introduced to use terrestrial networks on the return channel. An example is the Hugh's Direct PC which uses a high speed satellite receive channel, and a low speed terrestrial return channel.

A critical element of VSAT networks is the network availability. The VSAT system availability is affected by three major components: effects of rain attenuation, equipment availability, and software availability. The effects of rain attenuation for Ku-band networks are significant. While link availability is usually specified at 99.5%, link performance can be optimized to nearly any desired value through the use of energy dispersion techniques or large antenna sizes. The network hardware must be highly reliable. Hub hardware should provide for optional redundancy and the ability to achieve better than 99.9% availability. The use of hub diversity and uplink power control can also be used to improve the network availability. The VSAT hardware availability is less catastrophic; the loss of one VSAT does not constitute network failure but may require a service call to rectify the problem. Hence, it is common to use nonredundant but highly reliable VSAT units. Software availability needs to be improved since software failures dominate the overall availability of interactive networks in existing VSAT products.

Interactive networks have been by far the most popular for data communication and audio/video overlays. The remaining portion of this section is devoted to these networks. An interactive VSAT system consists of a

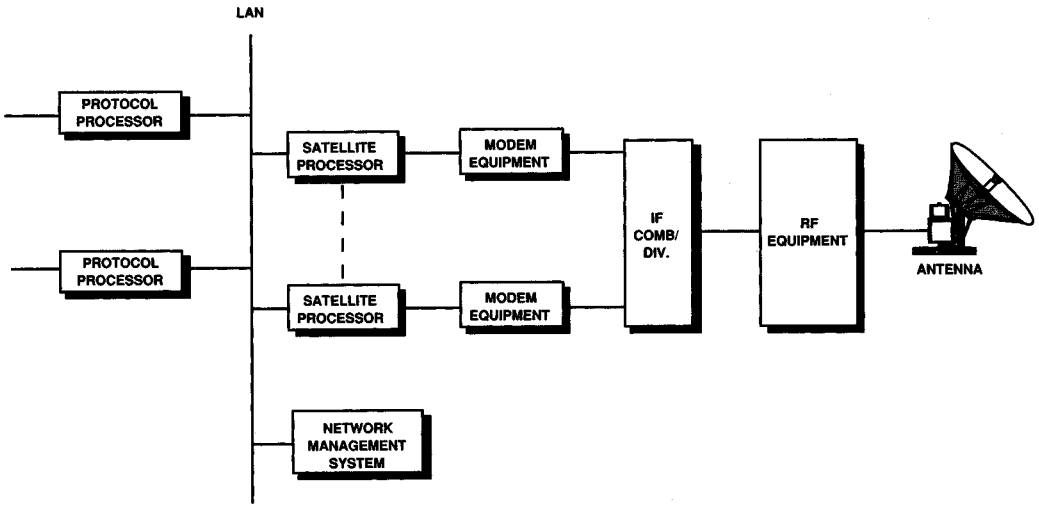


FIGURE 102.7 Block diagram of a hub.

hub, VSAT, network management system, and associated transmission and processing subsystems. These subsystems along with sophisticated **satellite access protocols** and terrestrial protocol interfaces make interactive networks a flexible and powerful communication medium.

Hub

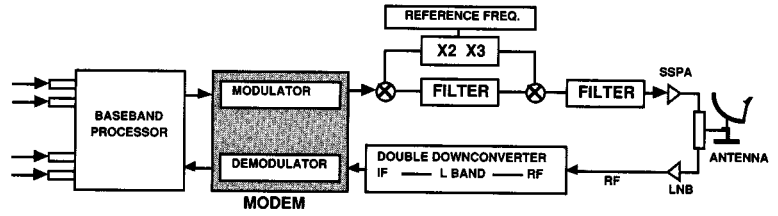
The hub performs all functions that are necessary to establish and maintain virtual connections between the central location and VSATs. In private dedicated networks, the hub is co-located with the user's data processing facility. In shared hub networks, the hub is connected to the user equipment via terrestrial backhaul circuits. Since the hub is a single point for failure in a star network, it is typically configured with 1:1 or 1:*N* redundancy. The hub consists of antenna, RF, and baseband equipment (Fig. 102.7). It will handle multiple channels of inbound and outbound data and often one or more channels of audio or video broadcast.

The hub antenna consists of a parabolic reflector and associated electrical and mechanical support equipment. The RF subsystem converts the modulated carrier to RF frequency, provides the necessary signal amplification, and transmits the resulting RF carrier to the antenna subsystem. It also receives RF signals from the antenna subsystem, provides low-noise amplification, RF/IF conversion, and passes the resulting IF carriers to the baseband equipment subsystem. The hub baseband equipment consists of the modem equipment and the processing equipment. The hub modems employ continuous modulators and burst demodulators. The processing equipment interfaces to the modem equipment and provides the satellite access processing and protocol processing for interface to the customer host.

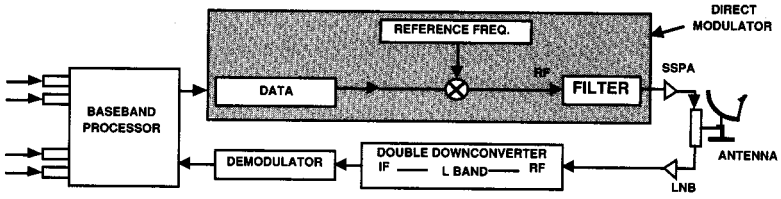
VSAT

The VSAT consists of an antenna, outdoor unit (ODU), interfacility link (IFL), and indoor unit (IDU). The IFL connects the IDU and ODU subsystems, providing the transmit and receive lines, monitor and control signals, and dc power for the ODU electronics. A single-cable IFL, in which all signals are multiplexed on the same cable, is usually used to reduce the cost of IFL. VSATs nominally use a 1.2- or 1.8-m offset feed parabolic antenna. Smaller antenna sizes are preferable to reduce the installation cost. Options for small antennas include the use of either a submeter parabolic reflector or a flat-plate antenna. The choice of antenna is a tradeoff among performance, installation cost, and aesthetic considerations.

The ODU consists of a solid-state power amplifier (SSPA), a low-noise amplifier, upconverter, and a downconverter. VSAT SSPA modules are usually between 1.0 to 3.0 W. The ODU cost can be significantly lowered by utilization of a low-power SSPA (0.1 to 0.5 W) consistent with obtaining the required output power. The VSAT receive side front end can be economically configured using an LNB. Low-cost HEMT LNBs are currently available with 50–60 dB gain and noise figures lower than 1.3 dB.



(a) Conventional VSAT



(b) VSAT using direct modulation.

FIGURE 102.8 Simplified VSAT block diagram.

Direct modulation of the RF carrier may lower the cost of the VSAT IF and RF electronics while consolidating modulation and upconversion functions. Direct modulation allows the design of a VSAT with fewer parts, smaller size, and lower weight than with traditional outdoor units. Figure 102.8 is a block diagram showing a conventional VSAT and a VSAT using direct modulation. An L-band receive interface between the ODU and IDU is preferable in order to receive audio and video overlays.

The IDU is located near the user terminal equipment. Major IDU functions include outbound carrier signal acquisition, tracking, demodulation, bit synchronization, burst modulation, and protocol processing. It also controls the operation of the ODU, monitors VSAT health, and responds to hub commands. The baseband processing system performs satellite channel access and protocol and customer interface processing functions. A video/audio port can be provided with an RF splitter at the IDU to separate the received audio/video signal for the optional video/ audio receiver.

Network Management System

The network management system (NMS) is a critical element of a VSAT network. Through the NMS, the user can have full control of his network, which is usually not possible in the case of terrestrial network facilities. The NMS generally provides a centralized management tool for hub and VSAT equipment configuration control, assignment of inbound and outbound satellite channels, network monitor and control, switchover to back-up equipment, network statistics collection, downline loading of new software, and report generation. In the shared hub environment, the hub operator controls the allocation of resources among various users and controls the RF transmission facility. The user must have the ability to manage his portion of the network transparent to other users. In the case of a dedicated hub, a single management entity can exert full control over the network, including RF transmission facilities.

The network management system standards community has defined five functional areas as requirements for network management systems. These areas are fault management, accounting management, configuration management, performance management, and security management. The VSAT network management system should be capable of interfacing with other network management systems by supporting a standard network management protocol. The protocol standard most widely accepted is the Simple Network Management Protocol (SNMP).

Transmission System

Most VSAT systems employ BPSK or QPSK modulation with rate R = 1/2, K = 7 convolutional coding and soft-decision Viterbi decoding on both the inbound and outbound channels. Differential phase shift keying (DPSK) modulation may be used to reduce the demodulator complexity and cost. DPSK is relatively insensitive

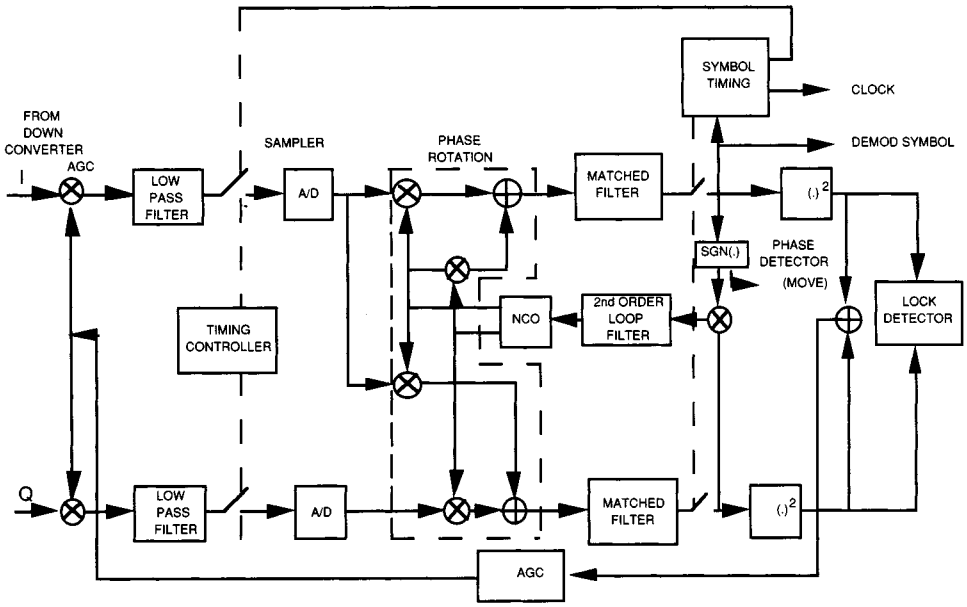


FIGURE 102.9 DSP demodulator functional diagram.

to phase noise and frequency offset, thus allowing the use of lower-cost LNBs in the VSAT terminals. However, as compared to BPSK, convolutionally encoded DPSK requires about 2 dB greater Eb/No at a BER of 10^{-5}. In addition, if operation is required below 10 dB, some form of low-level interleaving may be required.

In lieu of performing the VSAT demodulation function via the traditional analog circuit techniques, an all-digital implementation using digital signal processing (DSP) techniques may be considered. The merits of DSP include the development of a more testable, producible, maintainable, configurable, and cost-effective demodulator. Figure 102.9 presents an illustration of the DSP demodulator functions to be implemented using the DSP processor(s). The functions of the major blocks are as follows: phase locked loop (PLL) for carrier acquisition, narrowband Costas loop for data detection, external automatic gain control (AGC), dynamically advance/retard sampling to achieve optimum data sampling, and A/D converters for signal analog-to-digital conversion.

A VSAT system must employ frequency agility in the remote terminal to use an assigned block of frequencies within a transponder. Within the assigned frequency band, one or more outbound carriers and a number of inbound carriers are precisely located. On the VSAT receive or outbound side, the LNB output can be demodulated directly using a synthesizer-controlled local oscillator, or further downconversion can be used under synthesizer control to obtain the demodulator input signal at a standard IF frequency such as 70 or 140 MHz. In the inbound direction, channel selection can be accomplished by two methods. First, the carrier frequency of the modulator can be shifted to select the appropriate channel and a fixed upconverter may be used to obtain the RF signal. Second, the synthesizer output frequency may be multiplied up to RF to obtain the carrier, which may then be modulated directly with the data as described in Cannistraro and McCarter [1990].

Satellite Access Protocols

The multiple satellite access protocol is one of the most critical elements to the performance of a VSAT network. VSAT systems tend to be used in applications where message delay is critical and this protocol is the controlling element to the delay-throughput performance of the system. During the past 15 years, there have been numerous multiple-access protocols developed and simulated in the context of satellite packet communications [Raychaudhuri and Joseph, 1988]. Table 102.4 provides a comparison of throughput vs. delay for various satellite access protocols.

In the outbound or hub-to-VSAT direction, a TDM channel is employed. This channel may be regarded as a point-to-multipoint or broadcast channel with node selectivity being achieved by the use of addressing

TABLE 102.4 Random Multi-Access Protocols Comparison

	Throughput	Comments
Pure ALOHA	0.13–0.18	Low cost, good for variable-length messages
Slotted ALOHA	0.25–0.37	Good for fixed-length messages
Selective reject ALOHA	0.20–0.30	Variation of pure ALOHA with a modified algorithm
Tree CRA	0.40–0.49	Sensing capability for collision resolution, good for fixed-length messages
Announced retransmission random access (ARRA)	0.50–0.60	Uses modified algorithm of slotted ALOHA by announcement of transmission
Random access with notification	0.45–0.55	Uses partition for new and retransmitted message
CDMA	0.10–0.40	Used in spread spectrum systems, low delay

information embedded in the modulated data stream. The delay performance of this channel is essentially controlled by the queuing behavior of the hub. In the VSAT-to-hub direction, a large number of VSATs share the channel to conserve space segment. Most VSAT networks utilize a combination of slotted ALOHA protocol for the interactive component of the inbound traffic and a reservation TDMA scheme for any bulk data transfers. Most protocols are adaptive in the sense that as the channel traffic increases, they automatically evolve into reservation TDMA systems. Code division multiple access (CDMA) has been used in VSATs operating at C-band. CDMA permits more than one signal to simultaneously utilize the channel bandwidth in a noninterfering manner. This makes it possible to significantly increase the utilization and throughput of the channel.

Interface Capabilities

Most VSAT systems support common data communications protocols such as SDLC, X.25, Async, Bisync, TCP/IP, etc. Coexistence of different protocols is allowed in a network. A VSAT supports multiple ports with common interfaces such as RS232C, RS422, V.35, etc. VSAT networks typically must provide **protocol spoofing** to provide acceptable delay and throughput performance to the end-user application. To minimize the effect of satellite delay, the host computer front-end processor is emulated at the VSAT location, and multiple cluster controllers are emulated at the hub location. The polling associated with the front-end processor to cluster controller communication is not carried on the satellite link, but is instead emulated locally.

Video

Satellites are an excellent medium for video transmission since they can provide a broadcast capability with wide bandwidth. Video on satellites is ideal for applications such as videoconferencing, business TV, distance learning, satellite news gathering, etc.

Video Teleconferencing

Satellite communications provides a cost-effective and flexible means of interactive videoconferencing. Technological improvement in videocompression has resulted in low-cost codecs at data rates less than T1, and good quality videoconferencing is possible at data rates as low as 56 Kbps. Low-cost satellite terminals coupled with low-cost codecs are making videoconferencing via satellite affordable and practical for many organizations. Applications include all types of business meetings and technical information exchange such as management and staff meetings, new product introductions and updates, sales meetings, training, and market presentations. Videoconferencing allows people at different locations to meet with almost as much ease as being in the same room, providing benefits of increased productivity, reduced travel time and cost, and increased management visibility.

A generic videoconference system is presented in Fig 102.10. The system consists of a specially designed room, video/audio equipment, transmission equipment, monitor and control computer, and space segment. The video and audio feeds from the meeting room pass through the codec and are compressed. From the codec, the signal passes to the satellite modem for modulation. The radio frequency/terminal (RFT) upconverts the modulated carrier and amplifies it for transmission to the satellite. At the other site, the process is reversed.

A videoconferencing network features point-to-point, broadcast, or point-to-multipoint architectures. In a point-to-point system, two sites are configured for interactive conference with duplex audio and full motion video transmission. Videoconferencing broadcast is appropriate for formal presentations where the presenter

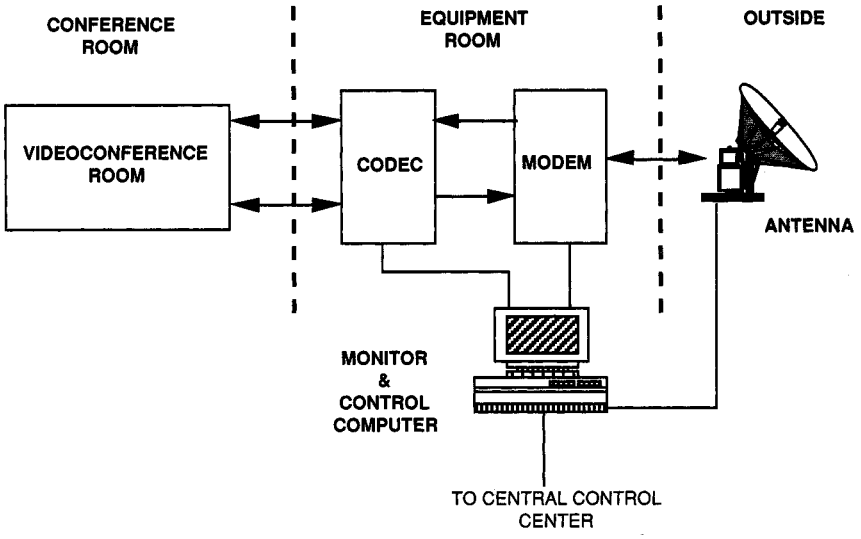


FIGURE 102.10 A generic videoconference system.

does not need to see the audience, such as a speech from a senior corporate executive. In a point-to-multipoint conference, multiple sites can receive a transmitting site. Two of the primary sites are fully interactive with each other. A feature called multipoint switching has been implemented in some commercial systems. This feature allows switching of receive and transmit sites during the conference. The multipoint switching feature can be provided using either a TDMA or SCPC system. A TDMA system allows multiple sites to transmit and receive in a mesh configuration. An economical multipoint switching system is possible with SCPC using only two transmit frequencies. In a "chair" controlled conference, the chair is assigned one of these frequencies for the duration of the conference. Dynamic allocation of the second frequency is controlled by the chair to any of the participating sites at any time during the conference.

Video Broadcast

Video broadcast over satellite is attractive for industry segments such as educational TV, distance learning, business television, and television receive-only (TVRO) applications. Business television allows users to transmit broadcast-quality video programming from a studio to any number of specified locations equipped with TVROs. A video broadcast capability, as an overlay to interactive data networks, is becoming increasingly popular for corporate presentations, education, and training.

A video uplink consists of a video exciter, HPA, antenna, and optionally an encryption system such as B-MAC (multiplexed analog component, version B) encoder, for business video broadcasts. Each remote VSAT must be configured to receive the video transmission. This involves adding a video receiver at each VSAT location that plugs into the VSAT IDU. Audio/video signals from the video receiver can be presented directly or through a B-MAC decoder to a standard TV monitor.

The digital compressed video signal can be used as a replacement for an analog video distribution. Digital coding technology can be used to compress video signals to reduce data rates to 2 Mbps or even lower and reproduce near broadcast-quality video. Distribution of digital video signals at such rates requires less transponder bandwidth and a smaller antenna at remote terminals. Compression techniques used are based on one or a combination of the following: inter/intra-frame prediction, adaptive differential transform, conditional replenishment, discrete cosine transform, adaptive prediction, motion compensation, and vector quantization [Patterson and Delp, 1990].

Satellite News Gathering

Satellite news gathering (SNG) is used for live, on-the-spot coverage and news exchanges with other commercial broadcast stations. This is made possible by the availability of occasional-use space segment and transportable earth stations on news trucks. An SNG systems consists of a compact earth station and video/audio transmission

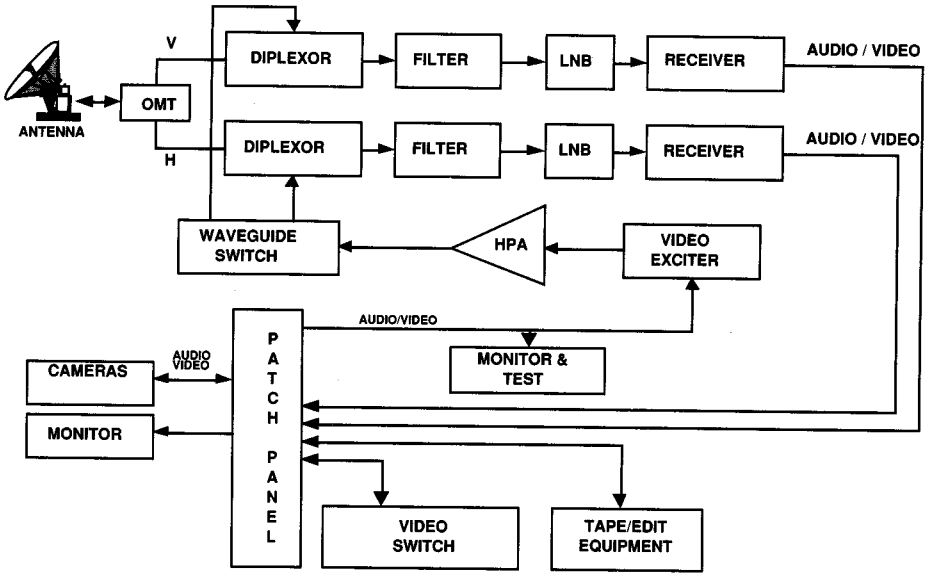


FIGURE 102.11 SNG vehicle video/audio system.

system on a truck. A duplex voice channel is used to coordinate between the space segment provider, studio, and the SNG truck.

Figure 102.11 presents a block diagram of a typical SNG system. The RF subsystem has a transmit path and two independent receive paths. The transmit path consists of an HPA and a frequency agile video exciter which modulates and upconverts the video signal to the satellite's RF frequency. A waveguide switch is used to select transmit polarization. Camera signals go simultaneously to tape for storage and for transmission over the satellite. A receive path is typically provided for both receive polarizations. Each path consists of a transmit reject filter and an LNB which downconverts to L-band. The received L-band signal passes through a video satellite receiver, from which point it can be routed to various monitor or test points or be routed to a tape device for recording and storage.

Audio

The use of commercial broadcast audio transmission via satellite began in the late 1970s with National Public Radio and Mutual Broadcasting using Western Union's WESTAR I satellite. The main application was to send high-quality audio to radio broadcast stations to transmit programming information. This type of system makes use of single channel per carrier (SCPC) satellite transmission, where each satellite channel corresponds to one audio channel. The entire satellite channel is FM modulated. Pre-emphasis is used over the channel to provide additional noise reduction. A variation of this technique, called multiple channel per carrier (MCPC), can be used to transmit multiple channels over a single satellite carrier. Figure 102.12 is a block diagram of the MCPC system.

As the marketplace searched for lower-cost systems, the FM^2 (or FM/FM) modulation technique evolved, allowing the use of low-cost FM receivers. Through a high-powered FM modulated carrier on the satellite, a low-cost audio and data broadcast receiver can be built. This FM/FM modulation technique is widely used to distribute audio and data on a low-cost basis.

In addition to audio broadcasts, satellite-based voice applications include point-to-point voice, multinode interactive voice, and voice over data VSATs. Point-to-point voice is most prevalently used for high-volume voice trunking for long-distance connectivity or transoceanic connectivity. A multinode, interactive voice architecture is ideal in providing voice connectivity to remote locations that are not serviced by terrestrial voice facilitates. Both mesh and star configurations are used to provide multinode voice connectivity. Automated satellite access control and resource allocation techniques are used to allow for granting requested on-demand

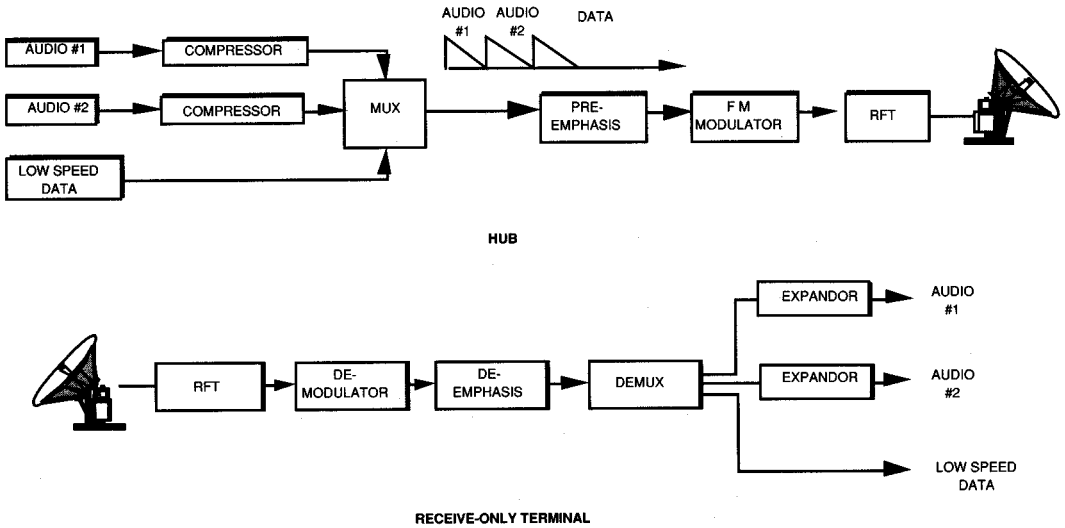


FIGURE 102.12 Block diagram of the MCPC system.

availability of voice connectivity. To support voice over data VSATs, an audio encoder is used to accept an analog voice signal, digitize and packetize it, and format it for transmission through the VSAT data network. A voice port may either be implemented as part of a "baseline" data/voice card or as an add-on stand-alone box.

The integrated data/voice system employs a TDM outbound carrier and shared inbound carriers for data and voice transmission. Two types of voice network communications alternatives may be implemented for voice channel communications: a poll/response access scheme and a reservation TDMA access scheme. With the poll/response access scheme, the hub polls the VSAT voice ports on a cyclic basis. The VSATs return their responses in the form of call requests or status updates. The number of sites in the voice network determines the rate at which VSATs are polled. Thus, this scheme is suitable for a small network. Excessive polling delays will be encountered for a network with a relatively large number of remotes.

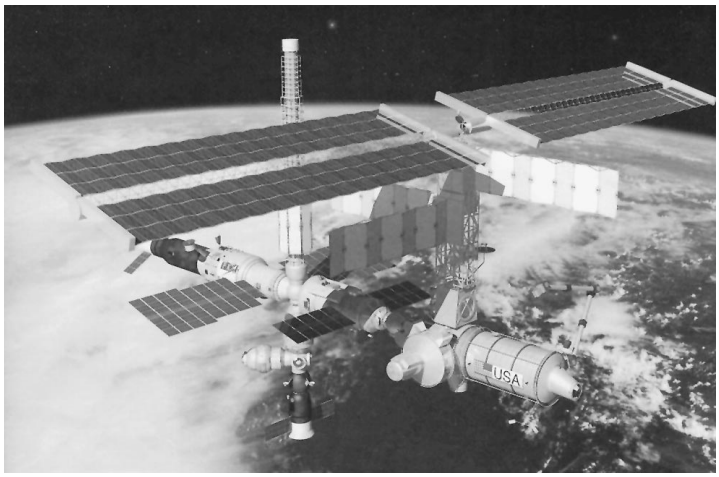
In reservation TDMA, on the other hand, voice call requests are serviced by the assignment (reservation) of a logical channel for inbound voice traffic. Although various means are implemented to avoid collisions on the satellite link, the time needed to reserve capacity on an inbound carrier may be lengthy, depending on traffic conditions. Therefore, call setup times are not as predictable as they are with the poll/response access scheme.

The VSAT design is ideally suited for digital compressed voice. Coding rates of 32, 16, and 9.6 kbps and lower can presently be achieved, depending on the compression technique employed. There are two classes of digitizing voice signals: waveform coding and vocoding. In waveform coding, the analog voice curve is coded and then reproduced by modeling its physical shape. Data rates are relatively high, i.e., higher than 9.6 kbps. Vocoding attempts to reproduce the analog voice curve by abstractly "identifying" the type and shape of the curve. Only a set of parameters is transmitted, describing the nature of the curve. Achieved data rates can be as low as 1.2 kbps.

Second-Generation Systems

The recent wave of satellites have much higher power than their predecessors. The Intelsat K satellite, for example, is equipped with 60-W TWTAs and serves increasing worldwide traffic, video, and VSAT services. Another example is the Telstar 4 satellite which has variable power up to 120 W for Ku-band transmissions and is being promoted to be HDTV compatible in preparation for expected widespread use of HDTV. Other trends in satellite design, i.e., NASA's advanced communication technology satellite (ACTS), include the use of multiple spot beams and onboard IF and/or baseband switching. Onboard switching coupled with electronically hopped spot beams and laser intersatellite links have been proposed. Spot beams provide higher satellite EIRP which permits small, low-cost VSATs to accommodate higher bit rate transmissions. The use of multiple-beam architectures also increases bandwidth availability through frequency reuse. Advances in multibeam satellites

International Space Station



The interim International Space Station will look like this. In the right foreground is the U.S. laboratory module and the station's airlock. In the center of the horizontal string of modules is the FGB energy block. The solar power array at the top is one of four that will provide power for the complete station. Below the tower is the Russian-built universal docking module and, at bottom, one of two crew transfer vehicles. (Photo courtesy of National Aeronautics and Space Administration.)

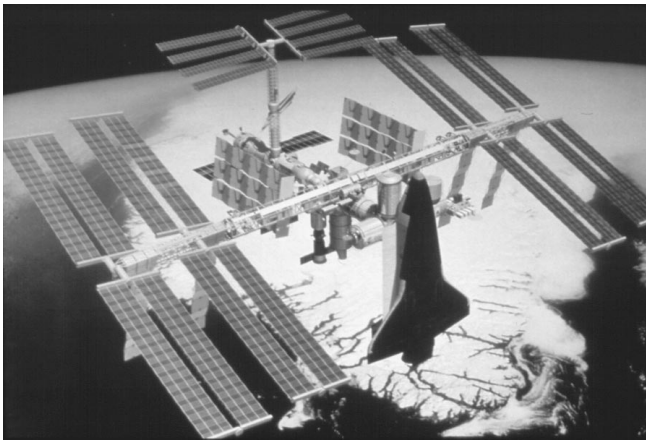
The International Space Station will be a permanent laboratory for human-monitored long term research in the unique environment of Earth-orbital space, an environment that cannot be duplicated on Earth for long duration experiments. This space station program draws upon the resources and scientific and technological expertise of 13 cooperating nations.

This project is being constructed in three phases. Phase I included the 1995 construction of two Boeing-built nodes (Node 1 and Node 2). The nodes will serve as connecting passageways between modules. Phase I was completed in early 1996 with the production of the U.S. laboratory module where astronauts will perform continuous scientific research.

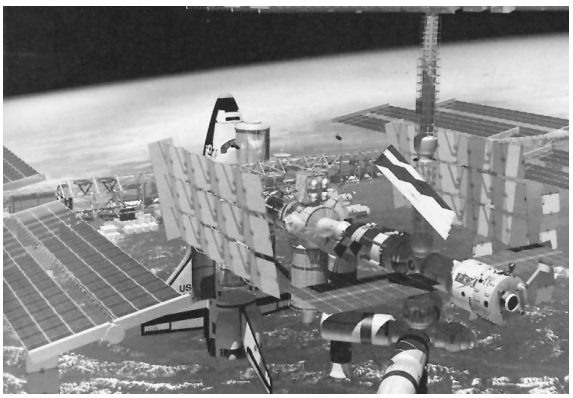
Phase II of the space station program begins in November 1997 with the launch of the FGB functional cargo block on a Russian *Proton* vehicle. The FGB is a 21-ton element that will provide altitude control and propulsion during the early assembly operations, plus solar power and berthing ports for additional modules. In May of 1998, the embryo space station will grow with the addition of the *Proton*-boosted Russian service module, which provides life support and habitation facilities, utilities, and thrusters. Then the crew transfer vehicle, a Russian Soyuz TM capsule, will be joined to the station. By June 1998, the first three-person crew will begin its orbital stay. Phase II will be completed in Spring of 1999.

In Phase III, the International Space Station will progress gradually to its ultimate status as a fully operational permanent orbital research facility. Among key additions to the core configuration are the remaining modules of the U.S.-built solar array; the Japanese experiment module, to be delivered in 2000; and the U.S. habitation module which contains the galley, toilet, shower, sleep stations, and medical facilities. With the delivery of a second Russian crew transfer vehicle in June 2002, the station will be virtually complete.

The completed station will measure 361 feet from tip to tip of the solar arrays. The pressurized living and working space is roughly equivalent to the passenger cabin volume of two Boeing 747 jetliners. The



A concept view of the International Space Station in its final configuration with a space shuttle orbiter docked at the fore port. The cylinder near the orbiter's nose is the U.S. centrifuge accommodation module. Below it, hidden by the orbiter, is the U.S. laboratory module, flanked by the European (left) and Japanese laboratories. (Photo courtesy of National Aeronautics and Space Administration.)



This concept view is of the station from the opposite (aft port) end. In the foreground (lower right) is the Russian service module, with living and working room for three crew members. Next, toward the center of the photo, is the FGB energy block, then (near the orbiter) the U.S. laboratory module. The vertically mounted cylinder below it is the U.S. habitation module. (Photo courtesy of National Aeronautics and Space Administration.)

space station will contain seven laboratories. In addition, the Japanese experiment module has an exposed "back porch" with 10 mounting spaces for experiments that require long duration with the space environment.

Beginning in 1997, there will be a total of 73 assembly and service flights until the station becomes fully operational in midyear 2002. (Courtesy of National Aeronautics and Space Administration.)

with onboard baseband processing allows some of the intelligence in the central hub and VSAT equipment to be moved to the satellite. The result is expected to be improved VSAT-to-VSAT communications and a platform to provide dynamic bandwidth allocation [Naderi and Wu, 1988].

The trend in deploying higher-power satellites has an inverse effect on the size of the earth station antenna. The earth stations are becoming smaller, less complex, and more cost effective. Private hubs are now typically in the range of 3.5 to 7.6 m and are not required to be staffed. Two-way VSATs antennas originally deployed in sizes from 1.2 to 1.8 m are now using elliptical or rectangular-shaped antennas with apertures equivalent to 1.0 m or less. Two-way ultra-small aperture terminals are also emerging. These lower-cost, lower-functionality earth stations are designed for thin route, niche-type applications such as point-of-sale and credit card transaction processing. The advances in DSP technology will continue to enhance the capabilities and performance while at the same time lowering the cost of VSATs. The advances in MMIC technology continue to miniaturize the RF components while increasing reliability.

With advances in digital signal processing and compression techniques, analog video and audio transmission will increasingly be converted to digital transmissions. The advanced compression techniques reduce the bandwidth requirements and allow for smaller and lower-cost VSAT antennas to be used. The continued technological advances in satellite technology and the emerging demand for more flexible communication services will generate new satellite communications applications, such as LAN interconnections and ISDN support [Murthy and Gordon, 1989]. Satellite communications will also play an increasing role in mobile communications on land, air, and on sea. In addition to telephony services, new services such as global distress and safety applications, global positioning, navigation, voice messaging, and data transmissions are now possible.

Defining Terms

Earth station: The interface point for communications to and from a satellite. An earth station (also known as a hub) consists of an antenna and transmit and receive subsystems.

Geosynchronous orbit: An orbit 22,753 miles above the earth in which an object will orbit the earth once every 24 hours above the equator and will appear to be stationary from the earth's surface.

Protocol spoofing: A technique used by VSAT networks to reduce the network delay. The satellite network emulates the host computer front-end processor at the VSAT location and emulates the multiple cluster controllers at the hub location.

Satellite access protocol: A set of rules by which a number of distributed VSATs communicate reliably over a shared satellite channel.

VSAT: Very small aperture terminal. A small earth station suitable for installation at a customer's premises. A VSAT typically consists of an antenna less than 2.4 m, an outdoor unit to receive and transmit signals, and an indoor unit containing the satellite and terrestrial interface units.

Related Topics

74.1 Introduction • 78.1 Introduction

References

J.C.L. Cannistraro and S. McCarter, "Direct modulation lowers VSAT equipment costs," *Microwaves and RF,* pp. 99–102, August 1990.

T.T. Ha, *Digital Satellite Communications,* New York: MacMillan, 1986.

M. Long, *World Satellite Almanac,* 3rd ed., Winter Beach, Fla.: MLE, Inc. 1991.

K.M. Murthy and K.G. Gordon, "VSAT networking concepts and new applications development," *IEEE Communications Magazine,* pp. 43–49, May 1989.

F.M. Naderi and W.W. Wu, "Advanced satellite concepts for future generation VSAT networks," *IEEE Communications Magazine,* vol. 26, pp. 13–22, July 1988.

H.A. Patterson and E.J. Delp, "An overview of digital image bandwidth compression," *Journal of Data and Computer Communications*, pp. 39–49, Winter 1990.

W. Pritchard and J.A. Sciulli, *Satellite Communication Systems Engineering,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.

A.H. Rana, J. McCoskey, and W. Check, "VSAT technology, trends, and applications," *IEEE Proc.,* vol. 78, no. 7, pp. 1087–1095, July 1990.

D. Raychaudhuri and K. Joseph, "Channel access protocols for Ku-band VSAT networks: A comparative evaluation," *IEEE Communications Magazine,* vol. 26, no. 5, pp. 34–44, May 1988.

Further Information

The *World Satellite Almanac* provides a tutorial of the satellite communications industry. It includes the technical characteristics and footprint maps for geosynchronous satellites worldwide. Contact: MLE Inc., P.O. Box 159, Winter Beach, FL 32971.

World Satellite Communications and Earth Station Design is a text which provides an analytical presentation of communication satellites and their applications. Contact: CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, FL 33431.

The monthly *IEEE Communications Magazine* investigates VSAT communications in a special series spanning several issues between 1988 and 1989. Contact: IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854-4150.

Clapp, G., Sworder, D. "Command, Control and Communications (C^3)"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

103

Command, Control, and Communications (C^3)

G. Clapp
Naval Command, Control and Ocean Surveillance Center

D. Sworder
University of California, San Diego

103.1 Scope
103.2 Background
103.3 The Technologies of C^3
103.4 The Dynamics of Encounters
103.5 The Role of the Human Decisionmaker in C^3
103.6 Summary

103.1 Scope

The focus of this chapter is not a detailed profile of a current or planned military C^3 **system** but it is rather on the issues and the technologies of the C^3 mission. Evolving technology, an evolving world order, and constant programmatic reorderings render such express descriptions to become rapidly outdated. Thus block diagrams of specific military systems (and listings of their acronyms) are de-emphasized. Of paramount interest is not electronics technology in isolation, but rather technology integrated into systems and analysis of these systems operating under complex real world environments that include technologically capable adversaries. The human commander or **decisionmaker,** as the principal action element in a C^3 system, is included explicitly in the system analysis.

103.2 Background

Electronics technology is nowhere more intensively and broadly applied than in military systems. Military systems are effective only through their command and control (C^2) and this is recognized by the fact that C^3 is a critical discipline within the military. Frequently systems will be denoted C^2I or C^3I rather than command and control. This adds to C^2 the essential area of **intelligence** and intelligence products derived from surveillance systems. All variants of these acronyms are to be considered equal, whether or not communications, intelligence, or surveillance have been left implicit or made explicit. Likewise the superscript notation is considered optional and interchangeable. The formal discipline of C3 within the military has not been matched by focused technical journals or university curricula due to its highly multidisciplinary nature.

Two definitions from a Joint Chiefs of Staff (JCS) publication [JCS, Pub. 1] capture the breadth of C2. This reference defines command and control as "The exercise of authority and direction by a properly designated commander over assigned forces in the accomplishment of his mission. Command and control functions are performed through an arrangement of personnel, equipment, communications, facilities, and procedures which are employed by a commander in planning, directing, coordinating and controlling forces and operations in the accomplishment of his mission."

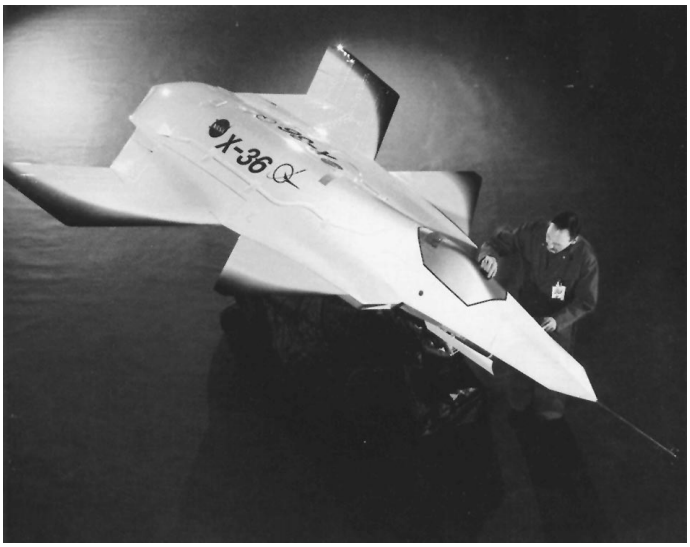
C2 systems are defined, with almost equal breadth, as "An integrated system comprised of doctrine, procedures, organizational structure, personnel, equipment, facilities, and communications which provides authorities at all levels with timely and adequate data to plan, direct and control their operations."

A Look Toward Future Flight

On March 19, 1996, NASA and McDonnell Douglas Corporation unveiled to the public a new subsonic flight vehicle designated X-36, a remotely piloted tailless research aircraft. The X-36 is designed to demonstrate the feasibility of future tailless military fighters that can achieve agility levels superior to those of today's aircraft.

In the absence of a tail, control of the X-36 is accomplished by a combination of thrust vectoring and innovative aerodynamic control features. Tailless fighter configurations offer reduced weight, increased range, and improvement in survivability. The X-36 is "flown" by a pilot located in a van at the flight test facility; a camera in the XX-36 cockpit relays instrument readings and displays to a console in the van. With a wing span of only 10.4 feet and a gross weight under 1,300 pounds, the X-36 is powered by a single turbofan originally designed as a cruise missile power plant.

The X-36 program is intended to establish confidence to incorporate these technologies in future piloted vehicles. This project exemplifies one aspect of a NASA aeronautical research and technology program that seeks to improve the performance, efficiency, and environmental characteristics of all types of planes and, additionally, addresses such infrastructure factors as air traffic control, navigation, and communications. (Courtesy of National Aeronautics and Space Administration.)



Designed jointly by NASA and McDonnell Douglas Corporation, the X-36 is a subscale, remotely piloted tailless vehicle for demonstrating technologies that could lead to lighter, longer-ranging, more survivable, more agile military fighter aircraft. (Photo courtesy of National Aeronautics and Space Administration.)

Though general, two points emerge from these definitions: (1) C^3 is multidisciplinary and (2) C^3 is a process which, to this point, includes only implicit roles for electronics technology. One military service, however, often refers to C^4 and C^4I or has even used C^4I^2 where the final C and second I refer to computers and interoperability, respectively, as acknowledgment of the increasing reliance on technology.

A C3 system can be visualized as shown in Fig. 103.1. Within the constraints imposed by organization, **doctrine,** and the skills of the personnel of the military unit, the commander plans and controls his forces. At a basic level, command and control is a resource allocation problem, which often must be solved under much tighter time horizons and subject to greater uncertainty levels than exist in civil applications.

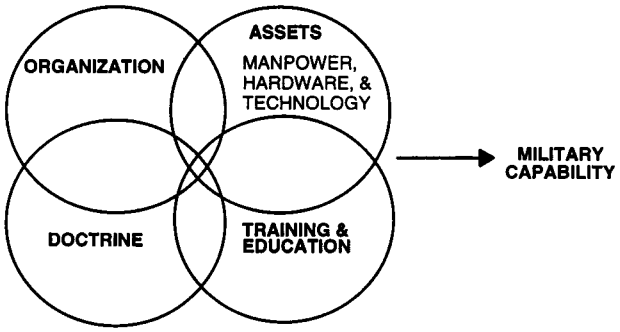


FIGURE 103.1 Components of C^3.

The four basic components display overlapped regions to indicate their inseparability. A portion of each category can be designed in isolation; a new antenna or a new radio with decreased size, weight, or power consumption has minimal impact on the other components. However, insertion of a broad new technology (e.g., a radio relay combined with a remotely piloted vehicle (RPV) or the networking of radios) has wide reaching consequences and it may take years to fully integrate into doctrine, training, and organization. The conjunction of the four areas, when specified with some detail, represents or contains an architecture. If the assets, the doctrine, and so on are limited to just one military function, then the aggregation is referred to as a mission architecture. Figure 103.2 depicts two approaches to achieving C3 architectures. The first [Fig. 103.2(a)] is essentially an aggregation and combination of existing assets and is referred to as a "bottom-up" architecture. The "top-down" version of architecture development [Fig. 103.2(b)] begins with earlier and high order perspective (and higher order oversight). Interfaces and interface standards become more important in top-down architectures; instead of numerous custom and unique interfaces, a minimal set of interface standards is desired. When new or updated equipment is designed or acquired it can be integrated without new interface developments, a key property of an **"open system" architecture.** A developing architecture of this type is entitled, at the Joint Chiefs of Staff level, "C4I for the Warrior." Service-specific top-down architectures are Copernicus (Navy), AirLand 2000 (Army), MTACCS (Marine Corps Tactical Command and Control System) and a yet unnamed Air Force architecture. Each of these are to be considered as evolving architectures and all reflect the impact and importance of scenarios with highly mobile nodes. The open system or top-down approach promotes interoperability between the developments of each service.

Capital investment constraints limit strict adherence to either architectural approach. MTACCS is a meta-system of seven independently developed systems and is best described as a hybrid architecture. Most communication systems within any of the above architectures existed prior to an architecture and thus have a hybrid nature.

Doctrine is a formalized description of military mission definitions and often includes the procedures to accomplish those missions. Doctrine will also often specify the organizational structure appropriate to the specific missions. Some military establishments adhere to strong doctrinal orientation, even down to strict dictation of technology developments. Other establishments treat doctrine as a loose guideline that can be liberally modified. One foreign military analyst observed that U.S. commanders did not seem to read their own doctrinal publications, and even if they did, would not feel compelled to follow them. A flexible military organization with flexible doctrine, however, can be constrained by inflexible hardware and software. Thus an emerging C3 emphasis is a technical focus on modular equipments, standard interfaces between equipments, "open system" architectures, and (software) programmable equipments.

The best way to understand military C3 is to view it as a set of adaptive control loops. The basic variable is information and most of the effort in C3 synthesis is devoted to information handling and management. The resource allocation problem with feedback found in C3 has obvious similarities to those found in corporate operations and public safety service operations. Each is characterized by multiple priorities, limited resources, timelines, and deadlines for performance. Measures of the consequences of a given action tend to be obscured both by its antecedent actions and by changing external **environments.** The external environment contains both continuous events (i.e., tracking of targets) and discontinuous events (i.e., an equipment failure or the onset of communications jamming).

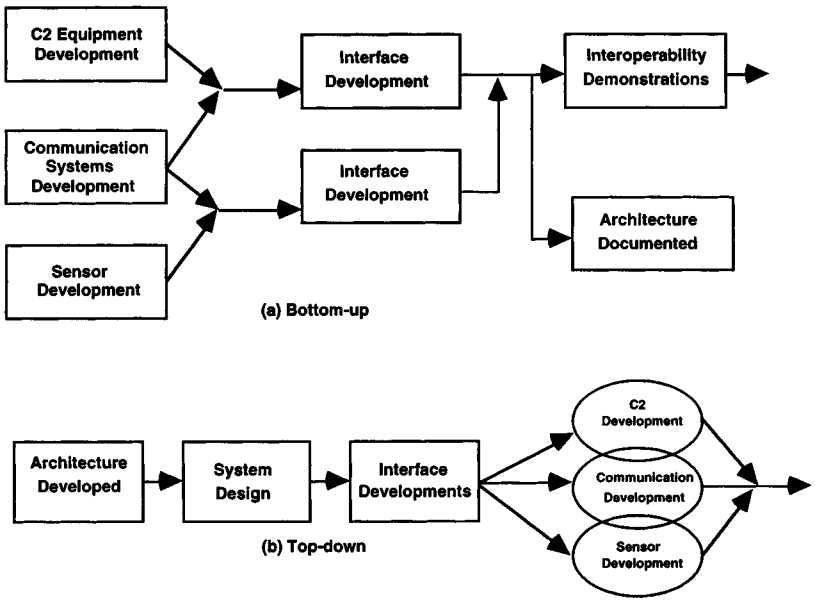


FIGURE 103.2 Architectural processes.

Command and control systems are examples of perhaps the most complex adaptive systems. In its *static* state, C3 assets are aggregates of sensors, processors, databases, humans (with their attributes and organizations), computer hardware/software, mobile platforms, weapons, and communication equipments distributed over wide areas. In the *dynamic* state these assets must be mapped into capabilities in the presence of uncertain or unexpected threats, evolving missions, changing environments, mixed with unreliable communications and possible deception. All can be expected to occur over extended geographic regions and at high tempos. In short, **C³ maps assets into capabilities**. The control processes require rapid and accurate decisionmaking; from this has come the need for heavy reliance on computer-based data systems and high-reliability communications. Despite the existence of fielded weapon systems capable of autonomous operation, the principal action element in the system is still human.

C3 system complexity arises primarily from the magnitude and mobility of the forces involved; forces that can be composed of up to thousands of mobile platforms and hundreds of thousands of personnel. To this is added the large amount of uncertainty present; uncertainty borne of the adversary, of human attributes, dynamics, hostile environments, and communications. Hundreds of radio frequency channels may be in simultaneous use supporting command, surveillance, intelligence, personnel, and logistics functions.

103.3 The Technologies of C³

The general scenario outlined in the previous sections is no longer accommodated by last generation technology of grease pencils, maps, and visual signaling. Technology covered in nearly every other chapter of this handbook is rapidly being incorporated into military C3 systems. Defense departments world wide continue to support technology developments from sub-micron microprocessing devices to global information systems.

Technologies with recent major impact on C3 are

a. Digital communications/data links/networking. The newer and critical role of digital (computer-computer) communications initially became possible through satellite communication systems. Tactical data links (short-range digital communications) have been enhanced by error control techniques such as coding, automatic repeat requests, and spread spectrum radios. Networking, a well-established commercial technique, is being developed for tactical applications. Networking offers survivability through alternate

routing, more efficient (shared) use of channel capacity, and interoperability between interconnected users. Commercial Integrated Services Digital Networks (ISDN) and Asynchronus Transfer Mode (ATM) technology is appearing in both global and nodal military applications. Traditional voice communications remain important; Department of Defense directives require all voice circuits to be secure or encrypted. Digitized voice techniques offer advantage in digital encryption and compression.

b. Space surveillance, terrestrial surveillance, data fusion. The quantity and quality of surveillance systems continues rapid growth utilizing sensors from ground-based, airborne, and space-based vantages. Remote sensing requirements continue to expand the need for real time digital data communications. Unmanned Airborne Vehicles (UAV) and Unmanned Underwater Vehicles (UUV) platform developments continue as a response to a broad range of C3I needs. Two classes of surveillance are active surveillance systems (radar, sonar, and optical) and passive systems (electronic surveillance measuring (ESM), acoustic, infrared and visual imagery). Passive techniques are preferred as they do not leave a signature that can be exploited by adversaries. A plethora of new sensor systems challenges the currently available communications, processors, and processing systems. Particularly challenging is both the fusion of the outputs of multiple similar sensors and also of dissimilar sensor systems. Fusion protocols and tracking algorithms, software intensive, claim an increasing fraction of available resources. With multiple new sensor systems, a technology challenge is the processing, correlation, and fusing of surveillance data into intelligence products and their distribution in a timely and usable form.

c. Computer-based data and information systems. From the communications and surveillance capabilities above, the objective has become formation of a consistent tactical picture throughout the operations theatre. Rapidly evolving processing technology allows vast amounts of data handling and management with corresponding shortening of control decisionmaking times. The ability to match computer processing capability with high data rate, reliable, and survivable computer-grade communications on a global basis to small mobile platforms is an ongoing challenge. Military information systems, in order to retain trusted functioning, require procedures for input data that may have been delayed, omitted, partial, inaccurate, or irrelevant (DOPII). Expanding amounts of software-based systems are needed as a response to increased tempo, data volume, and quality while reducting staff and manpower functions.

d. Architectures and architectural thinking. C3 assets, especially communications, are evolving as assets to be shared, controlled, and rapidly reallocated rather than be dedicated to a specific user. Joint and combined operations, requiring improved interoperability, are becoming common as operations become more regionalized. Two functions of focus, Battle Damage Assessment (BDA) and Indicators and Warnings (I&W), are best implemented when surveillance, communications, and intelligence are architecturally integrated. Integrated systems are also best for timely response to deception and false alarms. A current Navy direction is not to inundate the afloat commander with volumes of unsolicited data but rather have him request what is needed. This style, called information pull, represents a significant change from traditional information push. The impact on supporting communications is to give it a more "bursty" character, driven by external events.

e. Digital signal processing, programmable systems. Single-function C3 hardware is evolving to multifunction capability. Each node or platform will emerge with new capabilities that permits rapid and flexible reallocation. Current generation tactical military aircraft, as delivered, have virtually no additional space or weight allowance for new equipments. A desire is to evolve from costly retrofitting to a state of software insertion and integration. Traditional single-band radio systems will be replaced with **programmable multiband, multiwaveform systems.** Near real-time management and control of highly flexible, programmable systems will become a growing research and development thrust. Next generation cellular technology involving hybrids of frequency hopping, direct sequence spread, and time division spread spectrum techniques invokes new digital signal processing efforts. Also receiving development is Direct Satellite Broadcast (DSB) to tactical military units.

f. Interoperability and standards. C3I systems, with many dispersed nodes, rely heavily on computer-computer communications. Standards are being promoted by industry and government to simplify the development, acquisition, and insertion of new technology as well as to promote interoperability between

independently developed systems. Significantly the Department of Defense has edicted that commercial standards for electronics and telecommunications are to be utilized in preference to military standards in order to promote more rapid and lower cost acquisition of state-of-the-art technology. Two additional motivations for new standards are increased traffic requirements and increased system complexity. C3 applications and users have found significant benefit in increasing communication with programmatically unrelated data sources such as databases and sensors. There is an increase in internal communications as well. Also systems have become more complex, forcing programs to develop modularized architectures. Software is replacing hardware as the most complicated component of communications and C2 systems to design, build, and maintain. Modularized architectures are required to simplify development and enable insertion of new technologies.

The primary computer-to-computer communications architecture has been the Open Systems Interconnection (OSI) Reference Model. The OSI Reference Model has been successful as a layered architecture with well-defined interfaces and specified division of functions. The Department of Defense has committed to adopting an enhanced version of the OSI protocols, called the Government OSI Profile (GOSIP). OSI/GOSIP integration into C3 systems is lagging because of delays in accredited vendor implementations and the cost of upgrading the existing communications infrastructure. NATO is also adopting standards for their joint procurement policies; to a significant degree they overlap commercial standards.

OSI brings to C3 a set of application services that had not been previously available. For example, the OSI electronic mail standards (usually called X.400) provide message forwarding, distribution list creation and distribution, and obsolete message extraction among other services to users. In addition to the security protocols contained in the lower layers of the OSI stack, X.400 has its own security services such as message origin authentication, message flow confidentiality, message content integrity, and nonrepudiation of delivery, services that are highly desirable in C3 environments. OSI also has enhanced file transfer and management capabilities, systems management, directory, and transaction processing, among other application functions, all providing enhanced capability to C3 users.

g. Precision timing and position location (GPS). Navigation/position location historically is important and becomes more so in high dynamic maneuver warfare. With the introduction of the Global Positioning System (GPS), 3-dimensional positioning is available to the smallest of high-mobility nodes. Even with a less than complete satellite constellation, position accuracies can become less than 100 m.
h. Displays and workstations. High-resolution displays combined with programmable workstations and software lead to flexible node functions and consequently to flexible architectures. A C3 workstation could, in principle, support any of a number of C3I functions; a relocation of operators may be the only requirement to physically relocate a command node. Numerous decision aids are now being included within workstations and with their more comprehensive capability are now often described as decision support systems (DSS). Man-machine interface (MMI), as a result, grows in importance.
i. Software techniques. With the growing computational power and memory capability of microprocessor systems, C3 system performance will increasingly be determined by software performance. The cost and complexity of software appears to expand in proportion to host computer capability and is more frequently becoming a system limiting factor. ADA is dictated to be the common programming language of the Defense Department; however, exceptions can be approved. Verification and validation (V&V) of generated software and software maintainence have grown to necessitate organizational changes within the military. Software standards have also increased in importance in new C3 systems. POSIX standards (published as IEEE 1003) govern the software interfaces to operating system services in various computing platforms [NIST, 1990]. As such, they allow application programs written according to the standards to be reused. POSIX standardizes interfaces to security, networking, and diverse system services, including file management, memory and process management, and system administration services. POSIX.5 provides bindings for the ADA programming language.
j. Simulation and modeling. Both techniques are employed with the objective of designing or analyzing the performance of a C3I system. With the advent of faster computation, complex scenarios can be "gamed" in near real time, and modeling will then be within the decision aid realm.

103.4 The Dynamics of Encounters

Within dynamic systems, and the C3 systems that support them, it is important to identify and clarify time scales involved. Military engagements range from sub-second events such as local missile point defense to the long-term development and implementation of global strategy. Each involves basic aspects of decision and control theory: objectives, observations, and feedback and control. In the military environment, the observation aspect is especially complex, requiring the placement, collection, transmission, and aggregation of data from numerous dispersed sources. Control and decision techniques derived for one echelon level may be inappropriate for others, primarily due to the time available for the assessment and feedback process. Often, the impact of a decision will not be measurable before yet another control decision is required. Thus, the relative roles of automation and humans will be different at different levels. The human may have to project a decisionmaking consequence long before the system hardware/software can obtain measures of it.

As an example of encounter space-time domains, surface Navy echelon levels have order-of-magnitude scales as shown:

Organization Level	Time Scale of Interest	Geographic Extent (km)
Platform	seconds-minutes	10's
Battle Group	minutes-hours	100's
Fleet	hours-days	1000's
Theater	days-weeks	1000's +
Service/National	weeks-years	Global

At the platform level, the time scale range reflects engagement times which may include limited or local amounts of tracking. At the Battle Group level, the time scale corresponds to tasks such as maneuver, coordinated engagement, and track management.

Any of the organizational levels may additionally have planning functions that precede the operational time scales by up to months or years. The planning side includes events such as logistics, maintainence, training, and exercises, all of which contribute toward becoming a more capable combatant. Figure 103.3 portrays the planning and the operational or execution phases as well as portraying the adaptive control loop approach to C3. The lighter shaded feedback path is employed when it is required to compare status with the current plan. It is also available for adjustment when plans or objectives are modified. The execution phases are represented by the Stimulus-Hypothesis-Options-Response (SHOR) Paradigm suggested by Wohl [1981]. The control theoretic implications are apparent in the figure; the Stimulus-Hypothesis is a representation of situation assessment with its implicit uncertainty. Quickness and accuracy with which a military command organization can transverse the execution loop is a general measure of performance (MOP). Qualitatively it is generally accepted that the side with the best ability to transverse the SHOR execution loop will have a significant military advantage. In this light, attributes of the execution loop become a measure of effectiveness (MOE) of the C3 system in terms of operational outcomes. Rules of Engagement (ROE) impact tempo by reducing uncertainty or options available to the decisionmaker. Some scenarios develop with such quickness that the C3 system must react nearly reflexively (e.g., without consideration of possible options). One class of rules is made known to all the participants; if a particular manuever is observed, then a specified response will result.

The SHOR paradigm illustrates why counter-communications and counter-command and control are increasingly important operational and technical areas. Counter-C3 need only delay the process rather than disrupt or destroy it in order to be an effective technique. The Navy, for example, is now incorporating **electronic warfare** (EW) as a warfare area on equal status to the traditional anti-submarine (ASW), anti-aircraft warfare (AAW), and anti-surface warfare (ASUW) areas.

Control of the electromagnetic spectrum is becoming as critical as the control of the physical battlefield. Electronic counter-measures (ECM) such as jamming and deception are technical options available to the commander. Either adversary may elect to respond to the ECM threat by a series of electronic counter-counter measure (ECCM) techniques. Anti-jam (AJ) communications can employ a variety of techniques such as spread spectrum, power control, adaptive coding and feedback, multiple routes, and adaptive antenna arrays. A signal

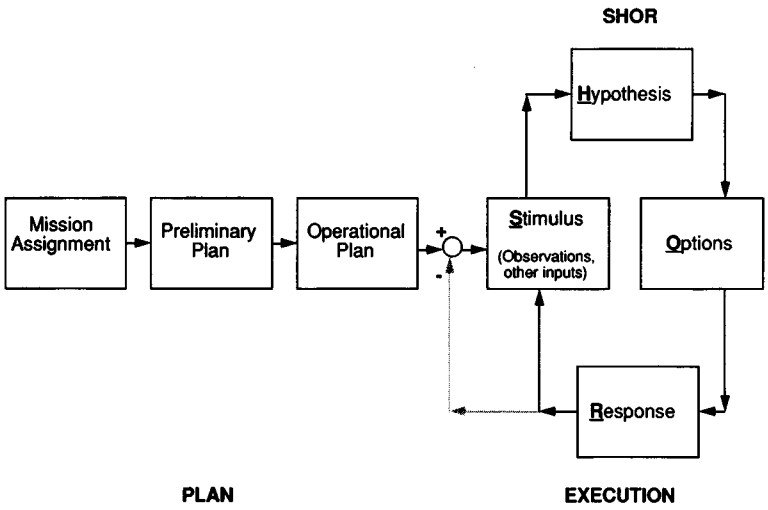


FIGURE 103.3 The planning and execution phases of operations.

may also be protected by making it difficult to intercept; some low probability of intercept (LPI) methods are again spread spectrum, directive antennas, power control, EM propagation strategies, and message brevity.

The SHOR paradigm has important advantages. First, it is generally applicable to all military echelon levels. Second, it represents a control process with its explicit dynamics rather than a relational or physical interconnection of system components. Finally, it puts focus on the roles of controlling and decisionmaking without a pre-bias on whether that function should be performed by humans or computers. The remaining challenge is to be able to describe both human and computer performance with a common type of representational framework.

103.5 The Role of the Human Decisionmaker in C^3

Designers of C3 systems often fail to acknowledge the fact that the "central, essential ingredients in any command and control system are not the things which they plan and design; rather they are the commanders and decisionmakers themselves" [Wohl, 1981]. Despite its centrality, designation of human roles is seemingly arbitrary and often controversial. In most system studies, the human decisionmaker is not thought of as an integral part of the system, but is instead given an external position as a "user" of data or an "input" to the rest of the system. Without a means of integrating the behavior of interrelated decisionmakers into a comprehensive description of system response, the proper hominal role is difficult to determine. To justify and support human action, a clear understanding of the benefits and limitations of human intervention is required.

The complexity and unpredictability of a C3 environment prompt the inclusion of hominal blocks. The ability to respond to changing operational conditions requires "intelligence," and in a C3 system this intelligence is distributed between people and algorithms. The human has a marvelous capacity for coping with vague and confusing data, making sense out of information so fragmentary that it would paralyze a computer. A computer information processing algorithm has, in turn, an unexcelled capability to process and display data at a rate that would bewilder a person. Proper marriage of humans and computers yields a robust system, quick to adapt to changes and capable of handling high data rates. For example, for the various subtasks found in the network management component of a C3 system, the relative roles of people and algorithms might be that shown in Fig. 103.4. With the advent of open system architectures, network management appears as a crucial resource allocation function. High speeds and large databases are best left within the domain of the computer, while those nodes demanding insight appropriately have a corporeal flavor.

A comprehensive C2 model is created by bringing together models of subsidiary elements. The form of these submodels should be as compliant as possible within constraints imposed by tractability. In any event, the model should display:

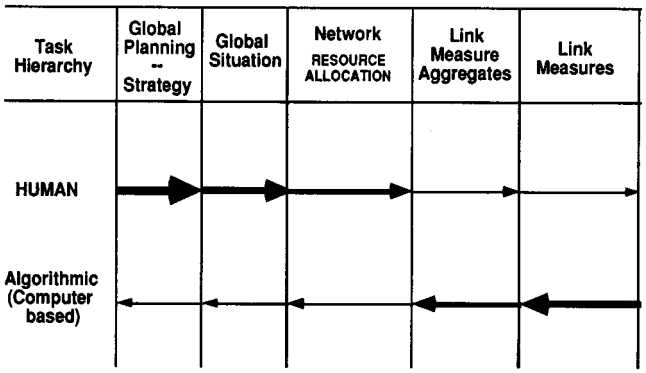


FIGURE 103.4 Control hierarchies.

1. An analytical structure permitting the evaluation of influence functions
2. Explicit communication dependence
3. Amenability to aggregation and disaggregation

Each of these desiderata arises in studies within the field of System Science, and this discipline would appear to provide the natural formalism for quantitative investigations of command and control systems. Athans articulates this view by observing that C3 systems "are characterized by a high degree of complexity, a generic distribution of the decision-making process among several decision making 'agents,' the need for reliable operation in the presence of multiple failures, and the inevitable interaction of humans with computer-based decision support systems and decision aids" [Athans, 1987]. It needs to be emphasized, however, that a C2 system differs from those commonly encountered in system theory in at least three primary ways:

1. Because command and control is at its essence a human decisionmaking activity, it is not sufficient to model only the sensors, computers, displays, etc. The hominal dynamics must be integrated with those of the electromechanical elements.
2. Any effective C3 system must have the capacity to evolve over time. Such systems are frequently established with a limited set of elements. Either for a specific operation or during their lifetime a subset of these elements will be modified or replaced, and their roles expanded or constricted as changing demands are placed upon the system. Hence, the system description must be more flexible than those in common use.
3. In contrast to conventional system design problems, there is no single nominal operating condition about which the system is maintained. Indeed, the critical attribute of a C3 system is its ability to respond to major changes in condition or state. In two Middle East naval events (USS *Stark*, *Vincennes*) the missile defense systems were set for a state that had just immediately changed. Furthermore, the system is often used in environments quite different from those envisioned in its design. Hence, the uncertain circumstances within which the decisionmakers must accomplish their tasks must be properly reflected in any system architecture.

A commander brings special skills to such a system, but some of them are difficult to quantify. For example, people have singular competence in:

1. Decisionmaking in semantically rich problem domains
2. Analogical reasoning and problem structuring
3. Information processing and application of heuristics

To properly identify a specific function for a human decisionmaker, the advantage accruing to his inclusion must be shown. Quantitative models of human responses have been developed in various ways, from ad hoc to purely normative. In the most promising of these, the form of the response dynamics of an individual commander is determined from the solution to an optimization problem. The optimization problem is framed by supposing that the decisionmaker strives to act in the most effective way, but is constrained by both cognitive

limitations and temporal pressures. When the decisionmaker's milieu and motivation are expressed in an analytical framework containing both the exogenous influences of the conflict and the endogenous predispositions generated by training and personal inclination, the input-output relation for the commander is, in principle, expressible as a set of differential equations with logical branching.

This fundamental modeling philosophy has been used successfully by several investigators. Wohl developed the SHOR model of decisionmaker action using the ideas from modern systems theory. The SHOR, in conjunction with planning models (see Fig. 103.3) can be phrased in analytical terms compatible with those of the electromechanical subsystems. With their common form, all of the submodels can be combined to create a comprehensive system description, integrating people with hardware and software algorithms. This model is useful in system architecture studies because it is applicable to all military echelon levels; it represents the fast dynamics of the system explicitly rather than by implicit relational blocks or physical interconnection of subsystem elements, and there is flexibility to allow whether a function is best performed by a human or by an algorithm.

A decisionmaker views a dynamic encounter as a temporally varying, geographically dispersed system subject to unpredictable events, both continuous and discrete. Because critical command decisions have an extended period of influence, the actions taken at different time scales cannot be isolated from each other. This issue of scale interaction comes to the fore particularly when hominal modeling is considered. In contrast to inanimate objects which usually have a single, natural time scale, the demands on a commander transcend the time scale divisions. A trained decisionmaker exhibits a wide spectrum of behaviors as both his tasks and operating environments change; the commander is the truly adaptive block in a command and control architecture. Athans referred to C3 systems as "event driven" because major changes in an engagement occur at isolated times and modulate the more frequent local irregularities [Athans, 1987]. He suggested that the proper model would be a hybrid in which "the state variables are both continuous and discrete." In this metapartitioning of the comprehensive state space, the discrete states represent global (or macro) occurrences that modulate the local (or micro) aspects. This decomposition is useful in formulating the human response model because people react differently in different time scales. The reaction to local phenomena has a reflexive quality. It is in this reaction to the infrequent, but pivotal, macroevents that the idiosyncracies thought to be particularly human are manifest.

To capture hominal behavior analytically, a framework delineating the intrinsic features of a C3 environment is required. At the macrolevel, the important attributes of a command and control environment are tempo, uncertainty, and complexity. The mission directed decisionmaker model (MDDM) described in Clapp and Sworder [1992] decomposes the C2 model in the hybrid form suggested by Athans. One block in the MDDM, the stimulus-hypothesis evaluation model (SHEM), quantifies relevant features of an engagement while representing the observation and situation assessment tendencies of the decisionmaker in terms of a few natural parameters. Because of its simple structure, the SHEM lends itself to the analysis of systems containing human decisionmakers.

To be more specific, the C2 environmental model must be flexible enough to portray the sudden, large-scale variations in circumstances which occur in operations. It is advantageous to phrase the model in such a way as to make explicit its dependence on events of macroscopic scale as well as the decisionmaker's response. The engagement model used in the MDDM has the form:

$$(d/dt)x_p = f(x_p, u_p, r_t) + g(x_p, u_p, r_t)w_t$$

where x_p is the "global" system state vector representing the external environment to which the decisionmaker seeks to respond. The decisionmaker's action variable is u_p. The process $\{w_t\}$ represents only one portion of the primitive randomness in the encounter—that associated with high-frequency uncertainty and various local disturbances. The supplementary process, $\{r_t\}$, indicates the mode of evolution of the encounter. Transitions in $\{r_t\}$ thus signify extensive events. These macro-events tend to have more temporal structure than that displayed by $\{w_t\}$, but the times of occurrence are typically unpredictable. Different values of r_t (sometimes called supervariables) are identified with different hypotheses delineating the macrostatus of the encounter. It is usually assumed that the number of modal hypotheses is finite.

Even with the aggregation implicit in the engagement model, the encounter dynamics are complex and nonlinear. A decisionmaker mentally converts the engagement dynamics into a hybrid equation with separate descriptions of local and global aspects. The input-output dynamics of the commander are expressed as an ordinary differential equation with updates at observation times. In Sworder et al. [1992], the ability of the SHEM to predict the response of a trained decisionmaker was investigated. An experiment measuring the proficiency of trained air-defense officers in differentiating hostile from friendly targets confirmed the utility of the SHEM.

103.6 Summary

C3I systems, commanders/decisionmakers, and decision aids all have a common performance objective. They must contribute to accurate and timely *situation assessments and responses* in scenarios that have a wide range of tempos, noise, clutter, uncertainty, and complexity.

The C3 system necessarily has the ability to rapidly acquire, process, and transfer large volumes of data over extended regions. Trained, experienced human decisionmakers excel at assessing complex patterns in highly cluttered environments and determining appropriate responses. Decision aids perform as a "smart" interface between these two dissimilar players. Electronics technology provides the means for designing increasingly capable C3 systems and is at its most effective when the system architecture allows flexible and dynamic interoperation of the various hardware "devices" with their trained and motivated decisionmakers.

Defining Terms

Command, control, communications (C^3): The process of mapping assets (resources available to the military commander) into capabilities. This control process is impacted by tempo, noise/clutter, and scenario complexity.

Decisionmaking: A commander's or operator's action that changes the status of his information or other assets under his control.

Doctrine: A formalized description of military mission definitions to include the procedures to accomplish those missions. Doctrine will also often specify the organizational structure appropriate to the specific mission.

Electronic warfare: Contention for the control of the electromagnetic (EM) spectrum, to allow active and passive EM sensing and communications while denying the same ability to adversaries. Includes deceptive EM techniques.

Environment: A set of objects outside the system, a change in whose attributes affects, and is affected by, the behavior of the system.

Information warfare: The protection, manipulation, degradation, and denial of information to include the traditional electronic warfare.

Intelligence: The aggregated and processed information about the environment, including potential adversaries, available to commanders and their staff.

Open system architecture: A layered architectural design that allows subsystems and/or components to be readily replaced or modified; it is achieved by adherence to standardized interfaces between layers.

Programmable radio system: Radios based on digital waveform synthesis and digital signal processing to allow simultaneous multiband, multiwaveform performance.

System: A set of objects with relations between them and their attributes or properties. It is embedded in an environment containing other interrelated objects.

Related Topics

70.1 Coding • 102.2 Communications Satellite Systems: Applications

References

M. Athans, "Command and control (C2) theory: A challenge to control science," *IEEE Trans. on Automatic Control,* vol. AC-32, pp. 286–293, April 1987.

G.A. Clapp and D.D. Sworder, "Command, control and communications: The human role in military C3 systems," in *Control and Dynamic Systems, Advances in Theory and Applications*, vol. 52, New York: Academic Press, 1992, pp. 513–541.

S. Johnson and M. Libicki, Eds., *Dominent Battlespace Knowledge: The Winning Edge,* Washington, D.C.: National Defense University Press, U.S. Government Printing Office, 1995.

Joint Chiefs of Staff (JCS), Publication 1, "Definitions," undated.

M.C. Libicki, *What is Information Warfare?,* Washington, D.C.: National Defense University Press, 1995.

National Institute of Standards and Technology [NIST], FIPS 151-1, POSIX: Portable Operating System Interface for Computer Environments (IEEE 1003.1–1988) March 1990.

D.D. Sworder, G.A. Clapp, and R. Vojak, "A Dynamic Input-Output Model of the Decisionmaking Process," Proceedings of the 1992 Symposium on Command and Control Research, Monterey, Calif., June 1992.

J.W. Wohl, "Force management requirements for air force tactical command and control," *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-11, pp. 618–639, Sept. 1981.

Further Information

W. Stallings, *Handbook of Computer-Communications Standards*, vol. 1, The Open Systems Interconnections (OSI) and OSI-Related Standards, New York: Macmillan, 1987.

W. Stallings, *Handbook of Computer-Communications Standards*, vol. 3, Department of Defense (DOD) Protocol Standards, New York: Macmillan, 1988.

Information Technology for Command and Control, S. Andriole and S. Halpern, Eds., IEEE Press, New York, 1991.

SIGNAL, a monthly (trade) magazine published by the Armed Forces Communications-Electronics Association (AFCEA), Annandale, Va. Contains numerous brief articles on current C3I topics of interest.

T.P. Coakley, *Command and Control for War and Peace*, National Defense University, U.S. Government Printing Office, Washington, D.C., 1992.

A.D. Hall, *Metasystems Methodology,* Oxford, England: Pergamon Press, 1989.

Consistency

Consistency is a concept that can be easily understood if we first define the set of all possible models that could have produced the *a posteriori* information **y**, in accordance with the class of measurement noise:

$$\mathcal{S}(\mathbf{y}) \triangleq \left\{ g \in \mathcal{S} \,\middle|\, \mathbf{y} = E(g, \eta),\ \eta \in \mathcal{N} \right\} \tag{104.40}$$

Therefore, $\mathcal{S}(y) \subset \mathcal{S}$ is the smallest set of models, according to all the available input data (*a priori* and *a posteriori*), that are indistinguishable from the point of view of the input information. This means that with the knowledge of ($\mathbf{y}$, $\mathcal{S}$, $\mathcal{N}$) there is no way to select a smaller set of candidate models. The "size" of set $\mathcal{S}(\mathbf{y})$ places a lower bound on the identification error, which cannot be decreased unless we add some extra information to the problem. This lower bound on the uncertainty error holds for any identification algorithm and represents a type of *uncertainty principle* of identification theory.

The *a priori* and *a posteriori* information are consistent if and only if the set $\mathcal{S}(\mathbf{y})$ is non-empty; otherwise, there is no model in $\mathcal{S}$ that could have possibly generated the experimental output.

Identification Error

The *a priori* knowledge of the real system and measurement noise present in the experiment **y** is stated in terms of sets $\mathcal{S}$ and $\mathcal{N}$. The statement of the problem does not assign probabilities to particular models or noise; therefore, it is deterministic in nature. In addition, the modeling error should be valid no matter which model $g \in \mathcal{S}$ is the real plant (or $\eta \in \mathcal{N}$ the real noise vector) that induces a worst-case approach. In this deterministic worst-case framework, the identification error should "cover" all models $g \in \mathcal{S}$ that combined with all possible noise vectors $\eta \in \mathcal{N}$, are consistent with the experiments, i.e., $\mathcal{S}(\mathbf{y})$. In practice, however, the family of models conservatively covers this "tight" uncertainty set. Hence, it provides an upper bound for the distance from a model to the real plant. In this framework, the worst-case error is defined as follows:

$$d(\mathcal{A}) \triangleq \sup_{\eta \in \mathcal{N},\, g \in S} m\left\{ g,\ \mathcal{A}\left[\mathrm{E}(g, \eta), S, \mathcal{N} \right] \right\} \tag{104.41}$$

where $m(\cdot,\cdot)$ is a specific metric.

The identification algorithm $\mathcal{A}$ maps both *a priori* and *a posteriori* information to a candidate nominal model. In this case, the algorithm is said to be *tuned* to the *a priori* information; otherwise, if it only depends on the experimental data, it is called *untuned*. Almost all classical parameter identification algorithms ([6]) belong to the latter class.

The identification error (Eq. (104.41)) can be considered as *a priori*, in the sense that it takes into account all possible experimental outcomes consistent with the classes $\mathcal{N}$ and $\mathcal{S}$ *before* the actual experiment is performed. Since it considers all possible experimental data **y**, it is called a *global* identification error. A *local* error that applies only to a specific experiment **y** can be defined as follows:

$$e(\mathcal{A}, \mathbf{y}) = \sup_{g \in S(\mathbf{y})} m\left[g, \mathcal{A}(\mathbf{y}, S, \mathcal{N}) \right] \tag{104.42}$$

Clearly, we always have $e(\mathcal{A}, \mathbf{y}) \le e(\mathcal{A})$. To decrease the local error more experiments need to be performed, whereas to decrease the global error new *types* of experiments, compatible with new *a priori* classes, should be performed, for example, reducing the experimental noise and changing $\mathcal{N}$ accordingly.

Convergence

Now, what happens with the family of models when the amount of information increases? It is desirable to produce a "smaller" set of models as input data increases, i.e., model uncertainty should decrease. The set of models are expected to tend to the real system when the uncertainty of the input information goes to zero. Hence, an identification algorithm $\mathcal{A}$ is said to be convergent when its worst-case global identification error

$e(\mathcal{A})$ in Eq. (104.41) goes to zero as the input information tends to be "completed." The latter means that the "partialness" and "corruption" of the available information, both *a priori* and *a posteriori*, tend to zero simultaneously.

Input information is corrupted by measurement noise. Thus, "corruption" tends to zero when the set $\mathcal{N}$ is a singleton $\mathcal{N} = \{0\}$. On the other hand, partialness of information can disappear in two different ways. By *a priori* assumptions when the set $\mathcal{S}$ tends to have only one element (the real system) or *a posterior* measurements when the amount of experimental information is completed by the remaining (usually infinite) data points. This can be unified as follows. The available information (*a priori* and *a posteriori*) is completed when the consistency set $\mathcal{S}(\mathbf{y})$ tends to only one element: the real system. Hence, an identification algorithm $\mathcal{A}$ converges if and only if

$$\lim_{size\left[S(\mathbf{y})\right]\to 0} e\left(\mathcal{A}\right) = 0 \tag{104.43}$$

Note that as the consistency set $\mathcal{S}(\mathbf{y})$ reduces to a single element, the experiment operator tends to be invertible. Since the identification error is defined in a worst-case sense, its convergence is uniform with respect to the *a priori* sets $\mathcal{N}$ and $\mathcal{S}$.

Algorithms and Further Research Topics

There are robust identification algorithms that consider frequency domain experiments, called $\mathcal{H}_\infty$–identification, this being the norm that measures the identification error. The two main ones are the two-stage and the interpolation algorithms. From time-domain measurements, several ℓ_1-identification procedures are available.

Due to the fact that robust identification is a currently active research area, there are yet many theoretical and computational aspects that have not been fully developed. Among others, there are problems related to identifying unstable plants and nonuniformly spaced experimental samples. Also, sample complexity is a recent research direction, as well as the mixture of time and frequency experiments and parametric and nonparametric models.

A complete description of both frequency ($\mathcal{H}_\infty$) and time (ℓ^1) domain identification algorithms and a discussion of the issues mentioned above can be found, for example in [9].

Defining Terms

BIBO stable: A system is Bounded Input Bounded Output stable if for all bounded inputs and zero initial conditions, the corresponding output is also bounded. In the case of finite-dimensional linear time invariant systems, this definition is equivalent to having all the poles of the system in the open left half plane $Re(s) < 0$.

Control oriented identification: A deterministic identification procedure that starting from experimental data generates a model consistent with both this data and some *a priori* assumptions on the class of systems under consideration.

Robust stability and performance: A given property of a system (such as stability or performance) is *robust* if it holds for a *family of systems* that represents (and contains) the nominal plant.

Robustness margin: A quantitative measure of stability, given by the distance from the nominal model representing the system, to the nearest model lacking the property under consideration. Examples are the classical gain and phase margins.

References

1. Barmish, B.R., *New Tools for Robustness Analysis*, Macmillan, 1994.
2. Bhattacharyya, S.P., Chapellat, H., Keel, L.H., *Robust Control: The Parametric Approach*, Prentice-Hall, 1995.
3. Doyle, J.C., Glover, K., Khargonekar, P., Francis, B., State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems, *IEEE Transactions on Automatic Control*, Vol. 34, 1989.

4. Gahinet, P., Apkarian, P., A linear matrix inequality approach to $\mathcal{H}_\infty$ control, *International Journal on Robust and Nonlinear Control*, 4, 421–448, 1994.
5. Iwasaki, T., Skelton, R., A complete solution to the general $\mathcal{H}_\infty$ control problem: LMI existence conditions and state-space formulas, *Automatica*, 1994.
6. Ljung, L., *System Identification: Theory for the User*, Prentice-Hall, 1987.
7. Mäkilä, P.M., Partington, J.R., Gustafsson, T.K., Worst-case control-relevant identification, *Automatica,* 31, 1799–1819, 1995.
8. Scherer, C., The Riccati Inequality and State-space $\mathcal{H}_\infty$ Optimal Control, Ph.D. Dissertation, Universitat Wurzburg, Germany, 1990.
9. Sánchez Peña, R., Sznaier, M., *Robust Systems Theory and Applications,* John Wiley & Sons, 1998.
10. Zhou, K., Doyle, J.C., Glover, K., *Robust and Optimal Control*, Prentice-Hall, 1996.

Further Information

Classical Identification:

Ljung, L., *System Identification: Theory for the User*, Prentice-Hall, 1987.
Söderström, T., Stoica, P., *System Identification*, Prentice-Hall, 1989.

ℓ_1 Optimal Control:

Dahleh, M.A., Díaz-Bobillo, I.J., *Control of Uncertain Systems: A Linear Programming Approach*, Prentice-Hall, 1995.

LQG Optimal Control:

Dorato, P., Abdallah, C., Cerone, V., *Linear-Quadratic Control: An Introduction*, Prentice-Hall, 1995.
Kwakernaak, H., Sivan, R., *Linear Optimal Control Systems,* Wiley Interscience, 1972.
Anderson, B.D.O., Moore, J.B., *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, 1990.

Robust Control:

Doyle, J.C., Francis, B., Tannembaum, A., *Feedback Control Theory*, Maxwell Macmillan, 1992.
Green M., Limebeer, D., *Linear Robust Control*, Prentice-Hall, 1995.
Morari, M., Zafirou, E., *Robust Process Control*, Prentice-Hall, 1989.
Sánchez Peña, R., Sznaier, M., *Robust Systems Theory and Applications,* John Wiley & Sons, 1998.
Zhou, K., Doyle, J.C., Glover, K., *Robust and Optimal Control*, Prentice-Hall, 1996.

Parametric Uncertainty:

Ackermann, J., *Robust Control: Systems with Uncertain Physical Parameters*, Springer-Verlag, 1993.
Barmish, B.R., *New Tools for Robustness Analysis*, Macmillan, 1994.
Bhattacharyya, S.P., Chapellat, H., Keel, L.H., *Robust Control: The Parametric Approach*, Prentice-Hall, 1995.

Software Packages:

Balas, G., Doyle, J.C., Glover, K., Parkard, A., Smith R., μ-*Analysis and Synthesis Toolbox*, The MathWorks Inc., Musyn Inc., 1991.
Gahinet, P., Nemirovski, A., Laub, A., Chilali, M., *LMI Control Toolbox*, The MathWorks Inc., Natick, MA, 1995.
Safonov, M., Chiang, R., *Robust Control Toolbox*, The MathWorks Inc., 1988.

D. McRuer "Man-Machine Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

105

Man-Machine Systems

105.1 Introduction
105.2 Several Natures of Man-Machine Control—A Catalog of Behavioral Complexities
105.3 Full-Attention Compensatory Operations—The Crossover Model
Crossover Frequency for Full-Attention Operations • Remnant • Effects of Changes in the Task Variables • Effects of Divided Attention

Duane McRuer
Systems Technology, Inc.

105.1 Introduction

In principle the dynamic behavior of the human element in man-machine systems can be described in terms similar to those used to describe other system elements. There are, however, major complications in quantification because of the enormous versatility of the human engaged, simultaneously, as the on-going *architect* and modifier of the man-machine system itself and as an operating entity within that system. In other words, the adaptive and learning capabilities of the human permit both set-up and modification of the effective system structure and the subsequent self-improvement and tuning of the human dynamic characteristics within that structure.

The situations which are simplest to quantify are those in which the *machine* has time-stationary dynamic properties and the human has, after architectural, learning, and adaptation phases, achieved a similar state. Under these circumstances human dynamic operations can be characterized by quasi-linear describing functions and a remnant [Graham and McRuer, 1961] or operator-induced noise. This is the context here.

105.2 Several Natures of Man-Machine Control—A Catalog of Behavioral Complexities

Figure 105.1 [McRuer and Krendel, 1974] shows a general quasi-linear man-machine system with time-stationary properties. This diagram is suitable for the description of human behavior in an interactive man-machine system wherein the human responds to visually sensed inputs and communicates with the machine via a manipulator of some sort (e.g., control stick, wheel, pedal, etc.). This block diagram indicates the minimum needed number of major functional signal pathways internal to the human operator to characterize different behavioral features. The constituent human sensing, data processing, computing, and actuating elements are connected as internal signal processing pathways which can be "reconfigured" as the situation changes. Such reconfiguration is an aspect of human behavior as a system architect. Functional operations on internal signals within a given pathway may also be modified.

The specific internal signal organizational possibilities depicted in Fig. 105.1 have been discovered by manipulating experimental situations (e.g., by changing system inputs and machine dynamics) to isolate different combinations of the specific blocks shown [McRuer and Jex, 1967; McRuer and Krendel 1974; McRuer 1980].

To describe the parts of the figure start at the far right with the *controlled element.* This is the machine being controlled by the human. To its left is the actual interface between the human and the machine—the neuromuscular

Hazardous Environment Robotics

Deneb Robotics, Inc. is an internationally known leader in 3D graphics-based factory simulation, telerobotics, and virtual reality software used widely in the aerospace, automotive, defense, environmental, medical, nuclear, and research communities.

Among the company's broad software product line is TELEGRIP™, which provides 3D graphical interface for previewing, interactive programming, and real-time bilateral control of remote robotic devices. It provides operators a system for safe, quick, and efficient remediation of hazardous environments from a single point of control and input that is isolated from virtually all operator hazards.

Accurate 3D kinematic models of the robot and work space components allow the operator to preplan and optimize robot trajectories before the program is automatically generated. Control commands are monitored when running in autonomous, teleoperational, or shared control modes to assure procedural safety.

A video camera provides a Deneb Robotics engineer with a view of the robot. (Photo courtesy of National Aeronautics and Space Administration.)

actuation system, which is the human's output mechanism. This in itself is a complicated feedback control system capable of operating as an open-loop or combined open-loop/closed-loop system, although that level of complication is not explicit in the simple feedback control system shown here. In the diagram the neuromuscular system comprises limb, muscle, and manipulator dynamics in the forward loop and muscle spindle and tendon organ ensembles as feedback elements. Again, many more biological sensors and other elements are actually involved; this description is intended only to be generally indicative of the minimum level of complexity associated with the *human actuation elements.* All of these elements operate within the human at the level from the spinal cord to the periphery.

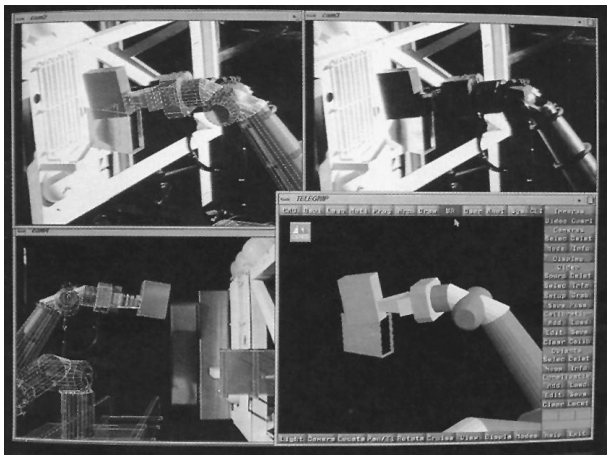
There are other sensor systems, such as joint receptors and peripheral vision, which indicate limb output position. These operate through higher centers and are subsumed in the *proprioceptive* feedback loop incorporating a block at the perceptual level further to the left in the diagram. If motion cues are present, these too can be associated in similar proprioceptive blocks with feedbacks from the controlled element output.

The other three pathways shown at the perceptual level correspond to three different types of control operations on the visually presented system inputs. Depending on which pathway is effectively present, the

A key feature of TELEGRIP is a video overlay option that utilizes video to calibrate 3D computer models with the actual environment. The video overlay technique is especially useful for on-line planning applications or teleoperations in remote, hazardous, or complex environments such as space, undersea, or nuclear sites.

A virtual reality calibration technique was developed for reliable and accurate matching of a graphically simulated environment in 3D geometry with actual video camera views. The system was designed for predictive displays with calibrated graphics that overlay in live video for telerobotics applications. For example, the system allows an operator to designate precise movements of a robot arm before sending the command to execute.

Following successful test of the video overlay techniques, an agreement was concluded with Deneb Robotics that allows the company to integrate video overlay into the commercially available TELEGRIP to expand its use in hazardous environment robotics. (Courtesy of National Aeronautics and Space Administration.)



The operator can view the video image of the real world environment (upper right) and the computer's interpretation of the same scene using TELEGRIP. (Photo courtesy of National Aeronautics and Space Administration.)

control structure of the man-machine system can appear to be *open-loop*, or *combination open-loop/closed-loop*, or totally *closed-loop* with respect to visual stimuli.

When the *compensatory* block is appropriate at the perceptual level, the human controller acts in response to errors or controlled-element output quantities only. Only the Y_{pe} block "exists", with Y_{pi} and the precognitive block both equal to zero. With the compensatory pathway operational, continuous closed-loop control is exerted on the machine so as to minimize system errors in the presence of commands and disturbances. **Compensatory behavior** will characteristically be present when the commands and disturbances are random-appearing and when the only information displayed to the human controller consists of system errors or machine outputs. In the simple case where the describing function Y_{pe} is defined so as to account for the perceptual and neuromuscular components, the system is single-input/single-output, and the operator-induced noise is neglected, the closed-loop system output/input dynamics will be

$$\frac{m}{i} = \frac{Y_{pe}Y_c}{1 + Y_{pe}Y_c} \tag{105.1}$$

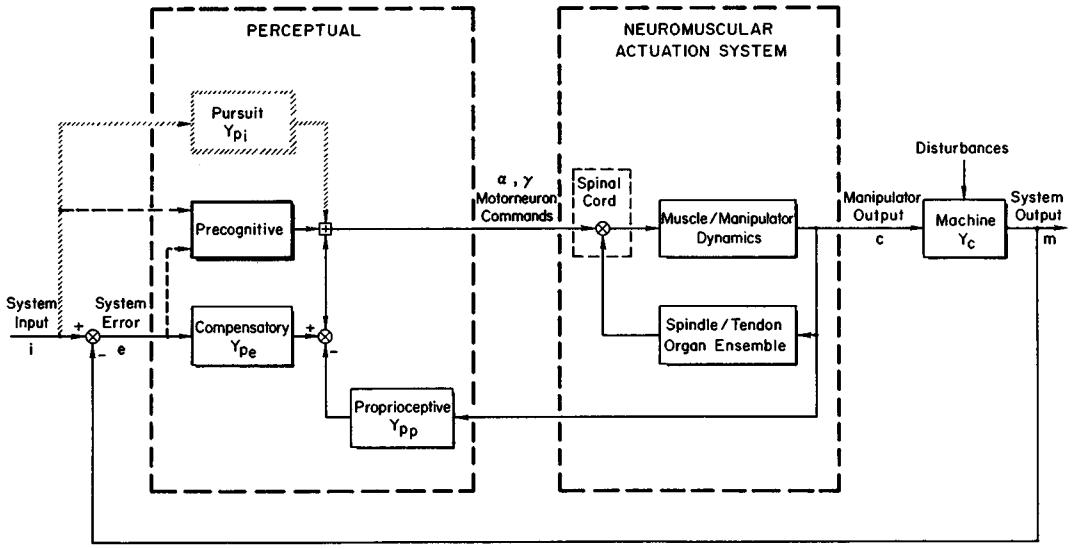


FIGURE 105.1 Major human operator pathways in a man-machine system.

and the error/input

$$\frac{e}{i} = \frac{1}{1 + Y_{pe}Y_c} \tag{105.2}$$

Thus, for compensatory situations, the man-machine system emulates the classic single-input/single-output feedback system. The output can be made to follow the input and the error can be reduced only by making the open-loop describing function large compared to 1 over the operating bandwidth of the system.

When the command inputs can be distinguished from the system outputs by virtue of the display (e.g., *i* and *m* are shown or detectable as separate entities relative to a reference) or preview (e.g., as in following a curved course) the *pursuit* block in Fig. 105.1 comes into play and joins the compensatory. The introduction of this new signal pathway provides an open-loop control in conjunction with the compensatory closed-loop error correcting action. The output/input dynamics of the man-machine system will then become

$$\frac{m}{i} = \frac{(Y_{pi} + Y_{pe})Y_c}{1 + Y_{pe}Y_c} \tag{105.3}$$

and the error/input describing function is

$$\frac{e}{i} = \frac{1 - Y_{pi}Y_c}{1 + Y_{pe}Y_c} \tag{105.4}$$

With the pursuit system organization the error can be reduced by the human's operations in two ways: by making the open-loop describing function large compared with 1 and by generating a pursuit path describing function which tends to be the inverse of the controlled element. This can, of course, only be done over a limited range of frequencies. The quality of the overall control in the pursuit case can, in principle, be much superior to that where only compensatory operations are possible.

An even higher level of control is possible. When complete familiarity with the controlled element dynamics and the entire perceptual field is achieved, the highly skilled human operator can, under certain conditions, generate neuromuscular commands which are deft, discrete, properly timed, scaled, and sequenced so as to result in machine outputs which are almost exactly as desired. These neuromuscular commands are selected

from a repertoire of previously learned control movements. They are conditioned responses which may be triggered by the situation and the command and control quantities, but they are not continuously dependent on these quantities. This pure open-loop programmed-control-like behavior is called **precognitive**. Like the pursuit pathway, it often appears in company with compensatory follow-up or simultaneous operations. This forms a dual-mode form of control in which the human's manual output is initially dominated by the precognitive action, which does most of the job, and is then completed when needed by compensatory error-reduction actions.

The above description of human action pathways available in man-machine systems has emphasized the visual modality. Similar behavior patterns can be exhibited to some extent in other modalities as well. Thus the human's interactions with machines can be even more extraordinarily varied than described here and can range completely over the spectrum from open-loop to closed-loop in character in one or more modalities.

105.3 Full-Attention Compensatory Operations—The Crossover Model

The compensatory pathways with manual control operations using the visual modality have been extensively studied. Thousands of experiments have been performed, and most of the adaptive features of human behavior associated with these kinds of operations are well understood. There are both classical control [e.g., McRuer and Krendel, 1974; and McRuer et al., 1990] and optimal control [e.g., Baron and Kleinman, 1969; Kleinman et al., 1970; Curry et al., 1976; and Thompson, 1990] theoretical formulations available to predict steady-state and dynamic performance.

By far the simplest human behavioral "law" for compensatory systems is the *crossover model*. This states that, for a particular controlled element transfer function, Y_c, the human operator adopts a describing function, Y_{pe}, such that the open-loop man-machine transfer characteristics appear as

$$Y_p Y_c = \frac{\omega_c e^{-\tau j\omega}}{j\omega} \tag{105.5}$$

The two parameters in the crossover model are the crossover frequency, ω_c, and an effective pure time delay, τ. The model applies only in the immediate region of the crossover frequency. The typical data shown in Fig. 105.2 illustrate how well this relationship is obeyed for a variety of subjects and a particular controlled element. The agreement with the amplitude ratio is excellent over a broad range of frequencies. The phase agreement is good in the region of the crossover frequency, ω_c, but departs somewhat at lower frequencies. Figure 105.2 also shows the *extended crossover model*. Here the effects in the crossover region of a potentially large number of low-frequency lags and leads (in the machine and/or the operator) are represented by a phase contribution given by $\exp(-j\alpha/\omega)$. Here the time constant $1/\alpha$ is a lumped-constant representation of myriad low-frequency phase characteristics. It is an appropriate approximation *only* in the general region of crossover and is not intended to extend to extremely low frequencies.

Fundamentally, the crossover model states that the human's transfer characteristics will be different for each set of machine dynamics, but that the form of the composite total open-loop dynamics will be substantially invariant. The effective time delay in Eq. (105.5) is a low-frequency approximation to the combination of all manner of high-frequency pure delays, lags, and leads, including a component representing the effects of

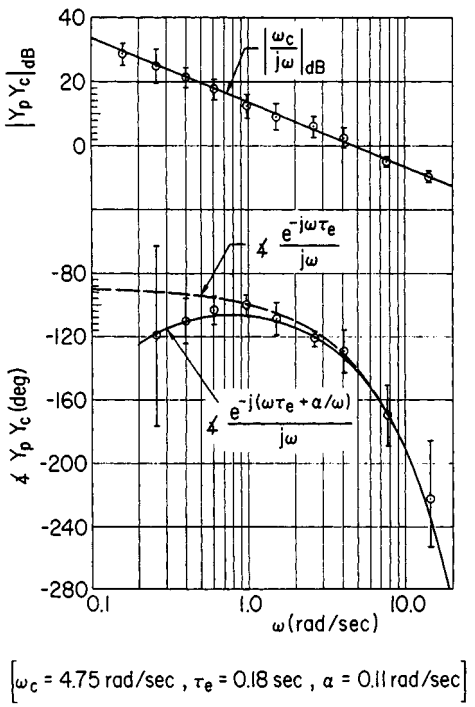


FIGURE 105.2 Data and crossover models for a simple rate-control-like controlled element.

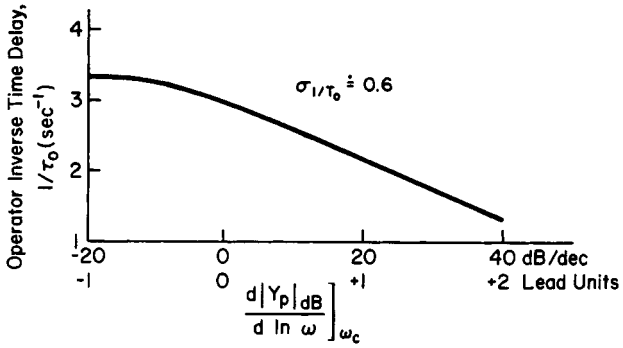


FIGURE 105.3 Variation of crossover model dynamic stimulus-response latency with degree of operator lead equalization.

the neuromuscular actuation system reflected to the crossover region. It follows that the effective time delay, τ, is not a constant. Its two major components are (1) the effective composite time delay of the controlled element (including manipulator effects)—the sum of the machine's lags minus leads at frequencies well above crossover and (2) the high-frequency dynamics of the human operator approximated by a pure delay which has an equivalent phase shift at frequencies within the crossover region. The latter includes a minimum of 0.1 second for the neuromuscular system and an additional increment which depends on the amount of lead generation required of the human to offset the controlled element deficiencies in order to make good the crossover model form. Figure 105.3 [McRuer and Krendel, 1974] shows this variation for a wide range of controlled elements (the neuromuscular delay component is included). More refined estimates are available [e.g., McRuer et al., 1990], but the above description is suitable for first-order estimates of behavior and dynamic performance.

Crossover Frequency for Full-Attention Operations

The crossover frequency tends to be constant for a given set of task variables (controlled-element form, inputs, disturbances, etc.). For example, as a controlled-element gain is changed, the human will change gain to compensate, resulting in the same crossover frequency. The maximum attainable crossover frequency, ω_u, will be

$$\omega_u = \frac{\pi}{2\tau} \tag{105.6}$$

This corresponds to zero phase margin. The nominal crossover frequency and associated pilot gain can be estimated from the condition to provide minimum mean-squared error in the presence of the appropriate form of continuous attention remnant. "Remnant" is operator-induced noise; as described below it depends on the nature of the operator's equalization and is larger when low-frequency lead is required to make good the crossover model. Thus, the need to generate lead impacts both the effective time delay and the remnant and, accordingly, the crossover frequency for which the minimum mean-squared error is obtained. The nominal crossover frequency for full-attention operations can be estimated [McRuer et al., 1990] using

	ω_c/ω_u
No Operator Lead	0.78
Low-Frequency Operator Lead	0.66

Remnant

The second component of the operator's response is operator-induced noise or remnant. Remnant can, in principle, result from several sources, but in single-loop systems with ideal linear manipulator characteristics

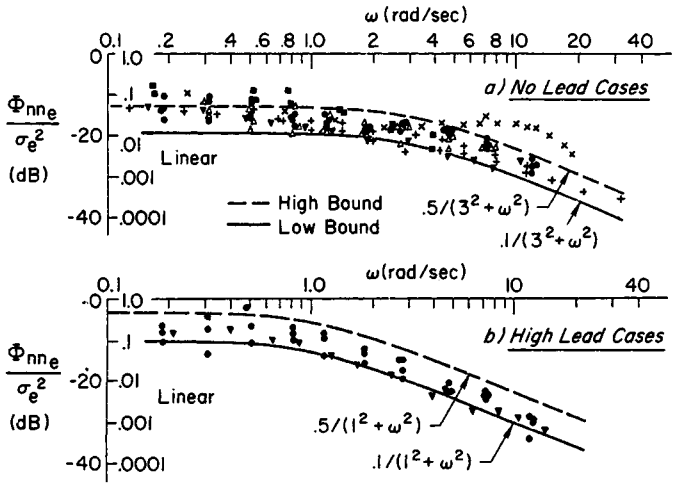


Figure 105.4 Normalized remnant spectra.

and no significant nonlinearities in the controlled element, the basic cause appears to be random time-varying behavior within the operator, which can be thought of as continuous random fluctuations in the effective time delay. The remnant can be described as a continuous, relatively broadband, power spectral density. Fig. 105.4 provides a cross-section of remnant data from several sources. It is very important to note that the magnitude of the power spectral density scales approximately with the mean-squared error.

Effects of Changes in the Task Variables

The task variable which has the most important effect on the trained operator's behavior is the controlled element dynamics. Indeed, the natures of human adaptive changes in adjusting to the controlled element is the main thrust of the crossover model and remnant discussion above. More generally, task variables other than the machine dynamics, as well as environmental and operator-centered variables, can change operator gain, and hence crossover frequency, effective time delay, and remnant. Accordingly, ω_c and τ variations become quantification measures of changes or differences in the task, environmental, and operator-centered variables expressed directly in terms of the operator's control actions.

A common example is the reduction of crossover frequency when the amplitude of the command or disturbance signals are very small. This reflects the human's indifference to small errors and constitutes the principal human behavioral nonlinearity in the crossover model context. Another example occurs in measuring the effects of training, where ω_c increases with trials until stable conditions are obtained for that particular subject and set of constant task and environmental conditions. Similarly, operator gain and remnant can be modified as a consequence of changes in operator-centered variables. A notable example is the decrease in gain and increase in remnant which accompanies alcohol ingestion.

Effects of Divided Attention

Human operators in man-machine systems are, in general, involved in two types of operations—control tasks and a diverse combination of monitoring/supervising/communicating/data-gathering/decision making activities referred to as "managerial tasks." While the operator's attention is "divided" between the control and managerial tasks, these are often performed nearly simultaneously as parallel processing operations.

By definition, control workload is highest when the operator's full attention is required for control purposes and when this attention is focused on only the most critical input information needed for closed-loop control. For this reason the full-attention crossover model and remnant for compensatory behavior treated above has received the major attention here. Estimates and considerations based on full-attention compensatory assumptions will generally be conservative. For instance, the dynamic performance of the overall man-machine system will typically be improved when additional cues and information provide the basis for the generation of pursuit behavior.

For a given situation the minimum divided attention level should be established by the demands of the control task. When divided attention conditions are present in compensatory situations the major effects on the control performance are reduced crossover frequency and increased system error. To a first order the divided attention effects on average crossover frequency are given in Fig. 105.5. Here the "control dwell fraction," is η, the proportion of the total time spent on the control task. There are many other complications and considerations [McRuer et al., 1990], but these require more than handbook treatment.

Figure 105.5 Effect of divided attention on processing bandwidth.

Defining Terms

Compensatory behavior: Human dynamic behavior in which the operator's actions are conditioned primarily by the closed-loop man-machine system errors.

Compensatory display: For the simplest case, a display which shows only the difference between the desired input command and the system output.

Precognitive behavior: Conditioned responses triggered by the total situation; essentially pure open-loop control.

Pursuit behavior: The human operator's outputs depend on system errors, as in compensatory behavior, but may also be direct functions of system inputs and outputs. The human response pathways make the man-machine system a combined open-loop, closed-loop system.

Pursuit display: In the simplest case, a display which shows input command, system output, and the system error as separable entities.

Related Topics

100.3 Frequency Response Methods: Bode Diagram Approach • 100.7 Nonlinear Control Systems

References

S. Baron, and D.L. Kleinman, "The Human As An Optimal Controller and Information Processor," NASA CR-1151, 1969.

R.E. Curry, W.C. Hoffman, and L.R. Young, "Pilot Modeling for Manned Simulation," AFFDL-TR-76-124, 1976.

D.Graham and D. McRuer, *Analysis of Nonlinear Control Systems,* New York: John Wiley & Sons, 1961 (also Dover, 1971).

D.L. Kleinman, S. Baron, and W.H. Levison, "An optimal control model of human response," *Automatica,* vol. 9, no. 3, 1970.

D.T. McRuer, "Human dynamics in man-machine systems," *Automatica,* vol. 16, no. 3, 1980.

D.T. McRuer, W.E. Clement, P.M. Thompson, and R.E. Magdaleno, "Pilot Modeling for Flying Qualities Applications," WRDC-TR-89-3125, vol. II, 1990.

D.T. McRuer, and H.R. Jex, "A review of quasi-linear pilot models," *IEEE Trans. Human Factors in Electronics,* vol. HFE-8, no. 3, 1967.

D.T. McRuer, and E.S. Krendel, "Mathematical Models of Human Pilot Behavior," AGARD-AG-188, 1974.

P.M. Thompson, "Program CC's Implementation of the Human Optimal Control Model," WRDC-TR-89-3125, vol. III, 1990.

Further Information

The references of the chapter, especially Kleinman et al. [1970], McRuer and Krendel [1974], and McRuer et al. [1990], comprise a good cross section of detailed information on modeling aspects of man-machine systems. An excellect general text is T.B. Sheridan and W.R. Farrell, *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*, Cambridge: MIT Press, 1974.

Encyclopedic coverage appears in K.R. Boff, L. Kaufman, and J.P. Thomas, *Handbook of Perception and Human Performance*, New York: Wiley, 1986, and K.R. Boff and J.E. Lincoln, "Engineering Data Compendium: Human Perception and Performance," Harry G. Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, 1988.

The aperiodic proceedings of the so-called "Annual Manual" contain a great deal of information about man-machine system developments. Since 1965 these have been published by NASA as SP's (NASA Special Publications) under the general heading of *NASA—University Conference on Manual Control.*

The text article emphasizes the dynamic behavior of the human, not the design of machine dynamics to achieve optimum characteristics in terms of man-machine system dynamic performance and human subjective approval. For these aspects of design, a comprehensive summary of models, references, and applications appears in "Advances in Flying Qualities," *AGARD Lecture Series LS-157,* 1988. Although the applications there are specifically for aerospace vehicle control, the principles illustrated apply to vehicles in general and to other machines subject to continuous control by a human operator.

As with other feedback control systems, system stability is a major consideration. In spite of the extraordinary adaptive properties intrinsic to human controllers, system instability is a rare but often unfavorable event. The nature of such man-machine oscillations and the design steps required to avoid them is treated extensively in Duane McRuer, *Pilot-Induced Oscillations and Human Dynamic Behavior,* NASA Contractor Report 4683, July 1995.

Boehmer, L.S. "Vehicular Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

106

Vehicular Systems

106.1 Introduction
106.2 Design Considerations
106.3 Land Transportation Classifications
106.4 Propulsion
106.5 Microprocessor Controls
106.6 Monitoring and Diagnostics

Linda Sue Boehmer
LSB Technology

106.1 Introduction

Vehicular systems have evolved and incorporated advances from many other fields of technology over the past decade. Instrumentation and controls for the various modes (aircraft, marine vessels, cars, trucks, buses, and rail vehicles) resemble each other more every year. Technology from one mode of transportation used to be of little interest to practitioners of any other mode. A technology historian might notice similarities among the functions of airport beacons, lighthouses, traffic lights, and railroad signals, but the specialists in each field had little to say to each other. This is no longer the case. Computers, microprocessor controls, electronics, GPS, and advanced networking and radio technologies are being applied in all forms of passenger and freight transportation, from aviation to marine, highway, and rail transport. The vehicles are now considered in context with an entire system within a particular mode, which is increasingly viewed as part of an overall transportation environment encompassing more than one mode.

Although "multimodal" is a term that was coined by policy makers to facilitate equitable distribution of funding among transportation modes and to facilitate interfaces among them, it applies equally well to the supporting technologies of the original modes. All modes now utilize microprocessor controls in their subsystems. With microprocessor control has come additional diagnostic capability and the use of system level intelligence, linking all intelligent subsystems and analog sensors and controls. Propulsion of vehicles now varies by mode less than it used to. Because of microprocessors, propulsion can be controlled more precisely, allowing vehicles to use non-traditional energy sources and to switch from one source to another easily, even automatically. We do not yet have the ideal multimodal vehicle, capable of navigating, either automatically or by a driver/operator, through air, in water, on roads, and on rails, but technology is no longer a limiting factor. We have not yet achieved the best balance among modes so that the most appropriate mode is utilized for passenger or freight transportation, but the tools exist to make those decisions possible. (This section deals primarily with land transportation. See the index for aviation and maritime applications.)

106.2 Design Considerations

If design could begin with a clean slate, the first step would be to decide which mode of transportation and which power source is best suited for the application, based on geography, priority of the passenger or cargo to be transported, energy efficiency, safety, cost per mile, and other factors. However, this is not really possible because the factors relating to funding sources and existing infrastructure often outweigh any technical considerations.

Road Engine

George B. Selden
Patented November 5, 1895
#549,160

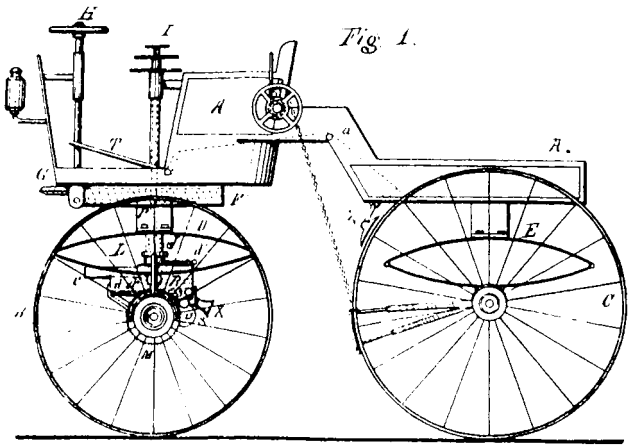
An excerpt from George Selden's patent application:

Be it known that I, George B. Selden, a citizen of the United States, residing in Rochester in the county of Monroe, in the state of New York, have invented an improved Road Engine, of which the following is a specification, reference being had to the accompanying drawings.

The object of my invention is the production of a safe, simple, and cheap road-locomotive light in weight, easy to control and possessed of sufficient power to overcome any ordinary inclination.

Perhaps one of the most famous patent lawsuits surrounded this patent applied for in 1879 by Selden, who was more of a patent attorney than an inventor. He sold the patent to Col. Allen Pope in 1899 who began to enforce the patent and won judgments against several car manufacturers—except one. They were forced to pay royalties on all vehicles they produced.

Henry Ford called the patent preposterous claiming that it didn't cover his vehicles that he had invented first. He refused to pay royalties and won a ruling in his favor in 1911, after millions had been paid out by his competitors. (Copyright © 1995, DewRay Products, Inc. Used with permission.)



Railroads, rail transit, buses, vans, and automobiles each have their own operating environments, although they are increasingly considered part of a system rather than simply as independent vehicles. Many of the underlying technologies are utilized across modal boundaries, but the basics of design remain highly dependent on mode.

Electrical systems are widely used in automobiles and trucks today and electronic systems currently make up about 10% of the value of a car. Electronic systems are currently used to control the engine, transmission, steering, braking, suspension, and traction. Many autos incorporate an integrated computer system for controlling the functions mentioned. In addition, electronic systems are used to display information such as speed and engine conditions.

Air bag inflation units use sensors and electronic controls to insure proper inflation within milliseconds after a collision.

As antilock brakes, **active suspensions**, and other computer-dependent technologies are fully utilized, electronic systems may constitute more than 20% of the value of a car. Much of the added computing power will be used for new technology for smart cars and smart roads, or ITS (intelligent transportation systems). The term refers to a varied assortment of electronics that provide real-time information on accidents, congestion, routing, and roadside services to drivers and traffic controllers. ITS encompass devices that would make vehicles more autonomous: collision-avoidance systems and lane-tracking technology that alert drivers to impending disaster or allow a car to drive itself. ITS also includes interfaces between personal traffic and mass transportation, particularly when rail traffic mixes with cars and at rail crossings, whether or not such crossings are protected by gates. Railroads and mass transit vehicles have many of the same internal subsystems as cars do, and also additional subsystems, for which cars may soon have analogous functions. The electrical and electronic systems and controls account for 15 to 30% of the cost of the vehicle.

The major vehicle subsystems include propulsion, braking, power conditioning, communication, passenger information (audio and visual), heating/air conditioning, door control, speed control, and monitoring and diagnostics. All of these subsystems interact with each other and many also interact with subsystems external to the vehicle.

Some of the initial design considerations for vehicles include:

- Will the vehicle interface with an existing fleet?
- Will the vehicle operate independently or as part of a consist, or both?
- What are the physical requirements, including dimensions, number of passengers (seated vs. standing)? Is the system (or portions of it) elevated, in tunnels, at grade, underground? Will the vehicles mix with or cross other types of traffic? Are ambient conditions exceptionally hot, cold, or dangerous?
- How closely must one vehicle (or consist) follow another and how fast must they be capable of traveling? Where, how often, and for how long will they stop?
- To what degree will the vehicles be automated; how much control will a human operator (and/or other crew) have and under what conditions?
- What kind of power is available, how will it be collected, and can energy generated during electric braking be returned to the power system?
- How will system and subsystem failures be handled, what kind of failures are acceptable, and to what degree (or for how long) will they be tolerated? How often and with what degree of expertise will the vehicles be serviced?

106.3 Land Transportation Classifications

Among land transportation vehicles, there are more subdivisions than the lay person might expect. In general, distinctions are based on vehicle size, weight, speed, and passenger or cargo capacity. Cars, trucks, mini-vans, vans, sport utility vehicles, etc. are familiar terms. There is some variation among them relative to the electronics embedded in the systems and the options available to the driver. The same is true for railroads, rail transit, and mass transit, although the distinctions among classifications are important to the manufacturers and operators of the equipment. The classifications include railroad, commuter rail, heavy rail, light rail, street car, trolley bus, bus, paratransit, and "people mover" or monorail.

106.4 Propulsion

By 1910 electric automobiles were commonplace. Nevertheless, they were replaced by gasoline-fueled automobiles by 1920 because electric cars operated at lower top speeds and over shorter ranges without recharging than gasoline cars could achieve. However, the availability of electric motive power remained a critical factor in the development of cities. Since the mid-1970s, when the electric vehicle reemerged as an appealing transportation option, many have recognized the potential of electric fleet vans. An **electric vehicle** uses electric energy storage, electric controls, and electric propulsion devices. Because the vans use batteries to drive their electric motors, they are well suited

to the short routes and regular schedules followed by vans in a company fleet. One such fleet van, the General Motors Griffon, is produced in England. Because the vans can be recharged regularly at night, they offer electric utilities a new off-peak demand. At the same time, electric vehicles run cleanly and burn no gasoline.

Increasing the distance an electric vehicle can travel on a single charge is the most significant factor in expanding the market for electric vans. The 60-mile (97-km) range of the Griffon makes it a replacement candidate for about 600,000 commercial fleet vehicles now operating in the United States. If advanced batteries doubled the range of a van to 120 miles, the potential market for these vehicles could top 2 million. A variety of electric cars has been introduced over the years, but none have enjoyed general use.

Products of combustion gradually were recognized as major air pollutants and fuels have been acknowledged as non-renewable resources, so alternatives have been sought with increasing diligence.

Today, electric motors are used more widely for rail vehicles than for cars, vans, or buses, although this is slowly changing (see above). Electric motors as a back-up mode for buses are becoming more common in certain areas where air pollution is considered a serious problem and in portions of systems, such as North American tunnels, where fumes from internal combustion can be hazardous.

Power distribution for rail vehicles ranges from three phase ac, at various voltages (usually collected from overhead wires), to several different dc voltages (usually collected from a "third rail"). Until recently (within the past 5 years in the U.S.), most electric traction motors utilized dc power. Today, most traction motors in new vehicles use ac power. Various techniques are used to cool the motors, depending on the operating environment. Collected power is conditioned continuously (see power "converters" and "inverters") to meet the motors' requirements and the power requirements of other vehicle systems, and also is stored to power critical on-board systems if power is lost.

Traction motors also are used for braking, which generates power that can be reconditioned and returned to the power system to power other vehicles or returned to the power grid. Power that cannot be returned or used elsewhere in the system is converted into heat by banks of large braking resistors. Electric braking is supplemented by mechanical braking systems which can be actuated pneumatically, hydraulically, and/or electrically. Coordination of propulsion and braking efforts, especially when traction surfaces are slippery, is an important design point.

Microprocessor controls have allowed optimization of automotive internal combustion processes to economize on fuel and minimize air pollution. Alternative fuels, such as natural gas, are becoming more common because the combustion process can be managed more uniformly, responsively, and safely than ever before.

Diesel electric locomotives have become the propulsion vehicle of choice for long haul freight railroads. Microprocessors control the combustion process which produces electricity to power electric motors when ac power is not available, such as on long sections of rail that are not yet electrified.

Dual mode buses utilize internal combustion when operating on the streets, but switch to electric power in tunnels. Dual mode rail vehicles or streetcars collect power from an overhead catenary or third rail, but can switch to battery power when they are not operating in electrified areas.

Improvements in battery technology (cost, life, power density, weight, and maintenance requirements) and solar power as a supplementary or primary source will improve the acceptance of alternatives to internal combustion and direct electric power.

106.5 Microprocessor Controls

All major vehicle subsystems and many minor ones are now microprocessor-controlled. Embedded microprocessors replace banks of relays and mechanical switches to perform functions on the vehicle and also to control functions that did not exist prior to the advent of microprocessors. Some major vehicle subsystems have several microprocessors handling different functions and coordinating analog and digital input and output signals.

Intelligent subsystems exchange information within a vehicle, among vehicles in a consist, and between the vehicle and its external environment. This information is exchanged through increasingly sophisticated networks, which may or may not use traditional wiring. There may also be a separate network or layers of error-checking to handle safety-critical data.

A human vehicle operator typically has status indicators, alarms, and controls. These have changed dramatically with the advances in microprocessor-controlled subsystems. The "glass cockpit" and "fly by wire" techniques

Power Applying Mechanism

Otto Zachow and William Besserdich
Patented December 29, 1908
#907,940

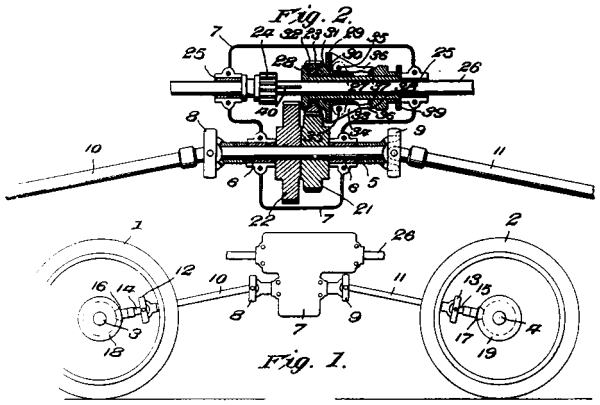
An excerpt from Otto Zachow and William Besserdich's patent application:

Our invention relates to new and useful improvements in power-applying mechanism and more particularly to that class adapted to be used in connection with motor-propelled vehicles, such as automobiles, or the like, and our object is to provide a mechanism of this class whereby the power may be applied to both the front and rear axles.

Early in this century, Otto Zachow and his friend, William Besserdich, had the idea of making all four wheels of their vehicle turn, when they slid into a ravine and got stuck. Out of frustration the four-wheel drive mechanism was invented.

At the beginning of World War I, they sold about 50 trucks equipped with the mechanism to Great Britain. When the United States entered the war, Zachow and Besserdich sold 3,750 trucks to the U.S. Army.

Today, four-wheel drive is available on a wide range of vehicles from small sport utilities to luxury imports. (Copyright © 1995, DewRay Products, Inc. Used with permission.)



developed for aviation generally transition to cars and trains as they are service-proven and as their cost decreases. A driver or train operator once had a few lights, a gauge or two, a throttle of some sort, and a brake handle or pedal. For vehicles that are not automated but are partially automated or allow manual operation at times, a human operator today is confronted by a dense array of dials, buttons, data and CCTV screens, LCD panels, microphones, and annunciators (audio and visual) of various types. In some cases there is far more information than an operator can use. In others, there is duplication between the old, analog indicators and controls and new, digital or "soft" controls on a touch screen.

There has been some resistance to technology advances, based on the perception that electronics are not as reliable as electromechanical devices (such as relays) and on a concern that they will be more complicated to troubleshoot and maintain than electro-mechanical and analog equipment.

106.6 Monitoring and Diagnostics

A great advantage of microprocessor-controlled systems is the degree to which they can be self-diagnosing. Each intelligent subsystem has internal self-diagnostics which include routines to perform initial tests on power-up, update checks for "hot" startup and continuous checking to assure that inputs and outputs are within expected ranges. Internal self-diagnostics are also capable of performing self-tests on request. When fault conditions are noted, typically there are internal resets to allow for inaccurate data, with faults being logged after a certain number of occurrences or duration of a fault condition.

The basis of any monitoring and diagnostics system is the underlying maintenance philosophy. Microprocessor controls make it possible to capture any combination of information from the intelligent subsystems. Information can then be processed and presented in a variety of ways, along with data from analog sensors which are not part of any intelligent system.

Information of interest includes operating status and existing or historical fault information. "Events" may include faults and also other expected actions that may or may not be considered faults. Historical information can include faults, events, and the status of parameters of interest associated with the fault. The amount of information that can be captured is limited only by the amount of memory provided and the speed at which it can be transferred. Typically, some (or all) of the memory is in a form that will allow the fault data to be preserved after power is lost or vehicles are shut down between operating periods.

Decisions on which information to capture, how often to sample, and how many samples to preserve are ideally based on what will be most valuable for troubleshooting existing faults and predicting future failures. A thorough understanding of each intelligent subsystem is needed in addition to an understanding of the environment in which it operates, including the other subsystems with which it interacts. It is also important to know what level of skill will be applied to interpreting the saved information. If the level of skill will be low, some degree of artificial intelligence can be designed into the diagnostics to guide a maintenance technician through the troubleshooting process.

The information needed to be collected is influenced by the target audience. The most detailed internal subsystem information is primarily used by engineers or specialized technicians to troubleshoot detailed failures or fine tune operation. A subset of less detailed information is used by maintenance staff to determine which components or sub-modules to replace or repair. A further subset of that information is used by a general troubleshooting staff to determine which subsystem is malfunctioning or which higher level modules to replace. An even smaller subset of operating status information and only a few major faults are needed by the operator, with a selection of that data being useful to a central control or maintenance dispatching facility if real time links are available. A variety of techniques are being used to present this information to the target audience(s) in ways and at times that are most appropriate.

"Event recorders" similar to the ones required on passenger airliners, capturing selected parameters, are now required by the FRA for railroads and are under consideration by industry standards groups for other modes.

Defining Terms

Active suspension: An electronically controlled suspension system for maintaining level suspension of a vehicle.

Consist: Two or more vehicles coupled together in a train. The vehicles may be identical or they may each lack a major subsystem (such as propulsion), whose functions may be handled by another vehicle in the consist.

Dual mode: Vehicles that are designed to switch manually or automatically from one type of propulsion to another, for instance from internal combustion to electric.

Electric braking: Use of traction motor to slow the vehicle.

Electric vehicle: Vehicle using electric energy storage, electric controls, and electric propulsion devices.

Traction motor: Electric motor that provides motive power to move vehicles.

Related Topics

66.2 Motors • 82.1 Practical Microprocessors

References

IEEE *Spectrum,* Technology Issue, January 1996.
R. K. Jurgen, "Putting electronics to work in the 1991 car models," *IEEE Spectrum,* pp. 75-78, December 1990.
Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA): FTA & FHA, Department of Transportation (see below).

Further Information

The following organizations can provide industry perspectives, applicable standards, and guidelines and technical information.

Vehicular Technology Society, Institute of Electrical and Electronics Engineers, 345 East 47th Street, New York, NY 10017-2394, 1-800-678-IEEE.

American Public Transit Association (APTA), 1201 New York Avenue, NW, Suite 400, Washington, D.C. 20005.

Federal Transit Administration (FTA), Federal Highway Administration (FHA), Federal Railroad Administration (FRA), Department of Transportation, 400 7th Street SW, Washington, D.C. 20590.

Society of Automotive Engineers, 400 Commonwealth Drive, Warrandale, PA 15096.

ITS America, 400 Virginia Avenue, SW, Suite 800, Washington, D.C. 20024-2730.

Chen, K. "Industrial Illuminating Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

107

Industrial Illuminating Systems

107.1 New Concepts in Designing an Industrial Illuminating System
Determination of Illuminance Levels • Illumination Computational Methods

107.2 Factors Affecting Industrial Illumination
Basic Definitions • Factors and Remedies • Daylighting

107.3 System Components
Light Sources • Ballasts • Luminaires

107.4 Applications
Types of Industrial Illuminating Systems • Selection of the Equipment

107.5 System Energy Efficiency Considerations
Energy-Saving Lighting Techniques • Lighting Controls • Lighting and Energy Standards

Kao Chen
Carlsons Consulting Engineers

107.1 New Concepts in Designing an Industrial Illuminating System

Determination of Illuminance Levels

Among the many new concepts for lighting design, the first to be discussed is the new method of determining **illuminance** levels. In the past when illuminating engineers wanted to find the recommended illuminance level for a given task, they would look in the lighting handbook to find a recommended level and then design an illuminating system for the task using the value as a minimum. This procedure provides very little latitude for fine-tuning an illumination design. In the new method, a more comprehensive investigation of required illuminance is performed according to the following steps:

1. Instead of a single recommended illuminance value, a category letter is assigned. Table 107.1 shows different category letters for a selected group of industries (partial only; for complete list see *IES Lighting Handbook* [1993]).
2. The category letters are used to define a range of illuminance. Table 107.2 details illuminance categories and illuminance values for generic types of activities in interiors.
3. From within the recommended range of illuminance, a specific value of illuminance is selected after consideration is given to the average age of workers, the importance of speed and accuracy, and the reflectance of task background.

The importance of acknowledging the speed and accuracy with which a task must be performed is readily recognized. Less obvious is the need to consider the age of workers and the reflectance of task background.

TABLE 107.1 Illuminance Categories for Selected Group of Industries

Area/Activity	Illuminance Category
Aircraft maintenance	[a]
Aircraft manufacturing	[a]
Assembly	
Simple	D
Moderately difficult	E
Difficult	F
Very difficult	G
Exacting	H
Automobile manufacturing	
Bakeries	
Mixing room	D
Face of shelves	D
Inside of mixing bowl	D
Fermentation room	D
Make-up room	
Bread	D
Sweet yeast-raised products	D
Proofing room	D
Oven room	D
Fillings and other ingredients	D
Decorating and icing	
Mechanical	D
Hand	E
Scales and thermometers	D
Wrapping	D
Book binding	
Folding, assembling, pasting	D
Cutting, punching, stitching	E
Embossing and inspection	F
Breweries	
Brew house	D
Boiling and keg washing	D
Filling (bottles, cans, kegs)	D
Candy making	
Box department	D
Chocolate department	
Husking, winnowing, fat extraction, crushing and refining, feeding	D
Bean cleaning, sorting, dipping, packing, wrapping	D
Milling	E
Cream making	
Mixing, cooking, molding	D
Gum drops and jellied forms	D
Hand decorating	D
Hard candy	
Mixing, cooking, molding	D
Die cutting and sorting	E
Kiss making and wrapping	E
Canning and preserving	
Initial grading raw material samples	D
Tomatoes	E
Color grading and cutting rooms	F
Preparation	
Preliminary sorting	
Apricots and peaches	D
Tomatoes	E
Olives	F
Cutting and pitting	E
Final sorting	E
Canning	
Continuous-belt canning	E
Sink canning	E
Hand packing	D
Olives	E
Examination of canned samples	F
Container handling	
Inspection	F
Can unscramblers	E
Labeling and cartoning	D
Casting (see Foundries)	
Central stations (see Electric generating stations)	
Chemical plants (see Petroleum and chemical plants)	
Clay and concrete products	
Grinding, filter presses, kiln rooms	C
Molding, pressing, cleaning, trimming	D
Enameling	E
Color and glazing—rough work	E
Color and glazing—fine work	F
Cleaning and pressing industry	
Checking and sorting	E
Dry and wet cleaning and steaming	E
Inspection and spotting	G
Pressing	F
Repair and alteration	F
Cloth products	
Cloth inspection	I
Cutting	G
Sewing	G
Pressing	F
Clothing manufacture (see Sewn Products)	
Receiving opening, storing, shipping	D
Examining (perching)	I
Sponging, decanting, winding, measuring	D
Piling up and marking	E
Cutting	G
Pattern making, preparation of trimming, piping, canvas and shoulder pads	E
Filling, bundling, shading, stitching	D
Shops	F
Inspection	G
Pressing	F
Sewing	G
Control rooms	
(see Electric generating stations—interior)	
Corridors (see Service spaces)	
Cotton gin industry	
Overhead equipment—separators, driers, grid cleaners, slick machines, conveyers, feeders and catwalks	D
Gin stand	D
Control console	D
Lint cleaner	D
Bale press	D
Dairy farms (see Farms)	
Dairy products	
Fluid milk industry	
Boiler room	D
Bottle storage	D
Bottle sorting	E

[a] Industry representatives have established a table of single illuminance values which, in their opinion, can be used. Illuminance values for specific operations can also be determined using illuminance categories of similar tasks and activities found in this table and the application of the appropriate weighting factors.

Source: IES Lighting Handbook, Application Volume.

TABLE 107.2 Illuminance Categories and Illuminance Values for Generic Types of Activities in Interiors

Type of Activity	Illuminance Category	Ranges of Illuminances		Reference Work-Plane
		Lux	Footcandles	
Public spaces with dark surroundings	A	20–30–50	2–3–5	
Simple orientation for short temporary visits	B	50–75–100	5–7.5–10	General lighting throughout spaces
Working spaces where visual tasks are only occasionally performed	C	100–150–200	10–15–20	
Performance of visual tasks of high contrast or large size	D	200–300–500	20–30–50	
Performance of visual tasks of medium contrast or small size	E	500–750–1,000	50–75–100	Illuminance on task
Performance of visual tasks of low contrast or very small size	F	1,000–1,500–2,000	100–150–200	
Performance of visual tasks of low contrast and very small size over a prolonged period	G	2,000–3,000–5,000	200–300–500	Illuminance on task, obtained by a combination of general and local (supplementary lighting)
Performance of very prolonged and exacting visual tasks	H	5,000–7,500–10,000	500–750–1,000	
Performance of very special visual tasks of extremely low contrast and small size	I	10,000–15,000–20,000	1,000–1,500–2,000	

Source: IES Lighting Handbook, Application Volume.

To compensate for reduced visual acuity, more illuminance is needed. Using the average age of workers as the age criterion is a compromise between the need of the young and the older workers and, therefore, a valid criterion.

Task background affects the ability to see because it affects **contrast**, an important aspect of visibility. More illuminance is required to enhance the visibility of tasks with poor contrast. Reflectance is calculated by dividing the reflected value by the incident value. The data given in Tables 107.3 and 107.4 are taken from the *IES Lighting Handbook* [1987] and are applied to provide a single value of illuminance from within the range recommended.

Illuminating system design can begin after the desired value of illuminance for a given task has been determined. Based on the *IES Handbook*, the zonal cavity method of determining the number of luminaires and lamps to yield a specified maintained luminance remains unchanged.

Illumination Computational Methods

Zonal Cavity Method. Introduced in 1964, the zonal cavity method of performing lighting computations has gained rapid acceptance as the preferred way to calculate number and placement of luminaires required to satisfy a specified illuminance level requirement. Zonal cavity provides a higher degree of accuracy than does the old lumen method, because it gives individual consideration to factors that are glossed over empirically in the lumen method.

Definition of Cavities. With the zonal cavity method, the room is considered to contain three vertical zones or cavities. Figure 107.1 defines the various cavities used in this method of computation. Height for luminaire to ceiling is designated as the ceiling cavity (h_{cc}). Distance from luminaire to the work plane is the room cavity (h_{rc}), and the floor cavity (h_{fc}) is measured from the work plane to the floor.

To apply the zonal cavity method, it is necessary to determine a parameter known as the "**cavity ratio**" (CR) for each of the three cavities. Following is the formula for determining the cavity ratio:

$$\text{cavity ratio} = \frac{5h\,(\text{room length} + \text{room width})}{(\text{room length} \times \text{room width})} \tag{107.1}$$

where h equals h_{cc} for ceiling cavity ratio (CCR), h_{rc} for room cavity ratio (RCR), h_{fc} for floor cavity ratio (FCR).

TABLE 107.3 Weighting Factors for Selecting Specific Illuminance Within Ranges A, B, and C

Occupant and Room Characteristics*	Weighting Factor		
	–1	0	+1
Workers' age (average)	Under 40	40 to 55	Over 55
Average room reflectance[1]	>70%	30 to 70%	<30%

Source: IES Lighting Handbook, Application Volume.

Note: This table is used for assessing weighting factors in rooms where a task is not involved.

1. Assign the appropriate weighting factor for each characteristic.
2. Add the two weights; refer to Table 107.2, Categories A through C:
   a. If the algebraic sum is –1 or –2, use the lowest range value.
   b. If the algebraic sum is 0, use the middle range value.
   c. If the algebraic sum is +1 or +2, use the highest range value.

*To obtain average room reflectance: determine the areas of ceiling, walls, and floor; add the three to establish room surface area; determine the proportion of each surface area to the total; multiply each proportion by the pertinent surface reflectance; and add the three numbers obtained.

TABLE 107.4 Weighting Factors for Selecting Specific Illuminance Within Ranges D through I

Task or Worker Characteristics	Weighting Factor		
	–1	0	+1
Workers' age (average)	Under 40	40 to 55	Over 55
Speed or accuracy*	Not important	Important	Critical
Reflectance of task background, %	>70%	30 to 70%	<30%

Source: IES Lighting Handbook, Application Volume.

Note: Weighting factors are based upon worker and task information.

1. Assign the appropriate weighting factor for each characteristic.
2. Add the two weights; refer to Table 107.2, Categories D through I:
   a. If the algebraic sum is –2 or –3, use the lowest range value.
   b. If the algebraic sum is –1, 0, or +1, use the middle range value.
   c. If the algebraic sum is +2 or +3, use the highest range value.

*Evaluation of speed and accuracy requires that time limitations, the effect of error on safety, quality, and cost, etc. be considered. For example, leisure reading imposes no restrictions on time, and errors are seldom costly or unsafe. Reading engineering drawings or a micrometer requires accuracy and, sometimes, speed. Properly positioning material in a press or mill can impose demands on safety, accuracy, and time.

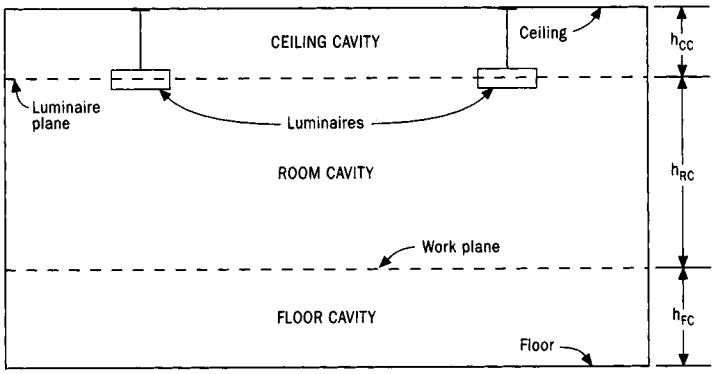


FIGURE 107.1 Basic cavity divisions of space.

Lumen Method Details. Because of the ease of application of the lumen method which yields the average illumination in a room, it is usually employed for larger areas, where the illumination is substantially uniform. The lumen method is based on the definition of a footcandle, which equals one lumen per square foot:

$$\text{footcandle} = \frac{\text{lumen striking an area}}{\text{square feet of area}} \tag{107.2}$$

In order to take into consideration such factors as dirt on the luminaire, general depreciation in lumen output of the lamp, and so on, the above formula is modified as follows:

$$\text{footcandle} = \frac{\text{lamps/luminaire} \times \text{lumens/lp} \times \text{CU} \times \text{LLF}}{\text{area/luminaire}} \tag{107.3}$$

In using the lumen method, the following key steps should be taken:

a. Determine the required level of illuminance.
b. Determine the **coefficient of utilization (CU)** which is the ratio of the lumens reaching the working plane to the total lumens generated by the lamps. This is a factor that takes into account the efficiency and the distribution of the luminaire, its mounting height, the room proportions, and the reflectances of the walls, ceiling, and floor. Rooms are classified according to shape by 10 room cavity numbers. The cavity ratio can be calculated using the formula given in Eq. (107.1). The coefficient of utilization is selected from tables prepared for various luminaires by manufacturers.
c. Determine the **light loss factor (LLF)**. The final light loss factor is the product of all the contributing loss factors. Lamp manufacturers rate filament lamps in accordance with their output when the lamp is new; vapor discharge lamps (fluorescent, mercury, and other types) are rated in accordance with their output after 100 hr of burning.
d. Calculate the number of lamps and luminaires required:

$$\text{no. of lamps} = \frac{\text{footcandles} \times \text{area}}{\text{lumens/lp} \times \text{CU} \times \text{LLF}} \tag{107.4}$$

$$\text{no. of luminaires} = \frac{\text{no. of lamps}}{\text{lamps/luminaire}} \tag{107.5}$$

e. Determine the location of the luminaire—luminaire locations depend on the general architecture, size of bays, type of luminaire, position of previous outlets, and so on.

Point-by-Point Method. Although currently light computations emphasize the zonal cavity method, there is still considerable merit in the point-by-point method. This method lends itself especially well to calculating the illumination level at a particular point where total illumination is the sum of general overhead lighting and supplementary lighting. In this method, information from luminaire **candlepower distribution** curves must be applied to the mathematical relationship. The total contribution from all luminaires to the illumination level on the task plane must be summed.

Direct Illumination Component. The angular coordinate system is most applicable to continuous rows of fluorescent luminaires. Two angles are involved: a longitudinal angle α and a lateral angle β. Angle α is the angle between a vertical line passing through the seeing task (point P) and a line from the seeing task to the end of the rows of luminaires. Angle α is easily determined graphically from a chart showing angles α and β

for various combinations of V and H. Angle β is the angle between the vertical plane of the row of luminaires and a tilted plane containing both the seeing task and the luminaire or row of luminaires. Figure 107.2 shows how angles α and β are defined. The direct illumination component for each luminaire or row of luminaires is determined by referring to the table of direct illumination components for the specific luminaire. The direct illumination components are based on the assumption that the luminaire is mounted 6 ft above the seeing task. If this mounting height is other than 6 ft, the direct illumination component shown in Table 107.5 must be multiplied by $6/V$, where V is the mounting height above the task. Thus the total direct illumination component would be the product of $6/V$ and the sum of the individual direct illumination components of each row.

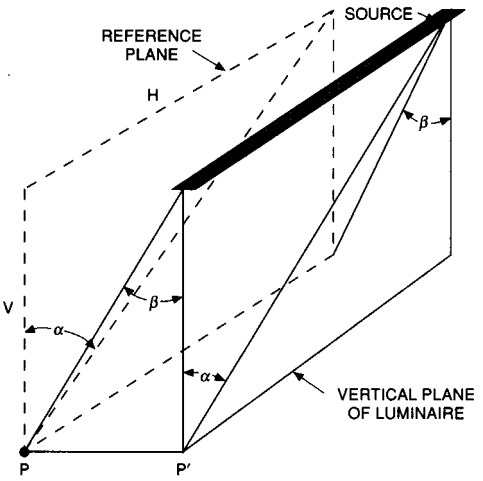


FIGURE 107.2 Definition of angular coordinate systems for direct illumination component.

Reflected Illumination Components on the Horizontal Surfaces. This is calculated in exactly the same manner as the average illumination using the lumen method, except that the reflected radiation coefficient (RRC) is substituted for the coefficient of utilization.

$$FC_{RH} = \frac{\text{lamps/luminaire} \times \text{lumens/lp} \times \text{RRC} \times \text{LLF}}{\text{area/luminaire}} \tag{107.6}$$

where RRC = LC_W + RPM ($LC_{CC} - LC_W$), LC_W = wall luminance coefficient, LC_{CC} = ceiling cavity luminance coefficient, and RPM = room position multiplier.

The wall luminance coefficient and the ceiling cavity luminance coefficient are selected for the appropriate room cavity ratio and proper wall and ceiling cavity reflectances from the table of luminance coefficients in the same manner as the coefficient of utilization. The room position multiplier is a function of the room cavity ratio and of the location in the room of the point where the illumination is desired. Table 107.6 lists the value of the RPM for each possible location of the part in the rooms of all room cavity ratios.

Figure 107.3 shows a grid diagram that illustrates the method of designating the location in the room by a letter and a number.

Reflected Illumination Components on the Vertical Surfaces. To determine illumination reflected to vertical surfaces, the approximate average value is determined using the same general formula, but substituting WRRC (wall reflected radiation coefficient) for the coefficient of utilization:

$$FC_{RV} = \frac{\text{lamps/luminaire} \times \text{lumens/lp} \times \text{WRRC} \times \text{LLF}}{\text{area/luminaire (on work plane)}} \tag{107.7}$$

where

$$\text{WRRC} = \frac{\text{wall luminance coefficient}}{\text{average wall reflectance}} - \text{WDRC} \tag{107.8}$$

where WDRC is the wall direct radiation coefficient, which is published for each room cavity ratio together with a table of wall luminance coefficients (see Table 107.5 for a specific type of luminance).

TABLE 107.5 Direct Illumination Components for Category III Luminaire (Based on F40 Lamps Producing 3100 Lumens)

Direct Illumination Components																
8	5	15	25	35	45	55	65	75	5	15	25	35	45	55	65	75
	Vertical Surface Illumination Footcandles at a Point on a Plane Parallel to Luminaires								Vertical Surface Illumination Footcandles at a Point on a Plane Perpendicular to Luminaires							
∝																
0–10	.9	2.6	3.6	3.9	3.3	1.9	.7	.1	.9	.8	.7	.5	.3	.1	—	—
0–20	1.8	5.0	7.0	7.7	6.6	3.8	1.5	.2	3.6	3.2	2.7	1.9	1.2	.5	.1	—
0–30	2.6	7.2	10.1	11.3	9.8	5.7	2.3	.3	7.7	7.0	5.8	4.3	2.7	1.1	.3	—
0–40	3.2	9.0	12.8	14.5	12.9	7.7	3.2	.5	12.6	11.6	9.7	7.5	4.9	2.1	.6	—
0–50	3.7	10.3	14.9	17.1	15.7	9.6	4.3	.7	17.8	16.6	14.2	11.2	7.7	3.4	1.1	.1
0–60	4.0	11.2	16.3	18.8	17.6	11.3	5.5	1.0	22.6	21.2	18.4	14.7	10.4	5.1	1.9	.2
0–70	4.1	11.6	17.0	19.8	18.9	12.7	6.8	1.4	26.2	24.7	21.8	17.8	13.1	7.2	3.2	.3
0–80	4.1	11.7	17.3	20.2	19.4	13.3	7.4	1.9	28.2	26.7	23.8	19.7	14.9	8.7	4.3	.8
0–90	4.1	11.7	17.3	20.2	19.4	13.4	7.5	2.0	28.6	27.1	24.2	20.1	15.3	9.1	4.7	1.1

F.C. at a Point on Work Plane								Category III
0–10	10.6	9.5	7.6	5.5	3.3	1.3	.3	—
0–20	20.6	18.5	14.9	10.9	6.6	2.6	.7	—
0–30	29.4	26.5	21.6	16.0	9.8	4.0	1.1	—
0–40	36.5	33.1	27.4	20.6	12.9	5.4	1.5	—
0–50	41.8	38.1	31.9	24.3	15.7	6.7	2.0	.1
0–60	45.2	41.3	34.8	26.8	17.6	7.9	2.6	.2
0–70	46.9	43.0	36.4	28.3	18.9	8.9	3.2	.3
0–80	47.4	43.6	36.9	28.8	19.4	9.3	3.5	.4
0–90	47.5	43.7	37.0	28.8	19.4	9.3	3.5	.4

2 T-12 Lamps—Any Loading
For T-10 Lamps—CU × 1.02

Luminance Coefficients for 20% Effective Floor Cavity Reflectance

Ceiling Cavity		Reflectances											
		80		50		10		80		50		10	
Walls		50	30	50	30	50	30	50	30	50	30	50	30
WDRC	RCR	Wall Luminance Coefficients						Ceiling Cavity Luminance Coefficients					
.281	1	.246	.140	.220	.126	.190	.109	.230	.209	.135	.124	.025	.023
.266	2	.232	.127	.209	.115	.182	.102	.222	.190	.130	.113	.024	.021
.245	3	.216	.115	.196	.105	.172	.095	.215	.176	.127	.105	.024	.020
.226	4	.202	.102	.183	.097	.161	.088	.209	.164	.124	.099	.023	.019
.212	5	.191	.097	.173	.090	.154	.082	.204	.156	.121	.094	.023	.018
.196	6	.178	.090	.163	.084	.145	.076	.200	.149	.118	.090	.022	.017
.182	7	.168	.083	.153	.078	.136	.071	.194	.144	.115	.087	.022	.017
.170	8	.158	.077	.145	.072	.130	.066	.190	.139	.113	.085	.021	.016
.159	9	.150	.072	.138	.068	.123	.062	.185	.135	.110	.082	.021	.016
.149	10	.141	.068	.130	.064	.116	.059	.180	.131	.107	.080	.020	.016

107.2 Factors Affecting Industrial Illumination

Basic Definitions

Illuminance. lluminance is the density of luminous lux on a surface expressed in either footcandles (lumens/ft^2) or lux (lx) (lux = 0.0929 fc).

Luminance (or photometric brightness). Luminance is the luminous intensity of a surface in a given direction per unit of projected area of the surfaces, expressed in candelas per unit area or in lumens per unit area.

Reflectance. Reflectance is the ratio of the light reflected from a surface to that incident upon it. Reflection may be of several types, the most common being specular, diffuse, spread, and mixed.

Glare. Glare is any brightness that causes discomfort, interference with vision, or eye fatigue.

TABLE 107.6 Room Position Multipliers

	A	B	C	D	E	F
			Room Cavity Ratio = 1			
0	.24	.42	.47	.48	.44	.48
1	.42	.74	.81	.83	.84	.84
2	.47	.81	.90	.92	.93	.93
3	.48	.83	.92	.94	.95	.95
4	.48	.84	.93	.95	.96	.97
5	.48	.84	.93	.95	.97	.97
			Room Cavity Ratio = 2			
0	.24	.36	.42	.44	.46	.46
1	.36	.51	.60	.63	.66	.68
2	.42	.60	.68	.72	.78	.83
3	.44	.63	.72	.77	.82	.85
4	.46	.66	.78	.82	.85	.86
5	.46	.68	.83	.85	.86	.87
			Room Cavity Ratio = 3			
0	.23	.32	.37	.40	.42	.42
1	.32	.40	.48	.51	.53	.57
2	.37	.48	.58	.61	.64	.67
3	.40	.51	.61	.65	.69	.71
4	.42	.53	.64	.69	.73	.75
5	.42	.57	.67	.71	.75	.77
			Room Cavity Ratio = 4			
0	.22	.28	.32	.35	.37	.37
1	.28	.33	.40	.42	.44	.48
2	.32	.40	.48	.50	.52	.57
3	.35	.42	.50	.54	.58	.61
4	.37	.44	.52	.58	.62	.64
5	.37	.48	.57	.61	.64	.66
			Room Cavity Ratio = 5			
0	.21	.25	.28	.31	.33	.33
1	.25	.29	.33	.36	.38	.42
2	.28	.33	.40	.42	.44	.48
3	.31	.36	.42	.46	.49	.52
4	.33	.38	.44	.49	.52	.54
5	.33	.42	.48	.52	.54	.56

	A	B	C	D	E	F
			Room Cavity Ratio = 6			
0	.20	.23	.26	.28	.29	.30
1	.23	.26	.29	.31	.33	.36
2	.26	.29	.35	.37	.38	.40
3	.28	.31	.37	.39	.41	.43
4	.29	.33	.38	.41	.43	.45
5	.30	.36	.40	.43	.45	.47
			Room Cavity Ratio = 7			
0	.18	.21	.23	.25	.26	.27
1	.21	.23	.26	.28	.29	.30
2	.23	.26	.30	.32	.33	.34
3	.25	.28	.32	.34	.35	.36
4	.26	.29	.33	.35	.37	.37
5	.27	.30	.34	.36	.37	.38
			Room Cavity Ratio = 8			
0	.17	.18	.21	.22	.22	.23
1	.18	.20	.23	.25	.26	.26
2	.21	.23	.26	.27	.28	.29
3	.22	.25	.27	.29	.30	.30
4	.22	.26	.28	.30	.31	.32
5	.23	.26	.29	.30	.31	.32
			Room Cavity Ratio = 9			
0	.15	.17	.18	.19	.20	.20
1	.17	.18	.20	.21	.22	.23
2	.18	.20	.23	.24	.25	.25
3	.19	.21	.24	.25	.26	.26
4	.20	.22	.25	.26	.26	.27
5	.20	.23	.25	.26	.27	.27
			Room Cavity Ratio = 10			
0	.14	.16	.16	.17	.18	.18
1	.16	.17	.18	.19	.19	.20
2	.16	.18	.19	.21	.22	.22
3	.17	.19	.21	.22	.23	.23
4	.18	.19	.22	.23	.23	.24
5	.18	.20	.22	.23	.24	.25

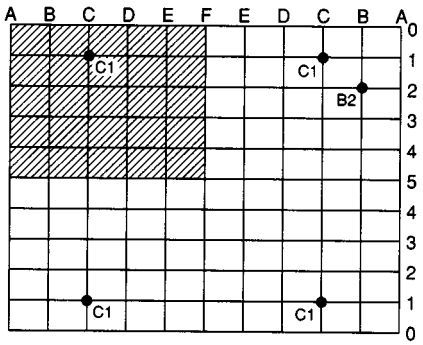


FIGURE 107.3 Grid diagram for locating points on the work plane.

Color Rendering Index (CRI). In 1964 the CIE (Commission Internationale de l'Eclairage) officially adopted the IES procedure for rating lighting sources and developed the current standard by which light sources are rated for their color rendering properties. The CRI is a numerical value for the color comparison of one light source to that of a reference light source.

Color Preference Index (CPI). The CPI is determined by a similar procedure to that used for the CRI. The difference is that CPI recognizes the very real human ingredient of preference. This index is based on individual preference for the coloration of certain identifiable objects, such as complexions, meat, vegetables, fruits, and foliage, to be slightly different than the colors of these objects in daylight. CPI indicates how a source will render color with respect to how we best appreciate and remember that color.

Equivalent Sphere Illumination (ESI). ESI is a means of determining how well a lighting system will provide task visibility in a given situation. ESI may be predicted for many points in a lighting system through the use of any of several available computer programs or measured in an installation with any of several different types of meters.

Visual Comfort Probability (VCP). Discomfort glare is most often produced by direct glare from luminances that are excessively bright. Discomfort glare can also be caused by reflected glare, which should not be confused with **veiling reflections**, which cause a reduction in visual performance rather than discomfort. VCP is based in terms of the percentage of people who will be expected to find the given lighting system acceptable when they are seated in the most undesirable location.

Factors and Remedies

Quality of illumination pertains to the distribution of luminaires in the visual environment. The term is used in a positive sense and implies that all luminaires contribute favorably to visual performance. However, glare, diffusion, reflection, uniformity, color, luminance, and **luminance ratio** all have a significant effect on visibility and the ability to see easily, accurately, and quickly. Industrial installations of poor quality are easily recognized as uncomfortable and possibly hazardous. Some of the factors are discussed in more detail below.

Direct Glare. When glare is caused by the source of lighting within the field of view, whether daylight or electric, it is defined as direct glare. To reduce direct glare, the following suggestions may be useful:

a. Decrease the brightness of light sources or lighting equipment, or both.
b. Reduce the area of high luminance causing the glare condition.
c. Increase the angle between the glare source and the line of vision.
d. Increase the luminance of the area surrounding the glare source and against which it is seen.

To reduce direct glare, luminaires should be mounted as far above the normal line of sight as possible and should be designed to limit both the luminance and the quality of light emitted in the 45–85 degree zone because such light may interfere with vision. This precaution includes the use of supplementary lighting equipment. There is such a wide divergence of tasks and environmental conditions that it may not be possible to recommend a degree of quality satisfactory to all needs. In production areas, luminaires within the normal field of view should be shielded to at least 25 degrees from the horizontal, preferably to 45 degrees.

Reflected Glare. Reflected glare is caused by the reflection of high-luminance light sources from shiny surfaces. In the manufacturing area, this may be a particularly serious problem where critical seeing is involved with highly polished sheet metal, vernier scales, and machined metal surfaces. There are several ways to minimize or eliminate reflected glare:

a. Use a light source of low luminance, consistent with the type of work in process and the surroundings.
b. If the luminance of the light source cannot be reduced to a desirable level, it may be possible to orient the work so that reflections are not directed in the normal line of vision.
c. Increasing the level of illumination by increasing the number of sources will reduce the effect of reflected glare by reducing the proportion of illumination provided on the task by sources located in positions causing reflections.

TABLE 107.7 Recommended Maximum Luminance Ratios for Industrial Areas

	Environmental Classification		
	A	B	C
(1) Between tasks and adjacent darker surroundings	3 to 1	3 to 1	5 to 1
(2) Between tasks and adjacent lighter surroundings	1 to 3	1 to 3	1 to 5
(3) Between tasks and more remote darker surfaces	10 to 1	20 to 1	*
(4) Between tasks and more remote lighter surfaces	1 to 10	1 to 20	*
(5) Between luminaires (or windows, skylights, etc.) and surfaces adjacent to them	20 to 1	*	*
(6) Anywhere within normal field of view	40 to 1	*	*

*Luminance ratio control not practical.

A—Interior areas where reflectances of entire space can be controlled in line with recommendations for optimum seeing conditions.

B—Areas where reflectances of immediate work area can be controlled, but control of remote surround is limited.

C—Areas (indoor and outdoor) where it is completely impractical to control reflectances and difficult to alter environmental conditions.

Source: *IES Lighting Handbook*, *Application Volume*.

d. In special cases, it may be practical to reduce the specular reflection by changing the specular character of the offending surface.

Distribution, Reflection, and Shadows. Uniform horizontal illuminance (maximum and minimum not more than one-sixth above or below the average level) is usually desirable for industrial interiors to permit flexible arrangements of operations and equipment and to assure more uniform luminance in the entire area.

Reflections of light sources in the task can be useful provided that the reflection does not create reflected glare. In the machining and inspection of small metal parts, reflections can indicate faults in contours, make scribe marks more visible, and so on.

Shadows from the general illumination systems can be desirable for accenting the depth and forms of various objects, but harsh shadows should be avoided. Shadows are softer and less pronounced when large diffusing luminaires are used or the object is illuminated from many sources. Clearly defined shadows are distinct aids in some specialized operations, such as engraving on polished surfaces, some type of bench layout work, or certain textile inspections. This type of shadow effect can best be obtained by supplementary directional lighting combined with ample diffused general illumination.

Luminance and Luminance Ratios. The ability to see details depends on the contrast between the detail and its background. The greater the contrast difference in luminance, the more readily the seeing task is performed. The eye functions most comfortably and efficiently when the luminance within the remainder of the environment is relatively uniform. In manufacturing, there are many areas where it is not practical to achieve the same luminance relationships as easily as in offices. Table 107.7 is shown as a practical guide to recommended maximum luminance ratios for industrial areas. To achieve the recommended luminance relationships, it is necessary to select the reflectances of all the finishes of the room surfaces and equipment as well as control of the luminance distribution of the lighting equipment. Table 107.8 lists the recommended reflectance values for industrial interiors and equipment. High-reflectance surfaces are desirable to provide the recommended luminance relationships and high utilization of light.

Color Quality of Light. In general, for seeing tasks industrial areas, there appears to be no effect upon visual acuity by variation in color of light. However, where color discrimination or color matching is a part of the work process, such as in the printing and textile industries, the color of light should be carefully selected. Color always has an effect on the appearance of the workplace and on the complexions of people. The illuminating system and the decorative scheme should be properly coordinated.

TABLE 107.8 Recommended Reflectance Values for Industrial Interiors and Equipment

Surfaces	Reflectance[1] (%)
Ceiling	80 to 90
Walls	40 to 60
Desk and bench tops, machines and equipment	25 to 45
Floors	not less than 20

[1]Reflectance should be maintained as near as practical to recommended values.

Source: *IES Lighting Handbook, Application Volume.*

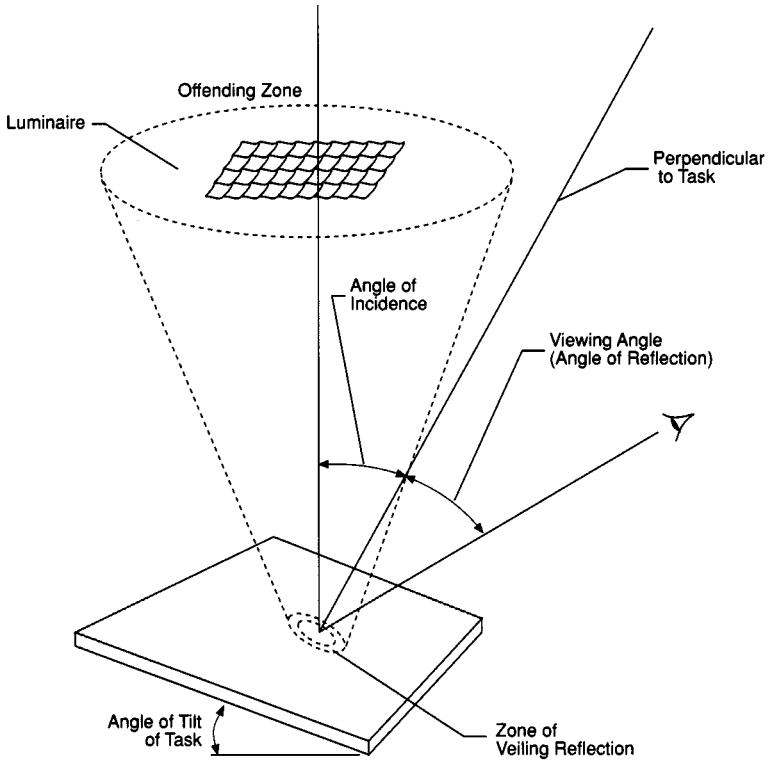


FIGURE 107.4 Diagram showing "offending zone" and zone of veiling reflection.

Veiling Reflections. Figure 107.4 shows that light would reflect into the eyes of the viewer from the "offending zone" and defines the zone of veiling reflection. Veiling reflection would diminish visibility, but the viewer would be unaware of it. The **contrast rendition factor (CRF)** can be applied as a measure of the amount of veiling reflection.

Another important factor is the **lighting effectiveness factor (LEF)**. An overall lighting system efficiency factor considers both the quality of light as reference to equivalent sphere illumination and the effects of veiling reflections. Light patterns such as "batwing" can help solve veiling reflection problems. Figure 107.5 shows the light distribution curve of a typical batwing luminaire.

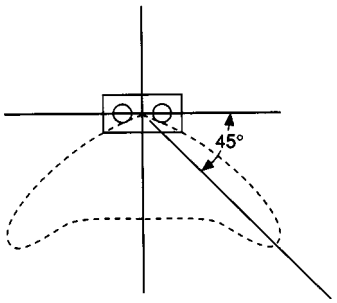


FIGURE 107.5 A typical "batwing" light distribution.

Daylighting

The daylight contribution should be carefully evaluated and should always be coordinated with a planned electric lighting system.

Fenestration. Fenestration has at least three useful purposes in industrial buildings:

a. For the admission, control, and distribution of daylight.
b. For a distant focus for the eyes, which relaxes the eye muscles.
c. To eliminate the dissatisfaction many people experience in completely closed-in areas.

An adequate electric lighting system should always be provided because of the wide variation in daylight.

Building Orientation. All fenestration should be equipped with control device appropriate to any luminance problems. Special attention should be given to glare control latitudes where fenestration frequently receives direct sunlight. Diffuse-glaring fixed or adjustable louvers are some of the control means that may be applied.

For an industrial building, windows in the sidewalls admit daylight and natural ventilation and afford occupants a view out. However, their uncontrolled luminance may be a problem. There are many control means to make daylight useful to workers' seeing tasks, resulting in energy savings as the ultimate goal.

107.3 System Components

Light Sources

Incandescent

a. Recent technology made possible a line of energy-saving incandescent lamps that use the rare gas krypton as a fill gas.
b. Reflector (R) lamps offer better utilization of the light provided by the lamp compared to a nonreflector type. In this family, there are R lamps, PAR (parabolic aluminized reflector) lamps, and a newer line ER (elliptical reflector) lamps which allow reduction of 50% or more in energy consumption.
c. Infrared Halogen (IR)—PAR lamps combine both the infrared heat-reflection technology and the regenerative halogen cleaning cycle to provide a dramatic increase in lamp efficacy (4% reduction in energy consumption). Available in 30, 60, and 100 W.

Fluorescent

a. Energy-efficient lamps are now available in all popular sizes and colors for most applications. Limitations of energy-saving reduced-wattage lamps are:
   - Ambient temperature must be above 60°F.
   - Used on high p.f. fluorescent ballasts only.
   - Not to be used where drafts of cold air are directed onto the lamp.
b. Typical energy savings are 6 W per lamp for the popular 4-ft 40-W replacement and 15 W per lamp for the 8-ft slimline 75-W replacement.
c. Compact fluorescent lamps are gaining popularity because they are energy efficient, fit into a small enclosed housing, and can be adapted for incandescent socket use.
d. Virtually all compact fluorescent lamps use the "rare earth" phosphors for good color rendition and lumen maintenance characteristics.
e. Utilizing advanced phosphor technology with the optimization of bulb diameter, 40-W lamps are now available that can be retrofitted in a F40 preheat or rapid-start circuit. The new lamp, which could save energy and improve color rendition requirement, has been legislated.
f. Refer to Tables 107.9 and 107.11 for the latest energy efficient lamps.

High-Intensity Discharge (HID)

Today HID lamps include mercury vapor, metal halide, high-pressure sodium, and low-pressure sodium lamps. Metal halide lamps offer the best opportunity from a color acceptability point of view. High-pressure sodium

lamps offer the highest luminous efficacy in an environment where color distinction is not critical. Since HID lamps have had very few problems in application, they are likely to experience further development in the coming years.

Ballasts

Fluorescent. Electronic ballasts are now available for the F40T12, the slimline, the new T8 lamps, and other energy-saving fluorescent lamps on both 120- and 277-V circuits. Using high-frequency ballasts, the efficacy can be raised by nearly 12%. Although electronic ballasts cost more than the standard core-coil ballasts, operating factors should reflect an appreciable reduction in life-cycle cost for a lighting system. There are two types of dimming ballasts: core and electronic. High-frequency ballasts can readily be used to dim fluorescent lamps over a wide range of light level. All external control wiring is low voltage or fiber-optic wiring.

A recent study indicates that 2 F40T12 lamps operated on an electronic ballast will attain an efficacy of 75–80 LPW versus 62 LPW for the same lamps if operated on a standard core-coil type ballast. With the dimmable electronically ballasted system, energy savings can be as high as 40% with respect to a core ballasted system.

High-Intensity Discharge. The choice of a ballast depends on economic considerations versus performance. A mercury lamp will operate from metal halide ballast, but the converse is not always true.

There are several different types of ballasts for high-pressure sodium lamps:

a. Reactor or lag ballast—Inexpensive, low power losses, and small in size.
b. Lead ballast —Fairly good regulation for both line and lamp voltage variation.
c. Magnetic regulated ballast—Provides best voltage regulation with change of either input voltage or lamp voltage. It is the most costly and has the greatest wattage loss.
d. Electronic ballast—Maintains a steady constant wattage output with changes in the source impedance as well as excellent regulation. During the life of a high-pressure sodium lamp, it can save 20% more energy by maintaining a constant wattage output in addition to the 15% intrinsic energy savings compared to an equivalent core-coil ballast.

Luminaires

Types of Industrial Luminaires. Selection of a specific type for an installation requires consideration of many factors: candlepower distribution, efficiency, shielding and brightness control, mounting height, lumen maintenance characteristics, mechanical construction, and environmental suitability for use in normal, hazardous, or special areas. In general there are five types in accordance with CIE classifications, namely, direct type, semi-direct type, direct-indirect type, semi-indirect type, and indirect type.

Figure 107.6 shows luminaire types with the percentage of total luminaire output emitted above and below horizontal.

Supplementary Luminaire Types. There are five major types based on the candlepower distribution and luminance:

Type S-I—directional
Type S-II—spread, high luminance
Type S-III—spread, moderate luminance
Type S- IV—uniform luminance
Type S-V—uniform luminance with pattern

High-Pressure Sodium. Proper luminaire design is the key to lighting efficiency. Newly developed luminaires use prismatic glass reflectors that are especially made for high-pressure sodium lamps. In addition to achieving maximum light utilization, they redirect the intense light source with excellent light cutoff and high-angle brightness control. Luminaire manufacturers recommend aluminum reflectors for all general-purpose industrial applications and glass-coated reflectors where maintenance practice is compatible with servicing glass.

Fluorescent. A new trend for lighting new buildings is the increased use of the reflectorized fixtures. This trend may be traced to an increase in the number of state and national lighting efficiency standards in recent

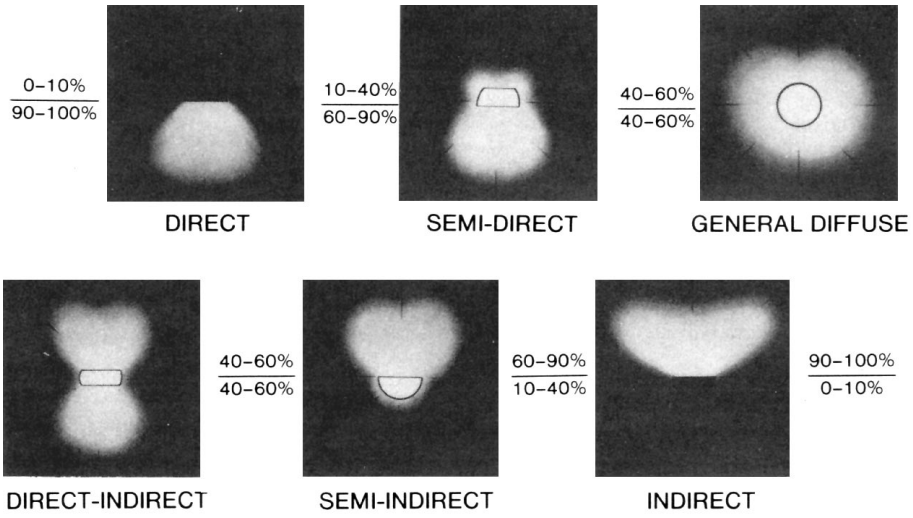


FIGURE 107.6 General lighting luminaire classifications.

years. However, these fixtures can create a "teardrop-like" distribution that may eliminate glare on a computer screen, but also reduces light to other areas.

107.4 Applications

Types of Industrial Illuminating Systems

Factory Illumination for Visual Tasks. The prime requirement for industrial illumination is to facilitate the performance of visual tasks through high-quality illumination. There are three types of lighting used in industrial areas.

- *General Lighting.* It should be designed to provide the desired level of illumination uniformly over the entire area. The variation of light level from point to point within the area should be within 17% of the selected level. A good general lighting system makes it possible to change the location of machinery without rearranging the lighting and also permits full utilization of floor space.
- *Localized General Lighting.* Within a general area there may be a few areas where tasks performed require a greater quantity of light and a different quality of light. When applied, care must be exercised to eliminate direct or reflected glare from the task and from other workers.
- *Supplementary Lighting.* Supplementary lighting is specified for different seeing tasks that require a specific amount or quality of light not readily obtained by standard general lighting methods. Supplementary lighting is a valuable industrial lighting tool. Typical problems arise where work is shielded from the general lighting system by an obstruction or its brightness is otherwise lowered where low contrast, such as scribe marks on steel, may lead to visual errors, and where the product moves too rapidly to be seen clearly by the unaided eye. To attain a good balance, it is important to coordinate the design of supplementary and general lighting with great care.

Security Lighting. Security lighting pertains to the lighting of building exterior and surrounding areas out to and including the boundaries of the property. Security lighting contributes to a sense of personal security and to the protection of property. It may be accomplished through:

- Surveillance lighting to detect and observe intruders.
- Protective lighting to discourage or deter attempts at entrance, vandalism, etc.
- Lighting for safety to permit safe movement of guards and other authorized persons.

Emergency Lighting. Emergency lighting is provided for use when the power supply for the normal lighting fails to ensure that escape routes can be effectively identified and used. Standby lighting is that part of emergency lighting that is sometimes provided to enable normal activities to continue.

The following are recommended minimum illumination requirements for exit signs and egress route:

- *Internally illuminated signs.* An illuminance of 54 lux (5 fc) on the face of the sign is usually specified.
- *Externally illuminated exit sign.* NFPA 101 requires 54 lux (5 fc) on the face of the sign.
- *Egress route.* The horizontal illuminance of any escape route should not be less than 1% of the average provided by the normal lighting, with a minimum average of 5 lux (0.5 fc) at floor level.
- *Location of egress luminaires.* A luminaire should be provided for each exit door and emergency exit door to provide sufficient light to a level of 30 lux (3 fc).

Summaries. In large industrial areas, all these lighting systems may be used. In small areas, localized general lighting may also serve as a substitute for general lighting. In this case, additional supplementary lighting may be required to increase the quantity or improve the quality of the illumination. Many factors must be considered in selecting a lighting system. It is not feasible to recommend one or two systems for all conditions. Because of the relationship of ceiling height to light utilization, most industrial applications call for either direct or semi-direct lighting systems.

Selection of the Equipment

In the selection of equipment, light sources, and luminaires, many variables must be considered. As with any list of variables, it is necessary for purpose of comparison to hold some factors constant. In industrial illumination that factor is usually mounting height and location.

High-Bay Areas. The work generally presents visual tasks that are not difficult because of large machinery and other objects. Illuminance levels for high-bay areas generally range from 50 to 150 fc, although more and more areas are being lighted with 200 and 300 fc. At a high mounting height, it is possible to obtain uniform illumination by using a few high-wattage sources rather than a larger number of low-wattage sources. For luminaires with medium and narrow distribution, greater mounting height or closer spacing is ordinarily required for uniform general illumination.

Regardless of mounting height, wide distribution luminaires are well suited for use in areas that are wide in respect to mounting height. Large machinery and objects tend to cut off light and cast shadows. Since this makes it difficult to see important vertical and angular surfaces, broad light distribution is essential.

High-intensity discharge or fluorescent luminaires for high-bay lighting may be enclosed, ventilated open, or nonventilated open. Enclosed luminaires are usually of a heavy-duty type with a gasketed glass cover to protect the reflector and light source from collection of dirt. The initial luminaire efficiency is lower and the equipment is more costly. Ventilated-open luminaires have largely replaced the nonventilated type.

As far as choices of lamps are concerned, metal halide and HPS are preferred over the mercury type. The use of fluorescent lamps in high-bay areas is limited. Only where the area proportions are such that the room cavity ratios are in the range of 1 to 3 may fluorescent lamps be acceptable. Only high or extra high output fluorescent in 8-ft sizes are recommended.

Medium- and Low-Bay Areas. Seeing tasks in medium- and low-bay areas are usually more difficult than those encountered in the high-bay areas. Increasing the size and reducing the brightness of the luminaires will improve visual comfort and will improve the visibility of specular objects. It may not improve the visibility of diffuse three-dimensional objects.

Luminaires used for general lighting in medium-bay areas are nearly always of the direct or semi-direct type, either fluorescent or wide distribution HID. They may be the ventilated or nonventilated type and the lamps may be shielded by louvers, baffles, or other devices. For lower mounting, the trend is toward the semi-direct type.

Some of the visual tasks involve specular or semi-specular objects, for which optimum lighting might be an indirect system. The quality of fluorescent sources, with their broad distribution of light, makes them a prime selection for medium- and low-bay lighting. When the proper quality control can be attained, low-wattage HID sources are finding an increasing number of low-bay applications.

107.5 System Energy Efficiency Considerations

Energy-Saving Lighting Techniques

Fluorescent Systems Considerations. Fluorescent lamps are sensitive to ambient temperatures. By using reduced wattage lamps or low-loss ballasts, less heat will be generated and the operating temperature point of the lamp will probably change. The critical area is the coldest spot on the bulb surface. Most fluorescent lamps will peak in light output at around a 100°F cold-spot temperature. For enclosed luminaire types that ordinarily operate the lamp at higher temperature, replacing standard lamps with high-efficacy, reduced wattage lamps may result in a net increase in luminaire output even though the reduced wattage lamps are rated for less output than are standard lamps.

Using Daylight. Daylight should be dealt with by first analyzing it and then establishing a design technique to integrate it with the electric lighting system. Daylight may be adequate in quantity and quality to reduce the electric lighting load and result in energy conservation. Poor quality of daylight may lead to discomfort and a loss in visibility that may result in a decrease in human performance and productivity.

Daylighting Design from Windows. The longhand design procedure involves two steps:

- Determine the quantity of illumination coming to the window surface.
- Use that quantity to determine the daylight contribution to the interior part of the space.

Once the contribution of illumination to the window surface has been calculated, two longhand methods are available to determine the illumination contribution to the space. The first method is to follow the point-by-point procedure, which makes two assumptions: (1) interreflected component is ignored and (2) the window is a uniform diffuse emitter. The second method is a lumen method that calculates illumination values at three points defined as the maximum, midway, and minimum. This method includes both the direct and interreflected components of illumination.

Task-Ambient Lighting. This is a particular form of nonuniform illumination that combines task illuminance and ambient illuminance. One advantage is improved energy efficiency. The task component of task-ambient lighting may take two forms: (1) furniture-mounted lighting built into a workstation or (2) floor-mounted fixtures that can be placed adjacent to a desk. The ambient lighting component may be supplied in two ways: (1) conventional luminaires on the ceiling or (2) indirect fixtures utilizing HID or fluorescent lamps with the output directed to the ceiling and adjacent walls. For ceiling-mounted troffers used for ambient lighting, a plug-in system of wiring should be considered so that luminaires can be relocated as task locations change.

Lighting Controls

In order to save energy, it is essential that minimum acceptable lighting levels be used during off-hours, cleaning periods, and for other nonpeak periods as is practical. The ultimate system of control would be to remotely control every fixture and to program the mode of operation, but this is hardly possible. Solid-state dimmers are available, or ballasts can be circuited in separate groupings. Solid-state controls are available for dimming entire areas of ballasted lights, but special ballasts are required and the controls could be expensive.

Manual control of a lighting system is often the least expensive, but also the least effective alternative. Automatic controls vary from a simple timer to a sophisticated computer system. Figure 107.7 shows a typical programmable lighting control scheme. A price versus benefit cost analysis will be required for each installation. The system should be programmed for normal operation and have a local manual override. A good convenient practice is to have lights switched in distributed groups so that areas can be lighted or darkened as conditions change.

Lighting and Energy Standards

In 1976, the Energy Research and Development Association (ERDA) contracted with the National Conference of States on Building Codes and Standards (NCSBCS) to codify ASHRAE 90-75. The resulting document was

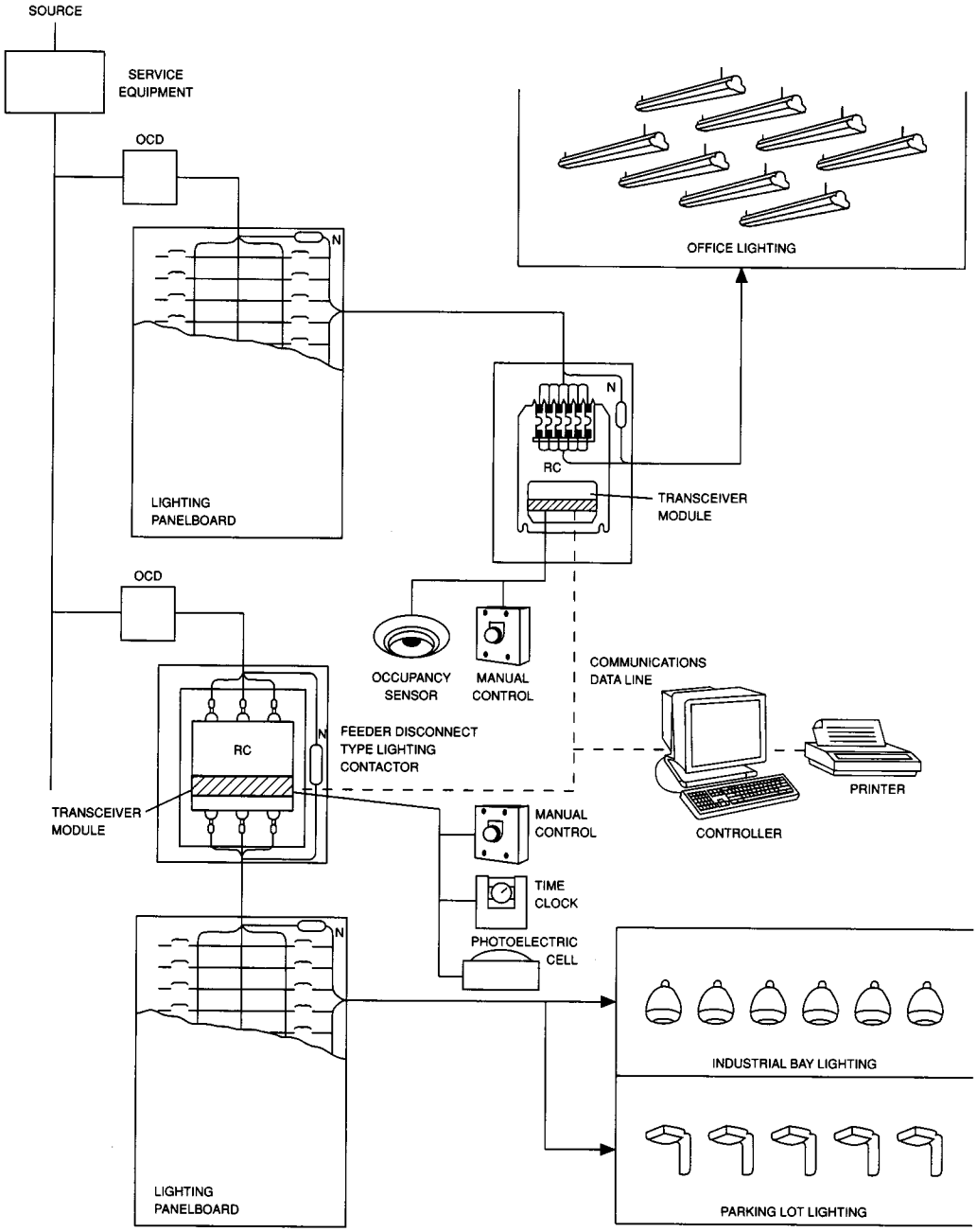


FIGURE 107.7 Programmable lighting control scheme.

called "The Model Code for Energy Conservation in New Buildings." The model code has been adopted by a number of states to satisfy the requirements of Public Laws 94-163 and 94-385.

There have been several revisions on the ANSI/ASHRAE/IES 90-75 since 1976. All were included in the lighting portion of ANSI/ASHRAE/IES 90A-1980, "Energy Conservation in New Building Design," and in EMS-1981, "IES Recommended Lighting Power Budget Determination Procedure."

ASHRAE/IES 90.1-1989, "Energy Efficient Design of New Buildings Except New Low-Rise Residential Buildings," is the third generation document on building energy efficiency since the first publication in 1975. This standard is intended to be a voluntary standard which can be adopted by building officials for state and local codes.

Lighting: The Next 10 Years

Many exciting developments are occurring in the lighting industry. However, one of the greatest challenges is the development of a replacement light source for the common cathode ray tube (CRT) found in televisions and other applications.

CRTs have gradually expanded from our primary information sources, televisions and computers, into scientific instrumentation, cars, and automatic teller machines. However, the inherent shortcomings of this technology are limiting the further development of existing and new applications that require display technology.

CRTs produce few lumens for the power they consume and are inherently large and heavy. Their many weaknesses are compounded in the area of big screen displays. The desire to view ever larger and higher resolution images is pushing the CRT beyond its practical limits. New lighting technologies applied to flat panel displays and projection systems could change the consumer television market if certain barriers are overcome.

One approach has been to adopt large area LCDs (which pose a significant engineering challenge) as the primary imaging device. However, these devices do not inherently emit light and would be useless in the dark without backlights. The light source of choice for this approach has been cold cathode fluorescent lamps, which are commonly applied in today's laptop computers.

We can expect to see unique lamp shapes and ingenious reflector systems developed to illuminate the LCD uniformly, without adding significantly to the overall depth of the display system. There is also a lot of work in the development of electroluminescent panels as an alternative to fluorescent lamps, although these devices present color difficulties as well as comparatively low efficacies.

Large-area LED image displays have already been fabricated since blue LEDs emerged in the marketplace. An array of red, green, and blue LEDs are mounted on a panel and individually addressed to generate all colors, including white. This approach will be redefined in the coming decade, but faces a number of challenges.

TABLE 107.9 The Proposed Efficiency Standards for Fluroescent Lamps

Lamp Type	Nominal Lamp Wattage	Minimum Average CRI	Minimum Average Lamp Efficacy
F40	>35 W	69	75
F40	≤35 W	45	75
F40/U	>35 W	69	68
F40/U	≤35 W	45	64
F96T12	>65 W	69	80
F96T12	≤65 W	45	80
F96T12/HO	>100 W	69	80
F96T12/HO	≤100 W	45	80

Note: The above excludes lamps designed for plant growth, cold temperature service, reflectorized/aperture, impact resistance, reprographic service, colored lighting, ultraviolet, and lamps with CRI of more than 82.

Another approach involves large-area plasma panels that are self-luminescent due to gas discharges. With this technology, the brightness improves considerably over that available from a CRT. Brilliant, high-resolution images have been achieved in medium screen sizes, but cost remains a problem and there is still the challenge of expanding the technology to even larger screen sizes.

The approach that is most likely to succeed during the next decade is the use of projection technology. These systems are fundamentally similar to slide or movie projectors where the film has been replaced by either a transmissive or reflective imaging device that provides a continuously variable image illuminated and projected onto a screen. The image can be front projected or rear projected.

The most pressing challenge remains: development of an illumination source that can provide a brighter image with better colors than CRT technology. If this can be achieved, other benefits will flow from the technology, with the potential to drastically reduce power consumption as well as cabinet size, weight, and cost. Furthermore, the inherent digital nature of the imaging panels would make the resulting product data compatible for the much touted merging of the Internet and television programming.

High-resolution rear projection televisions using lamps as the illumination source are already available in Japan from Sony and Sharp. We can expect to see this type of product in the U.S. market this year. However, the overall product cost, lamp life, and screen brightness all need improvement before this technology moves into the mainstream.

The technical challenge for the lighting industry is to produce a miniature point source that delivers high efficacy, high color temperature, and long lifetime. The challenge for the immediate future is to push the arc gap even smaller while extending the lamp life to be comparable to today's CRTs and maintaining lumen output and good color temperature. Furthermore, all these requirements have to be met at a very low cost.

The rewards for the successful manufacturer are immense, considering the size of the market, not to mention the spin-off markets that could pick up on this technology. During the next decade, we are sure to see a lot of exciting developments in this area, which will ultimately affect our daily activities. (Adapted from Ian Edwards, "Fundamentals of Lighting", *Optics & Photonics News,* Optical Society of America, 7(11), 20, 1996. With permission.)

Energy Policy Act

On October 25, 1992, the Energy Policy Act was signed into law by the President.

Among the many provisions, the act establishes energy efficiency standards for HVAC, lighting, and motor equipment; encourages establishment of a national window energy-efficiency rating system; and encourages state regulators to pursue demand-side-management (DSM) programs.

Under the bill, lighting manufacturers will have 3 years to stop making F96T12 and F96T12/HO 8 ft fluorescent lamps and some types of incandescent reflectors. Standard F40 lamps except in the SP and SPX or equivalent types of high color rendering lamps would also fade away. General service incandescent lamps to be axed would include those from 30 to 100 W, in 115 to 130 V ratings, having medium screw bases, of both reflector and PAR types, having a diameter larger than 2¾ in.

There are no immediate regulations impacting HID lamps. Within 18 months of the legislation's enactment, the Department of Energy (DOE) will determine the HID types for which standards could possibly save energy and publish testing requirements for these lamps.

As far as the general service lamps are concerned, the most common incandescent lamps — 40, 60, 75, 100, and 150 W — are not covered by an efficiency standard because there is no suitable method to ensure energy savings. These types, however, are covered by another provision of the law, namely the energy efficiency labeling standards.

TABLE 107.10 The Proposed Efficiency Standards for Incandescent Reflector Lamps

Nominal Lamp Wattage	Minimum Average Lamp Efficacy (LPW)
40–50	10.5
51–66	11.0
67–85	12.5
86–115	14.0
116–155	14.5

Note: The above excludes miniature, decorative, traffic signal, marine, mine, stage/studio, railway, colored lamps, and other special application types.

TABLE 107.11 1992 Energy Policy Act — Replacement Lamps

Present Type	W	Acceptable	W	Improved Type	W	Max. Savings	W
F96T12/CW	75	F96T12/CW/SS	60	F96T12/D41/SS	60	F096T8/741	59
F96T12/WW	75	F96T12/WW/SS	60	F96T12/D30/SS	60	F096T8/730	59

Effective April 28, 1994, the Federal Trade Commission (FTC) must provide manufacturers with labeling requirements for all lamps covered: fluorescent, incandescent, and reflector incandescent. Though not yet defined, the proposals include: an energy rating for the lamps, probably LPW (lumens per watt), and energy cost per year to operate the lamp. The energy efficiency label will then allow side-by-side comparison of two different lamp types, thus enabling consumers to make a more intelligent choice of lamps; taking into account not just the purchase price, but also the operating cost. Manufacturers must begin applying labels by April 28, 1995. Table 107.9 shows the proposed efficiency standards for the fluorescent lamps, and Table 107.10 shows the proposed efficiency standards for incandescent reflector lamps.

There is no requirement to replace all existing lamps in any installation. However, as these lamps burn out, the replacement must meet the new standards.

Replacement for popular fluorescent types includes reduced-wattage energy saving types. These lamps will meet the color and efficiency standards, as will the full wattage triphosphor lamps having a CRI over 69. Table 107.11 shows some types of replacement lamps. On the incandescent side, replacements for the standard incandescent spot and flood lamps will be lower wattage halogen type reflector lamps which do meet the LPW requirements. The halogen and halogen/infrared types of reflector lamps will remain the only type of such lamps on the market. ER and BR types, those intended for rough and vibration service will also be excluded here.

There is also a provision for lighting fixture manufacturers to come up with voluntary luminaire efficiency standards. If these standards are found to be inadequate, the DOE will come up with the mandatory efficiency standards.

The new Energy Policy Act is all-encompassing. It promises to change forever the way industries produce, distribute, and utilize the valued energy resources. The end result should be increased energy security, decreased environmental emissions, and cleaner air and water for all humankind.

Defining Terms

Candlepower distribution: A curve, generally polar, representing the variation of luminous intensity of a lamp or luminaire in a plane through the light center.

Cavity ratio (CR): A number indicating cavity proportions calculated from length, width, and height. It is further defined into ceiling cavity ratio, floor cavity ratio, and room cavity ratio.

Coefficient of utilization (CU): The ratio of the lumens reaching the working plane to the total lumens generated by the lamp. This factor takes into account the efficiency and distribution of the lumanaire, its mounting height, the room proportions, and the reflectances of the walls, ceiling, and floor.

Color preference index (CPI): Measure appraising a light source for appreciative viewing of colored objects or for promoting an optimistic viewpoint by flattery.

Color rendering index (CRI): Measure of the degree of color shift objects undergo when illuminated by the light source as compared with the color of those same objects when illuminated by a reference source of comparable color temperature.

Contrast: The relationship between the luminances of an object and its immediate background. It is equal to $(L_1 - L_2)/L_1$ where L_1 and L_2 are the luminances of the background and object. The ratio $\Delta L/L_1$ is also known as Weber's fraction where $\Delta L = L_1 - L_2$.

Contrast rendition factor (CRF): The ratio of visual task contrast with a given lighting environment to the contrast with sphere illumination.

Equivalent sphere illumination (ESI): The level of sphere illumination which would produce task visibility equivalent to that produced by a specific lighting environment.

Fenestration: Any opening or arrangement of opening (normally filled with media for control) for the admission of daylight.

Footcandle: The unit of illuminance when the foot is taken as the unit of length. It is the illuminance on a surface one square foot in area on which there is a uniformly distributed flux of one lumen.

Illuminance: The density of luminous flux on a surface expressed in either footcandles (lumens/ft^2) or lux (lx). (lux = 0.0929 fc)

Lighting effectiveness factor (LEF): The ratio of equivalent sphere illumination to ordinary measured or calculated illumination.

Light loss factor (LLF): The ratio of the illumination when it reaches its lowest level at the task just before corrective action is taken, to the initial level if none of the contributing loss factors were considered.

Luminance ratio: The ratio between the luminance of two areas in the visual field.

Veiling reflection: Regular reflections superimposed upon diffuse reflections from an object that partially or total obscure the details to be seen by reducing the contrast.

Visual comfort probability (VCP): This rating is based in terms of the percentage of people who will be expected to find the given lighting system acceptable when they are seated in most undesirable locations.

Related Topics

3.1 Voltage and Current Laws • 3.4 Power and Energy

References

ANSI/IES, "Recommended Practices for Industrial Lighting," Illuminating Engineering Society, New York, 1991.

K. Chen, *Energy Effective Industrial Illuminating Systems,* Lilburn, Ga.: The Fairmont Press, 1994.

K. Chen, *Industrial Power Distribution and Illuminating Systems*, New York: Marcel Dekker, 1990.

K. Chen, "New concepts in interior lighting design," *IEEE Trans. Industry Applications*, Sept./ Oct. 1984.

IES Lighting Handbook, Application Volume, Illuminating Engineering Society, New York, 1993.

Lighting Handbook, Westinghouse Electric Corporation, Bloomfield, N.J., 1976.

Further Information

L. Watson, *Lighting Design Handbook,* New York.: McGraw-Hill, 1991. It focuses on the art and process of lighting design and provides invaluable, up-to-date technical details on equipment, color use, scenic projection, lasers, holograms, fiber-optics, computers, and energy conservation.

C.L. Robbins, *Daylighting—Design and Analysis,* New York: Van Nostrand Reinhold, 1986. Organized to correspond to the building design process, the book contains data for calculation of annual cost and energy savings as well as many case studies.

Software—Lighting Calculations by Zonal Cavity Method, Orloff Computer Services, 1820 E. Garry Ave., Santa Ana, CA 92705.

Schmalzel, J.L.. “Instruments”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

108

Instruments

108.1 Introduction
108.2 Physical Variables
108.3 Transducers
108.4 Instrument Elements
108.5 Instrumentation System
108.6 Modeling Elements of an Instrumentation System
108.7 Summary of Noise Reduction Techniques
108.8 Personal Computer-Based Instruments
108.9 Modeling PC-Based Instruments
108.10 The Effects of Sampling
108.11 Other Factors

John L. Schmalzel
Rowan University

108.1 Introduction

Instruments are the means for monitoring or measuring physical variables. The basic elements of an instrumentation application are shown in Fig. 108.1. A physical system produces a *measurand, m*(*t*), shown as time-varying, which is transformed by a **transducer** into an electrical signal, *s*(*t*), that is then processed by an instrument to yield the desired output information variable, *i*(*t*). Producing meaningful information from physical variables requires conversion and processing. Electronic instruments require that physical variables be converted to electrical signals through a process of *transduction*, followed by signal *conditioning* and signal *processing* to obtain useful results.

108.2 Physical Variables

The measurand can be one of many physical variables; the type depends on the application. For example, in process control, typical measurands can include pressure, temperature, and flow. Representative physical variables with corresponding units are summarized in Table 108.1.

108.3 Transducers

Transducers convert one form of energy to another. To be useful for an electronic instrument, a transducer must produce an electrical output such as voltage or current to allow required signal conditioning and signal processing steps to be completed. A variety of transducers are available to meet a measurement requirement; some common examples are listed in Table 108.2. Transducers can be compared based on their operating principles, the measurand range, interface design, and reliability. Khazan [1994] gives a complete summary of transducer schemes.

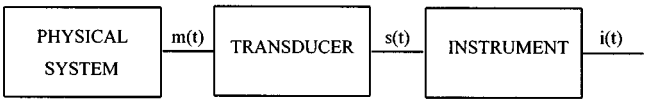


FIGURE 108.1 Generalized block diagram of an instrument applied to a physical measurement.

TABLE 108.1 Representative Physical Variables, Symbols, and Units

Physical Variable	Symbol	SI Units, Abbreviations
Current	I	ampere, A
Energy	E	joule, J
Force	F	newton, N
Flow	Q	volume flow rate, m^3/s
Frequency	f	hertz, Hz
Length	L	meter, m
Mass	m	kilogram, kg
Pressure	P	N/m^2
Power	P	Watt, W
Resistance	R	ohm, Ω
Temperature	T	Kelvin, K
Time	t	second, s
Velocity	V	m/s
Voltage	V	volt, V

TABLE 108.2 Representative Transducers

Measurand	Transducer	Operating Principles
Displacement (Length)	Resistive Capacitive Inductive	Change in resistance, capacitance, or inductance caused by linear or angular displacement of transducer element
Force	Strain gage	Resistance, piezoresistivity
Temperature	Thermistor	Resistance
	Thermocouple	Peltier, seebeck effect
Pressure	Diaphragm	Diaphragm motion sensed by a displacement technique.
Flow	Differential pressure	Pressure drop across restriction
	Turbine	Angular velocity proportional to flow rate

108.4 Instrument Elements

Signal conditioning consists of amplification, filtering, limiting, and other operations that prepare the raw instrument input signal for further operations. The signal may be the output of a transducer or it may be an electrical signal obtained directly from an electronic device or circuit. Signal processing applies some algorithm to the basic signal in order to obtain meaningful information. Signal conditioning and processing operations may be performed using *analog* or *digital* circuit techniques, or using a combination of methods. There are a variety of trade-offs between them. For example, analog methods offer bandwidth advantages, whereas digital techniques offer advanced algorithm support and long-term stability. The use of microprocessors within an instrument makes it possible to perform many useful functions including calibration, linearization, conversion, storage, display, and transmission. A block diagram of a representative microprocessor-based instrument is shown in Fig. 108.2.

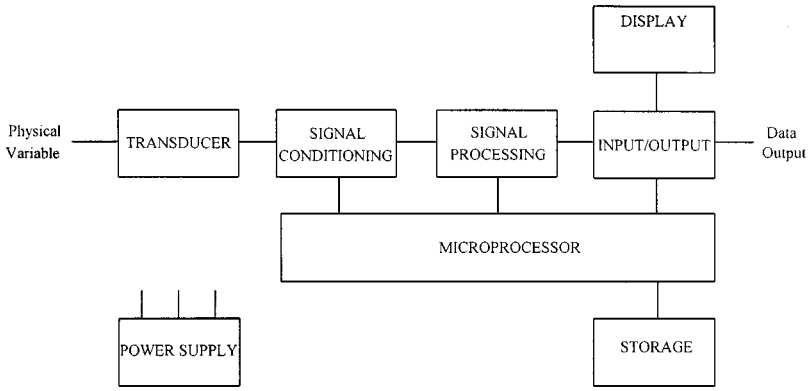


FIGURE 108.2 Block diagram of generalized, microprocessor-based instrument.

108.5 Instrumentation System

An instrument is never used in isolation. The instrumentation components contribute to an overall system response in a number of ways that are based on the **measurement system** elements present. These elements include: (1) sources, (2) interconnect, (3) device or system under test, (4) response measuring equipment, and (5) environmental variables. Figure 108.3 shows the elements of a typical instrumentation system.

108.6 Modeling Elements of an Instrumentation System

Best results are achieved when the instrumentation system is clearly understood, and its effects compensated for when practical. Lumped parameter modeling of the elements shown in Fig. 108.3 provides a means for determining the contribution each element makes to the overall system behavior. Of particular importance are the input and output impedances of each element. In addition, the effects of interconnect and environmental variables can also be modeled to determine their influence on the system. The relative dimensions of the measurement system with respect to the highest frequencies encountered—whether signal or noise—determine whether simplified circuit theory models, or generalized solutions to Maxwell's equations must be used. Generally, if measurement system dimensions are on the order of 1/20 of the shortest wavelength, simple circuit theory models can be used. Operation in this regime also allows impedance matching to be largely ignored; e.g., not requiring mandatory use of 50 Ω sources, 50 Ω transmission lines, and 50 Ω terminations which is commonly encountered in high-frequency systems. Table 108.3 summarizes several common instruments and input or output impedance models corresponding to Fig. 108.4. At low frequencies, interconnect can be modeled by ignoring the very low series resistance and inductance (Zs1, Zs2) terms, and considering only the shunt

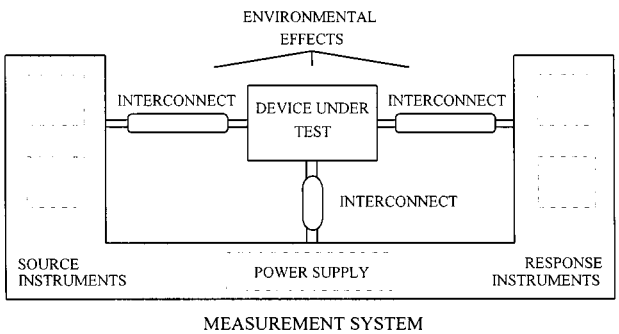


FIGURE 108.3 Fundamental elements of an instrumentation system.

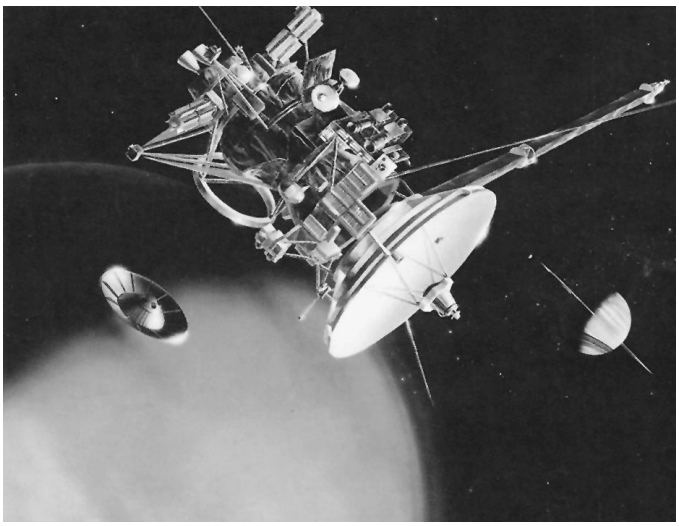
Cassini Spacecraft

One of NASA's latest planetary systems research segments is called the Discovery Program. This program is an effort to develop frequent, small planetary missions that perform high quality scientific investigations. Discovery missions planned for 1997 include the sending of a Mars lander to the planet and launching the Lunar Prospector to map the moon's surface composition.

The principle planetary mission of NASA's Discovery Program is the 1997 launch of Cassini. Cassini is a joint project of NASA, the European Space Agency (ESA), and the Italian Space Agency, and is managed by the Jet Propulsion Laboratory (JPL).

The flight vehicle consists of the main Cassini spacecraft and the ESA-built Huygens Probe, a 750-pound, six instrument package that will descend into the atmosphere of Saturn's moon Titan, which is believed to be chemically similar to the atmosphere of early Earth.

Launched towards the end of 1997, Cassini will make flybys of Venus and Jupiter en route to a rendezvous with Saturn in July 2004. Cassini will release the Huygens Probe during its first orbit, then make approximately 40 revolutions over a span of four years, while the spacecraft's 12 instruments conduct a detailed exploration of the whole Saturnian system, including Titan and the planet's other icy moons. (Courtesy of National Aeronautics and Space Administration.)



This artist's concept shows the Cassini spacecraft orbiting around Saturn, just after deploying a probe that will descend into the atmosphere of Saturn's moon Titan. Launched October 1997, Cassini will reach Saturn in July 2004 and orbit the planet for four years thereafter. (Photo courtesy of National Aeronautics and Space Administration.)

capacitance (Zp) which is in the range of 50 to 150 pF/m for different types of cable. At high frequencies, the characteristic impedance of the interconnect is used; e.g., 50 Ω or 75 Ω for commonly used coaxial cables; 120 Ω for twisted pair.

The response of an entire instrumentation system can be modeled by interconnecting the individual model elements. Figure 108.5 shows an example that was obtained by substituting models for an operational amplifier

TABLE 108.3 Summary of Common Instruments and Their Lumped-Parameter Models

Instrument Description, Model, Manufacturer	Input Impedance, Zi	Output Impedance, Zo
Function generator, FG501A, Tektronix		50 Ω
Multimeter, DM501A, Tektronix	10 MΩ (Volts mode)	
Oscilloscope, 54601A, Hewlett Packard	1 MΩ‖13 pF	

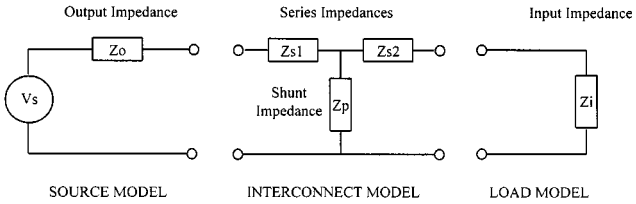


FIGURE 108.4 Simplified output and input models for instrument elements.

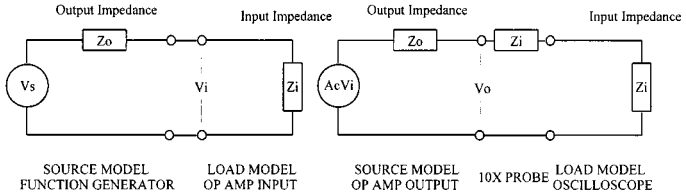


FIGURE 108.5 Model of representative instrumentation system. Each variable would be substituted as required. For example, Vs = 1.0sin2π1000t for a 0.707 Vrms, 1 kHz sine wave; Zo = 50 Ω for the FG501A; Zi = 1 kΩ for an op amp configured as an inverting amplifier with Ri = 1 kΩ; and gain of 10; AcVi = -10.0sin2π1000t; Zo = 1 Ω for low current output; Zi = 9 MΩ‖1.4 pF for a compensated 10X probe; and with Zi = 1 MΩ‖13 pF for the input model of the HP54601A oscilloscope.

(op amp) circuit (corresponds to the device under test in Fig. 108.3) that was driven by a function generator for the source, and that measured the response with an oscilloscope connected to the output of the op amp using a 10X probe. In this application, the impedance of the interconnect between the source and op amp can be neglected since the frequencies are low and the input impedance of the op amp is much greater than that of the cable. The circuit model of the compensated 10X probe contains a very high series impedance (9 MΩ‖1.14 pF) relative to the oscilloscope (1 MΩ‖13 pF) so it cannot be ignored.

The models can be used to determine the *frequency response* of the complete system which describes the magnitude and phase response of the system to sinusoidal, steady-state inputs. This can reveal the contribution of each element to the overall response and helps indicate which elements produce the dominant response. The graphical results of the frequency response analysis is termed a *Bode plot*. If each of N elements has an individual transfer function, $H_i(j\omega)$, $i = 1$ to N, then a composite transfer function can be found for the total system, $T(j\omega)$, which is generally not the simple product of each transfer function, $H_1(j\omega)\cdot H_2(j\omega)\cdot\ldots\cdot H_N(j\omega)$ due to loading effects between elements. The use of a circuit simulation program such as PSpice (MicroSim Corp.) simplifies the investigation into instrument behavior. A library of subcircuit models can be developed for each instrumentation and interconnect element to support measurement system loading effects analysis. For example, a PSpice subcircuit definition for the HP54601A oscilloscope is:

```
.SUBCKT HP54601A 1 2
Cin 1 2 13p
Rin 1 2 1MEG
.ENDS
```

TABLE 108.4 Noise Reduction Checklist

Source	Interconnect	Response
Shield enclosures	Shield leads	Shield enclosures
Filter inputs and outputs	Minimize loop area (twist leads)	Filter inputs and outputs
Limit bandwidth	Keep signal leads near ground	Limit bandwidth
Minimize loop areas	Separate low-, high-level signals	Minimize loop areas
	Keep signal and ground leads short	
	Low f: Use single ground	
	High f: Use multiple grounds	

Source: H.W. Ott, *Noise Reduction Techniques in Electronic Systems*, 2nd ed., New York: John Wiley & Sons, 1988. With permission.

This network model would be added as a load to the output of the device under test in order to predict its loaded behavior.

108.7 Summary of Noise Reduction Techniques

Elimination of undesired measurement errors benefits from a systematic approach to identifying and solving noise problems. Source, interconnect, and response elements of a measurement system can be treated individually. Some techniques, such as shielding, are applicable to all three. Various combinations of techniques should be tried to achieve best results. There are many choices of grounding techniques that vary depending on whether elements are floating or ground-referred, and based on bandwidth. In general, multiple ground connections that create *ground loops* should be avoided. Difficult ground loop problems may require isolation or other techniques to interrupt the ground connection between elements. Table 108.4 summarizes a checklist of noise reduction techniques.

108.8 Personal Computer-Based Instruments

Many instrument functions are available for interface to personal computer (PC) systems. These range from plug-in cards that reside on the PC backplane to standalone instruments that communicate with the PC over standard interfaces such as RS-232 or IEEE-488. Software to control *data acquisition*, analysis, and display completes the computer-based instrument. Examples of such software include *Lab Windows* or *Lab View* (National Instruments), *HP VEE* (Hewlett-Packard), and *Testpoint* (Keithley-Metrabyte). Figure 108.6 shows a block diagram of an output screen developed using *Lab Windows* for an acoustic measurement application. A

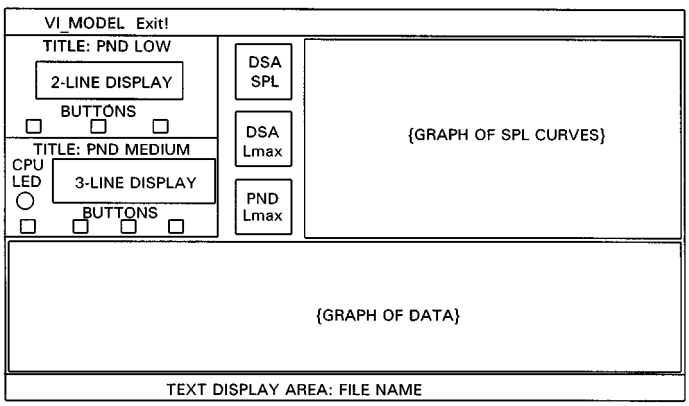


FIGURE 108.6 Example block diagram of a virtual instrument user interface.

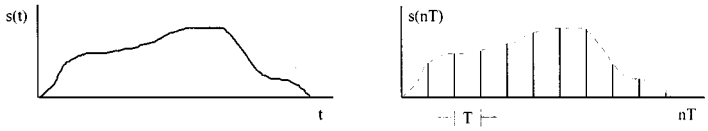


FIGURE 108.7 Sampling a continuous-time signal yields a discrete-time signal.

menu bar provides pull-down options. Several windows simultaneously display selection options and present results graphically and with text.

108.9 Modeling PC-Based Instruments

The approach outlined previously for modeling conventional measurement systems can be extended to PC-based instruments with one major difference: PC-based instruments by their nature are digital machines and perform functions in discrete time. Best performance of PC-based instrument systems must therefore consider *sampled data* effects. Figure 108.7 shows a data acquisition system modeled using an ideal sampler which instantaneously samples a continuous signal, $s(t)$, every T seconds. This yields a sequence, $s(nT)$, of discrete values that represent the value of the continuous signal at integer multiples of T seconds.

108.10 The Effects of Sampling

The Fourier transform of a sampled signal yields a frequency domain function that is periodic in frequency, with a period that is $1/f_s$. The *sampling theorem* states that in order to unambiguously preserve information, the sampling frequency, $f_s = 1/T$, must be at least twice the highest frequency present in the continuous-time signal. If f_s is less than twice the highest frequency, *aliasing* will occur. Aliased frequencies are indistinguishable from one another. A useful method for visualizing this result is through the use of an *aliasing diagram*. An example is shown in Fig. 108.8. Note that the *Nyquist frequency* is defined to be $f_s/2$.

108.11 Other Factors

Other important factors that should be considered when using PC-based instruments over manual counterparts are summarized in Table 108.5. Perhaps the most important choice is the selection of a minimum sampling rate for the data acquisition process. It must be chosen to meet the requirements of the Nyquist frequency. However, in order to ensure that no higher frequencies are present, an *anti-aliasing* low pass filter that eliminates energy above the Nyquist frequency should be employed. In order to provide sufficient transition bandwidth for the filter, a slightly higher sampling rate should generally be employed. A factor of 1.25 to 5 times the minimum f_s is a good compromise. Automated equipment may introduce substantial transients into the measurement system. Sufficient time must be provided for the resulting transients to settle to an acceptable error bound; for example, 1%.

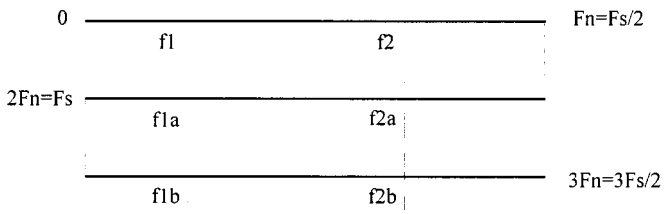


FIGURE 108.8 Aliasing diagram. The two baseband frequencies, f1 and f2, have aliases at frequencies that intersect the vertical dashed lines. For example, using a sampling frequency of 10 kHz (F_n = 5 kHz) with f1 = 1 kHz and f2 = 3.5 kHz, signals at 9 kHz (f_{1a}) and 11 kHz (f_{1b}) would be aliased to 1 kHz (f_1), while signals at 6.5 kHz (f_{2a}) and 13.5 kHz (f_{2b}) would be aliased to 3.5 kHz (f_2).

TABLE 108.5 Automated Measurement Factors

Factor	Consideration
Leveling	Frequency response measurements require use of a leveled generator. Alternatively, store a calibration curve.
Multiplexing	Measurements from multiple nodes require lead switching to shared instruments; consider these effects.
Sampling frequency	Must exceed the Nyquist frequency. Include an anti-aliasing filter. Manual instruments typically use integrating (dual-slope) analog-to-digital converters which give good noise rejection over integer numbers of line cycles. Faster sampling rates for ATE are achieved using successive-approximation or other techniques. User may have to perform averaging as a post-processing step in order to achieve acceptable signal-to-noise ratios.
Settling time	Allow sufficient time for transients to settle for both stimulus/response instruments and device under test.
Storage	Automatic measurements can produce large arrays of data at high speeds. Actual throughput to a hard disk may be much less than the maximum sampling rate of a data acquisition element (plug-in board, external instrument).
Triggering	Choices between free-running, external, and internal.

Defining Terms

Instrument: The means for monitoring or measuring physical variables. Usually includes transducers, signal conditioning, signal processing, and display.

Measurement system: The sum of all stimulus and response instrumentation, device under test, interconnect, environmental variables, and the interaction among the elements.

Transducer: A device that transforms one form of energy to an electrical output that can be processed by an instrument.

Virtual instrument: An instrument created through computer control of instrumentation resources with analysis and display of the data collected.

Related Topics

3.1 Voltage and Current Laws • 8.5 Sampled Data • 73.2 Noise • 112.1 Introduction

References

N. Ahmed and T. Natarajan, *Discrete-Time Signals and Systems,* Reston, Vir.:, Reston Publishing, 1983.

E.O. Doebelin, *Measurement Systems: Application and Design,* 4th ed., New York: McGraw-Hill, 1990.

J.P. Holman, *Experimental Methods for Engineers,* 6th ed., New York: McGraw-Hill, 1994.

A.D. Khazan, *Transducers and Their Elements: Design and Application,* Englewood Cliffs, N.J.: Prentice-Hall, 1994.

H.W. Ott, *Noise Reduction Techniques in Electronic Systems,* 2nd ed., New York: John Wiley & Sons, 1988.

W.J. Tompkins and J.G. Webster, Eds., *Interfacing Sensors to the IBM PC,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.

Further Information

The monthly journals, *IEEE Transactions on Instrumentation and Measurement,* and *IEEE Transactions on Biomedical Instrumentation,* report advances in instrumentation. For subscription information, contact: IEEE Service Center, 445 Hoes Lane, PO Box 1331, Piscataway, NJ 08855-1331. (800) 678-IEEE.

Information about automatic test equipment and software for data acquisition, analysis, and display, can be obtained from several vendors; for example, Hewlett-Packard, Englewood, CO, (800)-829-4444; Keithley-Metrabyte, Taunton, MA, (800) 348-0033; and National Instruments, Austin, TX, (512) 794-0100. Information about transducers can be obtained from Omega International, Stamford, CT, (203) 359-1660.

Kayton, M. "Navigation Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

109

Navigation Systems

109.1 Introduction
109.2 Coordinate Frames
109.3 Categories of Navigation
109.4 Dead Reckoning
109.5 Radio Navigation
109.6 Celestial Navigation
109.7 Map Matching Navigation
109.8 Navigation Software
109.9 Design Trade-Offs

Myron Kayton
Kayton Engineering Co.

109.1 Introduction

Navigation is the determination of the position and velocity of a moving vehicle on land, at sea, in the air, or in space. The three components of position and the three components of velocity make up a six-component **state vector** that fully describes the translational motion of the vehicle because the differential equations of motion are of second order. Surveyors are beginning to use the same sensors as navigators but are achieving higher accuracy as a result of longer periods of observation, a fixed location, and more complex, non-real-time data reduction.

In the usual navigation system, the state vector is derived on-board, displayed to the crew, recorded on-board, or transmitted to the ground. Navigation information is usually sent to other on-board subsystems; for example, to the waypoint steering, engine control, communication control, and weapon-control computers. Some navigation systems, called *position-location systems*, measure a vehicle's state vector using sensors on the ground or in another vehicle (Section 109.5). The external sensors usually track passive radar returns or a transponder. Position-location systems usually supply information to a dispatch or control center.

Traditionally, *ship navigation* included the art of pilotage—entering and leaving port, making use of wind and tides, and knowing the coasts and sea conditions. However, in modern usage, navigation is confined to the measurement of the state vector. The handling of the vehicle is called *conning* for ships, *flight control* for aircraft, and *attitude control* for spacecraft.

The term *guidance* has two meanings, both of which are different than *navigation*:

1. Steering toward a destination of known position from the vehicle's present position, as measured by a navigation system. The steering equations on a planet are derived from a plane triangle for nearby destinations and from a spherical triangle for distant destinations.
2. Steering toward a destination without calculating the state vector explicitly. A guided vehicle homes on radio, infrared, or visual emissions. Guidance toward a *moving* target is usually of interest to military tactical missiles in which a steering algorithm assures impact within the maneuver and fuel constraints of the interceptor. Guidance toward a *fixed* target involves beam riding, as in the Instrument Landing System, Section 109.5.

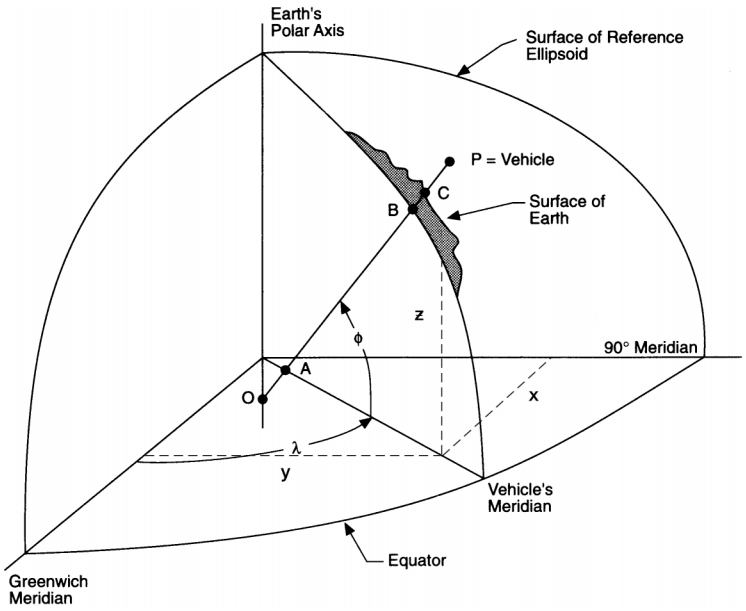


FIGURE 109.1 Latitude-longitude-altitude coordinate frame. ϕ = geodetic latitude; OP is normal to the ellipsoid at B; λ = geodetic longitude; h = $\overline{BP}$ = altitude above the reference ellipsoid = altitude above mean sea level.

109.2 Coordinate Frames

Navigation is with respect to a coordinate frame of the designer's choice. Short-range robots navigate with respect to the local terrain or a building's walls. For navigation over hundreds of kilometers (e.g., automobiles and trucks), various map grids exist whose coordinates can be calculated from latitude-longitude (Fig. 109.1). NATO land vehicles use a Universal Transverse Mercator grid. Long-range aircraft and ships navigate relative to an earth-bound coordinate frame, the most common of which are latitude-longitude-altitude and rectangular x, y, z (Fig. 109.1). The most accurate world-wide reference ellipsoid is described in [WGS-84, 1991]. Spacecraft in orbit around the earth navigate with respect to an earth-centered, inertially nonrotating coordinate frame whose *z* axis coincides with the polar axis of the earth and whose *x* axis lies along the equator. Interplanetary spacecraft navigate with respect to a sun-centered, inertially nonrotating coordinate frame whose *z* axis is perpendicular to the **ecliptic** and whose *x* axis points to a convenient star [Battin, 1987].

109.3 Categories of Navigation

Navigation systems can be categorized as:

1. *Absolute navigation systems* that measure the state vector without regard to the path traveled by the vehicle in the past. These are of two kinds:
   - Radio systems (Section 109.5). They consist of a network of transmitters (sometimes also receivers) on the ground or in satellites. A vehicle detects the transmissions and computes its position relative to the known positions of the stations in the navigation coordinate frame. The vehicle's velocity is measured from the Doppler shift of the transmissions or from a sequence of position measurements.
   - Celestial systems (Section 109.6). They measure the elevation and azimuth of celestial bodies relative to the land level and North. Electronic star sensors are used in special-purpose high-altitude aircraft and in spacecraft. Manual celestial navigation was practiced at sea for millennia (see Bowditch).
2. *Dead-reckoning navigation systems* that derive their state vector from a continuous series of measurements beginning at a known initial position. There are two kinds, those that measure vehicle heading and either

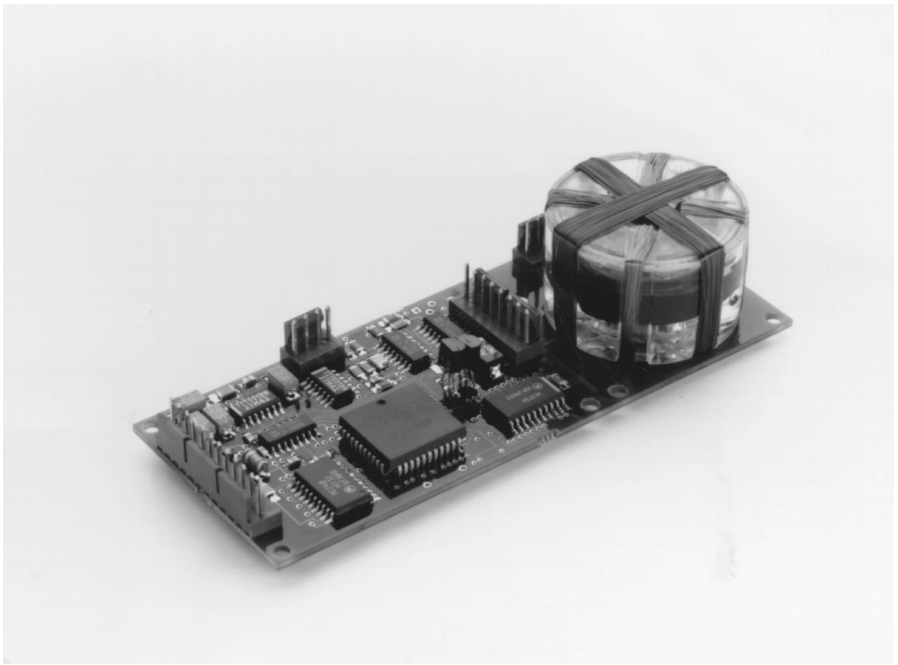


FIGURE 109.2 Saturated core ("flux-gate") magnetometer, mounted on a "compass engine" board. The two orthogonal sensing coils (visible) and the drive coil, wound on the toroidal core, measure two components of the magnetic field in the plane of the toroid. (Courtesy of KVH Industries, Inc.)

speed or acceleration (Section 109.4) and those that measure emissions from continuous-wave radio stations whose signals create ambiguous "**lanes**" (Section 109.5).

Dead reckoning systems must be reinitialized as errors accumulate and if power is lost.

3. *Mapping navigation systems* that observe and recognize images of the ground, profiles of altitude, sequences of turns, or external features (Section 109.7). They compare their observations to a stored database, often on compact disc.

109.4 Dead Reckoning

The simplest dead-reckoning systems measure vehicle heading and speed, resolve speed into the navigation coordinates, then integrate to obtain position (Fig. 109.3). The oldest heading sensor is the magnetic compass, a magnetized needle or electrically excited toroidal core (called a *flux* gate), as shown in Fig. 109.2. It measures the direction of the earth's magnetic field to an accuracy of 2 degrees at a steady velocity below 60-degrees magnetic latitude. The horizontal component of the magnetic field points toward *magnetic north.* The angle from true to magnetic north is called *magnetic variation* and is stored in the computers of modern vehicles as a function of position over the region of anticipated travel [Quinn, 1996]. *Magnetic deviations* caused by iron in the vehicle can exceed 30 degrees and must be compensated in the navigation computer or, in older ships, by placing compensating magnets near the sensor.

A more complex heading sensor is the *gyrocompass*, consisting of a spinning wheel whose axle is constrained to the horizontal plane (often by a pendulum). The ships' version points north, when properly compensated for vehicle motion, and exhibits errors less than a degree. The aircraft version (more properly called a *directional gyroscope*) holds any preset heading relative to earth and drifts at 50 deg/hr or more. Inexpensive gyroscopes (some built on silicon chips as vibrating beams with on-chip signal conditioning) are often coupled to magnetic compasses to reduce maneuver-induced errors.

The simplest speed-sensor is a wheel odometer that generates electrical pulses. Ships use a dynamic-pressure probe or an electric-field sensor that measures the speed of the hull through the conductive water. Aircraft

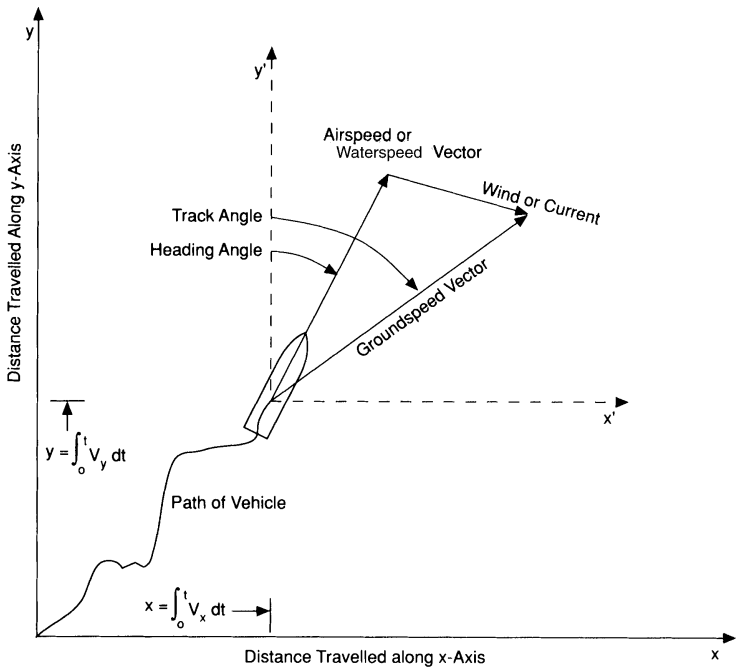


FIGURE 109.3 Geometry of dead reckoning.

measure the dynamic pressure of the air stream from which they derive airspeed in an *air-data* computer. The velocity of the wind or sea current must be vectorially added to that of the vehicle, as measured by a dynamic-pressure sensor (Fig. 109.3). Hence, unpredicted wind or current will introduce an error into the dead-reckoning computation. Most sensors are insensitive to the component of airspeed or waterspeed normal to their axis (*leeway* in a ship, *drift* in an aircraft). A Doppler radar measures the frequency shift in radar returns from the ground or water below the aircraft, from which speed is inferred. A Doppler sonar measures a ship's speed relative to the water layer or ocean floor from which the beam reflects. Multibeam Doppler radars or sonars can measure all the components of the vehicle's velocity. Doppler radars are widely used on military helicopters.

The most complex dead-reckoning system is an *inertial navigator* in which accelerometers measure the vehicle's acceleration while gyroscopes measure the orientation of the accelerometers. An on-board computer resolves the accelerations into navigation coordinates and integrates them to obtain velocity and position. The gyroscopes and accelerometers are mounted in either of two ways:

1. In servoed gimbals that angularly isolate them from rotations of the vehicle.
2. Fastened directly to the vehicle ("strap-down"), whereupon the sensors are exposed to the maximum angular rates and accelerations of the vehicle (Fig. 109.4).

Inertial-quality gyroscopes measure vehicle orientation within 0.1 degree for steering and pointing. Most accelerometers consist of a gram-sized proof-mass mounted on flexure pivots. The newest accelerometers, not yet of inertial grade, are etched into silicon chips. Older gyroscopes contained metal wheels rotating in ball bearings or gas bearings. The newest gyroscopes are evacuated cavities or optical fibers in which counter-rotating laser beams are compared in phase to measure the sensor's angular velocity relative to **inertial space** about an axis normal to the plane of the beams. Vibrating hemispheres and rotating vibrating tines are the basis of some navigation-quality gyroscopes (drift rates less than 0.1 deg/h).

Fault-tolerant configurations of cleverly oriented redundant gyroscopes and accelerometers (typically four to six) detect and correct sensor failures. Inertial navigators are used aboard naval ships, in airliners, in most military fixed-wing aircraft, in space boosters and entry vehicles, in manned spacecraft, in tanks, and on large mobile artillery pieces.

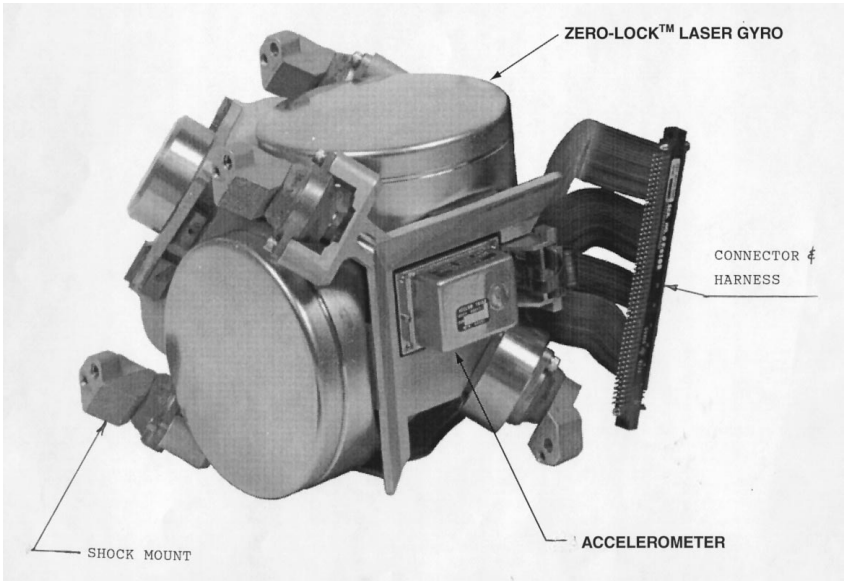


FIGURE 109.4 Inertial reference unit. Two laser gyroscopes (flat discs), an accelerometer, an electrical connector, and three shock mounts are visible. This unit is used in Airbuses and many military aircraft such as the F-18 and Comanche helicopter. (Courtesy of Litton Guidance and Control Systems.)

109.5 Radio Navigation

Scores of radio navigation aids have been invented and many of them have been widely deployed, as summarized in Table 109.1.

The most precise is the global positioning system (GPS), a network of 24 satellites and a half-dozen ground stations for monitoring and control. A vehicle derives its three-dimensional position and velocity from ranging signals at 1.575 GHz received from four or more satellites (military users also receive 1.227 GHz). The former Soviet Union deployed a similar system, called GLONASS. GPS offers better than 100-m ranging errors to civil users and 15-m ranging errors to military users. Simple receivers were available for less than $300 in 1997. They are used on highways, in low-rise cities, at sea, in aircraft, and in low-orbit spacecraft. GPS provides continuous worldwide navigation for the first time in history. It will make dead reckoning unnecessary on many vehicles and will reduce the cost of most navigation systems. Figure 109.5 is an artist's drawing of a GPS Block 2F spacecraft, scheduled for launch in the year 2002.

Differential GPS (DGPS) employs one or more ground stations at known locations, that receive GPS signals and transmit measured errors on a radio link to nearby ships and aircraft. DGPS improves accuracy (centimeters for fixed observers) and detects faults in GPS satellites. In 1997, the U.S. was conducting experiments with a nationwide DGPS system of about 25 stations. This *Wide Area Augmentation System* (WAAS) could eventually replace VORTAC and Category I ILS. A denser network of DGPS stations and GPS-emulating pseudolites, whose stations are located at airports, might replace ILS and MLS (below). In 1997, the cost, accuracy, and reliability of such a *Local Area Augmentation System* (LAAS) were still being compared to existing landing aids but marine LAAS were in operation for navigation into harbors in North America, the North Sea, and the Baltic Sea.

The most widely used marine radio aid in 1997 was Loran-C (see Table 109.1). The 100-kHz signals are usable within 1000 **nautical miles (nmi)** of a "chain" consisting of three or four stations. Chains cover the United States, parts of western Europe, Japan, Saudi Arabia, and a few other areas. The former Soviet Union has a compatible system called Chaika. The vehicle-borne receiver measures the difference in time of arrival of pulses emitted by two stations, thus locating the vehicle on one branch of a hyperbola. Two or more station pairs give a two-dimensional position fix whose typical accuracy is 0.25 nmi, limited by propagation uncertainties over the terrain between the transmitting station and the user. The measurement of 100-microsecond

TABLE 109.1 Worldwide Radio Navigation Aids

System	Frequency Hz	Band	Number of Stations	Number of Users in 1996 Air	Marine	Space	Land
Omega	10–13 kHz	VLF	8	15,000	10,000	0	0
Loran-C/Chaika	100 kHz	LF	50	120,000	550,000	0	25,000
Decca	70–130 kHz	LF	150	2,000	20,000	0	0
Beacons*	200–1600 kHz	MF	4000	130,000	500,000	0	0
Instrument Landing System (ILS)*	108–112 MHz 329–335 MHz	VHF UHF	1500	150,000	0	0	0
VOR*	108–118 MHz	VHF	1500	180,000	0	0	0
SARSAT/COSPAS	121.5 MHz 243,406 MHz	VHF UHF	5 satellites	200,000	200,000	0	100,000
Transit	150, 400 MHz	VHF	7 satellites	0	0	0	0
PLRS	420–450 MHz	UHF	None	0	0	0	2,000
JTIDS	960–1213 MHz	L	None	500	0	0	0
DME*	962–1213 MHz	L	1500	90,000	0	4	0
Tacan*	962–1213 MHz	L	850	15,000	0	4	0
Secondary Surveillance Radar (SSR)* Identification Friend or Foe (IFF)	1030, 1090 MHz	L	800	250,000	0	0	0
GPS-GLONASS	1227, 1575 MHz	L	24 + 24 satellites	120,000	275,000	4	125,000
Satellite Control Network (SCN)	1760–1850 MHz 2200–2300 MHz	S S	10	0	0	200	0
Spaceflight Tracking and Data Network (STDN)	2025–2150 MHz 2200–2300 MHz	S	3 satellites 10 ground	0	0	50	0
Radar Altimeter	4200 MHz	C	None	20,000	0	0	0
MLS*	5031–5091 MHz	C	30	100	0	0	0
FPQ-6, FPQ-16 radar	5.4–5.9 GHz	C	10	0	0	0	0
Weather/map radar	10 GHz	X	None	10,000	0	0	0
Shuttle rendezvous radar	13.9 GHz	Ku	None	0	0	4	0
Airborne Doppler radar	13–16 GHz	Ku	None	20,000	0	0	0
SPN-41 carrier-landing monitor	15 GHz	Ku	25	1600	0	0	0
SPN-42/46 carrier-landing radar	33 GHz	Ka	25	1600	0	0	0

*Standardized by International Civil Aviation Organization.

time difference is possible with low-quality clocks in the vehicles. Loran is also used by general aviation aircraft for en-route navigation and for nonprecision approaches to airports (in which the cloud bottoms are more than 200 feet above the runway). Loran service will probably be discontinued at the beginning of the 21st century.

The most widely used aircraft radio aid is VORTAC, whose stations offer three services:

1. Analog bearing measurements at 108 to 118 MHz (called VOR). The vehicle compares the phases of a rotating cardioid pattern and an omnidirectional sinusoid emitted by the ground station.
2. Pulse distance measurements (DME) at 1 GHz by measuring the time delay for an aircraft to interrogate a VORTAC station and receive a reply,
3. Tacan bearing information conveyed in the amplitude modulation of the DME replies from the VORTAC stations.

Omega is a worldwide radio aid consisting of eight radio stations that emit continuous sine waves at 10 to 13 kHz. Vehicles with precise clocks measure their range to a station by observing the absolute time of reception. Other vehicles measure the range differences between two stations in the form of phase differences between the received sinusoids. Differential Omega creates hyperbolic "lanes" that are 10 to 150 nmi wide. The lanes are indistinguishable from each other by measuring phase; hence the vehicle must count lanes from a point of known position. Errors are about 2 nmi due to radio propagation irregularities. Omega is used by submarines,

FIGURE 109.5 Global positioning satellite, Block 2F. (Courtesy of Rockwell.)

over-ocean general-aviation aircraft, and a few international air carriers. It was scheduled to be decommissioned in 1997.

Landing guidance throughout the western world, and increasingly in China, India, and the former Soviet Union, is with the Instrument Landing System (ILS). Transmitters adjacent to the runway create a horizontal guidance signal near 110 MHz and a vertical guidance signal near 330 MHz. Both signals are modulated such that the nulls intersect along a line in space that leads an aircraft from a distance of about 10 nmi to within 50 ft above the runway. ILS gives no information about where the aircraft is located along the beam except at two or three vertical *marker beacons*. Most ILS installations are certified to the International Civil Aviation Organization's (ICAO) *Category I*, where the pilot must abort the landing if the runway is not visible at an altitude of 200 ft. One hundred ILSs (in 1996) were certified to *Category II*, which allows the aircraft to descend to 100 ft before aborting for lack of visibility. *Category III* allows an aircraft to land at still lower weather ceilings. Category III landing aids are of special interest in Western Europe, which has the worst flying weather in the developed world. Category III ILS detects its own failures and switches to a redundant channel within one second to protect aircraft that are flaring-out (within 50 ft of the runway) and can no longer execute a missed approach. Once above the runway, the aircraft's bottom-mounted radar altimeter measures altitude and either the electronics or the pilot guides the flare maneuver. Landing aids are described by Kayton and Fried [1997].

Throughout the western world, civil aircraft use VOR/DME whereas military aircraft use Tacan/DME for en-route navigation. In the 1990s, China and the successor states to the Soviet Union were installing ICAO-standard navigation aids (VOR, DME, ILS) at their international airports and along the corridors that lead to them from the borders. Overflying western aircraft navigate inertially, with Omega, or with GPS. Domestic flights within the Soviet Union depended on radar tracking, non-directional beacons, and an L-band range-angle system called "RSBN". They will eventually upgrade to a satellite-based enroute and landing system.

U.S. Navy aircraft use a microwave scanning system at 15.6 GHz to land on aircraft carriers; NASA's space shuttle uses the Navy system to land at its spaceports. Another microwave landing system (MLS) at 5 GHz was supposed to replace the ILS in civil operations, especially for Categories II and III. However, experiments from

1990 to 1997 showed that differential GPS could achieve an accuracy better than 1 m as a landing aid. Hence, it is likely that a LAAS will replace or supplement ILS, which has been guaranteed to remain in service at least until the year 2010 (Federal Radionavigation Plan). NATO may use MLS or a LAAS as a portable landing aid for tactical airstrips.

All the space-faring nations operate worldwide radio networks that track spacecraft, compute their state vectors, and predict future state vectors using complex models of gravity, atmospheric drag, and lunisolar perturbations. NASA operates three tracking and data relay satellites (TDRS) that track spacecraft in low earth orbit with accuracies of 10 to 50 m and 0.3 m/s. Specialized ground-based tracking stations monitor and reposition the world's many communication satellites [Berlin, 1988]. Other specialized stations track and communicate with deep space probes. They achieve accuracies of 30 m and a few centimeters per second, even at enormous interplanetary distances, due to long periods of observation and precise orbit equations (see [Yuan, 1983]).

Position-location and position-reporting systems monitor the state vectors of many vehicles and usually display the data in a control room or dispatch center. Some vehicles derive their state vector from the ranging modulations; others merely report an independently derived position. Table 109.1 lists *Secondary Surveillance Radars* that receive coded replies from aircraft so they can be identified by human controllers and by collision-avoidance algorithms. The table also lists the U.S. NASA and military spacecraft-tracking networks (STDN and SCN). Tracking and reporting systems have long been in use at marine ports, for airplane traffic control and for space vehicles. They are increasingly being installed in fire trucks, police cars, ambulances, and delivery-truck fleets that report to a control center. The aeronautical bureaucracy calls them *Automatic Dependent Surveillance* (ADS) systems. The continuous broadcast of on-board-derived position (probably GPS-based) may become the basis of the worldwide air traffic control system of the early 21st century.

Several commercial communication satellites plan to offer digital-ranging services worldwide. The intermittent nature of commercial fixes would require that vehicles dead-reckon between fixes, perhaps using solid-state inertial instruments. Thus, if taxpayers insist on collecting fees for service, private comm-nav networks may replace the government-funded GPS and air-traffic communication network in the next century. Worldwide traffic control over oceans and undeveloped land areas would become possible.

Military communication-navigation systems measure the position of air, land, and naval vehicles on battlefields and report to headquarters; examples are the American Joint Tactical Information Distribution System (JTIDS) and the Position Location Reporting System (PLRS).

A worldwide network of SARSAT-COSPAS stations monitors signals from satellite-based transponders listening on 121.5, 243, and 406 MHz, the three international distress frequencies. Software at the listening stations calculates the position of Emergency Location Transmitters within 20 kilometers, based on the observed Doppler-shift history, so that rescue vehicles can be dispatched. Thousands of lives have been saved worldwide, from arctic bush-pilots to tropical fishermen.

109.6 Celestial Navigation

Human navigators use sextants to measure the elevation angle of celestial bodies above the visible horizon. The peak elevation angle occurs at local noon or midnight:

$$\text{elev angle (degrees)} = 90 - \text{latitude} + \text{declination}$$

Thus at local noon or midnight, latitude can be calculated by simple arithmetic. Tables of declination, the angle of the sun or star above the earth's equatorial plane, were part of the ancient navigator's proprietary lore. The declination of the sun was first publicly tabulated in the fifteenth century in Spain. When time became measurable at sea, with a chronometer in the nineteenth century and by radio in the twentieth century, off-meridian observations of the elevation of two or more celestial bodies were possible at any known time of night (cloud cover permitting). These fixes were hand-calculated using logarithms, then plotted on charts. In the 1930s, hand-held sextants were built that measured the elevation of celestial bodies from an aircraft using a bubble-level reference instead of the horizon. The accuracy of celestial fixes was 3–10 miles at sea and 5–20 miles in the air, limited by the uncertainty in the horizon and the inability to make precise angular measurements on a pitching, rolling vehicle. Kayton (1990) reviews the history of celestial navigation at sea and in the air.

The first automatic star trackers were built in the late 1950s. They measured the azimuth and elevation of stars relative to a gyroscopically stabilized platform. Approximate position measurements by dead reckoning allowed the telescope to point within a fraction of a degree of the desired star. Thus, a narrow field-of-view was possible, permitting the telescope and photodetector to track stars in the daytime. An on-board computer stored the right ascension and declination of 20–100 stars and computed the vehicle's position. Automatic star trackers are used in long-range military aircraft and on space shuttles in conjunction with inertial navigators. Clever design of the optics and of stellar-inertial signal-processing filters achieves accuracies better than 500 ft [Kayton and Fried, 1997].

Spacecraft use the line-of-sight to the sun and stars to measure orientation (for *attitude control*). Earth-pointing spacecraft usually carry horizon scanners that locate the center of the earth's carbon-dioxide disc. All spacecraft navigate by radio tracking from earth. When interplanetary spacecraft approach the target planet, the navigation computers (on earth) transform from sun-centered to planet-centered coordinates by observing star occultations and transmitting the images to earth for human interpretation. During the Apollo translunar missions, crews experimentally measured the angle between celestial bodies and the earth or moon with a specially designed manual sextant coupled to a digital computer which calculated the state vector. Other experiments have been made in which American and Soviet crews used manual sextants to observe the angle between celestial bodies and landmarks on earth, from which state vectors were calculated. Autonomous land vehicles on other planets and certain military spacecraft may need celestial navigation.

109.7 Map-Matching Navigation

As computer power grows, map-matching navigation is becoming more important. On aircraft, mapping radars and optical sensors present a visual image of the terrain to the crew. Automatic map-matchers have been built, since the 1960s, that correlate the observed image to stored images, choosing the closest match to update the dead-reckoned state vector. More commonly, aircraft and cruise missiles measure the vertical profile of the terrain below the vehicle and match it to a stored profile. Matching profiles, perhaps hourly, reduces the long-term drift of their inertial navigators. The profile of the terrain is measured by subtracting the readings of a baro-inertial altimeter (calibrated for altitude above sea level) and a radar altimeter (measuring terrain clearance). An on-board computer calculates the autocorrelation function between the measured profile and each of many stored profiles on possible parallel paths of the vehicle. The on-board inertial navigator usually contains a digital filter that corrects the drift of the azimuth gyroscope as a sequence of fixes is obtained. Hence the direction of flight through the stored map is known, saving the considerable computation time that would be needed to correlate for an unknown azimuth of the flight path. Marine versions profile the seafloor with a sonar and compare the measured profile to stored bottom maps.

GPS is adequate for automotive navigation except in high-rise cities, in tunnels, and on streets with heavy foliage. To fill coverage gaps, map-matching software can take advantage of the fact that the vehicle remains on roads. On the highway, dead-reckoning or GPS errors can be rectified to the nearest road. In cities, turns can be correlated with the nearest intersection of matching geometry. An accuracy of several meters is possible if all streets are included on the stored map (e.g., alleys, driveways, and parking garages).

The most complex mapping systems observe their surroundings, usually by digitized video, and create their own map of the surrounding terrain. Guidance software then steers the vehicle. In 1997, such systems were in development for hazardous sites such as nuclear plants, waste-disposal facilities, and battlefields, and for unmanned planetary exploration.

Delivery robots in buildings are furnished with a map and need only find their successive destinations while avoiding obstacles. They navigate by following stripes on the floor, by observing infrared beacons, or by observing the returns from on-board ultrasonic sonar or laser radar.

109.8 Navigation Software

Navigation software is sometimes embedded in a central processor with other avionic-system software, sometimes confined to one or more navigation computers. The navigation software contains algorithms and data

GPS Positioning System Delivers High Accuracy

A system for real-time differential GPS (DGPS) positioning will deliver submeter accuracy to Earth satellites and ground-based users worldwide. Developed at NASA's Jet Propulsions Laboratory, the system could improve real-time position accuracy to a few decimeters for single-frequency users and 10 cm or better for dual frequency users. In addition to high accuracy, the system provides nearly complete separation of GPS orbit and clock corrections and continuous determination of inter-frequency delay biases for all GPS satellites and reference receivers.

Key features include: the use of dynamic orbit estimation, which depends on high-accuracy satellite force models, signal models, geophysical models, and geometric models, in a Kalman filter formulation; use of real-time stochastic estimation to minimize orbit and clock errors arising from quasi-random variations in atmospheric propagation relays and solar radiation pressure; simultaneous processing of smoothed pseudorange and continuous carrier phase data; and use of the stable solar-magnetic reference frame, rather than an Earth-fixed frame, in computing the ionosphere corrections.

System operation began in January 1997. Early tests show approximate user differential range errors of less than 20 cm throughout the coverage area, with a North American reference network only. More comprehensive tests with additional global reference sites will be conducted. (Reprinted with permission of *NASA Tech Briefs,* 20(10), 30, 1996).

that process the measurements made by each sensor (e.g., inertial or air data). It contains calibration constants, initialization sequences, self-test algorithms, reasonability tests, and alternative algorithms for periods when sensors have failed or are not receiving information. In the simplest systems, the state vector is calculated independently from each sensor; most often, the navigation software contains multisensor algorithms that calculate the best estimate of position and velocity from several sensors. Prior to 1970, the best estimate was calculated from a least squares algorithm with constant weighting functions or from a frequency-domain filter with constant coefficients. Now, a *Kalman filter* calculates the best estimate from mathematical models of the dynamics of each sensor.

Digital maps, often stored on compact disc, are carried on some aircraft and land vehicles so position can be visually displayed to the crew. Military aircraft superimpose their navigated position on a stored map of terrain and cultural features to aid in the penetration of and escape from enemy territory. Civil operators had not invested in digital data bases as of 1996. Algorithms for waypoint steering and for control of the vehicle's attitude are contained in the software of the *flight management* and *flight control* subsystems.

Specially equipped aircraft (sometimes ships) are often used for the routine calibration of radio navigation aids, speed and velocity sensors, heading sensors, and new algorithms.

109.9 Design Trade-Offs

The designers of a navigation system conduct trade-offs for each vehicle to determine which navigation systems to use. Tradeoffs consider the following attributes:

- *Cost*, including the construction and maintenance of transmitter stations and the purchase of on-board electronics and software. Users are concerned only with the costs of on-board hardware and software.
- *Accuracy* of position and velocity, which is specified as a circular error probable (CEP, in meters or nautical miles). The maximum allowable CEP is often based on the calculated risk of collision on a typical mission.

- *Autonomy*, the extent to which the vehicle determines its own position and velocity without external aids. Autonomy is important to certain military vehicles and to civil vehicles operating in areas of inadequate radio-navigation coverage.
- *Time delay* in calculating position and velocity, caused by computational and sensor delays.
- *Geographic coverage*. Radio systems operating below 100 kHz can be received beyond line of sight on earth; those operating above 100 MHz are confined to line of sight. On other planets, new navigation aids—perhaps navigation satellites or ground stations—will be installed,
- *Automation*. The vehicle's operator (on-board crew or ground controller) receives a direct reading of position, velocity, and equipment status, usually without human intervention. The navigator's crew station disappeared in aircraft in the 1970s. Human navigators are becoming scarce, even on ships, in the 1990s, because electronic equipment automatically selects stations, calculates waypoint steering, and accommodates failures.

Defining Terms

Circular Error Probable (CEP): Radius of a circle, centered at the destination, that contains 50% of the navigation measurements from a large sample.

Ecliptic: Plane of earth's orbit around the sun.

Inertial Space: Any coordinate frame whose origin is on a freely falling (orbiting) body and whose axes are nonrotating relative to the fixed stars. It is definable within 10^{-7} degree/h.

Lanes: Hyperbolic bands on the earth's surface in which continuous-wave radio signals repeat in phase.

Nautical Mile (nmi): 1852 m, exactly. Approximately 1 min of arc on the earth's surface.

State vector: Six-component vector, three of whose elements are position and three of whose elements are velocity.

Update: The intermittent resetting of the dead-reckoned state vector based on absolute navigation measurements (see Section 109.3).

Related Topic

102.2 Communications Satellite Systems: Applications

References

R.H. Battin, *An Introduction to the Mathematics and Methods of Astrodynamics,* Washington: AIAA Press, 1987, 796 pp.

P. Berlin, *The Geostationary Applications Satellite,* Cambridge: Cambridge University Press, 1988, 214 pp.

N. Bowditch, *The American Practical Navigator,* Washington, D.C.: U.S. Government Printing Office, 1995, 873 pp.

M. Kayton, *Navigation: Land, Sea, Air, and Space,* New York: IEEE Press, 1990, 461 pp.

M. Kayton and W.R. Fried, *Avionics Navigation Systems,* 2nd ed., New York: Wiley, 1997, 773 pp.

R.A. Minzner, *The U.S. Standard Atmosphere 1976,* NOAA Report 76-1562, NASA SP-390, 1976 or latest edition, 227 pp.

NASA, *Space Network Users Guide,* Greenbelt, Md.: Goddard Space Flight Center, 1988 or latest edition, 500 pp.

B.W. Parkinson and J.J. Spilker, Eds., *Global Positioning System, Theory and Applications,* American Institute of Aeronautics and Astronautics, 1996, 1300 pp., 2 vols.

J. Quinn, "1995 revision of joint U.S./U.K. geomagnetic field models," *J. Geomagnetism and Geo-Electricity,* 1996.

U.S. Air Force, *NAVSTAR-GPS Interface Control Document,* Annapolis, Md.: ARINC Research, 1991, 115 pp.

U.S. Government, *Federal Radionavigation Plan,* Department of Transportation, 1996, 229 pp., issued biennially

WGS-84, U.S. Defense Mapping Agency, *World Geodetic System 1984,* Washington, D.C.: 1991.

J. Yuen, *Deep Space Telecommunication Systems Engineering,* New York: Plenum Press, 1983, 603 pp.

Y. Zhao, *Vehicle Location and Navigation Systems,* Massachusetts: Artech House, 1997, 345 pp.

Further Information

IEEE Transactions on Aerospace and Electronic Systems, bimonthly through 1991, now quarterly.
Proceedings of the IEEE Position Location and Navigation Symposium (PLANS), biennially.
Navigation, journal of the U.S. Institute of Navigation, quarterly.
Journal of Navigation, Royal Institute of Navigation (UK), quarterly.
AIAA Journal of Guidance and Control, bimonthly.
Commercial aeronautical standards produced by International Civil Aviation Organization (ICAO, Montreal), Aeronautical Radio, Inc. (ARINC, Annapolis, Md.), Radio Technical Commission for Aeronautics (RTCA, Inc., Washington) and European Organization for Civil Aviation Equipment (EUROCAE, Paris).

Ramakumar, R. "Reliability Engineering"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

110

Reliability Engineering[1]

110.1 Introduction
110.2 Catastrophic Failure Models
110.3 The Bathtub Curve
110.4 Mean Time To Failure (MTTF)
110.5 Average Failure Rate
110.6 *A Posteriori* Failure Probability
110.7 Units for Failure Rates
110.8 Application of the Binomial Distribution
110.9 Application of the Poisson Distribution
110.10 The Exponential Distribution
110.11 The Weibull Distribution
110.12 Combinatorial Aspects
110.13 Modeling Maintenance
110.14 Markov Models
110.15 Binary Model for a Repairable Component
110.16 Two Dissimilar Repairable Components
110.17 Two Identical Repairable Components
110.18 Frequency and Duration Techniques
110.19 Applications of Markov Process
110.20 Some Useful Approximations
110.21 Application Aspects
110.22 Reliability and Economics

R. Ramakumar
Oklahoma State University

110.1 Introduction

Reliability engineering is a vast field and it has grown significantly during the past five decades (since World War II). The two major approaches to reliability assessment and prediction are (1) traditional methods based on probabilistic assessment of field data and (2) methods based on the analysis of failure mechanisms and physics of failure. The latter is more accurate, but is difficult and time consuming to implement. The first one, in spite of its many flaws, continues to be in use. Some of the many areas encompassing reliability engineering are reliability allocation and optimization, reliability growth and modeling, reliability testing including accelerated testing, data analysis and graphical techniques, quality control and acceptance sampling, maintenance engineering, repairable system modeling and analysis, software reliability, system safety analysis, Bayesian analysis, reliability management, simulation and Monte Carlo techniques, Failure Modes, Effects and Criticality Analysis (FMECA), and economic aspects of reliability, to mention a few.

Application of reliability techniques is gaining importance in all branches of engineering because of its effectiveness in the detection, prevention, and correction of failures in the design, manufacturing, and operational

[1]Some of the material in this chapter was previously published by CRC Press in *The Engineering Handbook, R. C. Dorf, Ed.*, 1996.

Information Management System for Manufacturing Efficiency

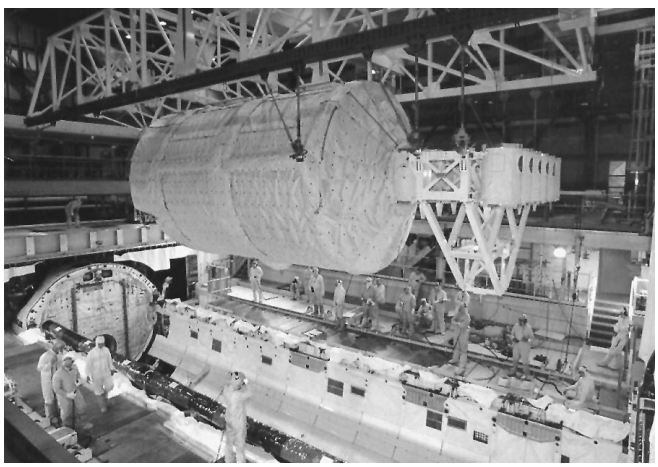
At current schedules, each of NASA's four Space Shuttle Orbiters must fly two or three times a year. Preparing an orbiter for its next mission is an incredibly complex process and much of the work is accomplished in the Orbiter Processing Facility (OPF) at Kennedy Space Center.

The average "flow" — the complete cycle of refurbishing an orbiter — requires the integration of approximately 10,000 work events, takes 65 days, and some 40,000 technician labor hours. Under the best conditions, scheduling each of the 10,000 work events in a single flow would be a task of monumental proportions. But the job is further complicated by the fact that only half the work is standard and predictable; the other half is composed of problem-generated tasks and jobs specific to the next mission, which creates a highly dynamic processing environment and requires frequent rescheduling.

For all the difficulties, Kennedy Space Center and its prime contractor for shuttle processing — Lockheed Space Operations Company (LSOC) — are doing an outstanding job of managing OPF operations with the help of a number of processing innovations in recent years. One of the most important is the Ground Processing Scheduling System, or GPSS. The GPSS is a software system for enhancing efficiency by providing an automated scheduling tool that predicts conflicts between scheduled tasks, helps human schedulers resolve those conflicts, and searches for near-optimal schedules.

GPSS is a cooperative development of Ames Research Center, Kennedy Space Center, LSOC, and a related company, Lockheed Missiles and Space Company. It originated at Ames, where a group of computer scientists conducted basic research on the use of artificial intelligence techniques to automate the scheduling process. A product of the work was a software system for complex, multifaceted operations known as the Gerry scheduling engine.

Kennedy Space Center brought Ames and Lockheed together and the group formed an inter-center/NASA contractor partnership to transfer the technology of the Gerry scheduling engine to the Space Shuttle program. The transfer was successfully accomplished and GPSS has become the accepted general purpose scheduling tool for OPF operations. (Courtesy of National Aeronautics and Space Administration.)



Kennedy Space Center technicians are preparing a Space Shuttle Orbiter for its next mission, an intricate task that requires scheduling 10,000 separate events over 65 days. A NASA-developed computer program automated this extremely complex scheduling job. (Photo courtesy of National Aeronautics and Space Administration.)

phases of products and systems. Increasing emphasis being placed on quality of components and systems, coupled with pressures to minimize cost and increase value, further emphasize the need to study, understand, quantify, and predict reliability and arrive at innovative designs and operational and maintenance procedures.

From the electrical engineering point of view, two (among several) areas that have received significant attention are electronic equipment (including computer hardware) and electric power systems. Other major areas include communication systems and software engineering. As the complexity of electronic equipment grew during and after World War II and as the consequences of failures in the field became more and more apparent, the U.S. military became seriously involved, promoted the formation of groups, and became instrumental in the development of the earliest handbooks and specifications. The great northeast blackout in the U.S. in November 1965 triggered the serious application of reliability concepts in the power systems area.

The objectives of this chapter are to introduce the reader to the fundamentals and applications of classical reliability concepts and bring out the important benefits of reliability considerations. Brief summaries of application aspects of reliability for electronic systems and power systems are also included.

110.2 Catastrophic Failure Models

Catastrophic failure refers to the case in which repair of the component is either not possible or available or of no value to the successful completion of the mission originally planned. Modeling such failures is typically based on life test results. We can consider the "lifetime" or "time to failure" T as a continuous random variable. Then,

$$P\left(\textit{survival up to time } t\right) = P\left(T > t\right) \equiv R\left(t\right) \tag{110.1}$$

where $R(t)$ is the **reliability** function. Obviously, as $t \to \infty$, $R(t) \to 0$ since the probability of failure increases with time of operation. Moreover,

$$P\left(\textit{failure at } t\right) = P\left(T \leq t\right) \equiv Q\left(t\right) \tag{110.2}$$

where $Q(t)$ is the unreliability function. From the definition of the distribution function of a continuous random variable, it is clear that $Q(t)$ is indeed the distribution function for T. Therefore, the failure density function $f(t)$ can be obtained as

$$f\left(t\right) = \frac{d}{dt} Q\left(t\right) \tag{110.3}$$

The **hazard rate function** $\lambda(t)$ is defined as

$$\lambda\left(t\right) \equiv \lim_{\Delta t \to 0} \frac{1}{\Delta t}\left[\begin{array}{l}\text{probability of failure in } (t, t+\Delta t), \\ \text{given survival up to t}\end{array}\right] \tag{110.4}$$

It can be shown that

$$\lambda\left(t\right) = \frac{f\left(t\right)}{R\left(t\right)} \tag{110.5}$$

The four functions, $f(t)$, $Q(t)$, $R(t)$, and $\lambda(t)$ constitute the set of functions used in basic reliability analysis. The relationships between these functions are given in Table 110.1.

TABLE 110.1 Relationships Between Different Reliability Functions

$f(t)$	$\lambda(t)$	$Q(t)$	$R(t)$
$f(t) = f(t)$	$\lambda(t) \exp\left[-\int_0^t \lambda(\xi)d\xi\right]$	$\frac{d}{dt} Q(t)$	$-\frac{d}{dt} R(t)$
$\lambda(t) = \frac{f(t)}{1 - \int_0^t f(\xi)d\xi}$	$\lambda(t)$	$\frac{1}{1 - Q(t)} \frac{d}{dt} (Q(t))$	$-\frac{d}{dt} [\ln R(t)]$
$Q(t) = \int_0^t f(\xi)d\xi$	$1 - \exp\left[-\int_0^t \lambda(\xi)d\xi\right]$	$Q(t)$	$1 - R(t)$
$R(t) = 1 - \int_0^t f(\xi)d\xi$	$\exp\left[-\int_0^t \lambda(\xi)d\xi\right]$	$1 - Q(t)$	$R(t)$

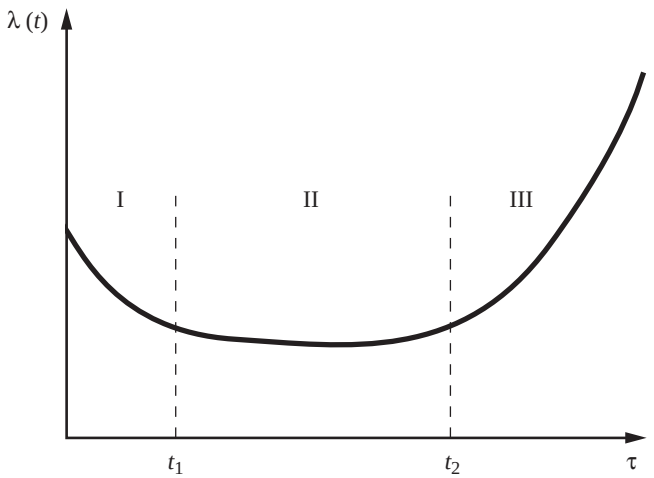


FIGURE 110.1 Bathtub-shaped hazard function

110.3 The Bathtub Curve

Of the four functions discussed, the hazard rate function $\lambda(t)$ displays the different stages during the lifetime of a component most clearly. In fact, typical $\lambda(t)$ plots have the general shape of a bathtub as shown in Fig. 110.1. The first region corresponds to wearin (infant mortality) or early failures during debugging. The hazard rate goes down as debugging continues. The second region corresponds to an essentially constant and low failure rate and failures can be considered to be nearly random. This is the useful lifetime of the component. The third region corresponds to wearout or fatigue phase with a sharply increasing hazard rate.

"Burn-in" refers to the practice of subjecting components to an initial operating period of t_1 (see Fig. 110.1) before delivering them to the customer. This eliminates all the initial failures from occurring after delivery to customers requiring high-reliability components. Moreover, it is prudent to replace a component as it approaches the wearout region, i.e., after an operating period of $(t_2\text{-}t_1)$. Electronic components tend to have a long useful life (constant hazard) period. Wearout region tends to dominate in the case of mechanical components.

110.4 Mean Time To Failure (MTTF)

The mean or expected value of the continuous random variable "time-to-failure" is the *MTTF*. This is a very useful parameter and is often enough to assess the suitability of components. It can be obtained using either the failure density function $f(t)$ or the reliability function $R(t)$ as follows:

$$MTTF = \int_0^\infty t\,f(t)dt \quad \text{or} \quad \int_0^\infty R(t)dt \tag{110.6}$$

In the case of repairable components, the repair time can also be considered as a continuous random variable with an expected value of *MTTR*. The mean time between failures, *MTBF*, is the sum of *MTTF* and *MTTR*. Since for well-designed components *MTTR*<<*MTTF*, *MTBF* and *MTTF* are often used interchangably.

110.5 Average Failure Rate

The average failure rate over the time interval 0 to T is defined as

$$AFR(0, T) \equiv AFR(T) = -\frac{\ln R(T)}{T} \tag{110.7}$$

110.6 *A Posteriori* Failure Probability

When components are subjected to a burn-in (or wearin) period of duration T, and if the component survives during $(0, T)$, the probability of failure during $(T, T+t)$ is called the *a posteriori* failure probability $Q_c(t)$. It can be found using

$$Q_c(t) = \frac{\int_T^{T+t} f(\xi)d\xi}{\int_T^\infty f(\xi)d\xi} \tag{110.8}$$

The probability of survival during $(T, T+t)$ is

$$R(t|T) = 1 - Q_c(t) = \frac{\int_{T+t}^\infty f(\xi)d\xi}{\int_T^\infty f(\xi)d\xi} = \frac{R(T+t)}{R(T)} = \exp\left[-\int_T^{T+t} \lambda(\xi)d\xi\right] \tag{110.9}$$

110.7 Units for Failure Rates

Several units are used to express failure rates. In addition to $\lambda(t)$ which is usually in number per hour, *%/K* is used to denote failure rate in percent per thousand hours and *PPM/K* is used to express failure rate in parts per million per thousand hours. The last unit is also known as *FIT* for "fails in time". The relationships between these units are given in Table 110.2.

TABLE 110.2 Relationships Between Different Failure Rate Units

	λ(#/hr)	%K	PPM/K (FIT)
λ =	λ	10^{-5} (%/K)	10^{-9}(PPM/K)
%/K =	10^{5} λ	%/K	10^{-4}(PPM/K)
PPM/K (FIT) =	10^{9} λ	10^{4} (%/K)	PPM/K

110.8 Application of the Binomial Distribution

In an experiment consisting of n identical independent trials, with each trial resulting in success or failure with probabilities of p and q, the probability P_r of r successes and $(n\text{-}r)$ failures is

$$P_r = {}_nC_r\, p^r\left(1 - p\right)^{n-r} \tag{110.10}$$

If X denotes the number of successes in n trials, then it is a discrete random variable with a mean value of (np) and a variance of (npq).

In a system consisting of a collection of n identical components with a probability p that a component is defective, the probability of finding r defects out of n is given by the P_r in Eq. (110.10). If p is the probability of success of one component and if at least r of them must be good for system success, then the system reliability (probability of system success) is given by

$$R = \sum_{k=r}^{n} {}_nC_k\, p^k\left(1 - p\right)^{n-k} \tag{110.11}$$

For systems with **redundancy**, $r < n$.

110.9 Application of the Poisson Distribution

For events that occur "in-time" at an average rate of λ occurrences per unit of time, the probability $P_x(t)$ of exactly x occurences during the time interval $(0\ ,\ t)$ is given by

$$P_x\left(t\right) = \frac{\left(\lambda t\right)^x e^{-\lambda t}}{x!} \tag{110.12}$$

The number of occurrences X in $(0,\ t)$ is a discrete random variable with a mean value μ of (λt) and a standard deviation σ of $\sqrt{\lambda t}$. By setting $X = 0$ in Eq. (110.12), we obtain the probability of no occurrence in $(0,t)$ as $e^{-\lambda t}$. If the event is failure, then no occurrence means success and $e^{-\lambda t}$ is the probability of success or system reliability. This is the well-known and often-used exponential distribution, also known as the constant-hazard model.

110.10 The Exponential Distribution

A constant hazard rate (constant λ) corresponding to the useful lifetime of components leads to the single-parameter exponential distribution. The functions of interest associated with a constant λ are:

$$f\left(t\right) = \lambda e^{-\lambda t},\ \ t > 0 \tag{110.13}$$

$$R(t) = e^{-\lambda t} \tag{110.14}$$

$$Q(t) = Q_c(t) = 1 - e^{-\lambda t} \tag{110.15}$$

The *a posteriori* failure probability $Q_c(t)$ is independent of the prior operating time T, indicating that the component does not degrade no matter how long it operates. Obviously, such a scenario is valid only during the useful lifetime (horizontal portion of the bathtub curve) of the component.

The mean and standard deviation of the random variable "lifetime" are

$$\mu \equiv MTTF = \frac{1}{\lambda} \qquad \text{and} \qquad \sigma = \frac{1}{\lambda} \tag{110.16}$$

110.11 The Weibull Distribution

The Weibull distribution has two parameters, a scale parameter α and a shape parameter β. By adjusting these two parameters, a wide range of experimental data can be modeled in system reliability studies.

The associated functions are

$$\lambda(t) = \frac{\beta t^{\beta-1}}{\alpha^{\beta}}; \ \alpha > 0, \ \beta > 0, \ t \geq 0 \tag{110.17}$$

$$f(t) = \frac{\beta t^{\beta-1}}{\alpha^{\beta}} \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{110.18}$$

$$R(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{110.19}$$

With $\beta = 1$, the Weibull distribution reduces to the constant hazard model with $\lambda = (1/\alpha)$. With $\beta = 2$, the Weibull distribution reduces to the Rayleigh distribution.

The associated *MTTF* is

$$MTTF = \mu = \alpha\Gamma\left(1 + \frac{1}{\beta}\right) \tag{110.20}$$

where Γ denotes the gamma function.

110.12 Combinatorial Aspects

Analysis of complex systems is facilitated by decomposition into functional entities consisting of subsystems or units and by the application of combinatorial considerations and network modeling techniques.

A **series** or **chain structure** consisting of n units is shown in Fig. 110.2. From the reliability point of view, the system will succeed only if all the units succeed. The units may or may not be physically in series. If R_i is the probability of success of the ith unit, then the series system reliability R_s is given as

$$R_s = \prod_{i=1}^{n} R_i \tag{110.21}$$

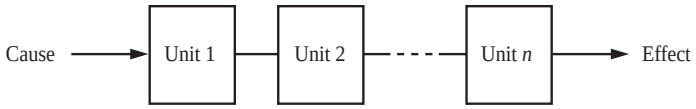


FIGURE 110.2 Series or chain structure.

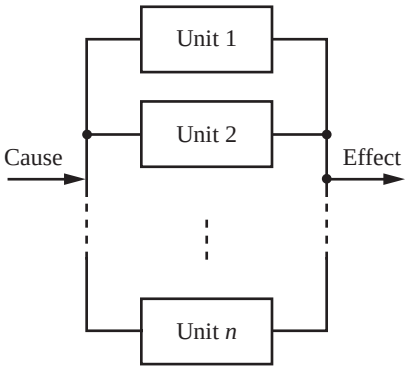


FIGURE 110.3 Parallel structure.

if the units do not interact with each other. If they do, then the conditional probabilities must be carefully evaluated.

If each of the units has a constant hazard, then

$$R_s(t) = \prod_{i=1}^{n} \exp(-\lambda_i t) \tag{110.22}$$

where λ_i is the constant failure rate for the ith unit or component. This enables us to replace the n components in series by an equivalent component with a constant hazard λ_s where

$$\lambda_s = \sum_{i=1}^{n} \lambda_i \tag{110.23}$$

If the components are identical, then $\lambda_s = n\lambda$ and the *MTTF* for the equivalent component is $(1/n)$ of the *MTTF* of one component.

A **parallel structure** consisting of n units is shown in Fig. 110.3. From the reliability point of view, the system will succeed if any one of the n units succeeds. Once again, the units may or may not be physically or topologically in parallel. If Q_i is the probability of failure of the ith unit, then the parallel system reliability R_p is given as

$$R_p = 1 - \prod_{i=1}^{n} Q_i \tag{110.24}$$

if the units do not interact with each other (meaning independent).

If each of the units has a constant hazard, then

$$R_p(t) = 1 - \prod_{i=1}^{n} \left[1 - \exp(-\lambda_i t)\right] \tag{110.25}$$

and we do not have the luxury of being able to replace the parallel system by an equivalent component with a constant hazard. The parallel system does not exhibit constant-hazard behavior even though each of the units has constant-hazard.

The *MTTF* of the parallel system can be obtained by using Eq. (110.25) in (110.6). The results for the case of components with identical hazards λ are: $(1.5/\lambda)$, $(1.833/\lambda)$, and $(2.083/\lambda)$ for $n = 2$, 3, and 4 respectively. The largest gain in *MTTF* is obtained by going from one component to two components in parallel. It is uncommon to have more than two or three components in a truly parallel configuration because of the cost involved. For two non-identical components in parallel with hazard rates of λ_1 and λ_2, the *MTTF* is given as

$$MTTF = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2} \tag{110.26}$$

An r-out-of-n structure, also known as a partially redundant system, can be evaluated using Eq. (110.11). If all the components are identical, independent, and have a constant hazard of λ, then the system reliability can be expressed as

$$R(t) = \sum_{k=r}^{n} {}_nC_k\, e^{-k\lambda t}\left(1 - e^{-\lambda t}\right)^{n-k} \tag{110.27}$$

For $r = 1$, the structure becomes a parallel system and for $r = n$, it becomes a series system.

Series-parallel systems are evaluated by repeated application of the expressions derived for series and parallel configurations by employing the well-known network reduction techniques.

Several general techniques are available for evaluating the reliability of complex structures that do not come under purely series or parallel or series parallel. They range from inspection to cutset and tieset methods and connection matrix techniques that are amenable for computer programming.

110.13 Modeling Maintenance

Maintenance of a component could be a scheduled (or preventive) one or a forced (corrective) one. The latter follows in-service failures and can be handled using Markov models discussed later. Scheduled maintenance is conducted at fixed intervals of time, irrespective of the system continuing to operate satisfactorily.

Scheduled maintenance, under ideal conditions, takes very little time (compared to the time between maintenances) and the component is restored to an "as new" condition. Even if the component is not repairable, scheduled maintenance postpones failure and prolongs the life of the component. Scheduled maintenance makes sense only for those components with increasing hazard rates. Most mechanical systems come under this category. It can be shown that the density function $f_T^*(t)$ with scheduled maintenance included can be expressed as

$$f_T^*(t) = \sum_{k=0}^{\infty} f_1(t - kT_M) R^k(T_M) \tag{110.28}$$

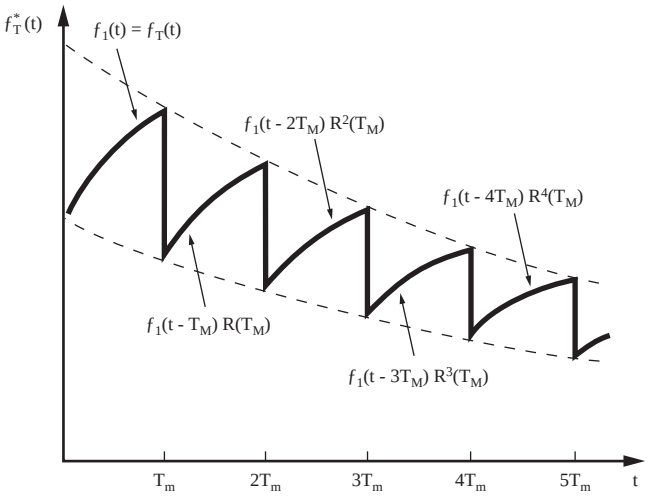


FIGURE 110.4 Density function with ideal scheduled maintenance incorporated.

$$\text{where } f_1(t) = \begin{cases} f_T(t) \text{ for } 0 < t \le T_M \\ 0 \qquad \text{otherwise} \end{cases} \tag{110.29}$$

$$R(t) = \text{component reliability function}$$

$$T_M = \text{time between maintenances, constant}$$

$$\text{and } f_T(t) = \text{original failure density function}$$

In Eq. (110.28), $k = 0$ is used only between $t = 0$ and $t = T_M$; $k = 1$ is used only between $t = T_M$ and $t = 2T_M$ and so on.

A typical $f_T^*(t)$ is shown in Fig. 110.4. The time scale is divided into equal intervals of T_M each. The function in each segment is a scaled-down version of the one in the previous segment, the scaling factor being equal to $R(T_M)$. Irrespective of the nature of the original failure density function, scheduled maintenance gives it an exponential tendency. This is another justification for the widespread use of exponential distribution in system reliability evaluations.

110.14 Markov Models

Of the different Markov models available, the discrete-state continuous-time Markov process has found many applications in system reliability evaluation, including the modeling of repairable systems. The model consists of a set of discrete states, called the state space, in which the system can reside and a set of transition rates between appropriate states. Using these, a set of first order differential equations are derived in the standard vector-matrix form for the time-dependent probabilities of the various states. Solution of these equations incorporating proper initial conditions gives the probabilities of the system residing in different states as functions of time. Several useful results can be gleaned from these functions.

110.15 Binary Model for a Repairable Component

The binary model for a repairable component assumes that the component can exist in one of two states—the UP state or the DOWN state. The transition rates between these two states, S_0 and S_1, are assumed to be constant

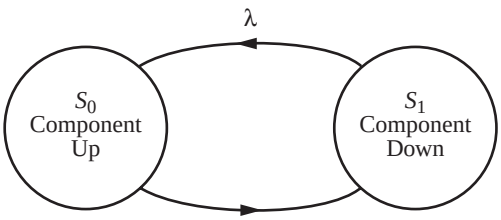


FIGURE 110.5 State space diagram for a single reparable component.

and equal to λ and μ. These transition rates are the constant failure and repair rates implied in the modeling process and their reciprocals are the *MTTF* and *MTTR*, respectively. Figure 110.5 illustrates the binary model.

The associated Markov differential equations are

$$\begin{bmatrix} P_0'(t) \\ P_1'(t) \end{bmatrix} = \begin{bmatrix} -\lambda & \mu \\ \lambda & -\mu \end{bmatrix} \begin{bmatrix} P_0(t) \\ P_1(t) \end{bmatrix} \tag{110.30}$$

with the initial conditions

$$\begin{bmatrix} P_0(0) \\ P_1(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{110.31}$$

The coefficient matrix of Markov differential equations, namely $\begin{bmatrix} -\lambda & \mu \\ \lambda & -\mu \end{bmatrix}$ is obtained by transposing the matrix of rates of departures $\begin{bmatrix} 0 & \lambda \\ \mu & 0 \end{bmatrix}$ and replacing the diagonal entries by the negative of the sum of all the other entries in their respective columns. Solution of (110.30) with initial conditions as given by (110.31) yields:

$$P_0(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t} \tag{110.32}$$

$$P_1(t) = \frac{\lambda}{\lambda + \mu}\left[1 - e^{-(\lambda+\mu)t}\right] \tag{110.33}$$

The limiting, or steady-state, probabilities are found by letting $t \to \infty$. They are also known as limiting **availability** A and limiting unavailability U and they are

$$P_0 \equiv \frac{\mu}{\lambda + \mu} \equiv A \quad \text{and } P_1 = \frac{\lambda}{\lambda + \mu} \equiv U \tag{110.34}$$

The time-dependent $A(t)$ and $U(t)$ are simply $P_0(t)$, and $P_1(t)$ respectively.

Referring back to Eq. (110.14) for a constant hazard component and comparing it with Eq. (110.32) which incorporates repair, the difference between $R(t)$ and $A(t)$ becomes obvious. Availability $A(t)$ is the probability

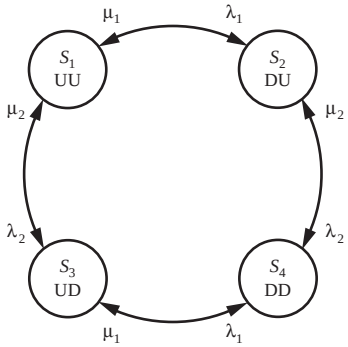


FIGURE 110.6 State space diagram for two dissimilar reparable components.

that the component is up at time t and reliability $R(t)$ is the probability that the system has continuously operated from 0 to t. Thus, $R(t)$ is much more stringent than $A(t)$. While both $R(0)$ and $A(0)$ are unity, $R(t)$ drops off rapidly as compared to $A(t)$ as time progresses. With a small value of $MTTR$ (or large value of μ), it is possible to realize a very high availability for a repairable component.

110.16 Two Dissimilar Repairable Components

Irrespective of whether the two components are in series or in parallel, the state space consists of four possible states. They are: S_1 (1 up, 2 up), S_2 (1 down, 2 up), S_3 (1 up, 2 down), and S_4 (1 down, 2 down). The actual system configuration will determine which of these four states corresponds to system success and failure. The associated state-space diagram is shown in Fig. 110.6. Analysis of this system results in the following steady-state probabilities:

$$P_1 = \frac{\mu_1 \ \mu_2}{\text{Denom}}\ ;\ P_2 = \frac{\lambda_1 \ \mu_2}{\text{Denom}}\ ;\ P_3 = \frac{\lambda_2 \ \mu_1}{\text{Denom}}\ ;\ P_4 = \frac{\lambda_1 \ \lambda_2}{\text{Denom}} \tag{110.35}$$

$$\text{where Denom} \equiv (\lambda_1 + \mu_1)(\lambda_2 + \mu_2) \tag{110.36}$$

For components in series, $A = P_1$ and $U = (P_2 + P_3 + P_4)$ and the two components can be replaced by an equivalent component with a failure rate of $\lambda_s = (\lambda_1 + \lambda_2)$ and a mean repair duration of r_s where

$$r_s \cong \frac{\lambda_1 \ r_1 + \lambda_2 \ r_2}{\lambda_s} \tag{110.37}$$

Extending this to n components in series, the equivalent system will have

$$\lambda_s = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad r_s \cong \frac{1}{\lambda_s} \sum_{i=1}^{n} \lambda_i \ r_i \tag{110.38}$$

$$\text{and system unavailability} = U_s \cong \lambda_s \ r_s = \sum_{i=1}^{n} \lambda_i \ r_i \tag{110.39}$$

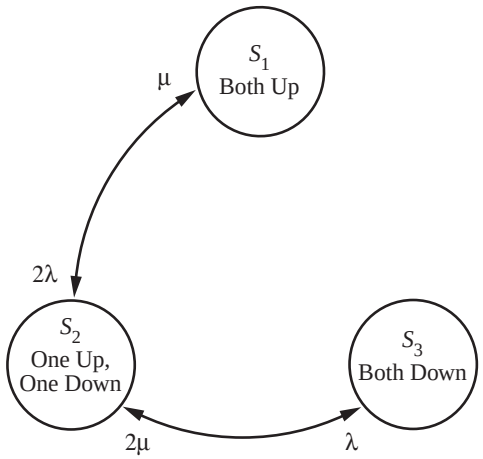


FIGURE 110.7 State space diagram for two identical reparable components.

For components in parallel, $A = (P_1 + P_2 + P_3)$ and $U = P_4$ and the two components can be replaced by an equivalent component with

$$\lambda_p \cong \lambda_1(\lambda_2 r_1) + \lambda_2(\lambda_1 r_2) \quad \text{and} \quad \mu_p = \mu_1 + \mu_2 \tag{110.40}$$

$$\text{and system unavailability} = U_p = \lambda_p(1/\mu_p) \tag{110.41}$$

Extension to more than two components in parallel follows similar lines. For three components in parallel,

$$\mu_p = (\mu_1 + \mu_2 + \mu_3) \quad \text{and} \quad U_p = \lambda_1\lambda_2\lambda_3 r_1 r_2 r_3 \tag{110.42}$$

110.17 Two Identical Repairable Components

In this case, only three states are needed to complete the state space. They are: S_1:Both UP; S_2: One UP and One DOWN; and S_3:Both DOWN. The corresponding state space diagram is shown in Fig. 110.7. Analysis of this system results in the following steady-state probabilities:

$$P_1 = \left(\frac{\mu}{\lambda + \mu}\right)^2; P_2 = \frac{2\lambda}{\mu}\left(\frac{\mu}{\lambda + \mu}\right)^2; P_3 = \left(\frac{\lambda}{\lambda + \mu}\right)^2 \tag{110.43}$$

110.18 Frequency and Duration Techniques

The expected residence time in a state is the mean value of the passage time from the state in question to any other state. Cycle time is the time required to complete an "in" and "not-in" cycle for that state. Frequency of occurrence (or encounter) for a state is the reciprocal of its cycle time. It can be shown that the frequency of occurrence of a state is equal to the steady-state probability of being in that state times the total rate of departure from it. Also, the expected value of the residence time is equal to the reciprocal of the total rate of departure from that state.

Under steady-state conditions, the expected frequency of entering a state must be equal to the expected frequency of leaving that state (this assumes that the system is "ergodic", which will not be elaborated for lack

of space). Using this principle, frequency balance equations can be easily written (one for each state) and solved in conjunction with the fact that the sum of the steady-state proabilities of all the states must be equal to unity to obtain the steady state probabilities. This procedure is much simpler than solving the Markov differential equations and letting $t \to \infty$.

110.19 Applications of Markov Process

Once the different states are identified and a state-space diagram is developed, Markov analysis can proceed systematically (probably with the help of a computer in the case of large systems) to yield a wealth of results useful in system reliability evaluation. Inclusion of installation time after repair, maintenance, spare, standby systems, and limitations imposed by restricted repair facilities are some of the many problems that can be studied.

110.20 Some Useful Approximations

1. For an r-out-of-n structure with failure and repair rates of λ and μ for each, the equivalent *MTTR* and *MTTF* can be approximated as

$$MTTR_{eq} = \frac{MTTR\ of\ one\ component}{n - r + 1} \tag{110.44}$$

$$MTTF_{eq} = \begin{pmatrix} MTTF \\ of\ one \\ component \end{pmatrix} \left(\frac{MTTF}{MTTR}\right)^{n-r} \left\{\frac{(n-r)!\,(r-1)!}{n!}\right\} \tag{110.45}$$

2. Influence of weather must be considered for components operating in an outdoor environment. If λ and λ' are the normal weather and stormy weather failure rates, λ' will be much greater than λ and the average failure rate λ_f can be approximated as

$$\lambda_f \cong \left(\frac{N}{N+S}\right)\lambda + \left(\frac{S}{N+S}\right)\lambda' \tag{110.46}$$

where N and S are the expected durations of normal and stormy weather.
3. For well-designed high-reliability components, the failure rate λ will be very small and $\lambda t << 1$. Then, for a single component,

$$R(t) \cong 1 - \lambda t \quad \text{and} \quad Q(t) \cong \lambda t \tag{110.47}$$

and for n dissimilar components in series,

$$R(t) \cong 1 - \sum_{i=1}^{n} \lambda_i t \quad \text{and} \quad Q(t) \cong \sum_{i=1}^{n} \lambda_i t \tag{110.48}$$

For the case of n identical components in parallel,

$$R(t) \cong 1 - (\lambda t)^n \quad \text{and} \quad Q(t) \cong (\lambda t)^n \tag{110.49}$$

For the case of an r-out-of-n configuration,

$$Q(t) \cong {}_nC_{(n-r+1)}(\lambda t)^{n-r+1} \tag{110.50}$$

The approximations detailed in (3) are called rare-event approximations.

110.21 Application Aspects

Electronic systems utilize large numbers of similar components over which the designer has very little control. Quality control methods can be used in the procurement and manufacturing phases. However, the circuit designer has no control over the design reliability of the devices except in cases such as custom-designed integrated circuits. In addition, electronic components cannot be inspected easily because of encapsulation. Although gross defects can be easily detected by suitable testing processes, defects that are not immediately effective (for example, weak mechanical bond of a lead-in conductor, material flaws in semiconductors, defective sealing, etc.) primarily contribute to unreliability. Temperature and voltage are the predominant failure-accelerating stresses for the majority of electronic components. As weaker components fail and are replaced by better ones, the percentage of defects in a population is reduced, resulting in a decreasing hazard rate. Wearout is rarely of significance in the failure of electronic components and systems. The designer should be careful to ensure that the loads (voltage, current, temperature) are within rated values and strive for a design that minimizes hot spots and temperature rises. Parameter drifts and accidental short circuits at connections can also lead to system failures. The circuit designer can follow a few basic rules to significantly improve electronic system reliability: reduce the number of adjustable components; avoid selection of components on the basis of parameter values obtained by testing; assemble components such that adjustments are easily accessible; and partition circuits into subassemblies for easy testing and diagnosis of problems.

Power systems are expected to provide all customers a reliable supply of electric power upon which much of modern life depends. Power systems are also very large, consisting of scores of large generators, hundreds of miles of high-voltage transmission lines, thousands of miles of distribution lines, along with the necessary transformers, switchgear, and substations interconnecting them. Reliability at the customer level can be improved by additional investment; the challenge is to balance reliability and the associated investment cost against the cost of energy charged to customers. This should be done in the presence of a number of random inputs and events: generator outages, line outages (which are highly weather dependent), random component outages, and uncertainties in the load demand (which is also weather dependent). Probabilistic techniques to evaluate power system reliability have been used effectively to resolve this problem satisfactorily. The system is divided into a number of subsystems and each one is analyzed separately. Then, composite system reliability evaluation techniques are employed to combine the results and arrive at a number of quantifiable reliability indices as inputs to managerial decisions. The major subsystems involved are generation, transmission, distribution, substations, and protection systems. Care should be taken to ensure that reliabilities of different parts of the system conform to each other and that no part of the system is unusually strong or weak. Obviously, different levels of reliability will be required for different parts of the system depending on the impacts of failures at different points on the interconnected power system.

110.22 Reliability and Economics

Reliability and economics are very closely related. Issues such as the level of reliability required, the amount of additional expenditures justified, where to invest the additional resources to maximize reliability, how to achieve a certain level of overall reliability at minimum cost, and how to assess the cost of failures and monetary equivalent of non-monetary items are all quite complex and not purely technical. However, once managerial decisions are made and reliability goals are set, certain well-proven techniques such as incorporating redundancy, improving maintenance procedures, selecting better quality components, etc. can be employed by the designer to achieve the goals.

Defining Terms

Availability: The availability $A(t)$ is the probability that a system is performing its required function successfully at time t. The steady-state availability A is the fraction of time that an item, system, or component is able to perfom its specified or required function.

Bathtub curve: For most physical components and living entities, the plot of failure (or hazard) rate vs. time has the shape of the longitudinal cross-section of a bathtub and hence its name.

Hazard rate function: The plot of instantaneous failure rate vs. time is called the hazard function. It cleary and distinctly exhibits the different life cycles of the component.

MTTF: The mean time to failure is the mean or expected value of "time to failure".

Parallel structure: Also known as a completely redundant system, it describes a system that can succeed when at least one of two or more components succeeds.

Redundancy: Refers to the existence of more than one means, identical or otherwise, for accomplishing a task or mission.

Reliability: The reliability $R(t)$ of an item or system is the probability that it has performed successfully over the time interval from 0 to t. In the case of non-repairable systems, $R(t) = A(t)$. With repair, $R(t) \leq A(t)$.

Series structure: Also known as a chain structure or non-redundant system, it describes a system whose success depends on the success of all of its components.

Related Topics

23.2 Testing • 98.5 Mean Time to Failure • 98.10 Markov Modeling • 98.12 Reliability Calculations for Real Time Systems

References

R. Billinton and R.N. Allan, Reliability Evaluation of Engineering Systems: Concepts and Techniques, 2nd ed., New York: Plenum, 1992.

E.E. Lewis, Introduction to Reliability Engineering, New York: John Wiley & Sons, 1987.

M.G. Pecht and F.R. Nash, "Predicting the reliability of electronic equipment", Proc. IEEE, 82(7), 992–1004, 1994.

R. Ramakumar, Engineering Reliability: Fundamentals and Applications, Englewood Cliffs, N.J.: Prentice–Hall, 1993.

M.L. Shooman, Probabilistic Reliability: An Engineering Approach, 2nd ed., Malabar, Fla.: R.E. Krieger Publishing Company, 1990.

For Further Information

R. Billinton and R.N. Allan, *Reliability Evaluation of Power Systems,* London, England: Pitman Advanced Publishing Program, 1984.

A.E. Green and A.J. Bourne, *Reliability Technology,* New York: Wiley-Interscience, 1972.

E.J. Henley and H. Kumamoto, *Probabilistic Risk Assessment—Reliability Engineering, Design, and Analysis,* New York: IEEE Press, 1991.

IEEE Transactions on Reliability, New York: Institute of Electrical and Electronics Engineers.

P.D.T. O'Connor, *Practical Reliability Engineering,* 3rd ed. New York: John Wiley & Sons, 1985.

Proceedings: Annual Reliability and Maintainability Symposium, New York: Institute of Electrical and Electronics Engineers.

D.P. Siewiorek, and R.S. Swarz, *The Theory and Practice of Reliabile System Design,* Bedford, Mass.: Digital Press, 1982.

K.S. Trivedi, *Probability and Statistics with Reliability, Queuing, and Computer Science Applications,* Englewood Cliffs, N.J.: Prentice-Hall, 1982.

A. Villemeur, *Reliability, Availability, Maintainability and Safety Assessment, Volumes I and II,* New York: John Wiley & Sons, 1992.

Blades, K., Allenby, B. "Environmental Effects"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

111

Environmental Effects

111.1 Introduction
111.2 Industrial Ecology
111.3 Design for Environment
111.4 Environmental Implications for the Electronics Industry
111.5 Emerging Technology
Integrated Circuits • Printed Wiring Boards
111.6 Tools and Strategies for Environmental Design
Design Tools • Design Strategies • Conclusion • Acknowledgements • Disclaimer

Karen Blades and
Braden Allenby
Lawrence Livermore National Laboratory

111.1 Introduction

The importance of electronics technology for consumers, and the electronics sector for the global economy, is already substantial and continues to grow rapidly. Such growth and innovation coupled with the global concerns for the environment and the need to better manage the resources of the earth pose many challenges for the electronics industry. While thought of as a "clean" industry, the technological advances made by the industry creates a significant demand on the earth's resources. As an example, the amount of water required in the production of semiconductors, the engines that motor most of today's electronic gadgets, is enormous—about 2000 gallons to process a single silicon wafer. Building silicon chips requires the use of highly toxic materials, albeit in relatively low volumes. Similarly, printed wiring boards present in most electronic products and produced in high volume use large amounts of solvents or gases which are either health hazards, ozone depleting, or contribute to the greenhouse effect and contain lead solder. The challenge for the industry is to continue the innovation that delivers the products and services that people want yet find creative solutions to minimize the environmental impact, enhance competitiveness, and address regulatory issues without impacting quality, productivity, or cost; in other words, to become an industry that is more "eco-efficient". Eco-efficiency is reached by the delivery of competitively priced goods and services that satisfy human needs and support a high quality of life, while progressively reducing ecological impacts and resource intensity, to a level at least in line with the earth's estimated carrying capacity.

Like sustainable development, a concept popularized by the Brundtland Report, *Our Common Future*, the notion of eco-efficiency requires a fundamental shift in the way environment is considered in industrial activity. Sustainable development—"development that meets the needs of the present without compromising the ability of future generations to meet their own needs" [World Commission on Environment and Development, 1987]—contemplates the integration of environmental, economic, and technological considerations to achieve continued human and economic development within the biological and physical constraints of the planet. Both eco-efficiency and sustainable development provide a useful direction, yet they prove difficult to operationalize and cannot guide technology development. Thus, the theoretical foundations for integrating technology and environment throughout the global economy are being provided by a new, multidisciplinary field known as "**industrial ecology**".

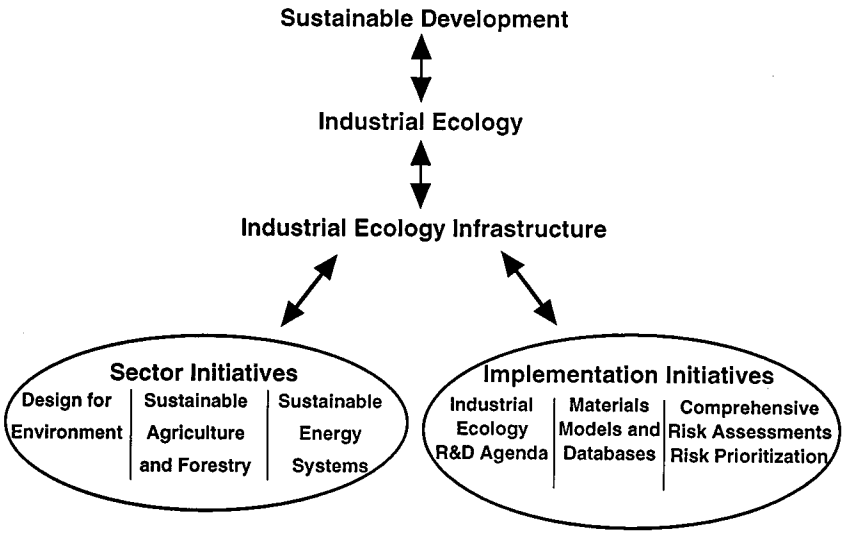


FIGURE 111.1 Industrial ecology framework.

The ideas of industrial ecology, which have begun to take root in the engineering community, have helped to established a framework within which the industry can move toward realizing sustainable development. The electrical, electronics, and telecommunications sectors are enablers of sustainability because they allow the provision of increasing quality-of-life using less material and energy, respectively, "**dematerialization**" and "**decarbonization**". This chapter will provide an introduction into industrial ecology and its implications for the electronics industry. Current activities, initiatives, and opportunities will also be explored, illustrating that the concomitant achievement of greater economic and environmental efficiency is indeed feasible in many cases.

111.2 Industrial Ecology

Industrial ecology is an emerging field that views manufacturing and other industrial activity including forestry, agriculture, mining, and other extractive sectors, as an integral component of global natural systems. In doing so, it takes a systems view of design and manufacturing activities so as to reduce or, more desirably, eliminate the environmental impacts of materials, manufacturing processes, technologies, and products across their life cycles, including use and disposal. It incorporates, among other things, research involving energy supply and use, new materials, new technologies and technological systems, basic sciences, economics, law, management, and social sciences.

The study of industrial ecology will, in the long run, provide the means by which the human species can deliberately and rationally approach a desirable long-term global carrying capacity. Oversimplifying, it can be thought of as "the science of sustainability". The approach is "deliberate" and "rational", to differentiate it from other, unplanned paths that might result, for example, in global pandemics, or economic and cultural collapse. The endpoint is "desirable", to differentiate it from other conceivable states such as a Malthusian subsistence world, which could involve much lower population levels, or oscillating population levels that depend on death rates to maintain a balance between resources and population levels. Figure 111.1 illustrates how industrial ecology provides a framework for operationalizing the vision of sustainable development.

As the term implies, industrial ecology is concerned with the evolution of technology and economic systems such that human economic activity mimics a mature biological system from the standpoint of being self-contained in its material and resource use. In such a system, little if any virgin material input is required, and little if any waste that must be disposed of outside of the economic system is generated. Energetically, the system can be open, just as biological systems are, although it is likely that overall energy consumption and intensity will be limited.

Although it is still a nascent field, a few fundamental principles are already apparent. Most importantly, the evolution of environmentally appropriate technology is seen as critical to reaching and maintaining a sustainable state. Unlike earlier approaches to environmental issues, which tended to regard technology as neutral at best, industrial ecology focuses on development of economically and environmentally efficient technology as key to any desirable, sustainable global state.

Also, environmental considerations must be integrated into all aspects of economic behavior, especially product and process design, and the design of economic and social systems within which those products are used and disposed. Environmental concerns must be internalized into technological systems and economic factors. It is not sufficient to design an energy efficient computer, for example; it is also necessary to ensure that the product, its components, or its constituent materials can be refurbished or recycled after the customer is through with it—all of this in a highly competitive and rapidly evolving market. This consideration implies a comprehensive and systems-based approach that is far more fundamental than any we have yet developed.

Industrial ecology requires an approach that is truly multidisciplinary. It is important to emphasize that industrial ecology is an objective field of study based on existing scientific and technological disciplines, not a form of industrial policy. It is profoundly a systems oriented and comprehensive approach which poses problems for most institutions—the government, riddled with fiefdoms; academia, with rigid departmental lines; and private firms, with job slots defined by occupation. Nonetheless, it is all too frequent that industrial ecology is seen as an economic program by economists, a legal program by lawyers, a technical program by engineers, and a scientific program by scientists. It is in part each of these; more importantly, it is all of these.

Industrial ecology has an important implication, however, of special interest to electronics and telecommunications engineers, and thus deserving of emphasis. The achievement of sustainability will, in part, require the substitution of intellectual and information capital for traditional physical capital, energy, and material inputs. Environmentally appropriate electronics, information management, and telecommunications technologies and services—and the manufacturing base that supports them—are therefore enabling technologies to achieve sustainable development. This offers unique opportunities for professional satisfaction, but also places a unique responsibility on the community of electrical and electronics engineers. We in particular cannot simply wait for the theory of industrial ecology to be fully developed before taking action.

111.3 Design for Environment

Design for Environment (DFE) is the means by which the precepts of industrial ecology, as currently understood, can in fact begin to be implemented in the real world today. DFE requires that environmental objectives and constraints be driven into process and product design, and materials and technology choices.

The focus is on the design stage because, for many articles, that is where most, if not all, of their life cycle environmental impacts are explicitly or implicitly established. Traditionally, electronics design has been based on a correct-by-verification approach, in which the environmental ramifications of a product (from manufacturing through disposition) are not considered until the product design is completed. DFE, by contrast, takes place early in a product's design phase as part of the concurrent engineering process to ensure that the environmental consequences of a product's life cycle are understood before manufacturing decisions are committed.

It is estimated that some 80 to 90% of the environmental impacts generated by product manufacture, use, and disposal are "locked-in" by the initial design. Materials choices, for example, ripple backwards towards environmental impacts associated with the extractive, smelting, and chemical industries. The design of a product and component selection control many environmental impacts associated with manufacturing, enabling, for example, substitution of no-clean or aqueous cleaning of printed wiring boards for processes that release ozone depleting substances, air toxics, or volatile organic compounds that are precursors of photochemical smog. The design of products controls many aspects of environmental impacts during use—energy efficient design is one example. Product design also controls the ease with which a product may be refurbished, or disassembled for parts or materials reclamation, after consumer use. DFE tools and methodologies offer a means to address such concerns at the design stage.

Obviously, DFE is not a panacea. It cannot, for example, compensate for failures of the current price structure to account for external factors, such as the real (i.e., social) cost of energy. It cannot compensate for deficiencies

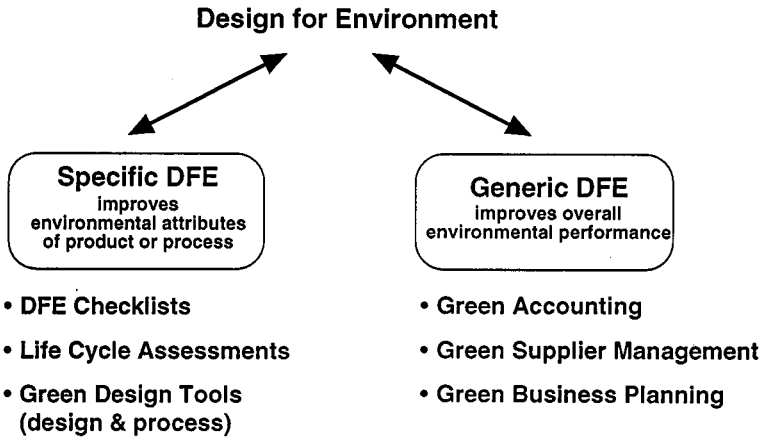


FIGURE 111.2 Examples of DFE activities within the firm.

in sectors outside electronics, such as a poorly coordinated, polluting, or even non-existent disposal and material recycling system in some areas of the world. Moreover, it is important to realize that DFE recognizes environmental considerations as on par with other objective and constraints—such as economic, technological, and market structure—not as superseding or dominating them. Nonetheless, if properly implemented, DFE programs represent a quantum leap forward in the way private firms integrate environmental concerns into their operations and technology.

It is useful to think of DFE within the firm as encompassing two different groups of activities as shown in Fig. 111.2. In all cases, DFE activities require inclusion of life-cycle considerations in the analytical process. The first, which might be styled "generic DFE", involves the implementation of broad programs that make the company's operations more environmentally preferable across the board. This might include, for example, development and implementation of "green accounting" practices, which ensure that relevant environmental costs are broken out by product line and process, so that they can be managed down. The "standard components" lists maintained by many companies can be reviewed to ensure that they direct the use of environmentally appropriate components and products wherever possible. Thus, for example, open relays might be deleted from such lists, on the grounds that they "can't swim", and thus might implicitly establish a need for chlorinated solvent, as opposed to aqueous, cleaning systems.

Contract provisions can be reviewed to ensure that suppliers are being directed to use environmentally preferable technologies and materials where possible. For example, are virgin materials being required where they are unnecessary? Do contracts, standards, and specifications clearly call for the use of recycled materials where they meet relevant performance requirements? Likewise, customer and internal standards and specifications can be reviewed with the same goal in mind.

The second group of DFE activities can be thought of as "specific DFE". Here, DFE is considered as a module of existing product realization processes, specifically the "Design for X", or DFX, systems used by many electronics manufacturers. The method involves creation of software tools and checklists, similar to those used in Design for Manufacturability, Design for Testability, or Design for Safety modules that ensure relevant environmental considerations are also included in the design process from the beginning. The challenge is to create modules which, in keeping with industrial ecology theory, are broad, comprehensive, and systems-based, yet can be defined well enough to be integrated into current design activities.

The successful application of DFE to the design of electronic systems requires the coordination of several design and data-based activities, such as environmental impact metrics; data and data management; design optimization, including cost assessments; and others. Failure to address any of these aspects can limit the effectiveness and usefulness of DFE efforts. Data and methodological deficiencies abound, and the challenge is great, yet experience at world class companies such as AT&T, Digital, IBM, Motorola, Siemens Nixdorf, Volvo, and Xerox indicate that it can be done. AT&T, for example, is testing a draft DFE practice; baselining the environmental attributes of a telephone at different life cycle stages to determine where meaningful environmental

improvements in design can be achieved; and developing software tools that can inform environmentally preferable design decisions [Seifert, 1995]. In Sweden, the government and Volvo have developed a relatively simple Environment Priority Strategies for Environmental Design, or EPS, system which uses Environmental Load Units, or ELUs, to inform materials choices during the design process. In Germany, Siemens Nixdorf has developed an "Eco-balance" system to help it make design choices that reflect both environmental and economic requirements. Xerox is a world leader in designing their products for refurbishment using a product life extension approach.

More broadly, the American Electronics Association (AEA) Design for Environment Task Force has created a series of White Papers discussing various aspects of Design for Environment and its implementation. The Microelectronics and Computer Technology Corporation (MCC) has published a comprehensive study [Lipp *et al.*, 1993] of the environmental impacts of a computer workstation, which is valuable not only for its technical findings, but for the substantial data and methodological gaps the study process identified. The Society of Environmental Toxicology and Chemistry (SETAC) and others, especially in Europe, are working on a number of comprehensive life-cycle assessment (LCA) methodologies designed to identify and prioritize environmental impacts of substances throughout their life cycle. The International Organization for Standards (ISO) is in the process of creating an international LCA standard. The IEEE Environment, Health and Safety Committee, formed in July, 1992, to support the integration of environmental, health, and safety considerations into electronics products and processes from design and manufacturing, to use, to recycling, refurbishing, or disposal has held a series of annual symposium on electronics and the environment. The proceedings from these symposia are valuable resources to the practitioners of DFE.

111.4 Environmental Implications for the Electronics Industry

Global concerns and regulations associated with environmental issues are increasingly affecting the manufacturing and design of electronic products, their technology development, and marketing strategies. No point illustrates this better than the German Blue Angel **Eco-Labeling** scheme for personal computers (the Blue Angel is a quasi-governmental, multi-attribute eco-labeling program). The Blue Angel requirements are numerous and span the complete life-cycle of the computers. Examples of some the requirements include: modular design of the entire system, customer-replaceable subassemblies and modules, use of non-halogenated flame retardants, and take back by manufacturers at the end of the product life. Market requirements such as these, focused on products and integrating as they do environmental and technology considerations, cannot possibly be met by continuing to treat environmental impact as an unavoidable result of industrial activity, i.e., as overhead. These requirements make environmental concerns truly strategic for the firm.

Perhaps the most familiar example of "a new generation of environmental management" requirements which will have enormous effects on electronics design is "**product take back**". These programs, such as the one mentioned in the Blue Angel labeling scheme, are being introduced in Germany and other countries for electronics manufacturers. They generally require that the firm take its products back once the consumer is through with them, recycle or refurbish the product, and assume responsibility for any remaining waste generated by the product. Other members of the European Union and Japan are among others considering such "take back" requirements. Similarly, the emergence of the international standard, **ISO 14000**, which includes requirements for environmental management systems, methodologies for **life cycle assessment** and environmental product specifications will have vast implications for the electronics industry. Though technically voluntary, in practice these standards in fact become requirements for firms wishing to engage in global commerce. These examples represent a global trend towards proactive management of business and products in the name of the environment.

111.5 Emerging Technology

New tools and technologies are emerging which will influence the environmental performance of electronic products and help the industry respond to the regulatory "push" and the market "pull" for environmentally responsible products. In the electronics industry, technology developments are important not only for the end-products, but

for components, recycling, and materials technology as well. Below is a brief summary of technology developments and their associated environment impacts as well as tools to address the many environmental concerns facing the industry.

The electronics industry has taken active steps toward environmental stewardship, evidenced by the formulation of the IEEE Environment, Safety and Health Committee, the 1996 Electronics Industry Environmental Roadmap published by MCC, and chapters focused on environment in The National Technology Roadmap for Semiconductors. Moves such as this, taken together with the technical sophistication of control systems used in manufacturing processes, have allowed the electronics industry to maintain low emission levels relative to some other industries. Despite the industry's environmental actions, the projected growth in electronics over the next 10 to 20 years is dramatic and continued technological innovation will be required to maintain historically low environmental impacts. Moreover, the rapid pace of technological change generates concomitantly high rates of product obsolescence and disposal, a factor that has led countries such as Germany and the Netherlands to focus on electronics products for environmental management.

Environmental considerations are not, of course, the only forces driving the technological evolution in the electronics industry. Major driving forces, as always, also include price, cost, performance, and market/regulatory requirements. However, to the extent that the trend is toward smaller devices, fewer processing steps, increased automation, and higher performance per device, such evolution will likely have a positive environmental impact at the unit production level, i.e., less materials, less chemicals, less waste related to each unit produced. Technology advances that have environmental implications at the upstream processing stage may well have significant benefits in the later stages of systems development and production. For example, material substitution in early production stages may decrease waste implications throughout the entire process. Since both semiconductors and printed wiring boards are produced in high volume and are present in virtually all electronic products ranging from electronic appliances, to computers, automotive, aerospace, and military applications, we will briefly examine the impact of these two areas of the electronic industry.

Integrated Circuits

The complex process of manufacturing semiconductor integrated circuits (IC) often consists of over a hundred steps during which many copies of an individual IC are formed on a single wafer. Each of the major process steps used in IC manufacturing involves some combination of energy use, material consumption, and material waste. Water usage is high due to the many cleaning and rinsing process steps. Absent process innovation, this trend will continue as wafer sizes increase, driving up the cost of water and waste water fees, and increasing mandated water conservation.

Environmental issues that also require attention include the constituent materials for encapsulants, the metals used for connection and attachment, the energy consumed in high-temperature processes, and the chemicals and solvents used in the packaging process. Here, emerging packaging technologies will have the effect of reducing the quantity of materials used in the packaging process by shrinking IC package sizes. Increasing predominance of plastic packaging will reduce energy consumption associated with hermetic ceramic packaging.

Printed Wiring Boards

Printed wiring boards represent the dominant interconnect technology on which chips will be attached and represents another key opportunity for making significant environmental advances. PWB manufacturing is a complicated process and uses large amounts of materials and energy (e.g., 1 MegaW of heat and 220 kW of energy is consumed during fabrication of prepeg for PWBs). On average, the waste streams constitutes 92%—and the final product just 8%—of the total weight of the materials used in PWB production process. Approximately 80% of the waste produced is hazardous and most of the waste is aqueous, including a range of hazardous chemicals. Printed wiring boards are not recycled because the removal of soldered subassemblies is costly and advanced chip designs require new printed wiring boards to be competitive. As a result, the boards are incinerated and the residual ash buried in hazardous waste landfills due to the lead content (from lead solder).

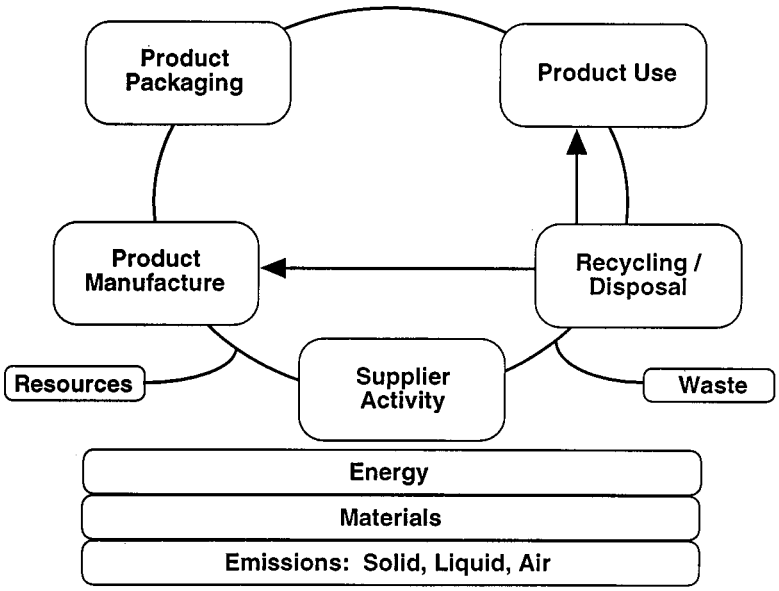


FIGURE 111.3 Design for environment: systems-based, life cycle approach.

111.6 Tools and Strategies for Environmental Design

The key to reducing the environmental impact of electronic products will be the application of DFE tools and methodologies. Development of CAD/CAM tools based on environmental impact metrics, materials selection data, cost, and product data management are examples of available or clearly foreseeable tools to assist firms in adopting DFE practices.

These tools will need to be based on life cycle assessment, the objective process used to evaluate the environmental impacts associated with a product and identify opportunities for improvement. Life cycle assessment seeks to minimize the environmental impact of the manufacture, use, and eventual disposal of products without compromising essential product functions. Figure 111.3 shows the life stages that would be considered for electronic products (i.e., the life cycle considered has been bounded by product design activities). The ability of the electronics industry to operate in a more environmentally and economically efficient mode, use less chemicals and materials, and reduce energy consumption will require support tools that can be used to evaluate both product and process designs. To date, many firms are making immediate gains by incorporating basic tools like DFE checklists, design standards and internal databases on chemicals and materials, while other firms are developing sophisticated software tools that give products environmental scores based on the product's compliance with a set of predetermined environmental attributes. These software tools rely heavily on environmental metrics (typically internal to the firm) to assess the environmental impact and then assign a score to the associated impact.

Other types of tools that will be necessary to implement DFE will include tools to characterize environmental risk, define and build flexible processes to reduce waste, and support dematerialization of processes and products. The following sections provide a brief review of design tools or strategies that can be employed

Design Tools

Environmental design tools vary widely in the evaluation procedures offered in terms of the type of data used, method of analysis, and the results provided to the electronic designer. The tool strategies range in scope from assessment of the entire product life cycle to the evaluation of a single aspect of its fabrication, use, or disposal. Today's DFE tools can be generally characterized as either life cycle analysis, recyclability analysis, manufacturing analysis, or process flow analysis tools.

The effectiveness of these design tools is based both on the tool's functionality as well as its corresponding support data. One of the biggest challenges designers face with regards to DFE is a lack of reliable data on materials, parts, and components needed to adequately convey the impact and trade-offs of their design decisions. To account for these data deficiencies, a number of environmental design tools attempt to use innovative, analytical methods to estimate environmental impacts: while necessary, this indicates they must be used with care and an understanding of their assumptions.

Although DFE provides a systems-based, life cycle approach, its true value to the system designer is lost unless the impact of DFE decisions on other relevant economic and performance measures (i.e., cost, electrical performance, reliability, etc.) can be quickly and accurately assessed. Trade-off analysis tools that have DFE embedded can perform process flow-based environmental analysis (energy/mass balance, waste stream analysis, etc.) concurrently with non-environmental cost and performance analysis so that system designers can accurately evaluate the impact of critical design decisions early.

Design Strategies

Design strategies such as lead minimization through component selection, and the reduction of waste resulting from rapid technological evolution through modular design, help to minimize the environmental impact of electronic products. Although at this time no suitable lead-free alternatives exist for electronic interconnections, designers can still minimize the lead content of electronic designs. Surface mount technology requires less solder than through-hole technology. New interconnection technologies, such as microball grid array and direct chip attachment, also require less solder. The environmental benefits increase with the use of these advanced interconnection technologies.

The rapid advancement of the electronics industry has created a time when many products become obsolete in less than five years' time. Electronic products must be built to last, but only until it is time to take them apart for rebuilding or for reuse of material. This means employing modular design strategies to facilitate disassembly for recycle or upgrade of the product rather than replacement. Designers must extend their views to consider the full utilization of materials and the environmental impact of the material life cycle as well as the product life cycle.

Conclusion

The diverse product variety of the electronics industry offers numerous opportunities to curtail the environmental impact of the industry. These opportunities are multidimensional. Services made possible through telecommunications technology enable people to work from home reducing emissions that would be generated by traveling to work. Smaller, faster computers and the Internet require less material usage, reducing the energy demand during processing and waste generated during fabrication. All these represent examples of how the electronics industry provides enablers of sustainability.

Global concerns and regulations associated with environmental issues are increasingly affecting the manufacturing and design of the electronics industry. Environmental management standards, "take back" programs, ISO 14000 standards development activity, and eco-label requirements represent a sample of the initiatives driving the industries move to more environmentally efficient practices. While the industry has initiated some activities to address environmental concerns, the future competitiveness of the industry will depend on improvements in environmental technology in manufacturing, accurate assessment of the environmental impact of products and process, and design products that employ design for environment, reuse, and recycleability. Industrial ecology offers a framework for analyzing the environmental effects of the electronics industry which is complicated by the rapid pace of change.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Defining Terms

Decarbonization: The reduction, over time, of carbon content per unit energy produced. Natural gas, for example, produces more energy per unit carbon than coal; equivalently, more CO_2 is produced from coal than from natural gas per unit energy produced.

Dematerialization: The decline, over time, in weight of materials used in industrial end products, or in the embedded energy of the products. Dematerialization is an extremely important concept for the environment because the use of less material translates into smaller quantities of waste generated in both production and consumption.

Design for environment (DFE): The systematic consideration of design performance with respect to environment over the full product and process life cycle from design through manufacturing, packaging, distribution, installation, use, and end of life. It is proactive to reduce environmental impact by addressing environmental concerns in the product or process design stage.

Eco-label: Label or certificate awarded to a product that has met specific environmental performance requirements. Some of the most widely known eco-labels include Germany's Blue Angel, Nordic White Swan, and U.S. Green Seal.

Industrial ecology: The objective, multidisciplinary study of industrial and economic systems and their linkages with fundamental natural systems.

ISO 14000: Series of international standards fashioned from the ISO 9000 standard which includes requirements for environmental management systems, environmental auditing and labeling guidelines, life cycle analysis guidelines, and environmental product standards.

Life cycle assessment (LCA): The method for systematically assessing the material use, energy use, waste emissions, services, processes, and technologies associated with a product.

Product take back: Program in which manufacture agrees to take back product at the end-of-life (typically at no cost to the consumer) and disposal of product in an environmentally responsible matter.

Related Topic

25.3 Application-Specific Integrated Circuits

References

B.F. Dambach and B.A. Allenby, "Implementing design for environment at AT&T," *Total Quality Environmental Management,* 4, 51–62, 1995.

T.E. Graedel and B.R. Allenby, *Industrial Ecology*, Englewood Cliffs, N.J.: Prentice-Hall, 1995.

S. Lipp, G. Pitts, and F. Cassidy, Eds. "A life cycle environmental assessment of a computer workstation," *Environmental Consciousness: A Strategic Competitiveness Issue for the Electronics and Computer Industry*, Austin, Tex.: Microelectronics and Computer Technology Corporation.

S. Pederson, C. Wilson, G. Pitts, and B. Stotesbery, Eds. *Electronics Industry Roadmap*, Austin, Tex.: Microelectronics and Computer Technology Corporation, 1996.

L. Seifert, "AT&T technology and the environment," *AT&T Tech. J.*, 74, 4–7, 1995.
World Commission on Environment and Development, *Our Common Future*, Oxford: Oxford University Press, 1987.

Further Information

The IEEE Environment, Health and Safety Committee annually sponsors and publishes the proceeding of the *International Symposium on Electronics and the Environment*. These proceedings are a valuable resource for practitioners of DFE.

The National Technology Roadmap for Semiconductors, published by the Semiconductor Industry Association contains information on the environmental impacts of semiconductor fabrication as well as initiatives begun to address these concerns.

The *AT&T Technical Journal* has a dedicated issue on Industrial Ecology and DFE entitled AT&T Technology and the Environment, volume 74, no. 6, November/December 1995.

Other suggested reading:

American Electronics Association, "The hows and whys of design for the environment," 1993.
B.R. Allenby and D.J. Richards, Eds., *The Greening of Industrial Ecosystems*, Washington, D.C.: National Academy Press, 1994.
P. Eisenberger, Ed., *Basic Research Needs for Environmentally Responsive Technologies of the Future*, Princeton, N.J.: Princeton Materials Institute, 1996.
T.E. Graedel and B.R. Allenby, *Design for Environment*, Englewood Cliffs, N.J.: Prentice-Hall, 1996.

Rimvall, C.M, Jobling, C.P. "Computer-Aided Control Systems Design"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

112

Computer-Aided Control Systems Design[1]

112.1 Introduction

112.2 A Brief History of CACSD
Technological Developments • User Interfaces • CACSD Packages of Note

112.3 The State of the Art in CACSD
Consolidation of CACSD • A critique of Matrix Environments for CACSD • "Open Systems" • Other Desirable Features

112.4 CACSD Block-Diagram Tools
Basic Block-Diagram System Representations • Architectures of Block-Diagram Systems • Open-Architectures of Block-Diagram Editors

C. Magnus Rimvall
F. L. Smith & Co. A/S

Christopher P. Jobling
University of Wales

112.1 Introduction

The use of computers in the design of control systems has a long and fairly distinguished history. It begins before the dawn of the modern information age with the analog computing devices that were used to create tables of ballistic data for artillery and anti-aircraft gunners and continues to the present day in which modern desktop machines have computing power undreamed of when the classical and modern control theories were laid down in the middle years of the twentieth century.

Modern computer-aided control system design (CACSD) has been made possible by the synthesis of several key developments in computing. The development and continued dominance of high-level procedural languages such as FORTRAN enabled the development and distribution of standard mathematical software. The emergence of fully interactive operating systems such as UNIX and its user "shells" influenced the development of CACSD packages which have been constructed along similar lines. The ready availability and cheapness of raster-graphic displays has provided the on-screen display of data from control systems analysis, the creation of tools for modeling control systems using familiar block diagrams and have the potential to make order-of-magnitude improvements in the ease-of-use, ease-of-manipulation, and efficiency of the interaction between the control designer, his model, analysis tools, and end-product—software for embedded controllers. The driving force of all these developments is the seemingly continual increase in computing power year-on-year and the result has been to make computers accessible to large numbers of people while at the same time making them easier to use.

A control engineer often describes systems through the use of block diagrams. This is not only the traditional graphical representation of a control system, it is also an almost discipline-independent, and thus universally understandable, representation for dynamic systems. The diagrams may also constitute a complete documentation

[1]Originally published as "Computer-Aided Control Systems Design", Chapter 23, pp 429–442, in Levine, W. S. (Ed.), *The Control Handbook,* CRC Press, 1995.

of the designed system. Block diagrams are self-documenting and, when appropriately annotated, may form complete and consistent specifications of control systems. It is, therefore, not surprising that a number of tools for modeling (control) systems through block diagrams have emerged on the market over the last 5 to 10 years.

In addition to serving as a documentation aid, the overall cost and cycle time for developing complex controllers is radically reduced if analysis/simulation code and/or real-time code is automatically generated from the block-diagrams. This eliminates time-consuming manual coding, and avoids the introduction of coding bugs.

In this chapter, we explore the state-of-the-art in CACSD. We begin with a brief survey of the tools that have been developed over the years. We then focus on the matrix environments that provide the current standard and attempt to explain why they are important. We also examine modern block-diagram editors, simulation and code generation tools, and finally allow ourselves to speculate on the future.

112.2 A Brief History of CACSD

The term computer-aided control system design may be defined as:

> The use of digital computers as a primary tool during the modeling, identification, analysis, and design phases of control engineering.

CACSD tools and packages typically provide well-integrated support for the analysis and design of linear plant and controllers although many modern packages also provide support for the modeling, simulation, and linearization of nonlinear systems and some have the capability of implementing a control law in software.

Figure 112.1 (adapted and updated from Rimvall [1987,1988]) illustrates the development of CACSD packages over the last four decades. In order to put events into proper context, other key influencing factors, chiefly hardware and software developments, are also shown. In this section we describe the background to the emergence of CACSD tools in more detail, starting with technological developments and then moving on to user interface aspects. The aim is to understand the current state-of-the-art by examining the historical context in which these tools have been developed.

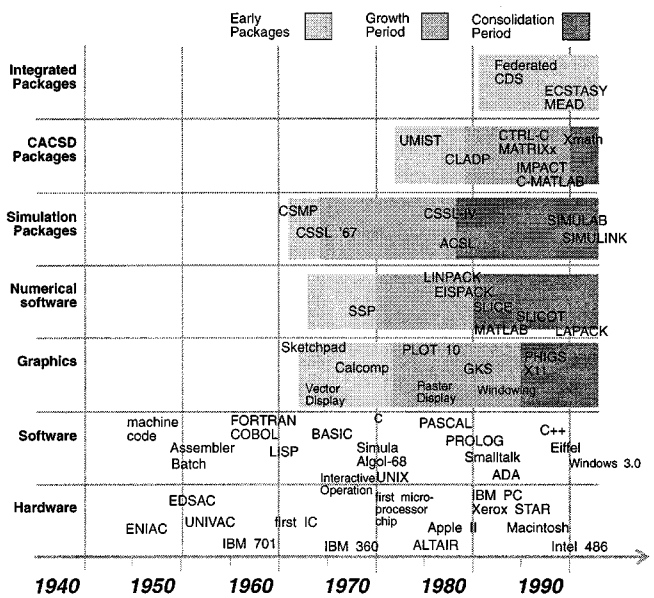


FIGURE 112.1 The historical development of interactive CACSD tools showing the availability of related hardware and software. Some actual products are included to indicate the state-of-the-art.

Technological Developments

Computing Hardware

Since 1953, there has been a phenomenal growth in the capabilities and power of computing hardware. Observers estimate that the power of computing devices (in terms of both execution speed and memory availability) has doubled every second to third year, whereas the size and cost (per computational unit) of the hardware has halved at approximately the same rate.

In terms of CACSD, the chief effect of these developments has been to widen the range of applications for computing and at the same time to make computers, and therefore the applications, widely available to practitioners in all branches of the subject. For example control engineers, control theorists, and control implementors all benefit as described below.

- Desk-top machines which are orders of magnitude more powerful than mainframe machines of two decades ago provide the means by which CACSD can be brought to the data analysis, model building, simulation, performance analysis and modification, control law synthesis, and documentation that is the day-to-day work of the control engineer.
- Without powerful computing hardware, many of the complex algorithms developed by control theorists for both analysis and implementation would otherwise be impractical.
- Embedded computer systems, which implement controllers, smart actuators, and smart sensors, are routinely used to implement the control laws developed by control engineers and control theorists.

System Software

The development of system software, such as operating systems, programming languages and program execution environments, has been slower than that of hardware, but is nonetheless impressive. Less impressive is the steadily increasing cost of application software, estimated at about 80% of the total installation cost of a computing system, which developments in computer science have been largely unable to reduce. We are, in fact, in the midst of a software crisis, dating from about 1968, which is the result of ever increasing improvements in hardware. Such improvements increase the possibilities for software, raise the expectations of users, and therefore raise the stakes in software production faster than improvements in software development technology has been made.

High Level Languages

The invention of FORTRAN was a major breakthrough in engineering computing. A high-level language, FORTRAN and the *compilers* that convert it into machine code, allowed engineers to write programs in a language that was sufficiently close to mathematical notation so as to be quite natural. Since its invention, numerous other high-level languages have been created, although FORTRAN continues to dominate engineering "number-crunching". For the implementation of control algorithms, assembly languages are still popular although high(er) level languages such as C, which is the predominant systems programming language, MODULA, and ADA are gaining acceptance in the market place.

Graphical Displays

Engineers are, in general, more comfortable with pictures than with text as a means of communicating their ideas. Hence, the wide availability of graphical displays is of prime importance to many areas of engineering computing. Indeed, the development of computer graphics has been the means by which certain control systems design techniques, such as multivariable control systems analysis, have been made practicable. Computer graphics have also been instrumental in providing improvements in human-machine interfaces such as schematic systems input and direct manipulation interfaces with windows, icons, pull-down menus, and pop-up dialog boxes. Further improvements in user interfacing techniques such as *hypermedia* will continue to rely on developments in display technology.

For modern CACSD, the most significant development in display technology has been the development of cheap, high-resolution *raster graphics displays,* although, historically, great strides were made with less well

known and more expensive *vector refresh* and *vector storage* display technology. The prime feature of raster-scan technology is that an area of the image may be made to appear to move on the screen by the application of simple logical operations. Raster graphics displays are, therefore, ideal for building direct manipulation graphics applications such as block diagram editors, which will be discussed later. They are not so well suited to the direct display and manipulation of vector images, which are a key part of many engineering graphics applications. For example, it is difficult to move part of a vector image such as a bode-plot without destroying the rest of the picture or to display sloping lines that look smooth at low resolutions. However, the dominance of the technology has been a factor in ensuring that the deficiencies in the technology can be overcome by clever software.

Quality Numerical Software

Following the invention of FORTRAN there was a gradual development of useful general purpose subroutines that could be archived into libraries, distributed, and shared. This lead eventually to the development of standard subroutine libraries such as EIS-PACK [Smith et al., 1977], LINPACK [Dongarra et al., 1979], and LAPACK [Anderson et al., 19789] (for solving eigenvalue problems and sets of linear equations) which have had a direct influence on the development of CACSD.

Simulation Languages

For many years before the predominance of digital computers, dynamic system behavior was simulated using analog and hybrid computers. *Digital simulation* began to take over from analog and hybrid simulation during the mid-1960s. Digital simulation programs can be used to model a wider range of nonlinear phenomena more reliably than analog or hybrid computers, at the cost of losing real-time and introducing quantization problems. However, the disadvantages of the technology are more than outweighed by improvements in modeling possibilities and increases in productivity. Digital simulation has superseded analog computation in all but a few specialized areas.

The first digital simulation systems were FORTRAN programs. Eventually, special purpose languages emerged which allowed statements written in a form close to state equation notation to be translated into FORTRAN which enabled the engineer to concentrate on the problem description. In 1967, a standard language called CSSL (Continuous Systems Simulation Language) [Augustin et al., 1967] was proposed by the U.S. Simulation Council and this forms the basis of most simulation languages in use today.

User Interfaces

Over the years, user interaction with computers has become progressively more direct. In the very early days, the user interface was another human being. These "operators" were gradually replaced by *operating systems* which provided communication first through the medium of punch-card and paper tape, then later by teletype machines, text-based visual display units, and, most recently, by windowed graphical user interfaces. Along with this change, there has been a corresponding change in style for CACSD tools. Batch mode programs were collected into "packages" and provided with question and answer or menued interfaces. These, in turn, have been largely superceded by command driven interfaces and direct-manipulation graphical user interfaces, currently used only for specialized tasks such as block-diagram input, will have a wider role in future CACSD packages.

CACSD Packages of Note

As the supporting technology has developed, control engineers mainly working in academia have been actively engaged in developing tools to support developments in control theory and in combining these tools into packages. Early pioneering work was carried out in Europe where the emphasis was on frequency response methods for multivariable control systems analysis and design. Some of the first CACSD packages were developed in the mid-1970s. In the U.S., control theory was concentrated in the time domain and made use of

state-space models. Several packages of tools for state-space design were created and reached maturity in the late 1970s. These packages were usually written in FORTRAN and made use of a question-and-answer interface. Some of the better packages made use of standard numerical libraries such as EISPACK and LINPACK, but many made use of home-grown algorithms with sometimes dubious numerical properties.

One of the earliest standardization efforts was concerned with algorithms and there have been several attempts to create standard CACSD libraries. One of these, SLICOT [van den Boom et al., 1991], is still ongoing. But it has to be admitted that such efforts have had little success in the marketplace. The real break-through came with the development of the "matrix environments", which are discussed in the next section. Currently, although many research groups continue to develop specialist tools and packages in conventional languages such as FORTRAN, most CACSD tool-makers now use these matrix environments as a high-level language for creating "toolboxes" of tools.

112.3 The State of the Art in CACSD

In this section we shall describe the matrix environments that have come to dominate CACSD, that is, the analysis, synthesis, and design of linear controllers for linear plants. We shall then move on to examine some of the requirements of CACSD which are less well served by the current generation of tools.

Consolidation of CACSD

As can be seen in Fig. 112.1, the 1980s was a decade of consolidation during which CACSD technology matured. Menu driven and Q&A dialogs were superseded by command languages. The matrix environment has become the *de facto* standard for CACSD.

The reasons for this are due to the simplicity of the data structures and the interface model and the *flexibility* of the package. We illustrate these properties using MATLAB (MATrix LABoratory) [Moler, 1980], the original matrix environment. Originally designed as a teaching program for graduate students, giving interactive access to the linear algebra routines EISPACK and LINPACK, MATLAB was released into the public domain in around 1980.

In MATLAB, matrices and matrix operations are entered into the computer in the straightforward fashion illustrated in Fig. 112.2.

This elegant treatment of linear algebra readily appealed to control scientists who realized that it was equally applicable to the solution of "modern control" problems based on linear state-space models (Fig. 112.3).

```
>>  a = [1 3 5
7 6 5; 0 0 5];

>>  [vec, val] = eig(a)

vec =

   -0.7408   -0.3622   -0.1633
    0.6717   -0.9321   -0.8981
         0         0    0.4082


val =
   -1.7202         0         0
         0    8.7202         0
         0         0    5.0000
```

FIGURE 112.2 Entering and manipulating matrices in MATLAB. In this example, a matrix is defined and its eigenvectors and eigenvalues are determined.

```
>>  A = [0,1,0;0,0,1;-2,-3,4];
>>  B = [0, 0, 1];
>>  C = [1, 0, 0];
>>  poles = eig(A)

poles =
   -0.4142
    2.0000
    2.4142

>>  stable = all(poles < 0)

stable =

     0
>>
```

FIGURE 112.3 Using state-space matrices. A simple stability test showing the power of the matrix functions built-in to MATLAB. The Boolean function 'all' returns the value TRUE (or 1) if all the elements of the argument are non-zero. The argument is itself a vector of Boolean values (that is, those values of the vector of the poles of the A matrix that are negative). By treating matrices as "first-class objects", MATLAB provides many such opportunities for avoiding loops and other control structures required to do similar tasks in conventional languages.

```
function qs = control(a, b)
% Returns the controllability matrix [b, ab, a^2b, ...]
% used as: qs = control(a, b)
[ma, na] = size(a); [mb, nb] = size(b);
if ma != na
        error('Non-square A matrix')
elseif ma != mb
        error('Unequal number of rows in A and B')
else
        qs = b; k = b;
        for i = 2:ma;
                k = a*k; qs = [qs, k];
        end
end
```

FIGURE 112.4 The extension of MATLAB by means of "macro" or M-files. Here is a routine for determining the controllability of a state-space model.

However, powerful though the basic "matrix calculator" capabilities of MATLAB are, its real flexibility is due to its support of *macro files*. A macro file (M-file), in its simplest form, is just a collection of ordinary MATLAB commands which ar stored in a file. When called, such a "script" of commands is executed just as if it had been typed by the user. MATLAB's real strength lies in its ability to use M-files to create new functions. Such a function is defined in Fig. 112.4. Once defined in this way, the new function can be executed as if it was a part of the language (Fig. 112.5).

```
>>  qs=control(A,B)

qs =

     0     0     1
     0     1     4
     1     4    13

>>
```

FIGURE 112.5 Using a user-defined function as an extension to MATLAB.

By creating a set of functions in this way, it is relatively easy to build up a "toolbox" of useful functions for a particular application domain. This is exactly what happened shortly after the release of the original MATLAB. Entrepreneurs quickly realized that if they cleaned up the code, added control oriented data types and functions and some graphics capability, MATLAB could be resold as a proprietary CACSD package. So, based mainly on the state-space methods in vogue in the U.S., several packages, such as MATRIXx and Ctrl-C, emerged and were a great success.

MATLAB itself underwent further development. It was rewritten in C for efficiency and enhanced portability and released as a commercial product in 1985. Like its competitors, the main market was initially the CACSD market, where, supported by two sets of toolbox extensions called the Control and Signal Processing Toolboxes, MATLAB made rapid inroads into academia and industry. A recent development has been the provision of add-on graphical input of system models, in the form of block diagrams, support for "point-and-click" nonlinear simulation, and enhanced graphical functionality. At least one package, MATRIXx, has evolved further by the addition of data structures and more sophisticated support for macro development. The result is the package X-Math described by Floyd et al. [1991].

A Critique of Matrix Environments for CACSD

MATLAB and similar matrix environments are far from completely ideal. Rimvall [1987] gave the following requirements for a CACSD environment which are largely still valid today.

- Software packages must support the same entities used by human specialists in the field.
- The basic commands of an interactive environment must be fast yet flexible.
- CACSD packages should support an algorithmic interface.
- The transition from basic use to advanced use must be gradual.
- The system must be transparent.
- Small and large systems should be equally treated by the user interface.
- The system must be able to communicate with the outside world.

Matrix environments do not meet all of these requirements. The following sections give a critical review of the state-of-the-art.

Support of Control Entities

For a control engineer, the entities of interest are

- numerical descriptions of systems (state-space models, transfer functions, etc.)
- symbolic elements for general system equations
- graphical elements for the definition of system topologies
- support of large-scale data management, e.g., in the form of a relational database
- support of small-scale data management, e.g., in the form of spreadsheets
- graphical displays of numerical computations, possibly together with graphical interaction for requirement specifications, etc.

MATLAB was developed by a numerical analyst for numerical analysts. Such people need, and MATLAB provides, only one data structure, the complex matrix. It is a credit to its flexibility that the package can be adapted to a control engineer's needs by the careful use of convention and toolbox extensions (Fig. 112.6), but the price paid is increased complexity.

Take, as a simple example, single-input single-output control systems design. For each element in the system model, i.e., plant, controller, feedback network, the user has to look after four matrices for a state-space model or two polynomials for a transfer function. He cannot simply refer to the "transfer function G", but must refer instead to the numerator and the denominator polynomials (see Fig. 112.7) that stand for G. These polynomials can, in turn, only be distinguished from row vectors by convention and context.

In MATRIXx, this problem was avoided by using packing techniques and a special data-structure so that, for example, the state-space model in Fig. 112.3, would have been stored as shown in Fig. 112.8 and additional data would be stored in the workspace of the program so that the A, B, C, D matrices could be later extracted when needed.

Such packing schemes are quite widely used by toolbox writers to overcome the limitations imposed by the two-dimensional matrix. One is usually advised, but not required, to manipulate such structures only through

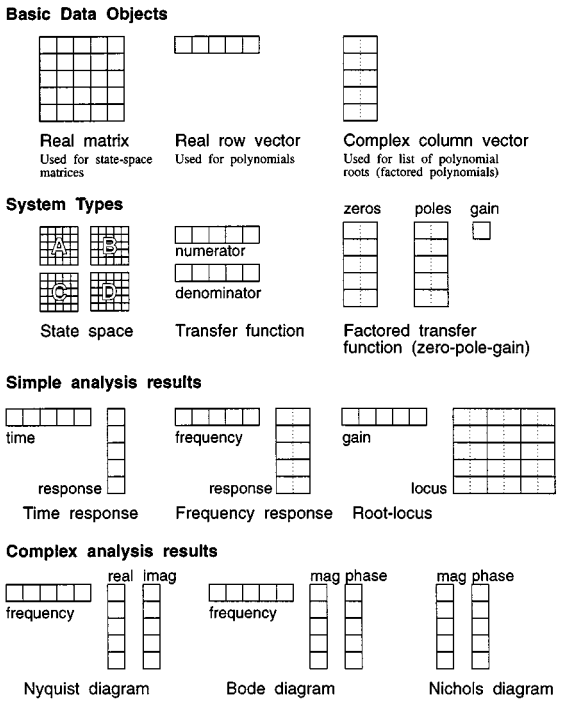


FIGURE 112.6 Some of the MATLAB conventions used to support control engineering data types.

```
>> % Plant: G(s) = 5/s(s^2 + 2s + 1)
>>  num_G = 5; den_G = conv([1 0],[1 2 1]);
>> % Controller: Gc(s) = 15(s + 1)/(s + 2)
>>   K_Gc = 15; Z_Gc = -1; P_Gc = -2;
>>   [num_Gc,den_Gc] = zp2tf(K_Gc,Z_Gc,P_Gc);
>> % Feedback: H(s) = 1/(s + 10)
>>   num_H = 1; den_H = [1 10];
```

FIGURE 112.7 Defining a control system in MATLAB.

```
>>  G = [ 0,  1, 0, 0
          0,  0, 0, 0
         -2, -3, 4, 1
          1,  0, 0, 0]
>> % size(A) = [3,3], size(B) = [3,1], size(C) = [1,3]
```

FIGURE 112.8 A packed "system matrix", additional values would have to be included to store the sizes of the relevant elements but these are not shown for clarity.

the packing and unpacking routines that usually accompany the toolbox code. For example, the packed state-space model might have a function `sstosys` to pack the data and `systoss` to unpack it into separate components as shown in Fig. 112.9. The advantage is that once packed, the state-space model *G* can be used in processing as if it was the single system object it represents. To see this, compare the code for simulation and analysis of a system given in Fig. 112.10(a) with the MATLAB Control System Toolbox code given in Fig. 112.10(b).

However, aside from the problem that packed data structures may be accidently used as ordinary matrices, there is a more severe problem that results from a lack of standardization. There are now a number of toolboxes

```
>>  G = sstosys(A,B,C,D);
>>  [a,b,c,d] = sstosys(G)

A =    1   0   0
       0   0   1
      -2  -3   4

B =    0
       0
       1

C =   1 0 0

D = 0
>>
```

FIGURE 112.9 Packing and unpacking system models.

```
>>  Go = series(Gc,G)
>>   rlocus(Go)
```

a) Using packed data

```
>>  [Go_A,Go_B,Go_C,Go_D] = ...
       series(Gc_A,Gc_B,Gc_C,Gc_D,G_A,G_B,G_C,G_D)
>>  rlocus(Go_A,Go_B,Go_C,Go_D)
```

b) Using non-packed data, the MATLAB control systems toolbox

FIGURE 112.10 Use of a packed datastructure to simplify interaction.

that are used in CACSD, and none of them takes a standard approach to packing data structures. Thus, the data structures used in the Multivariable Control Systems Toolbox are completely incompatible with those used in the Systems Identification Toolbox, which itself is incompatible with the standard Control Systems Toolbox. The consequence is that each toolbox must supply conversion tools and the situation is similar to the problems faced with integrating data from two different packages.

There is, therefore, an identified need for matrix environments to provide a wider range of data types, preferably user definable. These would be used in the same way as record datatypes are used in conventional programming systems and would be considerably safer to use since the types expected and returned by functions could be specified in advance and the scope for misuse would be much reduced. In addition, the need to invent new types for each application would be somewhat reduced. This approach has been taken in the matrix environment X-Math and similar features are planned for a future release of MATLAB. Some of the other requirements listed above, such as graphical systems input, graphical display of results, and spreadsheet data manipulation, are covered to a greater or lesser extent by the current generation of matrix environments. The others, namely symbolic data processing and database support, are not but are considered to be outside the scope of this article.

Fast Yet Flexible Command Language

MATLAB clearly satisfies this criterion as is evidenced by the natural interaction shown in Fig. 112.3. For CACSD use, it is debatable whether the principle still holds, mainly because of the way that the package entities needed for control have to be constructed and managed by the user. Nonetheless, no-one could complain that matrix environments are not flexible: the growing number of new control applications for them provides ample evidence of that.

Algorithmic Interface

The support of an algorithmic interface is simply a recognition of the fact that no package developer can anticipate the requirements of every user. So, the package must be extensible by provision of user-defined

macros and functions. MATLAB has these, and their provision is clearly important to the users of the package and developers of toolbox extensions. However, there is a limit to the software robustness of the mechanisms that MATLAB provides. MATLAB is an un-typed language, all data structures used in extensions to MATLAB are implemented in terms of collections of matrices and vectors. It is therefore up to the programmer to develop conventions for using these data items such that the algorithms work properly. A strongly typed language, in which the user must specify the nature of each data object before it is used, is a much safer basis on which to provide extensions that are to be used by many other people.

Transition From Basic to Advanced Use

The user of a CACSD package is faced with two different types of complexity: the complexity of the user interface and the complexity of the underlying theory and algorithms. In both cases extra guidance is needed for novice users. Typically, the designers of CACSD packages do not wish to stand in the way of the expert users, so they provide direct access to the whole package and interfere in the use of the package as little as possible. This creates problems for novice or infrequent users of the package—novices because they are coming to the package without any knowledge of it, infrequent users because they have probably forgotten most of what they learned the last time they used the package.

In MATLAB, the user interface is deceptively simple. One can take a short tutorial and learn the basic concepts and underlying principles in perhaps one hour. But what happens when one is finished with the tutorial and wants to do some actual work? The sheer number of commands in the system can be overwhelming. In basic MATLAB there are some two hundred commands, add a few toolboxes and the number quickly increases. The only way to find out how to use a command is to know its name. If you don't know the name you can list all the commands available, but since each command name is limited to eight characters, there is not necessarily going to be any relationship between command name and command function. Having found a command the next step is to learn how to use it. In a research prototype CACSD package called IMPACT, Rimvall and Bomholt [1985] provided a latent question and answer mode feature which switches from normal command entry to step by step elicitation of parameters when requested by the user. Other ways of overcoming some of these difficulties [Rimvall, 1988] include providing a means of loading toolboxes only when they are needed, thereby reducing the instantaneous "name-space", and providing operator overloading so that the same named procedure in X-Math [Floyd et al., 1991] and enables, for example, the multiplication operator '*' to mean matrix multiplication, series combination of systems, polynomial convolution, or time response evaluation depending on the types of the operands.

Transparency

This is a fundamental principle of software engineering that simply means that there should be no hidden processing or reliance on side effects on which the package depends for its correct operation. Everything the package does and the package itself should, at all times, be under the complete control of the user.

Scalability

This simply means that account should always be taken of the limitations of numerical algorithms. The package should warn the user when limits are reached and the algorithms should thereafter 'degrade gracefully'. It is surprising how many applications have been programmed with artificial limits set on various arrays which is fine so long as the user never presents the package with a problem that its designer never believed would ever be tackled (an inevitable event). Most matrix environments are limited only by the available memory.

"Open Systems"

The need to transfer data to other systems is simply a recognition that no one package can do all things equally well. In many applications, it makes sense to pass a processing task onto an expert. The ability of a package to be able to exchange data (both import and export) is the main feature of so-called open systems. At the very least it must be possible to save data in a form that can be retrieved by an external program. MATLAB and its

cousins provide basic file transfer capabilities, but the ideal CACSD package would have some link to a much more convenient data sharing mechanism such as could be provided by a database.

Other Desirable Features

- *Form or menu drive input* is often more useful than a functional command driven interface for certain type of data entry. A good example is the plotting of results where the selection of options and parameters for axis scaling, tick marks, etc. are more conveniently specified by means of a dialog box than by a series of function calls. Such a facility is provided in X-Math's graphics.
- *Graphical input* is useful for defining systems to be analyzed. Today, most of the major packages provide block diagram input, usually tied to nonlinear simulation. What is rarer is graphical input of more application-specific system representations such as circuit diagrams.
- *Strong data typing,* as already discussed, is useful for toolbox developers since it provides a robust means of developing extra algorithms within the context of the CACSD package. On the other hand, there is a fine balance between the needs of the algorithm developer and the algorithm implementor. The former is probably best served by a type-less environment in which it is easy and quick to try out new ideas (such an environment is often called a *rapid-prototyping environment*). The latter, who needs to ensure that the algorithms will work properly under all conditions, needs strong typing to ensure that this can be guaranteed. A similar dichotomy between inventors and implementors can be observed in software engineering.
- *Data persistence.* Unless explicitly saved, CACSD data is not maintained between sessions. Neither can data easily be shared between users. The evolution of models and results over time cannot be recorded. Hence, CACSD packages need database support.
- *Matrix environments only support numerical computation.* It is often useful to be able to manipulate a symbolic representation of a control system. Delaying the replacement of symbolic parameters for numerical values for as long as possible can often yield great insight into such properties as stability, sensitivity, and robustness.

112.4 CACSD Block-Diagram Tools

As we have discussed in the previous sections, the 1980s was an important decade for control engineering. Apart form new theories, better design methods, and more accurate numerical algorithms, this was the decade when powerful and easy-to-use interactive CACSD tools were put on the average control engineer's desk. Through the use of interactive and extendible programs, new methods and algorithms could be easily implemented and quickly brought to bear on real control engineering problems. Yet despite this tremendous improvement in the availability of good control design environments, the total cost and cycle time for a complex control design was still perceived by many groups and companies as being too high. One of the major remaining bottlenecks was the manual conversion of a control design into testable simulation code and, at a later stage, the conversion of the eventual design into the actual embedded real-time controller code.

A control engineer often describes a system through the use of block diagrams of different kinds. To bypass the bottleneck between theoretical design and actual real-time implementation, systems which took engineering block diagrams and automatically converted them into simulation and/or real-time code started to emerge in the middle of the 1980s. As an early example, already in 1984 General Electric decided to develop a block-diagram-based tool with automatic code generation capabilities. This program allowed draftspersons to enter control block diagrams and automatically convert the functionality of these diagrams into real-time code. Although it used limited graphics, this GE-Internal "Autocode" program successfully produced code at 50% of the cost of traditionally generated code, primarily due to error reduction of not hand coding. This reduction of costs provided the evidence that automatic translation of block diagrams is both feasible and desirable. However, due to advances in both computer graphics and code-generation techniques, the first tool was obsolete by the late 1980s. In recent years, several commercial block-diagram-based tools have become available. These

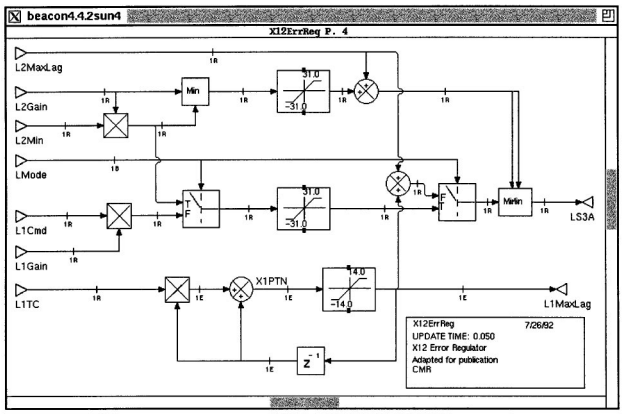


FIGURE 112.11 A signal-flow diagram in the BEACON system.

tools include System Build from Integrated Systems Incorporated, ModelC from Systems Control Technology, the PC-Based XAnalog from Xanalog, Simulab/Simulink from the Mathworks, and BEACON from General Electric. Some of these tools primarily serve as interfaces to analysis packages such as MATRIXx (System-Build), CTRL-C (Model-C), and MATLAB (Simulink). In some cases they can also be used to directly generate a computer language such as FORTRAN, ADA, or C. A summary of an early 1989 evaluation of the suitability of using System Build, CTRL-C, and Grumman's Protoblock for engine control is given in [Spang et al., 1993].

Basic Block-Diagram System Representations

Some basic user requirements fulfilled by most modern block-diagram oriented CACSD packages are

1. A simple-to-use graphical user-interface that can be used with little or no training. The graphical interface is usually based on the Macintosh, MS-Windows, and/or the X-Window System standard.
2. A set of rules for drawing controls-oriented diagrams, sometimes adhering to a standard diagram representations such as IEC-1331 or Petri Nets.
3. An object-based representation of the diagram entities and their graphical behavior. The underlying package must retain a semantical understanding of the diagram so that, for example, pertinent information such as signal types, dimensions, and ranges are propagated through the diagram, or connecting lines are retained when objects are moved.
4. Hierarchical structure which allows individual blocks to reference either other block diagrams or external modules (e.g., pre-coded system primitives).
5. Efficient internal simulation capabilities and/or real time code generation capabilities including optimization of execution speed and/or memory allocation.

As a consequence of the last two points, the block-diagram tools must have an open architecture so that the created modules can be associated with external code in a modular fashion. There are two main reasons for this:

- When the block-diagrams are used to simulate a physical system, the resulting models must frequently be interfaced with already existing submodels (e.g., from various FORTRAN libraries).
- When real-time controllers are implemented, the auto-generated code must be interfaced with operating system code and other "foreign" software.

All of today's block-diagram CACSD tools use hierarchical *signal-flow diagrams* as their main system representation. As illustrated in Fig. 112.11, a signal-flow diagram is a directed graph with the nodes representing standard arithmetic, dynamic and logic control blocks such as adders, delays, various filters, nonlinear blocks, and Boolean logic blocks. The connections between the blocks represent "signal" information which is

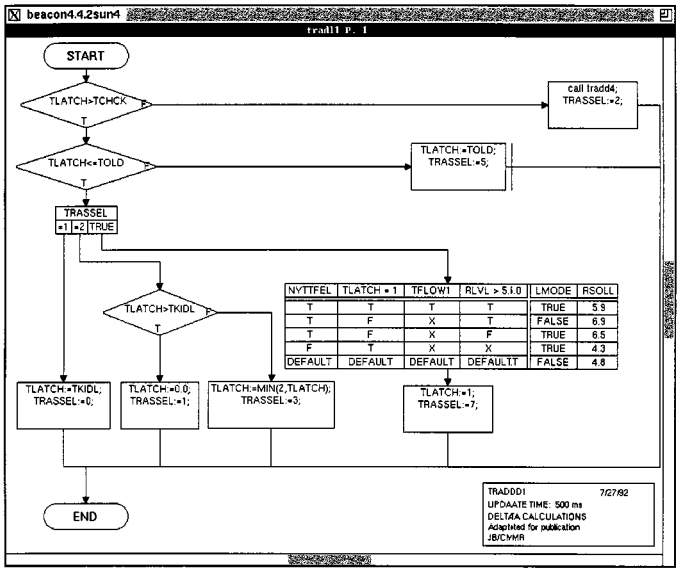


FIGURE 112.12 A BEACON control-flow block diagram.

transmitted from one block to another. The connections also indicate the order of execution of the various blocks. Signal flow diagrams are ideal for describing the dynamics of a system or controller.

Some CACSD packages also support some alternate system representation better suited for the logic and sequencing portion of a controller. Possible representations include *ladder-logic, dynamic truth-tables, flowcharts, Petri-nets,* or *state-transition diagrams.*

Figure 112.12 shows a typical control flow diagram or flowchart. The connections in this case represent the order of execution. The triangular blocks are decision blocks while the square blocks are variable assignment blocks written in a PASCAL-like language. Also shown are a multiway branch and a truth table. BEACON requires that the control flow diagrams produce structured code which equivalently means that a diagram can be implemented as a sequence of if-then-else statements without go-to's.

Hierarchies greatly facilitate the drawing and organization of diagrams. They provide appropriate levels of abstraction so that individual diagrams can be understood without clutter from details. Hierarchies simplify individual diagrams, making the resulting code easier to test. One can build up a set of subdiagram libraries which can be linked into possibly several higher level diagrams. Some block-diagram editors also allow the mixing of various diagram types in a hierarchical fashion (e.g., to call a low-level signal-flow diagram implementing a control-law scheme from a decision-making flow-chart diagram).

The graphical modeling environments cannot be viewed as replacements for the matrix environments described in the previous sections, as most of the block-diagram environments have very limited analytical capabilities (usually only simulation and linearization). However, many of today's block diagram tools have been developed as companion packages by the same commercial vendors that also sell matrix environments. Through linearization, it thus becomes possible to transform a non-linear block diagram to a linear representation which can then be analyzed and used for design in the matrix environment. Unfortunately, such automatic transformations are only available between tools from the same vendor, cross-translations between arbitrary tools are not possible.

Architectures of Block-Diagram Systems

To illustrate typical features and capabilities of a block-diagram oriented simulation or code-generation package, examples will be drawn from BEACON, a CACSD environment developed at GE between 1989 and 1995. There

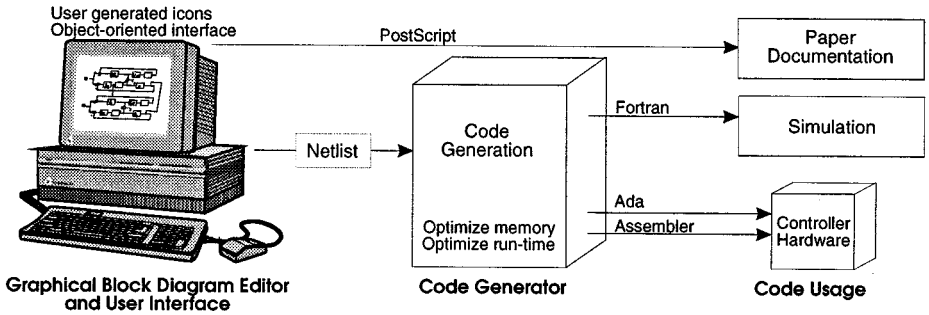


FIGURE 112.13 The BEACON architecture.

are, of course, many other block diagram systems, but being commercial products, the essential features are difficult to describe in detail. That said, another system that is well documented and worthy of study is the BlockEdit tool which was part of ECSTASY, a CACSD package developed in the UK in the late 1980s [Munro and Jobling, 1994]. BEACON has been in production use within GE since the first quarter of 1992. Through the use of BEACON, the company has been able to substantially reduce the overall cost and cycle time for developing complex controllers. The automatic generation of code not only eliminates the time-consuming manual coding, but also avoids the manual introduction of bugs into the code.

BEACON allows the user to graphically design a complete real-time controller as a series of hierarchical block diagrams. These diagrams can thereafter be automatically converted into a variety of computer languages for either control analysis, simulation, or real-time computer code, as illustrated in Fig. 112.13.

As shown in this figure, the BEACON system consists of three major components:

1. A graphical block-diagram editor with which the engineer designs the system to be simulated/coded [Spang et al., 1993]. Within this editor, the user may also create new graphical icons representing various numerical or logical blocks.
2. A netlist generated from the diagram and containing a full description of that diagram. The netlist format is keyword-oriented, it has a syntax resembling that of a higher-level language such as Pascal or Ada. To allow a variety of code generators and other uses such as the generation of I/O or termination lists or the automatic generation of test cases, all of the information except graphical location contained in the block diagram is written to the ASCII nestlist file.
3. An automatic code generator which translates the block diagrams into simulation and/or real-time computer code [Rimvall et al., 1993].

The BEACON architecture is one of the most open and extendible in the industry, allowing for straightforward extensions to the capability of the system and easy interfacing to other systems. Therefore, the architecture of other block diagram environments is often variants of that of BEACON. Some of the most common differences found in other systems are:

- *Built-in simulation capabilities.* Many of today's commercial systems have a non-linear simulation engine directly built into the system, avoiding BEACON's explicit translation step. Simulation results may then also be directly displayed on or accessed from the original diagram (e.g., in the form of time histories). This allows the user to see immediately the effects of any changes made to the diagram. One drawback of this approach is that these non-compiled approaches all have some kind of threaded-code or interpretative model execution, leading to much slower simulations than explicitly compiled simulation models such as those coming out of BEACON. Some systems allow for either of the two approaches.
- *The avoidance of an explicit netlist.* Many systems have a monolithic architecture with no direct access to the information in a modeled system. This prevents users from directly interfacing the block-diagram editor to other tools or filters (as often performed on a quite *ad-hoc* basis by the users within GE).
- *No code-generation.* Some older systems have built-in simulation capabilities only, with no generation of real-time or explicit simulation code.

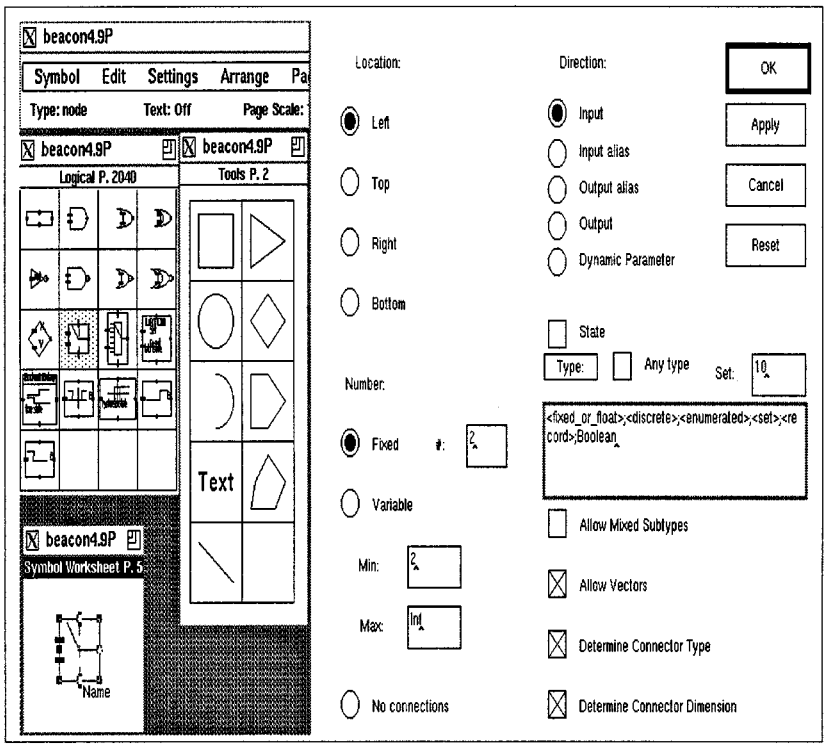


FIGURE 112.14 The BEACON symbol editor.

Open-Architectures of Block-Diagram Editors

Flexible block-diagrams have the capability of allowing users to develop or modify the graphical representation of symbols to meet the needs of various applications. In addition, it must be possible to add or modify the semantical meaning of the new or changed graphical symbols for simulation- or code-generation purposes.

The Editing of Block-Diagram Symbols

In BEACON, all symbols were developed using a Symbol Editor as shown in Figs. 112.14 and 112.15. This graphical editor is similar to most other object-oriented graphical editors, with the additional ability to describe diagram connectivity and the display of changing parameter values on the symbol itself. Each symbol is made up of a variety of separate objects (shapes) that are grouped together.

In Fig. 112.14, we see a Symbol Editor session, with the edited Switch symbol in the lower left window. The drawing primitives with its graphical shapes is the one in the middle. The large window to the right is an example of a block attributes window. In this case, it is the connectivity definition attributes for the left edge of the switch block; these attributes are used to define the sides and vertices which allow inputs or outputs, the allowed number of connections, vector dimension, and types.

Associated with most BEACON block symbols is a parameter form. These forms are unique for each individual block, allowing the user to define the parameters of the block and the specific function. For example, the integrator allows specification of the type of integration to be implemented as well as rate limits and initial conditions.

The forms are constructed during palette design using the Forms Editor shown in Fig. 112.15. To the left of the screen we see the actual parameter form of the Integrator block. In the middle we have the palette from which the primitive form elements may be picked. Each primitive forms object, such as text/value boxes and action buttons, have definable characteristics that will vary from element to element. To the right of Fig. 112.15 we see the characteristics of the data-input box for the parameter “lower limit”.

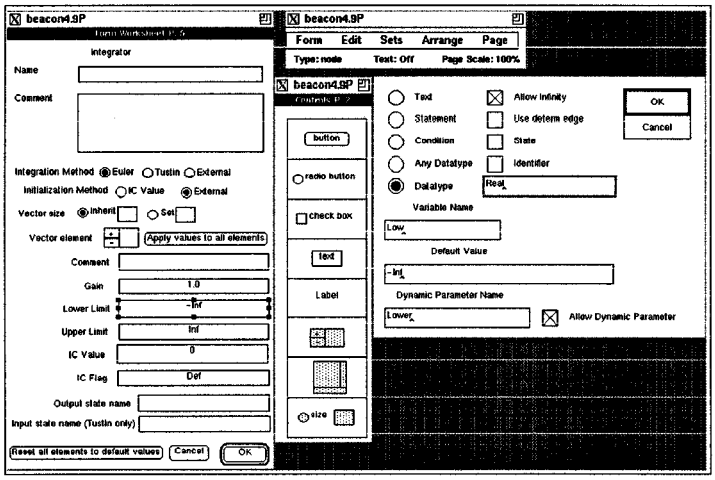


FIGURE 112.15 Examples of block parameter forms.

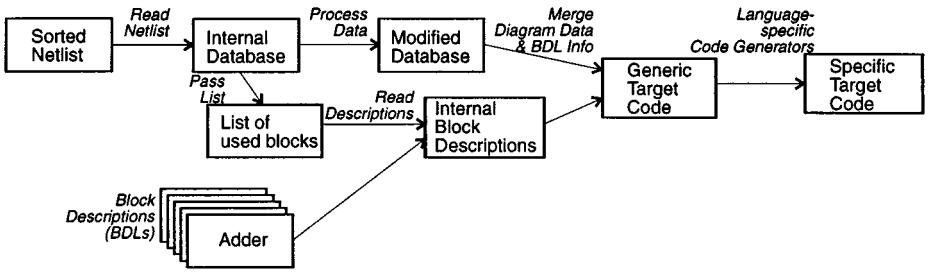


FIGURE 112.16 General principle of the workstation-based code generator.

The Functional Description of Symbols

The BEACON code generator will process a netlist into FORTRAN, Ada, C, or 68000 code. It accomplishes this by merging the block ordering information, the connectivity information, and the block-specific parameter values found in the netlist with block-specific functional descriptions of each block type. These block descriptions are stored separately from the netlist. This process is illustrated in Fig. 112.16.

Each block type supported by BEACON (e.g., adder, integrator, switch) will have a single block definition describing the functionality of the block. Whenever a new block symbol is added using the graphical Symbol Editor, a corresponding block definition file must be added to the system too. This block definition is written in "BEACON Block-Definition Language" (BDL), a special-purpose structured language that contains all the necessary elements for describing block connectivity, block parameters, and algorithms, as well as implementational detail such as fixed-point scaling.

Code From Signal-Flow Diagrams

The code resulting from a signal-flow diagram is a well-structured and yet locally optimized implementation of the diagram. It has the following characteristics:

- Through processing the sorted netlist, each block on the diagram is individually mapped onto the target language. Named blocks, e.g., `LIMAXLAG` in Fig. 112.11, are preceded by a comment stating that name in the code.

- Each connection on the diagram corresponds to a memory location in the code. To ensure readable code, each labeled connection, e.g., `X1PTN` in Fig. 112.11, is explicitly declared as a variable in the code. Unlabeled connections are mapped into reusable temporary variables or, in the case of assembler code, temporarily used registers. This ensures a locally optimized and yet fully readable code.
- States and other variables explicitly named on the diagram retain their name in the code. Unnamed states are automatically assigned unique names.
- Each numerical value is directly inserted into the code using the appropriate format of the target language and arithmetic type/precision used.

Code From Control-Flow Diagrams

Control-flow diagrams are processed in a similar manner to signal-flow diagrams. The main difference is that while a signal flow diagram uses a fixed set of blocks with well-defined semantics (the block interconnections and block parameters being the only variants between two blocks of the same type), the blocks in control-flow diagrams may contain arbitrary expressions, assignment statements, and/or procedure calls (as shown in Fig. 112.12). These BEACON language constructs must be translated into the primitives of each target language.

The BEACON graphical editor ensures that control-flow diagrams are well structured, i.e., that the diagram can be mapped into structured code. The automatic translation of large truth-tables into complex structured code is particularly time-saving.

Conclusions

In this chapter we have reviewed the tools available for the computer-aided design of control systems. The main features of the current state-of-the-art are analysis tools built around a "matrix environment" and modeling, simulation, and code generation tools constructed around the block diagram representation. For the most part, control systems analysis and design is done from a textual interface and modeling, simulation, and code generation rely on a graphical user interface. There are links between the two "environments", usually provided by some form of linearization.

Future CACSD environments will have to give equal emphasis to "control data objects" as they now do for matrices. This is becoming urgent as the number of specialist toolboxes being written for MATLAB and similar packages increases. Only by having a set of commonly approved data-types can the further development of incompatible data formats *within a single package* be prevented. Rimvall has defined an extended MATLAB-compatible command language to overcome such problems and the issues are discussed in [Rimvall and Wette, 1993].

As graphical user interfaces become more popular on computing devices, the possibilities for interactive manipulation of systems will have to be explored. We expect that graphical tools for control systems analysis and design will become common-place over the next few years and may eventually replace textual interfaces for most users.

A final important area for development of CACSD will be driven by the need to embed control systems design into information systems for enterprise integration. To some extent this is already happening with the need for multidisciplinary teams of engineers to work on common problems. The computer-based support of such projects requires facilities for the development and exchange of models, the storage of design data, version control, configuration management, project management, and computer-supported cooperative work. It is likely that CACSD will have to develop into a much more open set of tools supported by databases, networks, and distributed computation. The implications of some of these developments are discussed in [Barker et al., 1993].

Related Topics

100.1 Models • 100.2 Dynamic Response • 100.3 Frequency Response Methods: Bode Diagram Approach

References

E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Soresen, "LAPACK: A portable liner algebra library for supercomputers," technical report, Argonne National Laboratory, 1989.

D. Augustin, J. C. Strauss, M. S. Fineberg, B. B. Johnson, R. N. Linebarger, and F. J. Samson, "The SCi continuous system simulation language (CSSL)," *Simulation,* 9(6), 281–304, 1967.

H. A. Barker, M. Chen, P. W. Grant, C. P. Jobling, and P. Townsend, "Open architecture for computer-aided control engineering," *IEEE Control Systems Magazine,* 12(3), 17–27, 1993.

J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, "LINPACK users' guide," *Lecture Notes in Computer Science,* 1979.

M. A. Floyd, P. J. Dawes, and U. Milletti, "X-Math: a new generation of object-oriented CACSD tools," in *Proceedings European Control Conference,* 3, 2232–2237, 1991.

C. Moler, "MATLAB—user's guide," technical report, Alberquerque, N.M.: University of New Mexico, 1980.

N. Munro and C. P. Jobling, "ECSTASY: A control system CAD environment," in *CAD for Control Systems,* D. A. Linkens, Ed., New York: Marcel Dekker, pp. 449–467.

C. M. Rimvall, "CACSD software and man-machine interfaces of modern control environments," *Transactions of the Institute of Measurement and Control,* 9(2), 1987.

C. M. Rimvall, "Interactive environments for CACSD software," in *Preprints of 4th IFAC Symp. on Computer Aided Design in Control Systems CADCS '88,* pp. 17–26, Beijing, PRC, 1988.

C. M. Rimvall and L. Bomholt, "A flexible man-machine interface for CACSD applications," in *Proc. 3rd IFAC Symp. on Computer Aided Design in Control and Engineering,* Pergamon Press, 1985.

C. M. Rimvall, M. Radecki, A. Komar, A. Wadhwa, H. A. Spang III, R. Knopf, and M. Idelchik, "Automatic generation of real-time code using the BEACON CAE environment," in *Proceedings of the 12th IFAC World Congress on Automatic Control,* 6, 99–104, 1993.

C. M. Rimvall and M. Wette, "Towards standards for CACE command syntax and graphical interfaces," in *Proceedings of the 12th IFAC World Congress on Automatic Control,* 8, 87–390, 1993.

B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, and Y. Ikebe, "Matrix eigensystem routines—EISPACK guide extension," *Lecture Notes in Computer Science,* 51, 1977.

H. A. Spang III, C. M. Rimvall, H. A. Sutherland, and W. Dixon, "An evaluation of block diagram CAE tools," in *Proceedings of the 11th IFAC World Congress on Automatic Control,* 9, 79–84, 1990.

H. A. Spang III, A. Wadhwa, C. M. Rimvall, R. Knopf, M. Radecki, and M. Idelchik, "The BEACON block-diagram environment," in *Proceedings of the 12th IFAC World Congress on Automatic Control,* 6, 105–110, 1993.

A. van den Boom, A. Brown, F. Dumortier, A. Geurts, S. Hammarling, R. Kool, M. Vanbegin, P. van Dooren, and S. van Huffle, "SLICOT: A subroutine library in control and systems theory," in *Proceedings 5th IFAC Symposium on Computer Aided Design in Control Systems—CADCS'91,* pages 1–76, Swansea, UK, 1991.

Further Information

Keeping up to date with developments in CACSD is not always easy but the proceedings of the triennual IFAC symposium on Computer-Aided Design in Control Systems (CADCS) and the IEEE biennual workshop on CACSD are useful indicators of the latest trends. The proceedings of the last three of these meetings are given below. The other items give useful snapshots of the state-of-the-art at various points in the last 10 years or so. In addition to these sources, the *IEEE Control Systems Magazine* regularly publishes articles on CACSD and is a good place to look for other information.

M. Jamshidi and C. J. Herget, Eds., *Computer-Aided Control Systems Engineering,* North-Holland, 1985.

CADCS, *Proceedings of the 5th IFAC Symposium on Computer Aided Design in Control Systems,* Swansea, UK: Pergamon Press, 1991.

M. Jamshidi, M. Tarokh, and B. Shafai, *Computer-Aided Analysis and Design of Control Systems,* Englewood Cliffs, N.J.: Prentice-Hall, 1991.

CACSD, *Proceedings of the IEEE Control Systems Society Symposium on CACSD,* Napa, Calif.: IEEE, 1992.

M. Jamshidi and C. J. Herget, Eds., *Recent Advances in Computer-Aided Control Systems Engineering. Studies in Automation and Control,* Amsterdam: Elsevier Science Publishers, 1992.

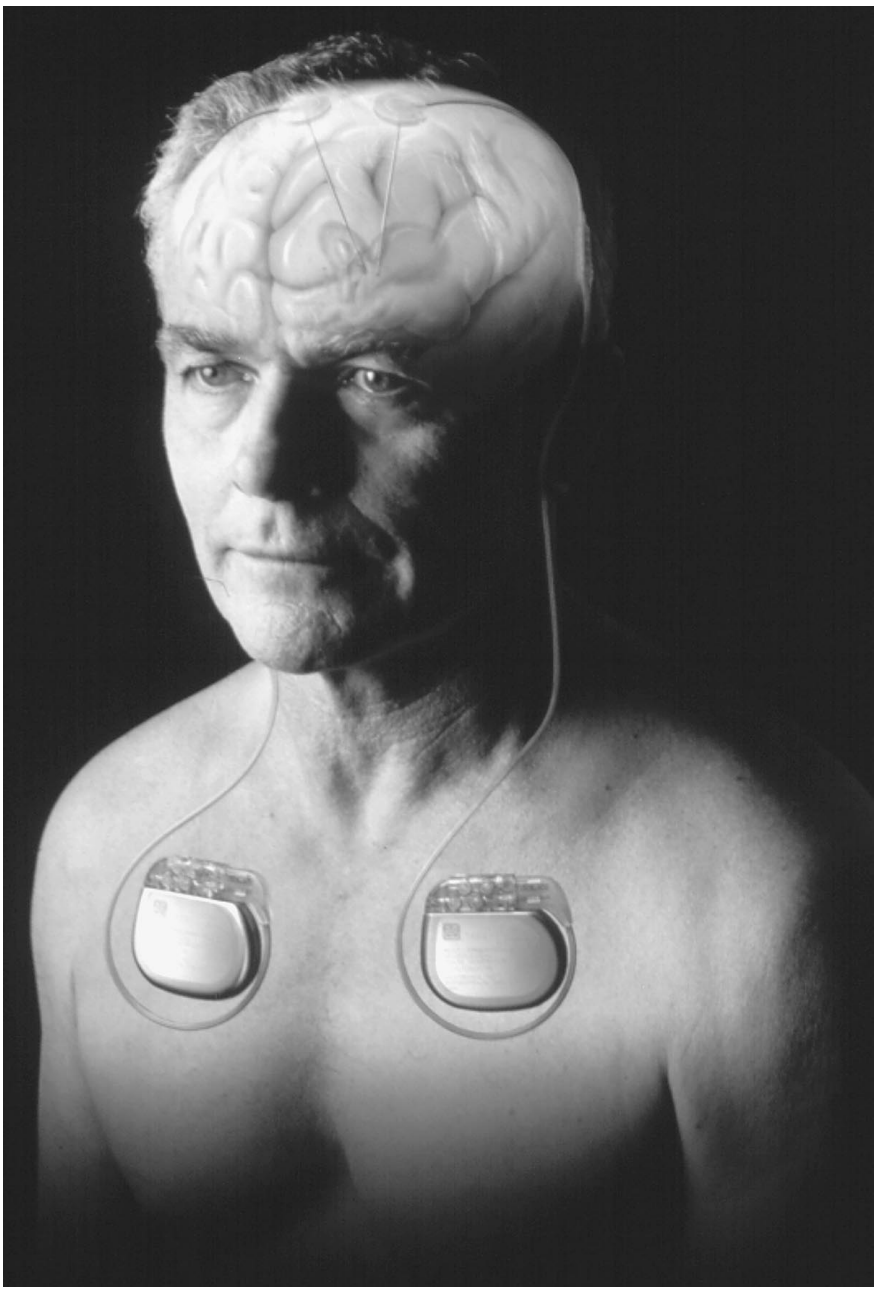
CACSD, *Proceedings of the IEEE/IFAC Joint Symposium on Computer-Aided Control System Design,* Tucson, Az., Pergamon Press, 1994.

D. A. Linkens, Ed., *CAD for Control Systems,* New York: Marcel Dekker, 1994.

CACSD, *Proceedings of the IEEE Symposium on Computer-Aided Control System Design,* Deerborn: IEEE, 1996.

IFAC CACSD, *Proceedings of the 1997 IFAC Symposium on Computer-Aided Control System Design,* Ghent, Belgium: IFAC, 1997.

Bronzino, J.D. "Section XI – Biomedical Systems"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



Essential tremor is the most common neurological movement disorder in the U.S. The condition afflicts more than one million people, usually age 45 and older, in the U.S. alone. Thanks to Medtronic, Inc., the world's leading medical technology company specializing in implantable and invasive therapies, help for this debilitating condition is in sight.

The Activa Tremor Control System, shown above, uses an implanted device, similar to a cardiac pacemaker, to deliver electrical stimulation to block or override brain signals that cause tremor. The system allows the stimulation level to be adjusted to the needs of each patient.

Patients in Canada, Europe, and Australia are already using the Activa Tremor Control System. In the U.S. it has been recommended unanimously for marketing clearance by the Neurological Device Panel Advisory Committee to the U.S. Food and Drug Administration. (Photo courtesy of Medtronic, Inc.)

XI

Biomedical Systems

113 Bioelectricity *J.P. Reilly, L.A. Geddes, C. Polk*
Neuroelectric Principles • Bioelectric Events • Application of Electrical and Magnetic Fields in Bone and Soft Tissue Repair

114 Biomedical Sensors *M.R. Neuman*
Physical Sensors • Chemical Sensors • Bioanalytical Sensors • Applications • Summary

115 Bioelectronics and Instruments *J.D. Bronzino, E.J. Berbari, P.L. Johnson, W.M. Smith*
The Electroencephalogram • The Electrocardiograph • Pacemakers/Implantable Defibrillators

116 Medical Imaging *M.D. Fox, L.A. Frizzell, L.A. Franks, L.S. Darken, R.B. James*
Tomography • Ultrasound • Semiconductor Detectors for Radiation Measurements

117 Biocomputing *L. Kun, M.F. Baretich*
Clinical Information Systems • Hospital Information Systems

118 Computer Design for Biomedical Applications *R. Luebbers*

119 Rehabilitation Engineering, Science, and Technology *C.J. Robinson*

Joseph D. Bronzino
Trinity College

TECHNOLOGICAL INNOVATION in the twentieth century has progressed at such an accelerated pace that it has permeated almost every facet of our lives. This is especially true in the field of medicine and the delivery of health care services. Although the art of medicine has a long history, the evolution of a health care system capable of providing a wide range of positive therapeutic treatments in the prevention and cure of illnesses is a decidedly new phenomenon. Of particular importance in this evolutionary process has been the establishment of the modern hospital as the center of a technologically sophisticated health care system. In the process, the discipline of *biomedical engineering* has emerged as an integrating medium for two dynamic professions, medicine and engineering, assisting in the struggle against illness and diseases by providing materials, tools, and techniques (such as signal and image processing and artificial intelligence) that can be utilized for research, diagnosis, and treatment by health care professionals.

Today, biomedical engineering is an interdisciplinary branch of engineering heavily based both in engineering and in the life sciences. It ranges from theoretical, nonexperimental undertakings to state-of-the-art applications. It can encompass research, development, implementation, and operation. Accordingly, like medical practice itself, it is unlikely that any single person can acquire expertise that encompasses the entire field. As a result, there are now a great number of biomedical engineering specialists to cover this broad spectrum of activity. Yet because of the interdisciplinary nature of this activity, there is considerable interplay and overlapping

of interest and effort between them. For example, biomedical engineers engaged in the development of biosensors may interact with those interested in prosthetic devices to develop a means to detect and use the same bioelectric signal to power a prosthetic device. Those engaged in automating the clinical chemistry laboratory may collaborate with those developing expert systems to assist clinicians in making clinical decisions based upon specific laboratory data. The possibilities are endless.

There are seven major career areas in biomedical engineering: (1) application of engineering system analysis and modeling (computer simulation) to biological problems; (2) measurement or monitoring of physiological signals; (3) diagnostic interpretation via signal processing techniques of bioelectric data; (4) therapeutic and rehabilitation procedures and devices; (5) prosthetic devices for replacement or augmentation of bodily functions; (6) computer analysis of patient-related data; and (7) medical imaging, i.e., the graphic display of anatomical detail or physiological function. Biomedical engineers, therefore, engage in the following pursuits:

- Design of instrumentation for human physiology research
- Monitoring astronauts and maintenance of life in space
- Research in new materials for implanted artificial organs
- Development of new diagnostic instruments for blood analysis
- Computer modeling of the function of the human heart
- Writing software for analysis of medical research data
- Analysis of medical device hazards for the U.S. government
- Monitoring the physiological functions of animals
- Development of new diagnostic imaging systems
- Design of telemetry systems for patient monitoring
- Design of biomedical sensors for measurement of human physiological systems variables
- Research on artificial intelligence (AI) and development of expert systems for diagnosis of diseases
- Design of closed-loop control systems for drug administration
- Modeling of the physiological systems of the human body
- Design of instrumentation for sports medicine
- Development of new dental materials
- Design of computers and communication aids for the handicapped
- Research in pulmonary fluid dynamics (biorheology)
- Study of the biomechanics of the human body

This list is not intended to be all-inclusive, for there are many other applications that utilize the talents and skills of the biomedical engineer. In fact, the list of activities of biomedical engineers depends upon the medical environment in which they work. This is especially true for the "clinical engineers," i.e., biomedical engineers employed in hospitals or clinical settings.

The utilization of biomedical engineers offers great potential benefit in the identification of problems and needs of our present health care delivery system that can be solved using existing engineering technology and systems methodology. Consequently, the field of biomedical engineering offers hope in the continuing battle to provide high-quality health care at reasonable cost. The purpose of this section, therefore, is to provide a broad overview of biomedical engineering topics of interest to electrical engineers.

Nomenclature

Symbol	Quantity	Unit
C	resistivity	Ω
C	proton density	
C_m	membrane capacitance	F
d	diameter	m
D	lateral beamwidth	
f_d	doppler shift	Hz
f_L	Larmour frequency	Hz
F	Faraday constant	
Γ	pressure reflection coefficient	
I_{mt}	minimum threshold current	A
I_t	threshold current	A
k	propagation constant	
L	internodal distance	m
m	mass	g
n	valence of ion	
p	pressure	N/m
Q_T	threshold charge	C
R	universal gas constant	
R	transmembrane resistance	Ω
σ	conductivity	S/m
t_d	decay time	s
t_r	rise time	s
T	absolute temperature	K
V	membrane potential	V
W	nodal gap width	
Z	valence of substance	
Z	acoustic impedance	W

Reilly, J.P., Geddes, L.A., Polk, C. "Bioelectricity"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

113

Bioelectricity

113.1 Neuroelectric Principles
Electrical Model for Nerve Excitation

113.2 Bioelectric Events
Origin of Bioelectricity • Law of Stimulation • Recording Action Potentials • The Electrocardiogram (ECG) • Electromyography (EMG) • Electroencephalography (EEG) • Magnetic (Eddy-Current) Stimulation

113.3 Application of Electric and Magnetic Fields in Bone and Soft Tissue Repair
History • Devices for Bone and Cartilage Repair • Soft Tissue Repair and Nerve Regeneration • Mechanisms and Dosimetry

J. Patrick Reilly
Metatec Associates

L. A. Geddes
Purdue University

C. Polk
University of Rhode Island

113.1 Neuroelectric Principles

J. Patrick Reilly

Natural bioelectric processes are responsible for nerve and muscle function. These processes can be affected by externally applied electric currents that are intentionally introduced through medical devices or unintentionally introduced through accidental exposure (electric shock). A thorough treatment of this topic is given in Reilly [1992].

Externally applied electric currents can excite nerve and muscle cells. Muscle can be stimulated directly or indirectly through the nerves that enervate the muscle. Thresholds of stimulation of nerve are generally well below thresholds for direct stimulation of muscle. An understanding of neuroelectric principles is a valuable foundation for investigation into both sensory and muscular responses to electrical stimulation.

Figure 113.1 illustrates functional components of sensory and motor (muscle) **neurons**. The illustrated nerve fibers are **myelinated**, i.e., covered with a fatty layer of insulation called *myelin* and having *nodes of Ranvier* where the myelin is absent. The conducting portion of the nerve fiber is a long, hollow structure known as an **axon**. The axon plus myelin sheath is frequently referred to as a nerve **fiber**, or neuron. Bundles of neurons are called *nerves*.

The body is equipped with a vast array of sensors (receptors) for monitoring its internal and external environment. Electrical stimulation generally involves the *somatosensory* system, i.e., the system of receptors found in the skin and internal organs. Other specialized receptors include those in the visual and auditory systems and chemical receptors by which neurons communicate with one another.

The somatosensory receptors can be classified as mechanoreceptors, thermoreceptors, chemoreceptors, and nociceptors. Numerous specializations of mechanoreceptors respond to specific attributes of mechanical stimulation. Thermoreceptors are specialized to respond to either heat or cold stimuli. Nociceptors are unresponsive until the stimulus reaches the point where tissue damage is imminent and are usually associated with pain. Many nociceptors are responsive to a broad spectrum of noxious levels of mechanical, heat, and chemical stimuli. The muscles are equipped with specialized receptors to monitor and control muscle movement and posture. Figure 113.1 illustrates a *pacinian corpuscle*, which responds to the onset or termination of a pressure stimulus applied to the skin.

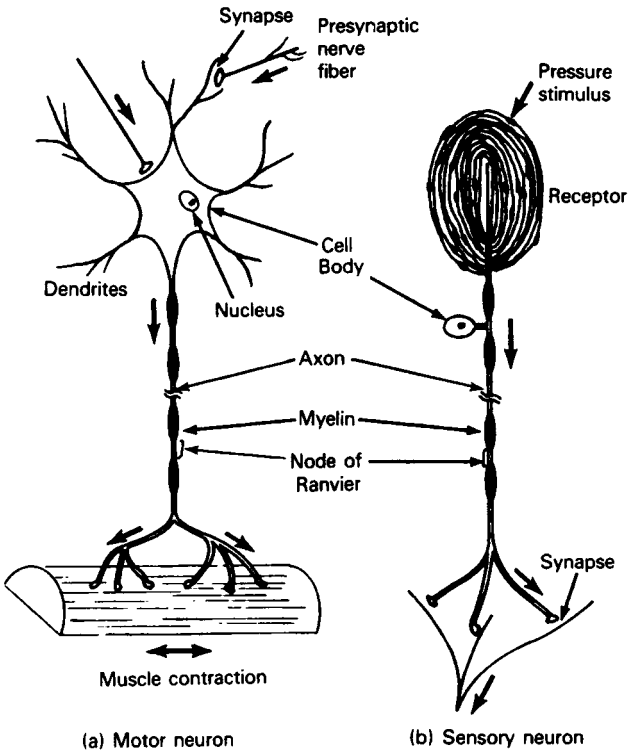


FIGURE 113.1 Functional components of (a) motor and (b) sensory neurons. Arrows indicate the direction of information flow. Signals are propagated across synapses via chemical neurotransmitters and elsewhere by membrane depolarization. Synapses are inside the spinal column. The sizes of the components are drawn on a distorted scale to emphasize various features.

When a sensory receptor is stimulated, it produces a voltage change called a *generator potential*. The generator potential is graded: if you squeeze a pacinian corpuscle, for example, it produces a voltage; if you squeeze it harder, it produces a greater voltage. The generator potential initiates a sequence of events that leads to a propagating **action potential** (a "nerve impulse" in common parlance).

The functional boundary of the biological cell is a thin (about 10 nm) bimolecular lipid and protein structure called a **membrane**. Electrochemical forces across the membrane regulate chemical exchange across the cell. The medium within the cell (the *plasm*) and outside the cell (the *interstitial fluid*) is composed largely of water containing various ions. The difference in the concentration of ions inside and outside the cell causes an electrochemical force across the cell membrane. The membrane is a semipermeable dielectric that allows some ionic interchange. Under conditions of electrochemical equilibrium (no net force in either direction), the membrane will attain a potential described by the *Nernst* equation

$$V_m = \frac{RT}{FZ} \ln \frac{[S]_o}{[S]_i} \tag{113.1}$$

where $[S]_i$ and $[S]_o$ represent the concentrations of ionic substance S inside and outside the cell, R is the universal gas constant, T is absolute temperature, F is the Faraday constant (number of coulombs per mole of charge), and Z is the valence of substance S. Using the values $R = 8.31$ J/mol K, $T = 310$ K, $F = 96{,}500$ C/mol, and $Z = +1$ (for a monovalent cation), converting to the base 10 logarithm, and expressing V_m in millivolts, we obtain

$$V_m = 61 \log \frac{[S]_o}{[S]_i} \tag{113.2}$$

In a quiescent state, nerve and muscle cells maintain a membrane potential typically around –60 to –90 mV, with the inside of the cell negative relative to the outside. Two ions that are involved in the electrical response of nerve and muscle are Na^+ and K^+. The concentration of these ions inside and outside the cell dictates the Nernst potential according to Eq. (113.2). Example concentrations in $\mu M/cm^3$ for a nerve axon would be $[Na^+]_i = 50$, $[Na^+]_o = 460$, $[K^+]_i = 400$, and $[K^+]_o = 10$. The Na^+ potential is found to be around +60 mV; the K^+ potential is found to be somewhat more negative than the resting potential. Obviously, the cell maintains in a state of electrochemical disequilibrium. The energy that maintains this force is derived from the metabolism of the cell—a dead cell will eventually revert to a state of equilibrium. Considering the transmembrane potential (≈100 mV), and its small thickness (≈10 nm), the electric field across the membrane is enormous (≈10 MV/m).

The membrane is semipermeable; that is, it is a lossy dielectric which allows the passage of certain ions. The ionic permeability varies substantially from one ionic species to another. The ionic channels in the excitable membrane will vary their permeability in response to the transmembrane potential; this property distinguishes the excitable membrane from the ordinary cellular membrane, and it supports propagation of nerve impulses.

The electrodynamics of the excitable membrane of unmyelinated nerves were first described in detail in the Nobel prize work of Hodgkin and Huxley [1952]. This work was later extended to the myelinated nerve membrane by Frankenhaeuser and Huxley [1964]. Figure 113.2 illustrates an electrical model of the Hodgkin-Huxley membrane, which consists of nonlinear conductances for Na^+ and K^+ and a linear leakage element. The potential sources shown in the diagram are the Nernst potentials for the particular ions as given by Eq. (113.2). The capacitance term C_m is formed by the dielectric membrane separating the conductive media on either side. The conductances g_{Na} and g_K apply to Na^+ and K^+ channels; the conductance g_L is a general "leakage" channel that is not specific to any particular ion. The g_{Na} and g_K conductivities are highly dependent on the voltage applied across the membrane as described by a set of nonlinear differential equations. When the membrane is in the resting state, $g_{Na} << g_K$, and the membrane potential moves toward the Nernst potential for Na^+. In this depolarized state, the membrane is said to be excited. The transition between the resting and excited condition of the membrane occurs rather abruptly when the membrane potential has been depolarized by roughly 15 mV. After excitation, the ionic channel conductances vary again, causing the membrane to revert back to its resting potential.

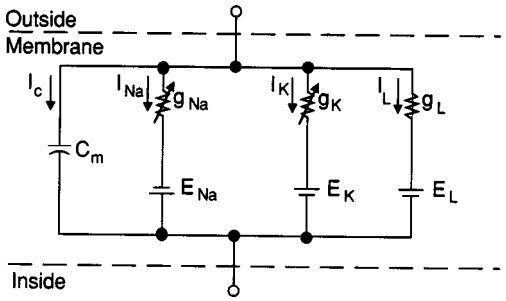


FIGURE 113.2 Hodgkin-Huxley membrane model.

The duration of the excited state lasts roughly 1 ms. The progression of the membrane voltage during the period of excitation and recovery is termed an *action potential*. After the membrane has been excited, it cannot be reexcited until a recovery period, called the **refractory period**, has passed.

Figure 113.3 illustrates the processes that support the propagation of an action potential. Consider that point A on the axon is depolarized. The local depolarization causes ionic transfer between adjacent points on the axon, thus propagating the region of depolarization. If depolarization were initiated from an external electrical source on a resting membrane at point A, an action potential would propagate in both directions away from the site of stimulation. The body's natural condition, however, is to initiate an action potential at the terminus of the axon, which then propagates in only one direction.

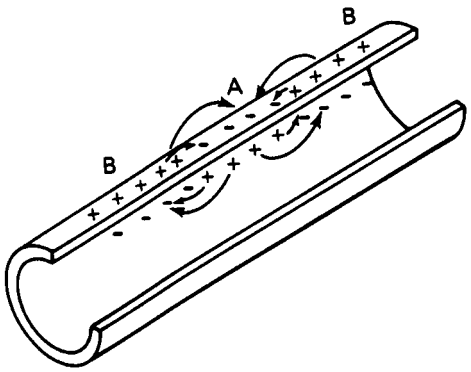


FIGURE 113.3 Spread of the depolarization wave front along an axon. Depolarization occurring in region A results in charge transfer from the adjacent regions.

Electrical Model for Nerve Excitation

Myelinated fibers have much lower thresholds of excitation than unmyelinated fibers. Accordingly, the myelinated fiber is an appropriate choice for electrical stimulation studies.

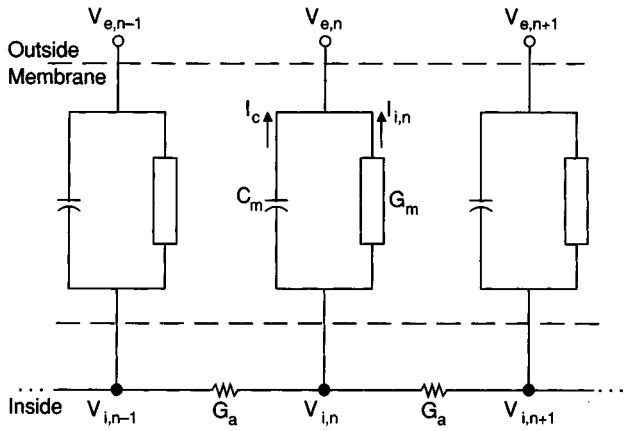


FIGURE 113.4 Equivalent circuit model for electrical excitation of myelinated nerve fiber. The membrane conductance G_m is described by nonlinear ionic conductances, similar to the representation in Fig. 113.2.

Figure 113.4 illustrates an electrical model for myelinated nerve as originally formulated by McNeal [1976]. The myelin internodes are treated as perfect insulators and the nodes as individual circuits consisting of capacitance C_m and an ionic conductance term. The nodes are interconnected through the internal axon medium by conductances G_a. The current flowing in the biological medium creates voltage disturbances $V_{e,n}$ at the exterior of the nodes.

The current emanating from the nth node is the sum of capacitive and ionic currents described by

$$C_m \frac{dV_n}{dt} + I_{i,n} = G_a(V_{i,n-1} - 2V_{i,n} + V_{i,n+1}) \tag{113.3}$$

where C_m is the membrane capacitance at the node, V_n is the transmembrane potential, $I_{i,n}$ is the total ionic current, and $V_{i,n}$ is the internal voltage. In this expression, V_n is taken relative to the resting potential, such that $V_n = 0$ applies to the membrane resting potential. The ionic current flux is the sum of individual ionic terms (similar to the representation in Fig. 113.4),

$$I_{i,n} = \pi dW(J_{\mathrm{Na}} + J_{\mathrm{K}} + J_L + J_P) \tag{113.4}$$

where the J terms are ionic current densities as described by a set of nonlinear differential equations developed by Frankenhaeuser and Huxley [1964] for a myelinated nerve membrane. Other relationships are

$$G_a = \frac{\pi d^2}{4\rho_i L} \tag{113.5}$$

$$C_m = c_m \pi dW \tag{113.6}$$

where d is the axon diameter at the node, ρ_i is the resistivity of the internal axon medium, L is the internodal distance, W is the nodal gap width, and c_m is the membrane capacitance per unit area. The relationship between the axon diameter d and the fiber diameter D (including myelin) is $d \approx 0.7D$. The voltage V_n across the membrane is

$$V_n = V_{i,n} - V_{e,n} \tag{113.7}$$

where $V_{i,n}$ and $V_{e,n}$ are the internal and external nodal voltages with reference to a distant point in the conducting medium outside the axon. Substituting Eq. (113.7) into (113.3) results in

$$\frac{dV_n}{dt} = \frac{1}{C_m}[G_a(V_{n-1} - 2V_n + V_{n+1} + V_{e,n} - 2V_{e,n} + V_{e,n+1}) - I_{i,n}] \tag{113.8}$$

For application to an unmyelinated fiber, Eq. (113.8) may be analogously expressed in continuous form as

$$\tau_m \frac{\partial V}{dt} - \lambda^2 \frac{\partial^2 V}{\partial x^2} + V = \lambda^2 \frac{\partial^2 V_e}{\partial x^2} \tag{113.9}$$

where V and V_e are membrane voltage and external voltage, respectively, at longitudal position x. Equation (113.9) can be derived from first principles, or can be obtained from (113.8) by substituting $C_m = c_m \pi d \Delta x$, $G_a = \pi d^2/(4\rho_i \Delta x)$, $G_m = g_m \pi d \Delta x$, where d is the fiber diameter, Δx is the longitudinal increment, ρ_i is the axoplasm resistivity (in Ωcm) internal to the fiber, c_m is capacitance per unit area, and g_m is conductance times unit area. Continuous and discrete spatial derivatives are connected by $\partial^2 V/\partial x^2 \approx (V_{n-1} - 2V_n + V_{n+1})/\Delta x^2$; $\partial^2 V_e/\partial x^2 \approx (V_{e,n-1} - 2V_{e,n} + V_{e,n+1})/\Delta x^2$; τ_m is the member time constant given by c_m/g_m; λ is the membrane space constant given by $\lambda = (r_m/r_i)^{1/2} = (d\rho_m/4\rho_i))^{1/2}$, and ρ_m is the membrane specific resistance (in Ωcm^2). An additional relationship is $I_{i,n} = V/G_m$.

If one treats λ as a constant, then (113.9) describes the membrane response only during its sub-threshold (linear) phase. For membrane depolarization approaching the threshold of excitation, membrane conductance of ionic constituents becomes highly nonlinear, as noted above — it is this nonlinear behavior that leads to nerve excitation.

The left-hand side of Eq. (113.9) is the so-called cable equation that was developed by Oliver Heaviside over 100 years ago in connection with the analysis of the first transatlantic telegraphy cable. The right-hand side is a driving function due to the external field in the biological medium. For additional information on cable theory as applied to the excitable membrane, the reader is directed to Jack et al. [1983].

One conclusion that can be drawn from Eqs. (113.8) and (113.9) is that a second spatial derivative of voltage (or equivalently a first derivative of the electric field) must exist along the long axis of an excitable fiber in order to support excitation. Nevertheless, excitation is possible in a locally constant electric field where the fiber is terminated or where it bends. The orientation change or the termination creates the equivalent of a spatial derivative of the applied field. Stimulation at "ends and bends" can be the dominant mode of excitation in many cases.

The external voltages in Eq. (113.8) are dependent on the distribution of current within the biological medium. For a point electrode in an isotopic medium, for instance, we can determine these voltages by

$$V_{e,n} = \frac{\rho_e I}{4\pi r_n} \tag{113.10}$$

where r_n is the distance between the stimulating electrode and the nth node and ρ_e is the resistivity of the external medium. For a uniform current density flowing in a direction parallel to the fiber axis, the external voltages are determined by

$$V_{e,n} = V_{e,1} + ELn \tag{113.11}$$

where $V_{e,1}$ is a reference voltage at the terminal node, L is the internodal distance, n is the node number, and E is the electric field in the medium. The electric field is related to current density by $J = E\sigma$, where $\sigma = 1/\rho_e$ is the conductivity of the medium and J is the current density. Since the response of the electrical model is independent of $V_{e,1}$, we may assume $V_{e,1} = 0$ for convenience in Eq. (113.11).

The internodal distance L is proportional to fiber diameter D through the relationship $L/D \approx 100$. Other fiber diameter relationships are expressed in Eqs. (113.5) and (113.6). Because of these relationships, thresholds of electrical stimulation will vary inversely with fiber diameter. The distribution of myelinated nerve diameters found in human peripheral nerve or skeletal muscle typically ranges from 5 to 20 μm.

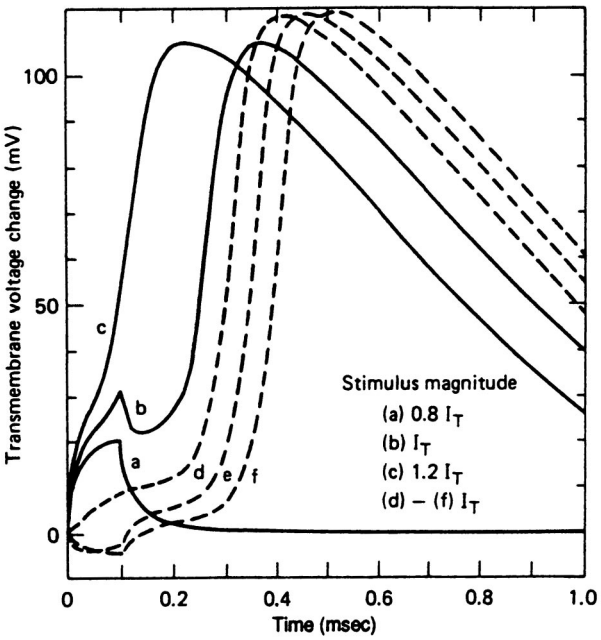


FIGURE 113.5 Response of myelinated nerve model to rectangular monophasic current of 100 ms duration, 20-μm diameter fiber, point electrode 2 mm from central node. Solid lines show response at node nearest electrode for three levels of current. I_T denotes threshold current. Dashed lines show propagated response at next three adjacent nodes for a stimulus at threshold. (*Source:* J. P. Reilly, V. T. Freeman, and W. D. Larkin, "Sensory effects of transient electrical stimulation—Evaluation with a neuroelectric model," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 12, pp. 1001–1011, © 1985 IEEE.)

Figure 113.5 illustrates the response of the myelinated nerve model of Fig. 113.4 to a rectangular current stimulus [Reilly et al., 1985]. The example is for a small cathodal electrode that is 2 mm radially distant from a 20-μm fiber and directly above a central node. The transmembrane voltage ΔV is scaled relative to the resting potential. The solid curves show the response at the node nearest the stimulating electrode. Response *a* is for a pulse that is 80% of the threshold current, *b* is at threshold, and *c* is 20% above threshold. The threshold stimulus pulse in this example has an amplitude I_T of 0.68 mA. Response *a* is similar to that of a linear network with a parallel resistor and capacitor and charged by a brief current pulse. Responses *b* and *c* demonstrate the highly nonlinear response of the excitable membrane. The dashed curves in Fig. 113.5 show the membrane response to a threshold stimulus at the three nodes adjacent to the one nearest the stimulating electrode. The time delay implies a propagation velocity of 43 m/s, which is typical of a 20-μm fiber. The membrane response seen in curves *b* through *f* illustrates the action potential described earlier. The action potential is typically described as an "all-or-nothing" response; that is, its amplitude is not normally graded—either the axon is excited, or it is not.

The threshold current needed for excitation is highly dependent on its duration and waveshape. A common format for representing the response of a nerve is through **strength-duration curves**, i.e., the plot of the threshold of excitation versus the duration of the stimulating current. We can determine the threshold of excitation by "titrating" the stimulus current between a threshold and no-threshold condition.

Figure 113.6 illustrates strength-duration curves derived from the myelinated nerve model described previously under the same conditions applying to Fig. 113.5. Three types of stimulus current apply to Fig. 113.6: a monophasic constant current pulse, a symmetric biphasic rectangular current, and a single cycle of a sine wave. The phase duration indicated on the horizontal axis applies to the initial cathodal half cycle for the two biphasic waves. Stimulus magnitude is given in terms of peak current on the right vertical axis and in terms of the charge in a single monophasic phase of the stimulus on the left vertical axis. The charge is computed by $Q = It_p$ for the rectangular waveforms and $Q = (2/\pi)It_p$ for the sinusoidal waveforms (I is threshold current and t_p is phase duration).

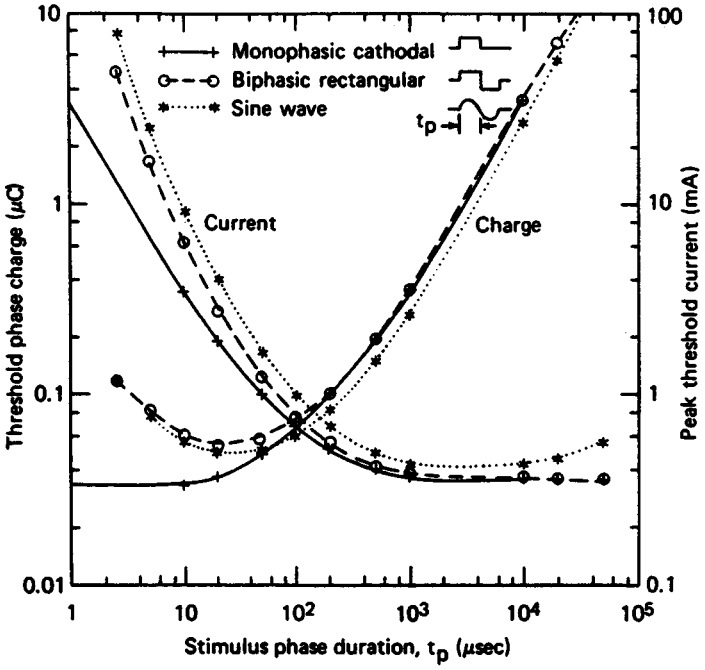


FIGURE 113.6 Strength/duration relationships derived from the myelinated nerve model: current thresholds and charge thresholds for single-pulse monophasic and for single-cycle biphasic stimuli with initial cathodal phase, point electrode 2 mm distant from 20 μm fiber. Threshold current refers to the peak of the stimulus waveform. Charge refers to a single phase for biphasic stimuli. (*Source:* J. P. Reilly, V. T. Freeman, and W. D. Larkin, "Sensory effects of transient electrical stimulation—Evaluation with a neuroelectric model," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 12, pp. 1001–1011, © 1985 IEEE.)

The solid curve labeled "current" is of the type that is most often represented as a strength-duration curve. For this curve, the minimum threshold current occurs for long-stimulus durations and is called the *rheobasic current*, or simply **rheobase**. The duration consistent with twice the rheobase is called the **chronaxie**. The solid curve in Fig. 113.6 labeled "charge" gives the area under the rectangular current pulse. The threshold charge is a minimum for short-duration stimuli.

Mathematical curve fits to the strength-duration curves for monophasic rectangular stimuli are

$$\frac{I_T}{I_o} = \frac{1}{1 - e^{-t/\tau e}} \tag{113.12}$$

and

$$\frac{Q_T}{Q_o} = \frac{t/\tau_e}{1 - e^{-t/\tau e}} \tag{113.13}$$

where I_T is threshold current, Q_T is threshold charge, I_o is the minimum threshold current for long-duration stimuli, Q_o is the minimum threshold charge for short-duration stimuli, and τ_e is an experimentally determined strength-duration time constant. It is readily shown that chronaxie = $\tau_e \ln 2 = 0.693\tau_e$ in this formulation.

Values of I_o and Q_o vary considerably with experimental parameters such as electrode size and location and the size of the neuron. Values of τ_e also vary considerably with experimental conditions: a value around 250 μs is typical for both sensory and motor nerve excitation via cutaneous electrodes, and values around 125 μs are observed for stimulation of axons by small electrodes. Much longer time constants are associated with direct stimulation of muscle cells.

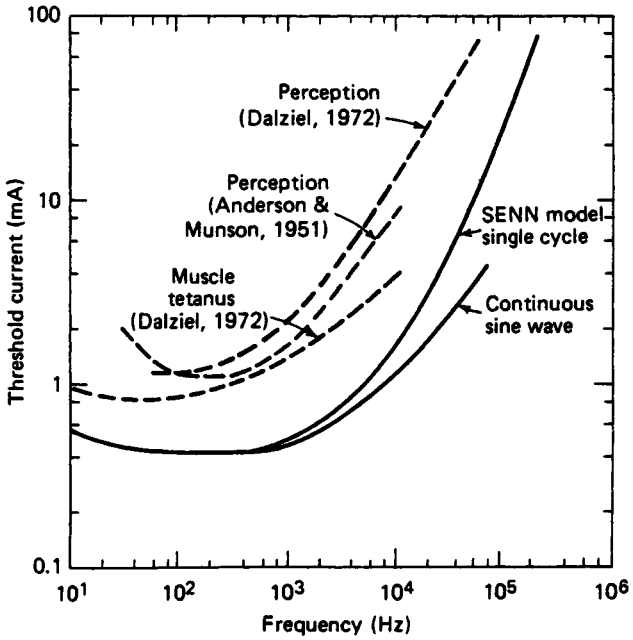


FIGURE 113.7 Strength-frequency curves for sinusoidal current stimuli. Dashed curves are from experimental data. Solid curves apply to myelinated nerve model. Experimental curves have been shifted vertically to facilitate comparisons.

The current reversal of a biphasic stimulus can reverse a developing action potential that was elicited by the initial phase. As a result, a biphasic pulse may have a higher threshold than a monophasic pulse as suggested by the biphasic thresholds in Fig. 113.6. The degree of biphasic threshold elevation is magnified as the stimulus duration is reduced.

A sinusoidal current is a special case of a biphasic stimulus. Sinusoidal threshold response can be represented by strength-frequency curves, as shown by the solid curves in Fig. 113.7 for the myelinated nerve model. Several experimental curves have been included in the figure; these have been shifted vertically to facilitate comparisons. Notice that the myelinated nerve model predicts a lower threshold for stimulation by a continuous sine wave as compared with a single cycle.

The strength-frequency curve follows a U-shaped function, with a minimum at mid frequencies and an upturn at both low and high frequencies. At low frequencies the slow rate of change of the sinusoid prevents the membrane capacitance from building up a depolarizing voltage because membrane capacitance is counteracted by membrane leakage. This process describes the neural property known as *accommodation*, i.e., the adaptation of a nerve to a slowly varying or constant stimulus. The high-frequency upturn occurs because of the canceling effects of a current reversal on the membrane voltage change. An empirical fit to strength-frequency curves is

$$I_t = I_o K_H K_L \tag{113.14}$$

where I_t is the threshold current, I_o is the minimum threshold current, and K_H and K_L are high- and low-frequency terms, defined, respectively, as

$$K_H = \left[1 - \exp\left(-\frac{f_e}{f}\right)\right]^{-a} \tag{113.15}$$

and

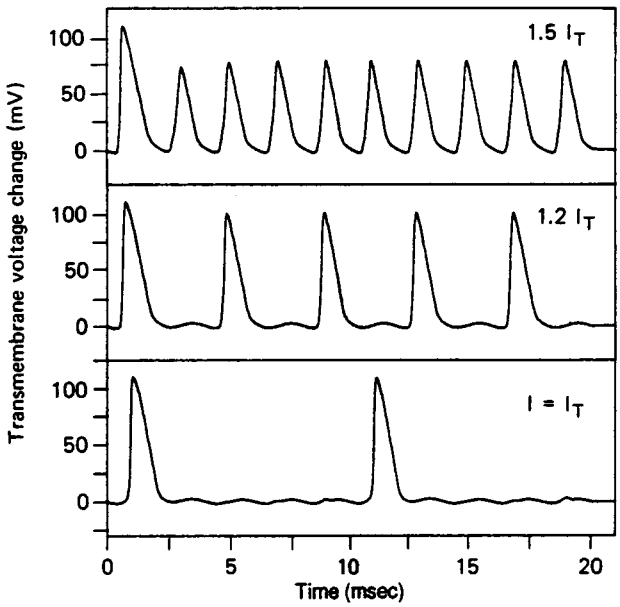


FIGURE 113.8 Model response to continuous sinusoidal stimulation at 500 Hz. The lower panel depicts the response to a stimulus current set at threshold level (I_T) for a single-cycle stimulus. Upper panels show responses for stimulation 20 and 50% above the single-cycle threshold. (*Source:* J. P. Reilly, V. T. Freeman, and W. D. Larkin, "Sensory effects of transient electrical stimulation—Evaluation with a neuroelectric model," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 12, pp. 1001–1011, © 1985 IEEE.)

$$K_L = \left[1 - \exp\left(-\frac{f}{f_o}\right)\right]^{-b} \tag{113.16}$$

where f_e and f_o are constants that determine the points of upturn in the strength-frequency curve at high and low frequencies, respectively. An upper limit of $K_L \leq 4.6$ is assumed for Eq. (113.16) to account for the fact that excitation may be obtained with finite dc currents. An empirical fit of Eqs. (113.15) and (113.16) to the mylinated nerve model thresholds indicates that $a = 1.45$ for a single-cycle stimulus and $a = 0.9$ for a continuous stimulus; $b = 0.8$ regardless of stimulus duration. The value of I_o will depend on various conditions of stimulation, including the size of the electrode, its location on the body, and the location of the stimulated nerve.

With continuous sinusoidal stimulation, it is possible to produce a series of action potentials that are phase-locked to the individual sinusoidal cycles, as noted in Fig. 113.8. This makes the sinusoidal stimulus much more potent than a single pulse of the same phase duration. This potency is a consequence of the fact that perceived magnitude for neurosensory stimulation and muscle tension for neuromuscular stimulation both increase with the rate of action potential production.

Defining Terms

Action potential: A propagating change in the conductivity and potential across a nerve cell's membrane; a nerve impulse in common parlance.

Axon: The conducting portion of a nerve fiber—a roughly tubular structure whose wall is composed of the cellular membrane and which is filled with an ionic medium.

Chronaxie: The minimum duration of a unidirectional square-wave current needed to excite a nerve when the current magnitude is twice rheobase.

Fiber, nerve: A single nerve cell; a neuron—classified on the presence or absence of myelin. Myelinated nerve cells have diameters typically in the range 2 to 20 μm and conduction velocities of 5 to 120 m/s; unmyelinated nerves have diameters from 0.3 to 1.3 μm and conduction velocities of 0.6 to 2.3 m/s. Fiber lengths may be up to 1 m. The term nerve usually refers to a bundle of nerve fibers.

Membrane: The functional boundary of a cell. Nerve cells possess membranes that are excitable by virtue of their nonlinear electrical conductance properties (see Action potential).

Myelinated nerve: A nerve fiber insulated with a fatty substance called myelin and having periodically exposed *nodes of Ranvier.*

Neuron: A nerve cell. Sensory neurons carry information from sensory receptors in the peripheral nervous system to the brain; motor neurons carry information from the brain to the muscles.

Refractory period: A period of time after the initiation of an action potential during which further excitation is impossible (absolute refractory period) or requires a greater stimulus (relative refractory period).

Rheobase: The minimum current necessary to cause nerve excitation—applicable to a long-duration current (e.g., several milliseconds).

Strength-duration curve: A curve expressing the functional relationship between the threshold of excitation of a nerve fiber and the duration of a unidirectional square-wave electrical stimulus.

References

B. Frankenhaeuser and A. F. Huxley, "The action potential in the myelinated nerve fiber of *Xenopus laevis* as computed on the basis of voltage clamp data," *J. Physiol.,* vol. 171, pp. 302–315, 1964.

A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.,* vol. 117, pp. 500–544, 1952.

J. J. B. Jack, D. Noble, and R. W. Tsien, *Electric Current Flow in Excitable Cells*, Oxford: Clarendon Press, 1983.

D. R. McNeal, "Analysis of a model for excitation of myelinated nerve," *IEEE Trans. Biomed. Eng.,* vol. BME-22, pp. 329–337, 1976.

J. P. Reilly, V. T. Freeman, and W. D. Larkin, "Sensory effects of transient electrical stimulation—Evaluation with a neuroelectric model," *IEEE Trans. Biomed. Eng.,* vol. BME-32, no. 12, pp. 1001–1011, 1985.

J. P. Reilly, *Electrical Stimulation and Electropathology*, New York: Cambridge University Press, 1992.

Further Information

For further information, the reader is directed to the references listed at the end of this chapter. Additional references are:

R. Plonsey and R.C. Barr, *Biolectricity—A Quantitive Approach,* New York: Plenum, 1988.

W. Agnew and D. McCreery, *Neural Prostheses,* Englewood Cliffs, N.J.: Prentice-Hall, 1990.

E.R. Kandel, J.H. Schwartz, and T. M. Jessell (Eds.), *Principles of Neural Science,* 3rd ed., New York: Elsevier, 1991.

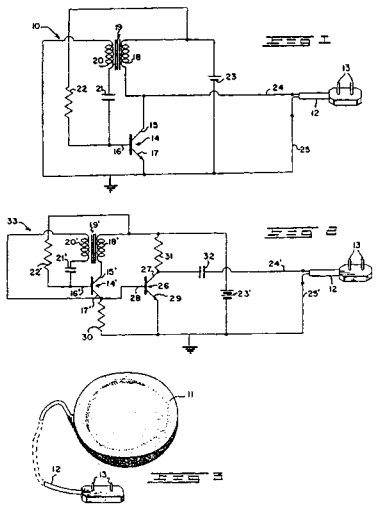
Several journals treat engineering applications of neuroelectric principles, such as *IEEE Transactions on Biomedical Engineering, Medical and Biological Engineering and Computing,* and *Annals of Biomedical Engineering.*

Of the many conferences treating bioelectric responses, one having a broad range of applications is the IEEE Annual Conference on Engineering in Medicine and Biology.

113.2 Bioelectric Events

L. A. Geddes

Bioelectric signals are exploited for the diagnostic information that they contain. Such signals are often used to monitor and guide therapy. Although all living cells exhibit bioelectric phenomena, a small variety produce potential changes that reveal their physiological function. The most familiar bioelectric recordings are the electrocardiogram, ECG (which reflects the excitation and recovery of the whole heart), the electromyogram, EMG (which reflects the activity of skeletal muscle), and the electroencephalogram, EEG (which reflects the activity of the outer layers of the brain, the cortex). The following paragraphs will describe (1) the origin of



Medical Cardiac Pacemaker

Wilson Greatbatch
Patented October 9, 1962
#3,057,356

An excerpt from Greatbatch's patent application:

The primary object of this invention is to provide an improved artificial cardiac pacemaker for restoring satisfactory heart rhythm to a heart which is functioning inadequately due to conduction defects in the auricular-ventricular bundle.

Another object of this invention is to provide an artificial cardiac pacemaker requiring low power consumption, so that battery operation is feasible for long uninterrupted periods without battery replacement.

Another object of this invention is to provide an artificial cardiac pacemaker which may be directly connected to the surface of the ventricle of the heart.

A still further object of this invention is to provide an artificial cardiac pacemaker which is constructed from materials compatible to the body environment and is of such an electrical and mechanical configuration, that permanent implantation of the device within the human body is both feasible and practical.

Greatbatch's pacemaker was the first to be compact enough and use such low power that it could be implanted within the body and run for five years before requiring battery replacement. Wilson Greatbatch, Inc. is a leading producer of pacemaker batteries and other medical products. (Copyright © 1995, DewRay Products, Inc. Used with permission.)

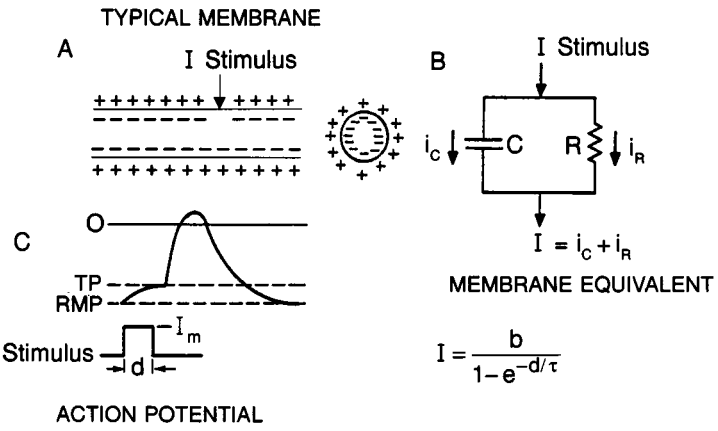


FIGURE 113.9 (A) Typical charged membrane, (B) its equivalent circuit, and (C) action potential resulting from a stimulus *I* of duration *d*.

all bioelectric phenomena; (2) the nature of the electrical activity of the heart, skeletal muscle, and the brain; and (3) the characteristics of instrumentation used to display these events.

Origin of Bioelectricity

Cell membranes resemble charged capacitors operating near the dielectric breakdown voltage. Assuming a typical value of 90 mV for the transmembrane potential and a membrane thickness of 100 Å, the voltage gradient across the membrane is 0.9×10^5 V/cm. A typical value for the capacitance is about 1 μF/cm^2.

The transmembrane charge is the result of a **metabolic process** that creates ionic gradients with a high concentration of potassium ions (K^+) inside and a high concentration of sodium ions (Na^+) outside. There are concentration gradients for other ions, the cell wall being a semipermeable membrane that obeys the Nernst equation (60 mV/decade concentration gradient for univalent ions). The result of the ionic gradient is the transmembrane potential that, in the cells referred to earlier, is about 90 mV, the interior being negative with respect to the exterior. Figure 113.9 illustrates this concept for a cylindrical cell.

The transmembrane potential is stable in inexcitable cells, such as the red blood cell. However, in excitable cells, a reduction in transmembrane potential (either physiological or induced electrically) results in excitation, characterized by a transmembrane ion flux, resulting from a membrane permeability change. When the transmembrane potential is reduced by about one-third, Na^+ ions rush in; K^+ ions exit slightly later while the cell depolarizes, reverse polarizes, then repolarizes. The resulting excursion in transmembrane potential is a propagated action potential that is characteristic for each type of cell. In Fig. 113.10 are shown the action potentials of (A) a single cardiac ventricular muscle cell, (C) a skeletal muscle cell, and (E) a nerve cell. In (B) and (D), the ensuing muscular contractions are shown. An important property of the action potential is that it is propagated without decrement over the entire surface of the cell, the depolarized region being the stimulus for adjacent polarized regions. In contractile cells it is the action potential that triggers release of mechanical energy as shown in Figs. 113.10(B) and (D).

Law of Stimulation

Although action potentials are generated physiologically, it should be obvious that excitable cells can be made to respond by the application of a negative pulse of sufficient current density (I) and duration (d) to reduce the transmembrane potential to a critical value by removing charge, thereby reducing the membrane potential to the threshold potential (TP), as shown in Fig. 113.9. The law of stimulation is $I = b/(1 - e^{-d}/\tau)$, where b is the threshold current density for an infinitely long-duration pulse and τ is the cell membrane time constant, being different for each type of excitable tissue. Figure 113.11 is a plot of the threshold current (I) versus duration (d) for mammalian cardiac muscle, sensory receptors, and motor nerve. This relationship is known as the *strength-duration curve.*

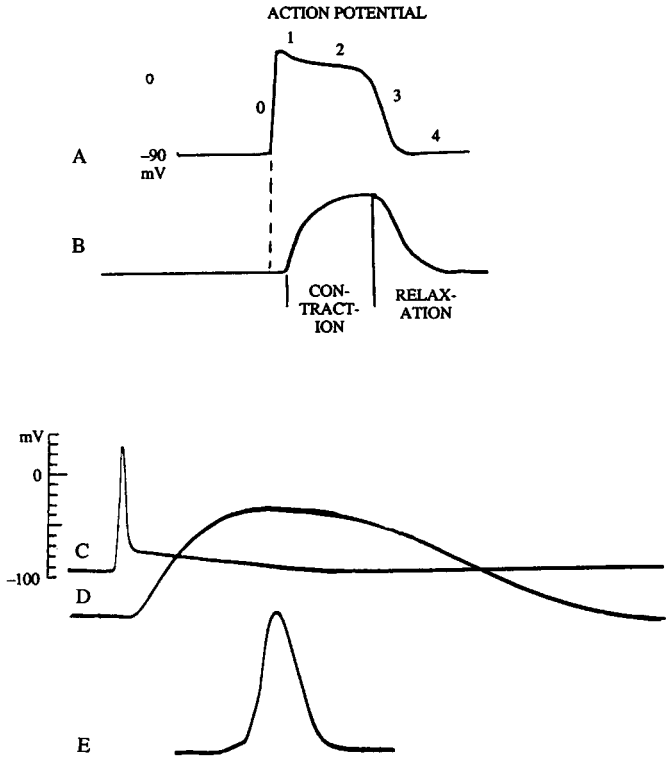


FIGURE 113.10 The action of (A) cardiac muscle and (B) its contraction, (C) skeletal muscle and (D) and its contraction. The action potential of nerve is shown in (E).

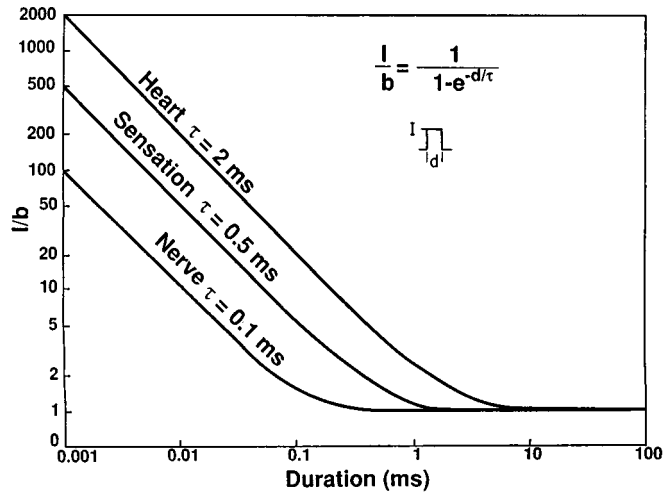


FIGURE 113.11 The strength-duration curve for heart, sensory receptors, and motor nerve. I is the stimulus current, b is the rheobasic current, and τ is the membrane time constant. The stimulus duration is d.

Recording Action Potentials

Action potentials of single excitable cells are recorded with transmembrane electrodes (micron diameter) only in research studies. When action potentials are used for diagnostic purposes, extracellular electrodes are used that are both large and distant from the population of cells which become active and recover. The depolarization

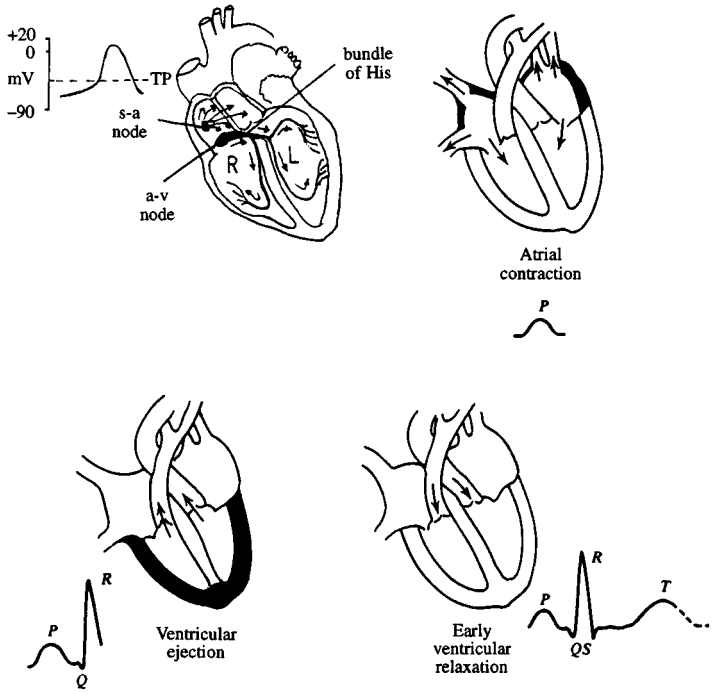


FIGURE 113.12 Genesis of the ECG. The SA node is the pacemaker, setting the rate. Excitation is propagated from the atria to the AV node, then to the bundle of His, and to the ventricular muscle via the Purkinje fibers. The SA node has a decreasing membrane potential that reaches the threshold potential (TP), resulting in spontaneous excitation (inset).

and repolarization processes send small currents through the conducting environmental tissues and fluids, resulting in a time-varying potential field. Appropriately placed electrodes allow recording the electrical activity of the bioelectric generators. However, the waveforms of such recordings are vastly different from those of the transmembrane action potentials shown in Fig. 113.10. By using cable theory, it is possible to show that such extracellular recordings resemble the second derivative of the excursion in transmembrane potential [Geddes and Baker, 1989]. Despite the difference in waveform, extracellular recordings identify the excitation and recovery processes very well.

The Electrocardiogram (ECG)

Origin

The heart is two double-muscular pumps. The atria pump blood into the ventricles, then the two ventricles contract. The right ventricle pumps venous blood into the lungs, and the left ventricle pumps oxygen-rich blood into the aorta. Figure 113.12 is a sketch of the heart and great vessels, along with genesis of the ECG.

The ECG consists of two parts: the electrical activity of the atria and that of the ventricles. Both components have an excitation wave and a recovery wave. Within the right atrium is a specialized node of modified cardiac muscle, the sinoatrial (SA) node, that has a spontaneously decreasing transmembrane potential which reaches the threshold potential (TP), resulting in self-excitation (Fig. 113.12, upper left). Therefore the SA node is the cardiac pacemaker, establishing the heart rate. The SA node action potential stimulates the adjacent atrial muscle, completely exciting it and giving rise to the first event in the cardiac cycle, the P wave, the trigger for atrial contraction. Atrial excitation is propagated to another specialized node of tissue in the base of the ventricles, the atrioventricular (AV) node, the bundle of His and the Purkinje fibers. Propagation of excitation over the ventricles gives rise to the QRS, or simply the R wave, which triggers ventricular contraction. Meanwhile during the QRS wave, the atria recover, giving rise to the T_p wave, following which the atria relax. The T_p wave

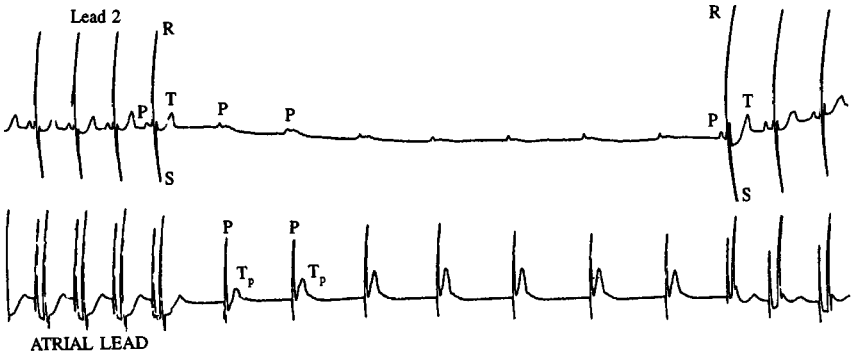


FIGURE 113.13 Lead 2 ECG and an atrial lead. In the center of the record AV block was produced, showing the P waves in lead 2 and the P and T_p waves in the atrial lead.

is not ordinarily seen in the ECG because it is obscured by the ventricular QRS wave. During the QRS wave the ventricles contract, then relax following their recovery potential, the T wave; Fig. 113.12 summarizes this sequence.

Ordinarily the T_p wave is not visible. However, if the propagation of excitation from the atria to the ventricles is blocked, the T_p wave can be seen. Figure 113.13 is a record of the ECG from a limb lead and a recording from a lead within the right atrium in a subject with transient AV block. Note that the sharp P wave in the atrial lead coincides with the P wave in the limb recording and that the atrial lead shows both P and T_p waves, easily identified during AV block.

Clinical Significance

From the foregoing it can be seen that the ECG is only a timing signal; there is no dynamic information in its amplitude. Nonetheless, by observing the orderly P-QRS-T sequence it is possible to determine if the excitatory and recovery processes in the heart are functioning normally.

Disturbances in the orderly timing of the cardiac cycle are elegantly displayed by the ECG. For example, each atrial excitation may not be delivered to the AV node. AV block exists when there is less than a 1/1 correspondence between the P and QRS complexes (Fig. 113.13).

Figure 113.14(1) shows a normal ECG and Fig. 113.14(2) illustrates a 2/1 AV block with two P waves for each QRS-T complex. Complete AV block exists when none of the atrial excitations reach the AV node, as shown in Fig. 113.14(3). In this case the ventricles developed their own rhythm, which was slow; cardiac output is low, and in such a situation an artificial pacemaker must be implanted.

For many reasons, the atria develop a rapid rate called *atrial tachycardia* or *supraventricular tachycardia.* A very rapid atrial rate is designated *atrial flutter* [Fig. 113.14(4)]. With both atrial tachycardia and flutter, the atrial contractions are coordinated, although the ventricular pumping capability is reduced owing to inadequate filling time. The ventricles are driven at a rapid rate, and cardiac output is low.

Atrial fibrillation is an electrical dysrhythmia in which all the atrial muscle fibers are contracting and relaxing asynchronously and there is no atrial pumping. This dysrhythmia [Fig. 113.14(5)] causes the ventricles to be excited at a very rapid and irregular rate. Cardiac output is reduced, and the pulse is rapid and irregular in force and rate.

If the propagation of excitation in the ventricles is impaired by damage to the bundle of His, the coordination of excitation and contraction is impaired and reveals itself by a widening of the QRS wave, and often a notch is present; Fig. 113.14(6) illustrates right (RBBB) and left (LBBB) bundle-branch block. These waveforms are best identified in the chest (V) leads.

All parts of the heart are capable of exhibiting rhythmic activity, there being a rhythmicity hierarchy from the SA node to the ventricular muscle. In abnormal circumstances the atria and ventricles can generate spontaneous beats. Such ectopic excitations do not ordinarily propagate normally, and therefore the ECG waveforms are different. Figure 113.14(7) illustrates ventricular **ectopic beats** in which the first (l) Q-S,T wave arose at the apex and the second (2) R-S,T wave arose at the base of the ventricles. The coupled (bigeminus)

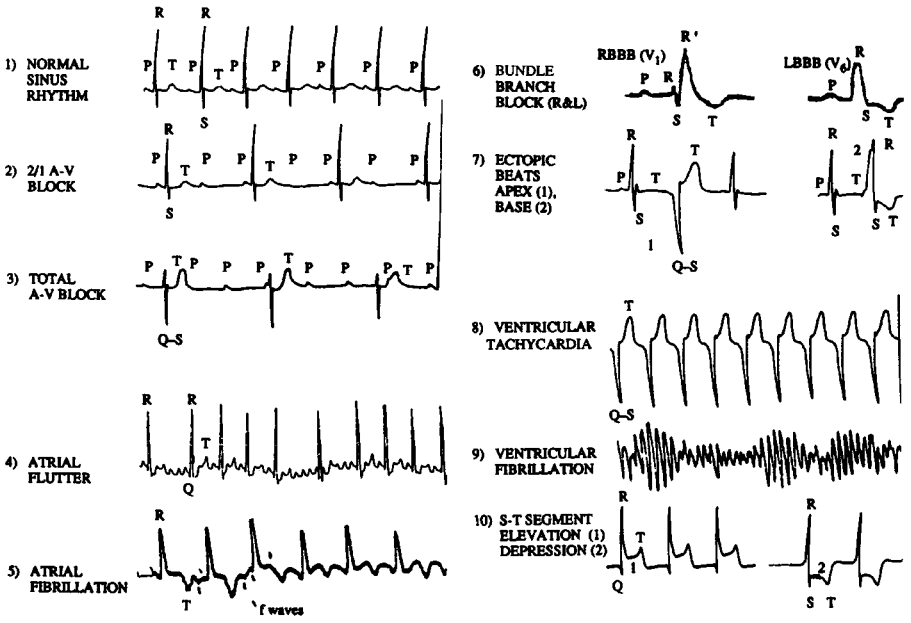


FIGURE 113.14 ECG waveforms.

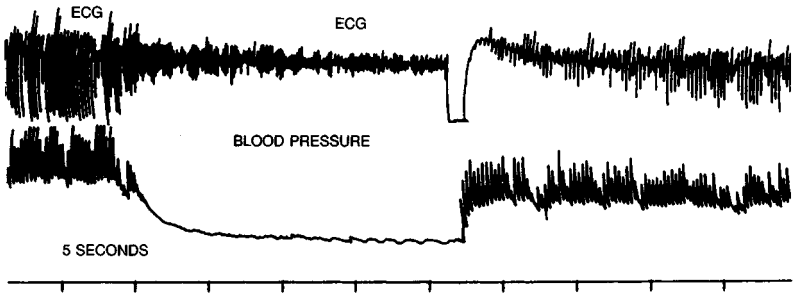


FIGURE 113.15 The electrocardiogram (ECG) and blood pressure during ventricular tachycardia (left), which progressed to ventricular fibrillation (center). A strong transchest shock was applied to defibrillate the ventricles that resumed pumping with the tachycardia returning (right).

beat usually produces no arterial pulse because of inadequate time for filling of the ventricles and poor coordination of the contraction.

The ventricles may become so excitable that they develop a rapid rhythm called *ventricular tachycardia,* as shown in Fig 113.14(8). In this situation the pumping capability is diminished owing to the high rate that impairs filling and to impaired coordination of contraction. Ventricular fibrillation is a condition in which all of the ventricular muscle fibers contract and relax independently and asynchronously. Pumping ceases and cardiac output falls to zero. The ECG [Fig. 113.14(9)] exhibits rapid oscillations of waxing and waning amplitude at a rate of 800 to 1500 per minute. Ventricular fibrillation is lethal unless the circulation is restored within a few minutes, first by cardiopulmonary resuscitation (CPR) and then by electrical defibrillation. The latter technique employs the delivery of a substantial pulse of current through the heart applied directly or with transchest electrodes. Figure 113.15 illustrates ventricular tachycardia (left), ventricular fibrillation (center), and defibrillation (right), with the restoration of pumping.

When a region of the ventricles is deprived of its coronary artery blood supply, the cells in this region lose their ability to generate action potentials and to contract. These cells remain depolarized while they are dying and do not contribute to genesis of the QRS-T complex. Instead, there appears a shift in the portion of the

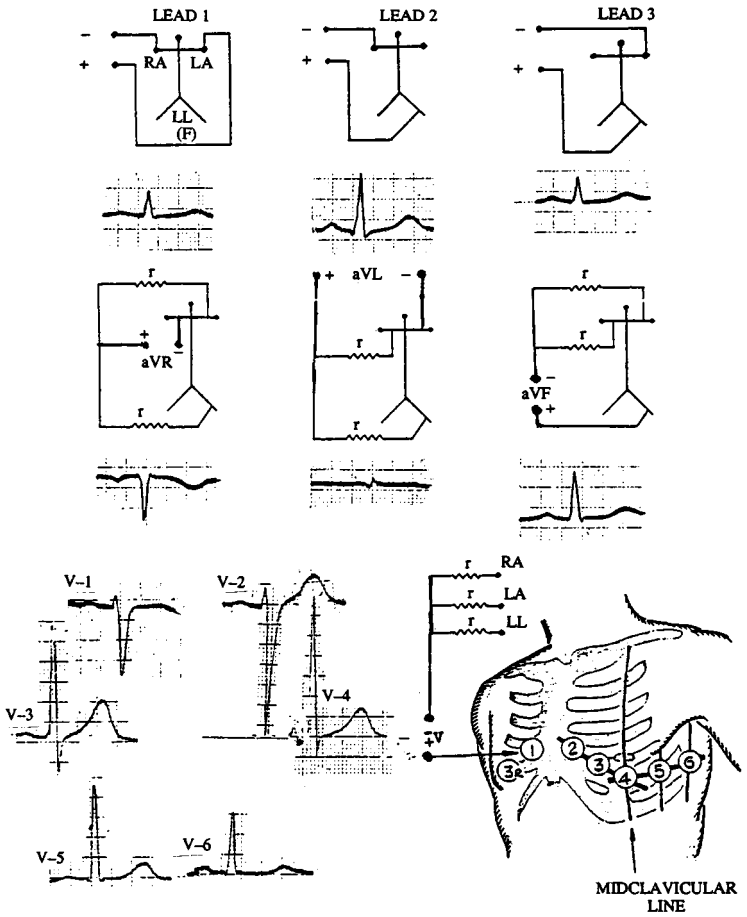


FIGURE 113.16 The limb and chest (V) leads.

ECG between the S and T waves, i.e., there is an S-T segment shift. This is the cardinal sign of a **myocardial infarction** (heart attack) and is almost always accompanied by chest pain (angina pectoris). Figure 113.14(10) illustrates the ECG in myocardial infarction. Whether the S-T segment displacement is up (1) or down (2) depends on the region of the ventricles injured, as well as the lead used to record the ECG.

ECG Leads

The spread of excitation and recovery over the atria and ventricles varies in direction with time. Therefore, excitation and recovery are vectors, and the location of body-surface electrodes is important. For this reason, standard electrode sites have been adopted, as shown in Fig. 113.16. There are three standard limb leads, three augmented (a) limb leads, and six chest (V) leads, the latter being monopolar. The reference for the monopolar chest leads is the centerpoint of three resistors (r), each joined to one of the limb electrodes. The right leg is used to ground the subject. Each lead "sees" a different region of the heart. The use of so many leads allows quick and easy identification of the direction of propagation of excitation (and recovery) by merely inspecting the amplitudes of the waveforms in the various leads. If excitation (or recovery) travels orthogonal to a lead axis, the net amplitude will be zero or very small. If excitation (or recovery) travels parallel to a lead axis, the amplitude of the wave will be maximum. Figure 113.16 illustrates the amplitudes of the P, QRS, and T waves for the 12 ECG leads. Note that leads 1, 2, 3, aVR, aVL, and aVF identify the vector projections in the frontal plane. Leads V_{1-6} identify the vector components in a quasi-horizontal plane. There are normal values for the amplitudes and durations for the P, QRS, and T waves as well as their vectors. The interested reader can find more on ECG in the many handbooks on this subject. Two good, recently published texts are by Chou [1991] and Phillips and Feeney [1990].

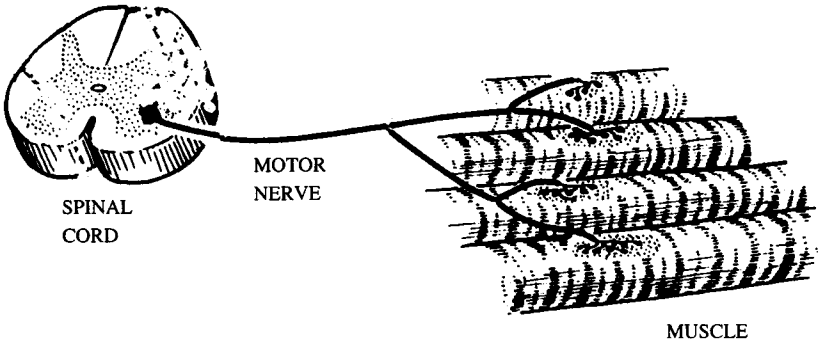


FIGURE 113.17 The functional unit of the muscular system, the motor unit, consisting of a nerve cell located within the spinal cord, its axon, and the muscle fibers that it innervates.

Instrumentation

Standards of performance evolved from recommendations by the American Medical Association in 1950 and the American Heart Association in 1954 and 1967. These recommendations have been collected and expanded into an American National Standard, published by the Association for the Advancement of Medical Instrumentation (AAMI) [1991]. The title of the document is "Diagnostic Electrocardiographic Devices." This document not only lists all the performance and labeling requirements, but also provides useful information on testing ECGs and should be consulted by those contemplating construction of an ECG. Only some of the highlights of the standard will be presented here.

The ECG is displayed by a direct-writing pen that employs a heated stylus writing on thermosensitive paper. Two chart speeds are used, 25 and 50 mm/s. The rulings on the paper represent 40 ms when the standard speed (25 mm/s) is used. The amplitude sensitivity is 10 mm for a 1-mV input signal. The sinusoidal frequency response extends from 0.05 to 100 Hz for the 30% attenuation points. The input stage is a differential amplifier with an input impedance in excess of 2.4 MΩ The common-mode rejection ratio (CMRR) is measured with a 20-V (rms) 60-Hz generator with an output impedance of 51,000 Ω connected in series with a 10-pF capacitor. The 60-Hz CMRR should be in excess of 5000. The maximum dc leakage current through any patient electrode is 0.2 μA.

Electromyography (EMG)

The electrical activity of skeletal muscle is monitored to assess the integrity of the motor nerve that supplies it and to evaluate recovery of the motor nerve following injury to it. The EMG is also characteristically altered in many degenerative muscle diseases. Although muscle action potentials can be detected with skin-surface electrodes, a monopolar or bipolar needle electrode is used in clinical EMG. The electrical activity is displayed on an oscilloscope screen and monitored aurally with a loudspeaker.

Contraction of Skeletal Muscle

The functional unit of the muscular system is the motor unit, consisting of a nerve cell located within the spinal cord, its axon (nerve fiber), and the group of muscle fibers that it innervates, as shown in Fig. 113.17. Between the nerve fiber and the muscle fibers is the myoneural junction, the site where acetylcholine is liberated and transmits excitation to the muscle fibers. The number of muscle fibers per nerve fiber is called the innervation ratio, which ranges from 1:1 to about 1000:1; the former ratio is characteristic of the extraocular muscles, and the latter is typical for the postural muscles.

A single stimulus received by the nerve fiber physiologically, or a single stimulus delivered to it electrically, will cause all the innervated muscle fibers to contract and relax; this response is called a *twitch*. Figure 113.10(C) and (D) illustrates the relationship between the muscle action potential and twitch. Note that the action potential is almost over before contraction begins and the contraction far outlasts the duration of the action potential.

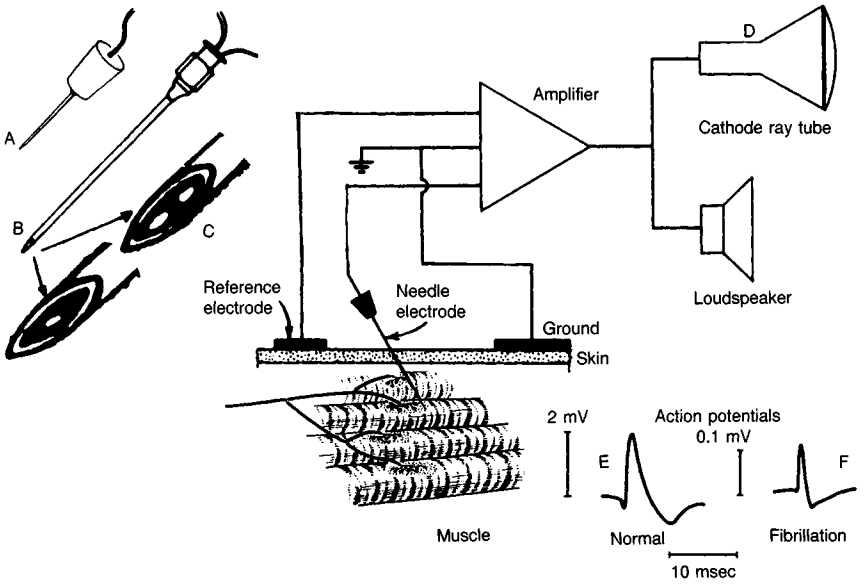


FIGURE 113.18 Equipment used for electromyography. (A) Needle electrode, (B) hypodermic monopolar and (C) bipolar electrodes, (D) the recording apparatus, (E) skeletal muscle action potential, and (F) fibrillation potential.

If multiple stimuli are delivered to a single motor-nerve fiber with an increasing frequency, the twitches fuse into a sustained (tetanic) contraction whose force is much more than that of a twitch. This occurs because each action potential liberates contractile energy. The critical fusion frequency depends on the type of muscle, but in general it is about 25 to 40 per second.

The force developed by a whole muscle consisting of thousands of motor units is graded in two ways: (l) by the frequency of nerve impulses in each nerve fiber and (2) by the number of motor units that are activated.

Clinical EMG

When the electrical activity of skeletal muscle is examined for diagnostic purposes, an insulated needle electrode, bare only at the tip [Fig. 113.18(A)], is inserted into the muscle and paired with a skin-surface electrode. Another skin-surface electrode is used to ground the subject. Occasionally a coaxial needle electrode [Fig. 113.18(B)] or a bipolar hypodermic needle electrode [Fig. 113.18(C)] is used. In the latter case the outer sleeve is used as the ground. A high-gain, differential amplifier, oscilloscope, and loudspeaker are used, as shown in Figure 113.18(D).

In a normal subject at rest, the electrical activity monitored during insertion of the needle electrode consists of a short burst of muscle action potentials displayed on the oscilloscope and heard in the loudspeaker. These action potentials are called *insertion* potentials and subside quickly in the normal muscle. When the muscle is at rest, there is no electrical activity (electrical silence). If the muscle is contracted voluntarily, the frequency of action potentials increases with the force developed by the muscle. However, there is no linear or constant relationship between these two events. Each action potential, called a *normal motor-unit* potential, lasts a few milliseconds to the first zero crossing, as shown in Fig. 113.18(E).

There is considerable art associated with placing the exploring electrode. If the electrode tip is not adjacent to contracting muscle fibers, the sound of the action potential is muffled and electrode adjustment is required. The same is true for detecting fibrillation potentials (see below).

If the nerve cell in the spinal cord or the nerve fiber supplying a muscle is damaged, the muscle cannot be contracted voluntarily or reflexly (by tapping the tendon) and is therefore paralyzed. In the absence of therapeutic intervention and with the passage of time, the nerve beyond the damaged site dies and the muscle fibers start to degenerate. In about 2 1/2 to 3 weeks in humans, the individual muscle fibers start to contract and relax spontaneously and randomly, producing short-duration, randomly occurring action potentials called

fibrillation potentials [Fig. 113.18(F)], which are displayed on the oscilloscope screen and heard as clicks in the loudspeaker. Although there is electrical activity, the muscle develops no net force. The fibrillation potentials persist as long as there are viable muscle fibers. In such a denervated muscle, insertion of the needle electrode elicits a vigorous train of short-duration insertion potentials that resemble fibrillation potentials with a frequency of about 1 to 10 per second. If the damaged ends of the nerve are brought together surgically, the central end of the nerve will start to grow slowly and reinnervate the muscle. Gradually the fibrillation potentials disappear, although the muscle is still not able to be contracted. Long before there is visible evidence of muscle contraction, if the subject is presented with the EMG display and asked to contract the affected muscle, primitive muscle action potentials can be elicited. With the passage of time, the fibrillation potentials disappear and there is electrical silence at rest and primitive (nascent) motor-unit activity occurs with voluntary contraction. Later when reinnervation is complete, only normal motor-unit potentials are present with voluntary contraction and electrical silence at rest.

The EMG is also used to diagnose some degenerative muscle and related nerve disorders. *Myotonia* is a degenerative disease of muscle fibers in which the muscle relaxes poorly. Insertion of the needle electrode elicits an intense burst of insertion potentials that sound like a thunderstorm in the loudspeaker. A similar response is obtained by tapping the muscle. When relaxation does occur, there is electrical silence. Voluntary contraction produces normal action potentials along with shorter-duration action potentials from the diseased muscle fibers.

Myasthenia gravis is a disease in which there is impairment of transmission of acetylcholine across the myoneural junctions to the muscle fibers. As a result, muscle contraction cannot be sustained. Because the muscle fibers are normally innervated, there are no fibrillation potentials. With voluntary contraction, normal action potentials occur, and if the disease is severe, the action potentials decrease in frequency as the force of contraction decreases and soon sustained muscle contraction cannot be maintained.

Muscular dystrophy is a degenerative disease of muscle fibers in which there is **atrophy** of some fibers, swelling in others, and an increase in sarcolemmal and connective tissue with the deposition of fat. Insertion of the needle electrode elicits a vigorous burst of short-duration, high-frequency action potentials. Typically at rest there are no fibrillation potentials. With voluntary contraction, the action potentials are short in duration, high in frequency, and produce a whirring sound in the loudspeaker. As fatigue supervenes, the frequency and amplitude decrease.

The reader who is interested in obtaining more information on EMG will find it in books by Cohen and Brumlik [1969] and Marinacci [1955]. Both contain a wealth of clinical information.

Instrumentation

As yet there is no American National Standard for EMG, although steps are being taken in this direction. As shown in Fig. 113.18, the EMG is displayed in two ways: (1) visually with an oscilloscope and (2) aurally with a loudspeaker. Both are needed to enable acquisition and analysis of the EMG.

Buchtal et al. [1954] stated that the principal frequency components for the human EMG require a bandwidth of 1 Hz to 10 kHz. It has been found that a time constant of about 50 ms is satisfactory, which corresponds to a low-frequency –3-db point of 3 Hz. For needle electrodes with a tip diameter of 0.1 mm or larger, the input impedance (one side to ground) should not be lower than that of a 500-kΩ resistor in parallel with less than 25-pF capacitance.

Smaller-area electrodes require a higher input impedance [Geddes et al., 1967]. The cable used to connect the needle electrode to the amplifier should not add more than 250 pF to the input capacitance. The common-mode rejection ratio (CMRR) should be in excess of 5000.

Electroencephalography (EEG)

The electrical activity of the brain can be recorded with electrodes on the scalp, on the exposed brain, or inserted into the brain. The latter method is used in research studies. When recordings are made with brain-surface (cortex) electrodes, the recording is called an electrocorticogram (ECoG). With scalp electrodes, the recording is designated an electroencephalogram (EEG) that is displayed by direct-inking pens using a chart speed of 3 cm/s. Typically 8 to 12 channels are recorded simultaneously.

Although the brain consists of about 10^{14} neurons, the EEG reflects the electrical activity of the outer layer, the cortex, which is the seat of consciousness. The type of electrical activity depends on the location of the electrodes and the level of alertness. The frequency and amplitude are profoundly affected by alertness, drowsiness, sleep, hyperventilation, anesthesia, the presence of a tumor, head injury, and epilepsy. The clinical correlation between cerebral disorders and the voltage and frequency spectra is well ahead of the physiological explanations for the waveforms.

Recording Technique

Both bipolar [Fig. 113.19(A)] and monopolar [Fig. 113.19(B)] techniques are used. With monopolar recording, one side of each amplifier is connected to a reference electrode, usually on the earlobe. With bipolar recording, the amplifiers are connected between pairs of scalp electrodes in a regular order. With both types of recording, one-half the number of channels is connected to electrodes on the opposite side of the head. In this way, the electrical activity from homologous areas of the brain can be compared at a glance.

With the bipolar method illustrated in Fig. 113.19(A), abnormal activity located under electrode X will be revealed as a phase reversal in adjacent channels. With monopolar recording using the earlobe reference electrode [Fig. 113.19(B)] the abnormal activity under electrode X will be largest in the channel connected to that electrode and smaller in the adjacent channels.

In clinical EEG, 21 electrodes are applied to the scalp in what is known as the 10-20 system. This array was established by the International Federation of EEG Societies in 1958. The 10-20 system employs skull landmarks as reference points to locate the electrodes.

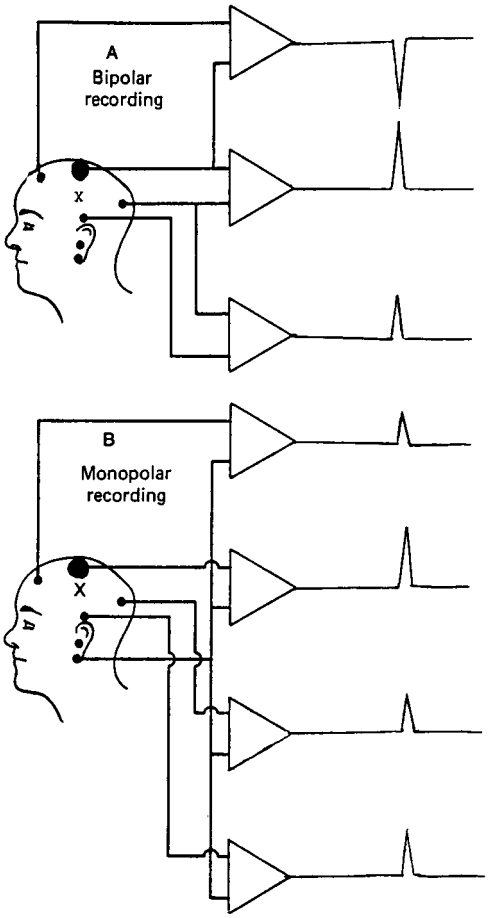


FIGURE 113.19 Methods of recording the EEG. (A) The bipolar and (B) the monopolar method. Note how abnormal activity under electrode X is revealed by the two techniques.

The Normal EEG

In the normal resting adult, the EEG displays a fluctuating electrical activity having a dominant frequency of about 10 Hz and an amplitude in the range of 20 to 200 μV. This activity is called the *alpha rhythm* and ranges in frequency from about 8 to 12 Hz, being most prominent in the **occipital** and **parietal** areas. It may occupy as much as half the record. The alpha rhythm increases in frequency with age from birth and attains its adult form by about 15 to 20 years. The alpha rhythm is most prominent when the eyes are closed and in the absence of concentration. Opening the eyes, engaging in patterned vision, or performing such cerebral activity as mental arithmetic diminishes or abolishes the alpha rhythm. Figure 113.20 presents a good example of this phenomenon.

Although the alpha rhythm is the most prominent electrical activity, other frequencies are present. For example, there is a considerable amount of low-voltage, high-frequency (beta) activity ranging from 18 to 30 Hz. It is usually found in the frontal part of the brain. However, the normal electroencephalogram contains waves of various frequencies (in the range of 1 to 60 Hz) and amplitudes, depending on the cerebral state. To establish communication among electroencephalographers, a terminology has been developed to describe waveforms and their frequencies; Table 113.1 presents a glossary of these terms.

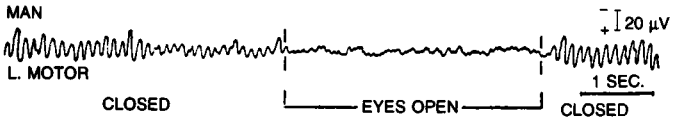


FIGURE 113.20 The EEG of a relaxed human subject with eyes closed and open. Note that the record is dominated with alpha rhythm (8–12 Hz) when the eyes are closed. (*Source:* Derived in part from M.A.B. Brazier, *The Electrical Activity of the Nervous System,* London: Sir Isaac Pitman & Sons, Ltd., 1951. With permission.)

TABLE 113.1 EEG Waveform Terminology

Waveform	Frequency, Hz	Conditions
Alpha	8–12	Parietal-occipital, associated with the awake and relaxed subject, prominent with eyes closed.
Beta	18–30	More evident in frontal-parietal leads, seen best when alpha is blocked.
Delta	1–3.5	Associated with normal sleep and present in children less than 1 year old, also seen in organic brain disease.
Theta	4–7	Parietal-temporal, prominent in children 2 to 5 years old.
Sleep spindle (sigma)	12–14	Waxing and waning of a sinusoidal-like wave having the envelope that resembles a spindle, seen during sleep.
Lambda	Transient	Visually evoked, low-amplitude, occipital wave, resulting from recognition of a new visual image.
Spike and wave	ca. 3	Sharp wave (spike) followed by rounded wave associated with petit mal epilepsy.

Drowsiness and sleep affect the normal EEG profoundly. Figure 113.21 illustrates the typical changes that occur as a subject goes to sleep. With drowsiness, the higher-frequency activity which is associated with alertness or excitement and the alpha rhythm that dominates the waking record in the relaxed state are replaced by a characteristic cyclic sequence of changes which constitute the focus of a new specialty devoted to sleep physiology, in which the EEG is used to identify different stages of sleep.

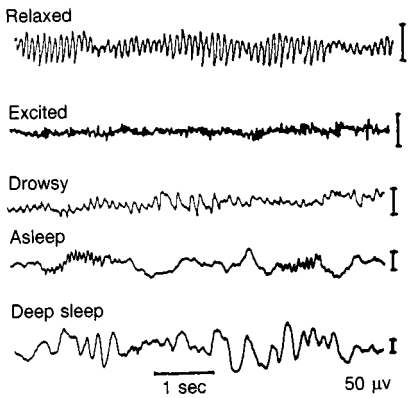


FIGURE 113.21 The EEG of a subject going to sleep.

Rapid, deep breathing (hyperventilation) at a rate of 30 per minute for about 3 min reduces the partial pressure of carbon dioxide in the blood which reduces cerebral blood flow. A typical EEG response consists of large-amplitude, bilaterally synchronous, frontally prominent waves with a frequency of 4 to 7 per second. The frequency usually decreases with increasing hyperventilation. The lack of bilateral symmetry is an indication of abnormality.

Anesthesia dramatically alters the EEG in a manner that depends on the type and amount of anesthetic given. Despite differences among anesthetic agents, some important similarities accompany anesthesia. The first change is replacement of the alpha rhythm with low-voltage high-frequency activity that accompanies the analgesia and delirium stages. Thus the EEG resembles that of an alert or excited subject, although the subject is not appropriately responsive to stimuli; usually the response is excessive and/or inappropriate. From this point on, the type of EEG obtained with deepening anesthesia depends on the type of anesthetic. However, when a deeper level of anesthesia is reached, the EEG waveform becomes less dependent on the type of anesthetic. Large-amplitude low-frequency waves begin to dominate the record, and with deepening anesthesia their frequency is reduced and they begin to occur intermittently. With very (dangerously) deep anesthesia, the record is flat (i.e., isoelectric). Complicating interpretation of the EEG in anesthesia are the effects of **hypoxia, hypercapnia**, and hypoglycemia, all of which mimic deep anesthesia.

Clinical EEG

The EEG plays a valuable role in identifying intracranial pathology. The clinical utility relies on recognition of patterns of frequency, voltage, and waveform. Localization of abnormal areas is provided by the multiple scalp electrodes and recording channels.

The EEG has its greatest value as an aid in the diagnosis and differentiation of the many types of epilepsy, a condition in which groups of neurons in the brain become hyperexcitable and, depending on their location, produce sensory, motor, and/or **autonomic** manifestations. The epilepsies associated with cortical lesions are often detected by the scalp EEG. The EEG in epileptics is usually abnormal between, as well as during, attacks. The EEG provides information on the location of the area (or areas) of abnormal neuronal activity.

Petit mal epilepsy is characterized by a transient loss (few to 20 s) of conscious thought, although motor activity may continue. Often there are eye movements and blinking. The EEG shows a characteristic 3 per second spike-and-wave pattern [Fig. 113.22(A)]. Psychomotor epilepsy is characterized by sensory hallucinations and abnormal thoughts, often with stereotyped behavior. During the attack, the subject is stuporous and the EEG [Fig. 113.22(B)] has a characteristic pattern. Jacksonian, or motor, epilepsy starts in a specific area of the motor cortex and is preceded by an aura, a characteristic sensation perceived by the subject. The convulsion starts with localized muscle twitching that often starts in the face, hand, arm, then spreads over the entire body as a generalized convulsion; Fig. 113.22(C) shows the onset of a convulsion. Consciousness is lost during and for a short time after the fit. The EEG provides information on the origin of the abnormal discharge in the motor cortex. Grand mal epilepsy is characterized by a contraction all the muscles (tonic phase), then jerking (clonic phase). Consciousness is lost, and the subject is in a coma for some time following the attack. The EEG [Fig. 113.22(D)] shows high-voltage, high-frequency waves that progress over the entire cortex.

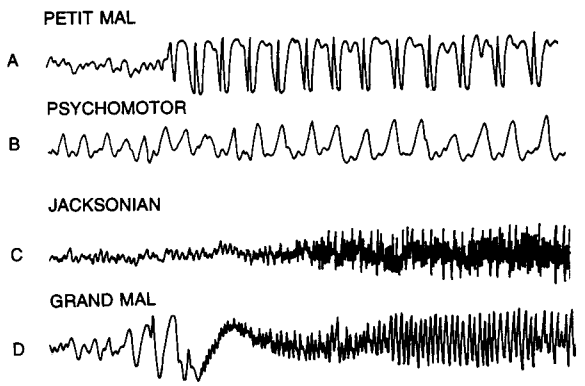


FIGURE 113.22 EEG waveforms in epilepsy: (A) petit mal, (B) psychomotor, (C) Jacksonian, and (D) grand mal.

Traumatic epilepsy results from injury to the brain. It is believed that contraction of scar tissue acts as a stimulus to adjacent nerve cells which discharge rhythmically, the excitation spreading to a grand mal convulsion. The EEG provides information on the origin of the abnormal discharge.

Tumors are associated with low-frequency (delta) waves. However, other intracranial lesions also produce slow waves. Although the EEG can identify the location of tumors, usually it cannot differentiate between brain injury, infection, and vascular accident, all of which produce low-frequency waves. Interpretation of the EEG always includes other clinical information.

For those wishing to delve deeper into EEG, additional information can be found in most textbooks of medical physiology. The three-volume *Atlas of EEG,* authored by Gibbs and Gibbs [1952], contains a wealth of information on EEG in epilepsy and includes a vast array of eight-channel EEGs.

Instrumentation

The American EEG Society [1986] published guidelines for the performance of EEG machines. The guidelines recommended a minimum of eight channels. Chlorided silver disks or gold electrodes, adhered to the scalp with collodion, are recommended; needle electrodes are not. A chart speed of 3 cm/s is standard, and a recording sensitivity of 5 to 10 μV/mm is recommended. The frequency response extends from 1 to 70 Hz for the –3-dB points.

Evoked Potentials

With the availability of signal averaging using a digital computer, it is possible to investigate the integrity of the neural pathways from peripheral sense organs to the cortex by using appropriate stimuli (e.g., clicks, light

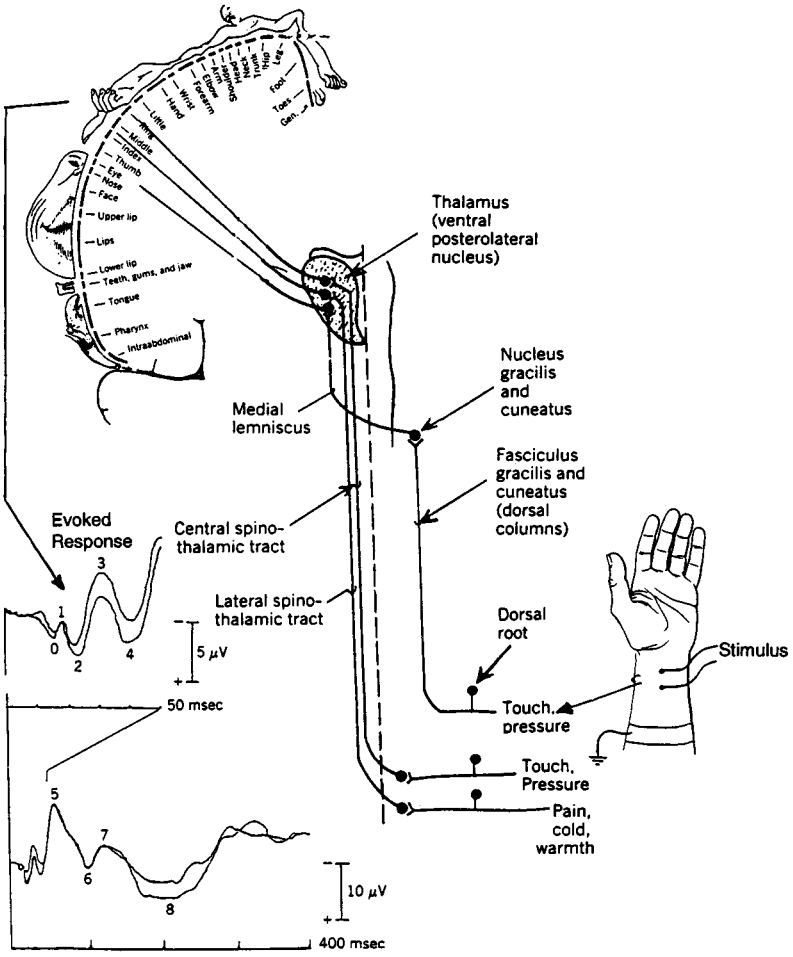


FIGURE 113.23 Pathways from the peripheral sense organs to the cortex and the topographical distribution of sensation along the cortex with Penfield's homunculus. Also shown are the stimulating electrodes on the wrist and the SSEPs recorded from the contralateral cortex. (*Source:* SSEPs redrawn from T. W. Picton, "Evoked cortical potentials, how? what? and why?," *Am. J. EEG Technol.*, vol. 14, no. 4, pp. 9–44, 1974. With permission.)

flashes, or current pulses). Usually the stimulus consists of a few hundred to about 1000 pulses, averaged to produce the somatosensory-evoked potential (SSEP). Likewise, it is possible to evaluate the integrity of the neural pathways from the motor cortex to peripheral muscles by applying multiple short-duration current pulses to scalp electrodes and recording nerve and/or muscle action potentials with skin-surface electrodes. Such recordings are called motor-evoked potentials (MEPs). With both SSEPs and MEPs, the largest responses appear on the opposite side of the body from the stimulus.

Because the responses are in digital form, they can be written out in hard-copy format. With SSEPs, the response consists of many waves occurring at various times after the stimulus. To permit close examination of the various waveforms, several displays are presented, each with a different time axis. Figure 113.23 presents a sketch of the neural pathways from the periphery to the cortex, showing the topographic distribution of sensation along the cortex using the homunculus created by Penfield, described in detail in 1968. Also shown in Fig. 113.23 is a typical SSEP obtained by stimulating the median nerve with skin-surface electrodes connected to an isolated (i.e., not grounded) stimulator output circuit. Note the remarkably low amplitude of the responses that were obtained by averaging the response to 240 stimuli. Note also that the first display showed the responses from 0 to 50 ms and the second display presented the responses in the 0- to 400-ms interval.

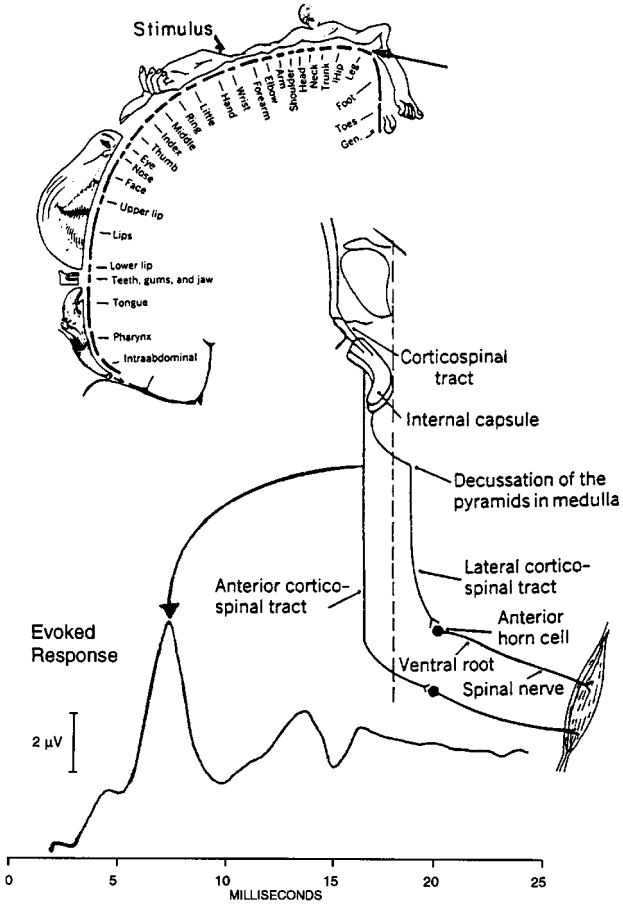


FIGURE 113.24 Neural motor pathways from the cortex to a typical muscle. The motor areas are represented by Penfield's homunculus. A train of 240 stimuli delivered to the motor cortex provided the average MEP detected with electrodes placed over the spinal column in this patient to whom a muscle paralytic drug was given; therefore the MEPs are from the spinal cord. (*Source:* Redrawn from Levy et al., 1984, and Penfield and Rasmussen, 1968.)

Figure 113.24 shows the motor pathways from the motor cortex to a typical muscle. The cortical motor areas are represented by Penfield's homunculus. A train of 250 stimuli were applied between electrodes on the scalp and in the mouth. The motor-evoked potentials were recorded with skin-surface electrodes over the spinal column. The MEP of a patient in whom the muscles were paralyzed is also shown in Fig. 113.24. Because the muscles were paralyzed, the MEPs shown in the figure represent action potentials in the spinal cord. Note the prominent peaks at 7 and 14 ms. These peaks provide information on the path taken by the nerve impulses initiated in the motor cortex.

Although there is no ANSI standard for evoked-potential recording, the American EEG Society [1986] published guidelines for equipment performance and recording techniques. This information should be consulted by those contemplating entry into this field.

Magnetic (Eddy-Current) Stimulation

When scalp electrodes are used to stimulate the brain, there is considerable skin sensation under the electrodes owing to the high perimeter current density [Overmeyer et al. 1979]. It has been found that sufficient eddy current can be induced in living tissue by discharging an energy-storage capacitor into an air-cored coil placed on the skin. This mode of stimulation is almost without skin sensation; by some it is called "ouchless stimulation"

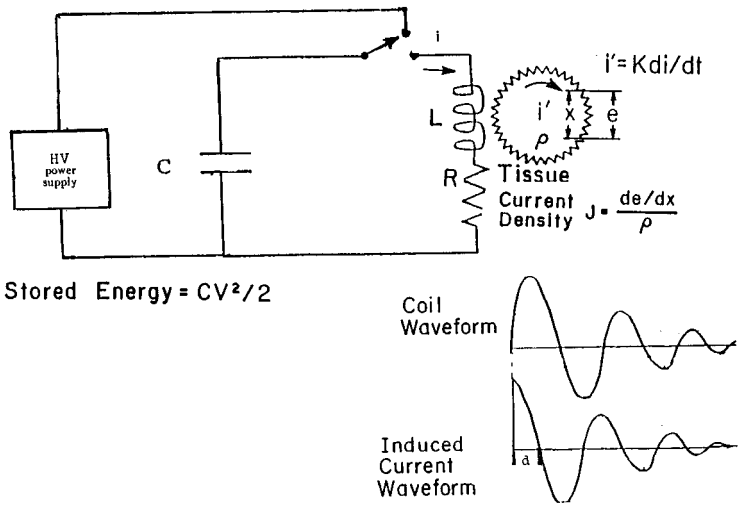


FIGURE 113.25 Simplest type of magnetic (eddy-current) stimulator and coil and induced current waveforms.

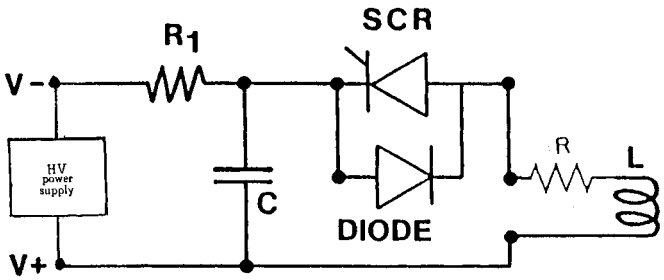


FIGURE 113.26 Magnetic (eddy-current) stimulator that recovers energy stored in the magnetic field.

and it can be used to stimulate the brain, peripheral nerves, and the heart. The parameters associated with eddy-current stimulation are kiloamperes, Teslas/sec, milliohms, microhenries, microseconds, and low damping. Because the forces on the coil conductors are very large, special care is required in fabricating such coils.

Stimulators

With **magnetic** (eddy-current) **stimulation**, three circuits and two coil geometries are used. The simplest circuit is illustrated in Fig. 113.25, which shows a capacitor (C) being discharged into the stimulating coil (L). The induced current (i) is proportional to the rate of change (di/dt) of the coil current (i). The resistance (R) in the entire circuit is low and the coil current is an underdamped sinusoid. The tissue is stimulated by the induced current density $J = k(de/dx)/\rho$, where de/dx is the induced voltage gradient, ρ is the tissue resistivity, and k is a constant that depends on coil geometry and target distance. The effective stimulus duration (d) is from the onset of the induced current to the first zero crossing as shown in Fig. 113.25. Typical durations (d) range from 20 to 200 μsec. When the damping is low, the multiple pulses of induced current can elicit responses, providing the period is longer than the refractory period of the tissue being stimulated. Note that the capacitor voltage falls to zero and the capacitor must be recharged to deliver the next stimulus.

The second type of eddy-current stimulator employs energy recovery and is shown in Fig. 113.26. The charged capacitor (C) is discharged into the stimulating coil (L) by triggering a silicon-controlled rectifier (SCR). By placing a diode across the SCR as shown, some of the energy stored in the magnetic field can be recovered to recharge the capacitor. In this way, the power supply need only deliver a modest amount of current to restore the charge on the capacitor.

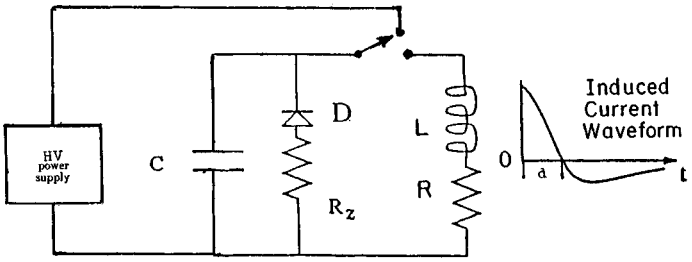


FIGURE 113.27 Method of using a diode (D) and current-limiting resistor (Rz) to avoid reverse polarity on the energy-storage capacitor (C).

With the circuits shown in Figs. 113.25 and 113.26, it is necessary to use non-polarized capacitors. When a long-duration pulse is desired, it is more practical to employ electrolytic capacitors because of their high energy storage per unit of weight. When employing electrolytic capacitors, it is necessary to prevent reverse charging. To achieve this goal, a diode (D) and a series current-limiting resistor (Rz) are placed across the capacitor, as shown in Fig. 113.27. The resulting induced current waveform is shown on the right.

Stimulating Coils

Air-cored coils are used because all known ferromagnetic materials saturate at the high flux densities used with eddy-current stimulation. Two types of coils are used: (1) annular and (2) coplanar, sometimes called figure eight, osculating, or butterfly. Most coils are annular in shape and the radius is chosen on the basis of the distance to the target tissue to be stimulated.

Nyenhuis et al. [1991] analyzed the factors pertaining to the optimum design of such annular coils and calculated the magnetic-field energy and ohmic heating. They found that the magnetic field energy exhibits a broad minimum when the coil outer radius is between two and five times the target distance. Ohmic heating in the coil decreases as the radius of the coil increases. Increasing annular width results in a small reduction in field energy and coil heating; thin coils (small height) reduce the field energy but increase heating. A reasonable compromise between efficiency and coil size is a coil with an outer diameter that is twice the distance between the coil surface and the underlying target tissue to be stimulated and whose height and annular width are 0.2 and 0.6 that of the mean radius, respectively.

The induced electric field gradient (*de/dx*) is maximum at the perimeter of an annular coil. Weissman and Epstein [1992] plotted the electric field induced in a saline volume conductor due to an annular coil; Fig. 113.28a is the result which shows a maximum over the perimeter of the coil.

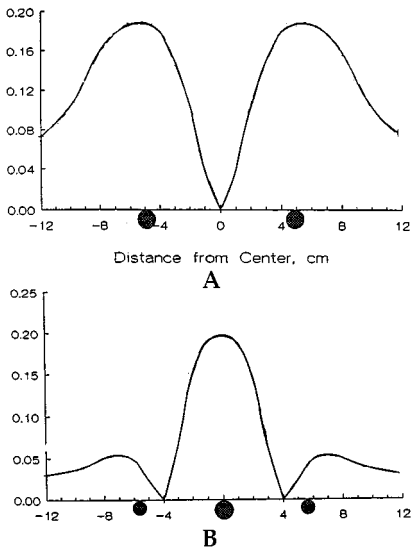


FIGURE 113.28 Induced electric field (de/dx) for a single annular coil (a) and for a pair of coplanar annular coils (b). (Redrawn from [Weissman and Epstein 1992]).

Eddy-current stimulation is very energy inefficient when compared to direct tissue stimulation with electrodes. To improve the efficiency, investigators have sought strategies to concentrate (focus) the magnetic field without the use of ferromagnetic material. Ueno et al. [1988] introduced what they called the figure-of-eight coil of the type shown in Fig. 113.29 in which the current in the conductors where the two coils touch is in the same direction, thereby concentrating the field. Note that the currents in the two coils are in opposite directions; hence the magnetic fields (B1, B2) are in opposite directions. With this coil configuration and current direction, the induced electric field is a maximum over the site where the coils touch, as shown in Fig. 113.28b. Although the use of a pair of coils increases the efficiency of magnetic stimulation, it makes coil placement more critical.

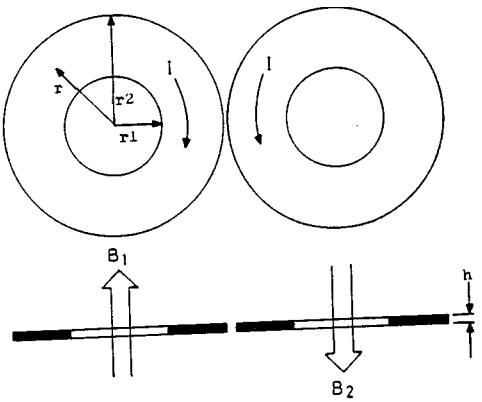


FIGURE 113.29 Current (I) and magnetic field (B) direction in the coplanar (figure-of-eight) coil.

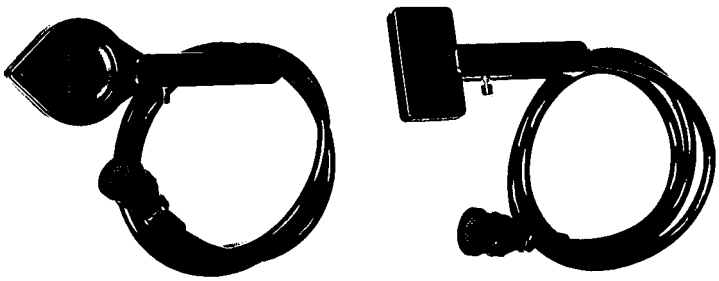


FIGURE 113.30 Commercially available coils for magnetic (eddy-current) stimulation. (Courtesy of Cadwell, Kennewick, WA 99336).

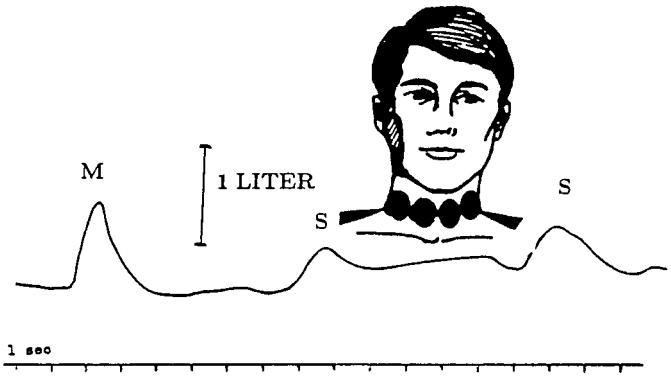


FIGURE 113.31 Magnetically induced inspiration (M) produced by a train of eddy-current stimuli applied to both phrenic nerves with two pairs of coplanar coils at the base of the neck. S = spontaneous breaths.

Figure 113.30a shows a commercially available annular coil and Fig. 113.30b shows the figure-of-eight (Butterfly™) coil. These coils can be used for stimulating the brain or peripheral nerve. Proper location of the coil with respect to the target tissue is essentials (see Fig. 113.28 for locations of maximum *de/dx*).

Many types of excitable tissue can be stimulated by eddy current. Geddes et al. [1991] investigated its use to produce inspiration in humans by stimulating both phrenic nerves at the base of the neck. The phrenic nerves cause the diaphragm to contract, thereby producing inspiration. By using two series-connected pairs of coplanar coils, each pair angled at 150°, current pulses at 25/sec were delivered from a 50 μF capacitor charged

to 1200 volts, the magnetically induced inspiration (M) of 900 mL shown in Fig. 113.31 was produced. To make this record, the subject first hyperventilated, then stopped breathing and the magnetic stimulator was turned on for about 1 sec. The breaths marked S are spontaneous breaths with a volume of 400 mL.

Cardiac muscle has been stimulated to contract with single pulses of eddy current. The first to achieve this feat was Bourland et al. [1990] who applied a coplanar coil to the left chest of an anesthetized dog. Blood pressure and the electrocardiogram were recorded and the heart was stopped by right vagal stimulation, during which time a pulse of current was delivered to the coplanar coils and a ventricular contraction was produced, showing an inverted QRS wave in the ECG and a blood-pressure pulse.

That it is difficult to stimulate the heart with a pulse of eddy current can be assessed by the parameters of the stimulator used by Bourland et al. [1990] which employed a 682 μF capacitor, charged to 9900 volts, and discharged into the 220 μH coplanar coil assembly. The peak current was 17,000 amps and the stored energy was 33,421 joules. The duration of the induced current pulse was 690 μsec. This stimulator was used by Mouchawar et al. [1992], who reported mean current and energy thresholds for cardiac stimulation in the range of 9200 amps and 11,850 joules.

From the foregoing it is clear that the ventricles can be caused to contract with pulses of eddy current. However, if it is desired to pace at 60/min. (1/sec), the power needed is 11.85 kilowatts, hardly a practical value for domestic or hospital use.

Summary and Conclusion

Although eddy-current stimulation of excitable tissue is quite popular now, the first report was by d'Arsonval [1896], who reported seeing bright flashes in the visual field (phosphenes) when the head was placed in a coil carrying 30 amperes of 42 Hz current (see Geddes translation [1991]). It is now known that stimulation of the retinal receptors in the eye produces such phosphenes.

Magnetic stimulation is largely used to excite nerve cells in the brain and spinal cord. The diagnostic information is contained in the time between the stimulus and the response (action potential). The same measure is used when peripheral nerve is stimulated.

A major advantage of magnetic stimulation is that no electrodes are required and the skin need not be exposed to apply the stimuli. However, the main advantage is that the skin sensation is very mild. A major disadvantage is the high energy needed to induce sufficient eddy current density to achieve stimulation. When repetitive pulses are required, the power drawn by the magnetic stimulator may require a 60-Hz AC energy source of 220 or 440 volts. Moreover, with repetitive stimulation, coil heating becomes a problem. The availability of magnetically permeable materials that saturate at several orders of magnitude above presently available materials would be of benefit to the field of magnetic (eddy-current) stimulation.

This section has focused only on the three most prominent bioelectric events, those of the heart, skeletal muscle, and brain. The eye, ear, sweat glands, and many types of smooth muscle produce action potentials that are used for their diagnostic value, as well as being the subject of on-going research. The reader interested in delving deeper into this field can find such information in a book by Geddes and Baker [1989].

Defining Terms

Atrophy: Wasting of cells deprived of nourishment.
Autonomic: That part of the nervous system which controls the internal organs.
Ectopic beat: A heart beat that originates from other than the normal site.
Hypocapnia: A condition of reduced carbon dioxide in the blood.
Hypoxia: A reduced amount of oxygen.
Magnetic stimulation: Eddy current stimulation.
Metabolic process: The method by which cells use oxygen and produce carbon dioxide and heat.
Myocardial infarction: A heart attack in which a region of the heart muscle is deprived of blood and soon dies.
Occipital: The back of the brain.
Parietal: The side of the brain.

Related Topic

2.1 Step, Impulse, Ramp, Sinusoidal, Exponential, and DC Signals

References

American EEG Society, "Guidelines in EEG and evoked potentials," *Amer. J. Clin. Neurophysiol.,* vol. 3 (Suppl.), 1986.

Association for the Advancement of Medical Instrumentation (AAMI), Diagnostic ECG Devices, ANSI-AAMI Standard EC-101-1991.

J.D. Bourland, G.A. Mouchawar, J.A. Nyenhuis, et al., "Transchest magnetic (eddy-current) stimulation of the dog heart," *Med. Eng. Comput.* March, pp. 196–198, 1990.

F. Buchtal, C. Guld, and P. Rosenflack, "Action potential parameters of normal human muscle and their dependence on physical variables," *Acta Physiol. Scand,*. vol. 32, pp. 200–220, 1954.

T-C. Chou, *Electrocardiography in Clinical Practice,* 3d ed. Philadelphia: W. B. Saunders, 1991.

H.L. Cohen and F. Brumlik, *Manual of Electromyography,* New York: Hoeber Medical Division, Harper & Row, 1969.

A. d'Arsonval, "Dispositifs pour la mesure des courants alternatifs de toutes frequences," *C. R. Soc. Biol.* (Paris), vol. 2, pp. 450–451, 1896.

L.A. Geddes, "The history of magnetic stimulation," *J. Neurophysiol.,* vol. 8, no. 1, pp. 3–9, 1991.

L.A. Geddes, L.E. Baker, and M. McGoodwin, "The relationship between electrode area and amplifier input impedance in recording muscle action potentials," *Med. Biol. Eng. Comput.*, vol. 5, pp. 561–568, 1967.

L.A. Geddes and L.E. Baker, *Principles of Applied Biomedical Instrumentations,* 3rd ed., New York: Wiley, 1989.

L.A. Geddes, G. Mouchawar, J.D. Bourland, and J. Nyenhuis, "Inspirations produced by bilateral electromagnetic, cervical phrenic nerve stimulation in man," *IEEE Trans. Biomed. Eng.,* vol. 30, no. 10, pp. 1047–1048, 1991.

F.A. Gibbs and E.L. Gibbs, *Atlas of Electroencephalography,* London: Addison-Welsey, 1952.

International Federation of EEGSocieties, J. Knott, Chairman, *EEGClin. Neurophysiol.*, vol. 10, pp. 378–380, 1958.

International Federation for Electroencephalography and Clinical Neurophysiology, *EEG. Clin. Neurophysiol.,* vol. 10, pp. 371–375, 1958.

W.J. Levy, D.H. York, M. McCaffery, and F. Tanzer, "Evoked potentials from transcranial stimulation of the motor cortex in humans," *Neurosurgery,* vol. 15, no. 3, pp. 287–302, 1983.

A.A. Marinacci, *Clinical Electromyography,* Los Angeles: San Lucas Press, 1955.

G.A. Mouchawar, J.D. Bourland, J.A. Nyenhuis, et al., "Closed-chest cardiac stimulation with a pulsed magnetic field," *Med. Biol. Eng. Comput.,* March, pp. 162–168, 1992.

J.A. Nyenhuis, G.A. Mouchawar, J.D. Bourland, and L.A. Geddes, "Energy considerations in the magnetic (eddy-current) stimulation of tissues," *IEEE Trans. Magn.,* vol. 27, no. 1, pp. 680–687, 1991.

K.M. Overmeyer, J.A. Pearce, and D.P. DeWitt, "Measurement of temperature distributions at electrosurgical dispersive electrode sites," *Trans. ASME, J. Biomech. Eng.,* vol. 101, pp. 66–72, 1979.

W. Penfield and T. Rasmussen, *The Cerebral Cortex of Man,* New York: Hafner, 1968.

R.E. Phillips and M.K. Feeney, *The Cardiac Rhythms: A Systematic Approach to Interpretation,* 3rd ed., Philadelphia: W. B. Saunders, 1990.

T.W. Picton, "Evoked cortical potentials, how? what? and why?," *Am. J. EEG Technol.,* vol. 14, no. 4, pp. 9–44, 1974.

S. Ueno, T. Tashiro, and K. Harada, "Localized stimulation of neural tissues in the brain by means of a paired configuration of time-varying magnetic fields," *J. Appl. Phys.,* vol. 64, no. 10, pp. 5862–5866, 1988.

J.D. Weissman and C.M. Epstein, "Magnetic stimulation of the nervous system," *Am. J. EEG Technol.,* vol. 32, pp. 127–146, 1992.

113.3 Application of Electric and Magnetic Fields in Bone and Soft Tissue Repair

C. Polk

History

As early as 1962 in the United States [Bassett and Becker, 1963], and even earlier—1957—in Japan [Fukada and Yasuda, 1957] it was shown that electric potential differences appear across both living and dead bone subjected to mechanical stress. Bassett and Becker observed that these stress-generated electrical signals decayed very slowly in comparison with similarly initiated signals in piezoelectric crystals and concluded [Bassett and Becker, 1963] that piezoelectric phenomena "while probably present, were not the sole cause of these potentials." Later analysis and experiments established that the observed signals were primarily due to ion displacement within the porous regions and multiple fluid-filled channels present in all bone. The early observations already suggested that direct application of an externally generated voltage might have an effect on bone development. This was shown to be the case by Bassett et al. [1964] who found that a dc current of the order of 1 μA (corresponding to a current density of approximately 0.01 A/m^2) produced massive osteogenesis near the cathode when electrodes were implanted into the femur of living dogs.

Having shown that application of a dc electric field to nonexcitable, connective tissue cells can produce effects similar to those elicited by mechanical stress, Bassett and his co-workers realized that clinical exploitation of these phenomena would require surgical implantation of electrodes with attendant danger of infection. They proceeded therefore to explore whether noninvasive, inductive coupling that gave waveforms similar to those endogenously produced by mechanical stress could lead to beneficial bone development. In 1974 they reported favorable results obtained with pulsed electromagnetic fields on dogs [Bassett et al., 1964]. Signals of this type have generally been identified as **PEMF** (pulsed electromagnetic field) in the orthopedics/electrical stimulation community for the last 20 years and have been applied successfully in a large number of cases for the repair of **nonunions** [Grossling et al., 1992]. In Germany relatively large-amplitude, low-frequency (<20 Hz) sinusoidal magnetic fields have been used for both bone repair and wound healing [Kraus, 1984].

Although the noninvasive PEMF treatment for nonunions (fractures that fail to heal) became—at least in the United States—the most widely used clinical application of subradio frequency fields, several investigators pursued the application of dc electric fields through implanted electrodes and the application of higher-frequency currents through electrode contacts placed on the skin surface to enhance bone repair [Brighton et al., 1979]. At the same time mostly laboratory investigations, *in vitro* and on animals, explored the application of all three modalities—PEMF, implanted dc electrodes, and higher-frequency coupling through skin electrodes—to produce blood vessel regeneration (**angiogenesis**), soft tissue healing, nerve repair or regeneration, and regression of tumors. Claimed to be useful in edema and pain management and for acceleration of wound repair is pulsed radio frequency (PRF), mainly the diathermy frequency of 27.12 MHz assigned by the FCC, at an average power level that should usually produce only very moderate tissue heating [Kloth and Ziskin, 1996; Markov, 1995]. Motivated by the suggested clinical applications, a large number of basic science investigations have been initiated, and are continuing today, with the object of understanding the mechanisms through which low-intensity electric and magnetic fields affect cells and living tissue. Some of these are reviewed briefly in the subsection Mechanisms and Dosimetry.

Devices for Bone and Cartilage Repair

In the United States medical devices are approved for clinical use only after it has been shown to the satisfaction of the U.S. Food and Drug Administration (FDA) that they are not only safe, but also effective. This is a much more stringent requirement than the mere demonstration of safety demanded presently in most European countries (although some EEC countries are considering moving toward approval criteria that are similar to those used in the United States). The United States laws governing the sale of medical devices for clinical use are also much more restrictive than the controls over implementation of new surgical procedures that involve only informal medical peer review. Even organized multi-patient clinical trials require not more than approval

TABLE 113.2 Electrical Bone Growth Stimulators Approved by the U.S. FDA as of July 1, 1996

Manufacturer FDA "PMA" Number* FDA Docket Number	Device	Indications	Technology	Approved	Text References
Electro-Biology, Inc. P790002 80M-0057	EBI Bi-OsteoGen/Bone Healing System**	Nonunion; congential pseudoarthroses; failed fusions	Noninvasive pulsed electro-magnetic field (PEMF)	November 1979	A B
American Medical Electronics, Inc. P850007 86M-013B	Physio-Stim	Nonunions (excluding vertebrae and flat bones)	Noninvasive PEMF	February 1986	C
American Medical Electronics, Inc P850007 90M-0067	Spinal-Stim	To promote spinal fusion as an adjunct to surgery or as nonop-erative treatment when 9 months have elapsed since the last surgery	Noninvasive PEMF	February 1990	D
Biolectron, Inc. P850022; 86M-0139	Orthopak BGS Device	Nonunions (excluding vertebrae and flat bones)	Noninvasive/cap-acitively coupled	February 1986	E
Electro-Biology, Inc. P790005; 80M-0254	Orthogen/Osteogen	Nonunion of long bones	Implantable dc	January 1980	F
Electro-Biology, Inc. P850035 87M-0174	SpF Implantable Spinal Fusion Stimulator† SpF-4 (original); SpF-2 (S5); SpF-2T, 4T (S6); SpF-XL (S13)	Spinal fusion adjunct	Implantable dc	April 1987	G
OrthoLogic P910066	OrthoLogic 1000	Nonunions (excluding vertebrae and flat bones)	Noninvasive PEMF	March 1994	

*These numbers are useful when calling the U.S. FDA (Dockets Management Branch) to obtain summary of safety and effectiveness of particular devices.

**Originally designed Bi-Osteogen Systems 204.

†Originally Osteostim HS 11.

by a local hospital-based institutional review board. The FDA Office of Device Evaluation does approve some new devices, after careful review, for clearly limited clinical trials. However, information on such limited, temporary approval is not made available. Table 113.2 therefore shows only those devices which are presently (July 1, 1996) approved and does not include experimental systems which may be undergoing currently limited clinical trials. Some of the latter are discussed further on, based only on information furnished by manufacturers or available from the published medical literature.

The devices listed in Table 113.2 are approved for one of three applications: the treatment of *nonunions* (fractures that have failed to heal after standard treatment involving setting and stabilization with casts), congenital **pseudoarthroses** [Bassett, 1984], and promotion of spinal fusion. Although many animal experiments (and possibly a few human trials, especially in Europe) have evaluated the application of electric or magnetic fields for acceleration of fresh fracture healing and for reversal of osteoporosis, no devices are currently approved in the United States for these purposes.

Classified by electrical and mechanical characteristics, the devices in Table 113.2 are either:

1. Noninvasive:
   a. Generating time-varying magnetic fields applied by coils to the affected body part (A, B, C, D, and H).
   b. Generating time-varying electric fields applied through skin-surface electrodes (**capacitively coupled**) (E).
2. Invasive: dc applied from an implanted battery (F, G).

A signal typical for some PEMF (A, C, D) devices is illustrated in Fig. 113.32. Part A shows the magnetic field versus time and Part B the corresponding electric field induced into a linear, isotropic medium. The waveform shown on Part B can be measured by a probe coil having a sufficiently large number of turns. The frequency spectrum of the electric field is shown in Fig. 113.33. Signals used by the different manufacturers are protected by patents, and FDA approval is for particular signal parameters within specified tolerances on time and amplitude. The pseudoarthrosis signal used by Electrobiology, Inc. (EBI) (B in Table 113.2) consists of single pulses repeated at a rate of 72 pps rather than the pulse bursts illustrated in Figs. 113.32 and 113.33. Each magnetic field pulse increases from zero to 3.5 mT in 380 μs and then decreases slowly to zero in approximately 4.5 ms. The signals that are now in use have evolved considerably from those employed in the initial studies, and some have little resemblance to the endogenous electrical signals elicited by mechanical stress.

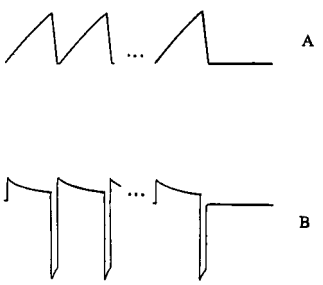


FIGURE 113.32 A typical PEMF signal (signal A of Table 113.2). (A) Magnetic field versus time. (B) Electric field ($\propto \partial B/\partial t$) versus time. Signal consists of 15 pulse bursts per second. Each burst is 4.5 ms long and contains 20 pulses. In each pulse the magnetic field increases from 0 to approx. 2 mT during 200 μs, decreases to 0 again during 23 μs, and is equal to 0 for 2 μs before the next 225-μs sequence begins.

The PEMF signals employed by the various manufacturers in the United States and Europe can have several different pulse shapes, rise and decay times, pulse widths, pulse repetition rates, and amplitudes. Since it has been shown (see Mechanisms and Dosimetry) that all these variables can have a profound effect on the biological action of a particular signal, it is essential that reports on effectiveness or lack of effectiveness of PEMF give an exact description of the signal which was used. Unfortunately the medical literature is replete with examples where this information is either incomplete or completely absent. Details of shape, orientation, and location of the application coil or coils are also important, since these parameters, together with pulse amplitude and shape, determine the nature of the magnetic and electric field at the location of the injured tissue. If the amplitude of the axially directed magnetic flux density B is constant over some region of radius R within the cross section of a circular cylinder, the induced electric field is

$$\bar{E} = -\left(\frac{\partial B}{\partial t}\right)\frac{r}{2}\,\hat{\phi} \tag{113.17}$$

provided the material of the cylinder is electrically homogeneous and isotropic. In Eq. (113.17) $r < R$ is the distance from the center of the cylinder and $\hat{\phi}$ is a unit vector in the circumferential direction. For magnetic fields varying sinusoidally as $B_0 \cos \omega t$, $\bar{E} = \omega B_0(r/2) \sin \omega t\, \hat{\phi}$. Since most biological objects are neither homogeneous nor isotropic, the actual induced electric fields at various points in the tissue or cells may deviate substantially from the values given by Eq. (113.17) [Polk and Song, 1990; Van Amelsfort, 1990]. Equation (113.17) is useful only for *estimating* the spatial *average* value of the induced electric field, which depends in the bone environment on the point-to-point variation of the electrical properties of muscle, fat, cartilage, periosteum (outer bone membrane), and bone marrow.

Current pulses of the PEMF devices are usually produced by the discharge of capacitor banks controlled by a timing network. The applicator coil cannot be interchanged among different devices because its inductance and resistance are a part of the discharge network. While earlier bone growth stimulators employed Helmholtz coil-pair arrangements, most present devices have single coils which can be custom-shaped for particular limbs. Figure 113.34 shows a typical system sold by EBI. This unit is driven by rechargeable batteries, and the control unit (shown in Fig. 113.34) includes an elapsed-time clock to measure the total time of stimulation of the fracture being treated. A typical treatment time can be between 2 and 10 h/day over a period of 6 months.

The so-called capacitively coupled device (E in Table 113.2) generates a continuous sine wave at a frequency of 60 kHz. The total current through the skin contains a not negligible conduction component since conductive contact is made between the applicator electrodes and the skin that represents a "leaky" capacitor. Electric fields produced at the tissue level by this device are between 1 and 50 V/m [Pollack and Brighton, 1989]. These levels

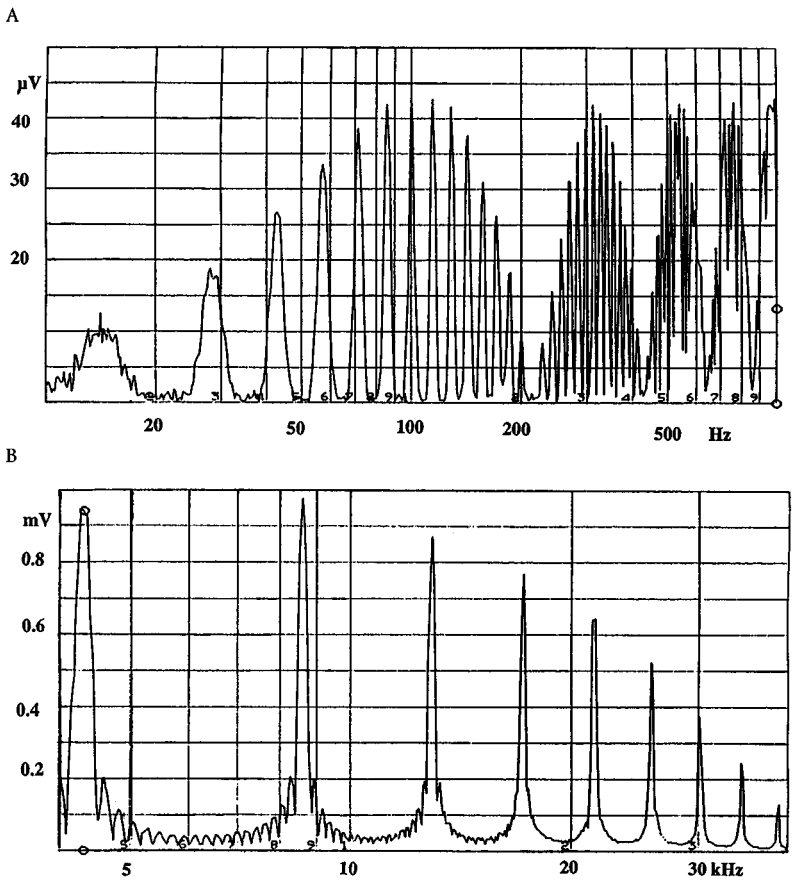


FIGURE 113.33 Electric field spectrum $|E(\omega)|$ of signal in Fig. 113.32 as measured by the output from an air-core coil (0.6 cm mean diameter, 65 turns). (A) 10 Hz to 1 kHz (50 μV full scale); (B) 4 to 40 kHz (1 mV full scale).

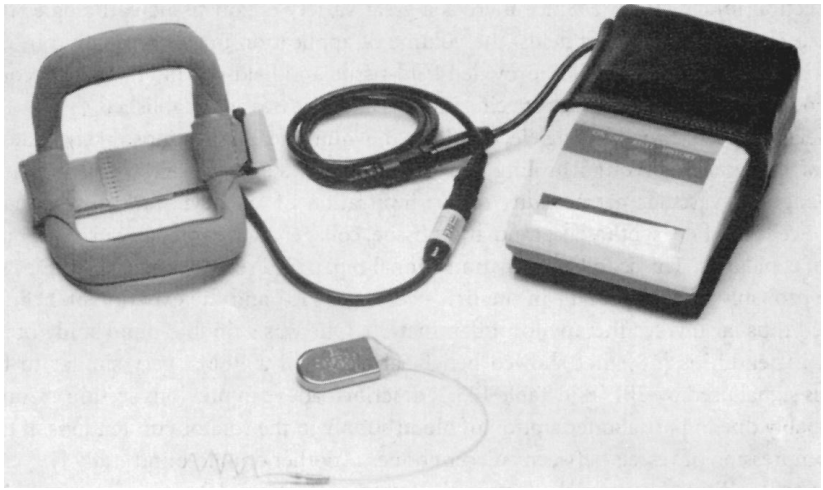


FIGURE 113.34 PEMF applicator and control unit and implantable dc stimulating device (battery with wire electrode) manufactured by Electrobiology, Inc. (systems A and G of Table 113.2). (Photograph courtesy of Electrobiology, Inc.)

are very much higher than the average amplitude of the electric fields produced in tissue by PEMF devices and also higher than the instantaneous peak values produced by some of the PEMF systems. It is interesting to note here that *in vitro* experiments with truly capacitively coupled fields showed enhancement of bone cell proliferation at very much lower amplitude (10^{-5} V/m) when the frequency of the continuous sine wave was 10 Hz rather than 60 kHz [Fitzsimmons et al., 1986]. The most recent addition to FDA approved non-invasive systems (H in Table 113.2) employs a sinusoidally time varying 76.6 Hz, 40 μT field superimposed on a DC field of 20 μT. Possible reasons for using this AC/DC "resonance" combination are discussed in the Mechanisms and Dosimetry section.

An invasive (implantable) dc device (F, G in Table 113.2) is also shown on Fig. 113.34. The small (approximately $4 \times 2 \times 0.5$ cm) titanium case contains a long-life battery that is connected to it. The case acts as the anode, and two (shown) or four titanium wires act as the cathode. The amplitude of the continuous current is between 5 μA (for some spinal fusion applications) and 20 μA (for nonunion of long bones). The cathodes are placed at the location where bone growth is to be enhanced, for example, at the vertebrae that are surgically fused, while the case is placed in a convenient location at some distance from the bone. Treatment details and success rates, in comparison with surgical procedures without use of dc stimulation, are discussed in the medical literature [Nerubay et al., 1986].

Although not used or approved for use in the United States, the German "Magnetodyn" system [Kraus, 1984] is interesting not only because it employs sinusoidal magnetic fields between 2 and 20 Hz but also because it relies on metallic implants that are used for fixation of the bone to act as the "secondary" of a "transformer" whose primary is the external applicator. Sometimes an implanted pickup coil ("secondary") is connected to fixation bars or to screws (electrically insulated from the bars) on each side of a pseudoarthrotic gap. With this system, peak electric fields of 40 V/m and current densities of 5 A/m^2 have been produced in the gap.

A PEMF signal (similar to B in Table 113.2) has also been used experimentally to arrest **osteonecrosis** (bone death possibly due to vascular impairment or toxic agents) of the femoral head. Application for 8 h/day over 12 months gave substantially better results than the standard surgical (decompression) treatment [Aaron and Steinberg, 1991].

Soft Tissue Repair and Nerve Regeneration

No electric or magnetic system to aid nerve regeneration or soft tissue repair is approved by the FDA at the present time for nonexperimental therapy. However, considerable animal and *in vitro* experimentation in this country and abroad suggests the clinical usefulness of electric currents for soft tissue repair [Canaday and Lee, 1991] and possibly also to enhance repair of nerve fibers that have sustained crush or transsection injury [Ito and Bassett, 1983; Siskin et al., 1990]. Since there is a great variety of soft tissue pathologies that could respond to electric or magnetic fields, the volume of application in this area could in the future become larger than in orthopedics, provided field-tissue and field-cell interactions become better understood and clinical benefits for specific injuries and diseases are established. A pulsed radio frequency (PRF) device sold in the U.S. by Electropharmacology, Inc., of Pompano Beach, FL, produces 27.12 MHz pulses of 65 microsecond duration. Pulse repetition rates can be selected between 80 and 600 pps and peak output power is adjustable between 174 and 363 watts. Thus, the highest possible average power delivered into the 9-in. diameter inductive applicator is 14.16 watt. ("Inductive" applicator means that the applicator is designed to make the ratio of electric to magnetic field in its immediate vicinity much less than the 377 Ω wave impedance of free space; thus, the magnetic field is maximized in the "near field" region). Coupling to living tissue at 27.12 MHz without impedance matching is rather poor [Polk, 1995]. If it is assumed that only 20% of the 14.16 W average power is absorbed by the tissue, application for the recommened treatment period of 30 min would correspond to an energy transfer of 1.2 kilocalories. If the mass of the local region to which this energy is applied is of the order of 2 kg, this would correspond to a temperature increase of 0.6°C in the absence of cooling by blood circulation or convection. In any case, although "the PRF modality was originally reported as a non-thermal biophysical treatment of infections" [Ginsberg, 1934; Markov and Pilla, 1995], the device is likely to produce at least moderate tissue heating when applied with maximum output settings. The Electropharmacology "MRT Softpulse Model 912" is marketed for clinical use in the U.S. under FDA rules which permit continued sale of

devices that are substantially equivalent to older equipment (in this case "Diapulse") in use before the enactment of present FDA rules.

Beneficial effects of time-varying electric fields in wound healing are most likely related to promotion of *angiogenesis*. Wound healing consists of several stages, the first being inflammation, when changes in vascular permeability occur; infiltration of leucocytes and macrophages takes place; and cells migrate, synthesize granulation tissue, collagen, and proteoglycans, and initiate formation of capillaries. This is followed by transitional repair and remodeling phases. Electrical currents are probably only important in the first two stages [Canaday and Lee, 1991] and the experimental clinical trials performed thus far involve therapy for inflammation. One was a double-blind study of persistent rotator cuff tendinitis [Binder et al., 1984] that showed beneficial effects of a PEMF very similar to the pseudarthrosis signal used by EBI (B in Table 113.2, described above in previous section). Cuff tendinitis is probably due to partial interruption of blood supply to the rotator cuff tendons of the shoulder by compression of vessels between adjacent bones. Another double-blind study [Ieran et al., 1990] employing a PEMF signal indicated beneficial effects in the treatment of skin ulcers.

Other clinical trials [Bentall, 1986] and animal experiments involved irradiation of wounds with pulsed high and very high frequencies between 3 and 44 MHz, at power levels from 73 μW to 15 W, employing pulse widths between 65 and 100 μs and pulse repetition frequencies between 200 Hz and 1 kHz. A commonly used animal model for wound healing is the McFarlane skin flap involving partial excision of a rectangular skin section on the back of the rat. Survival of the flap depends mainly on blood supply, with vascularization of the skin flap being an indirect measure of treatment success. The EBI signal shown on Fig. 113.33 (A in Table 113.2) is reported to have decreased skin flap necrosis when exposure was for 6 h/day for 3 days, while exposure for 18 h on the first day after injury had no observable effect [Luce and Bryant, 1986]. Exposure with a triangular, symmetric and almost continuous magnetic field (18-ms triangular pulse, followed by 2-ms pause) at a frequency of 50 Hz (8-mT peak to peak value) produced a significant increase in wound contraction in rats (in comparison with controls) who were exposed to the field for 30 min immediately after surgery and for the same period thereafter every 12 h. In an effort to determine what type of signals are most beneficial for the acceleration of wound healing, skin flap necrosis was observed under exposure to sinusoidal magnetic fields at constant $\partial B/\partial t$ of 0.5 T/s using frequencies of 20, 72, and 500-Hz [Sisken and Herbst, 1990]. While signals at 20 and 72 Hz significantly decreased necrosis after 7 days, the 500-Hz signal was ineffective.

Very soon after some types of bone injury and pathology were first treated with PEMF, the effects of PEMF on peripheral nerve regeneration became subjects of investigation. Improved neural function that appeared as an unintended "side effect" in the clinical treatment of nonunions led Kort et al. [1980] to a systematic investigation of PEMF effects on neural regeneration in rats. Other investigators employed other animal models and also compared PEMF with direct current as agents for neural regeneration. More recent work [Siskin et al., 1990] employed a PEMF signal consisting of 20-ms pulses at a repetition rate of 2 pulses per second with **exponential rise** and **decay times** of, respectively, 0.5 and 1 ms. Amplitudes were 0.3 mT for experiments with crush lesions of the sciatic nerve in rats and 0.05 mT for stimulation *in vitro* of neurite outgrowth in dorsal root ganglia. Estimated values for the mean induced electric field pulses were 5 mV/m in the animal experiments and 0.7 mV/m in the 60-mm-diameter culture dishes of the *in vitro* work. Stimulation for 2 h per day of the *in vitro* cultures produced approximately 50% enhancement of neurite outgrowth in comparison with controls after 2 days.

The *in vivo* experiments using the 0.3-mT pulse produced "a 22% increase in the rate of regeneration relative to controls" as measured by a standardized test of reflex response. A very interesting observation was that animals exposed to the 0.3-mT field for 4 h/day for 7 days *prior* to the crush injury, who received *no* treatment after injury, responded similarly to those treated postinjury. Analysis of extracts from sciatic nerve segments after sacrifice of the animals showed that treatment with PEMF changed the molecular weight distribution of synthesized polypeptides.

A report on a very limited (13 subjects) clinical trial [Ellis, 1987] of spinal nerve stimulation in para- and quadriplegics employing pulsed electric current introduced by needle electrodes produced "encouraging results" in terms of increased sensory perception and motor function. The signal obtained from a Chinese multipurpose therapy apparatus was described as follows: "The pulsed-wave generator produced a biphasic wave form of 2 ms duration with an initial slow positive deflection followed by an 80-μs rise time to its maximal negative deflection and subsequent asymptotic decay. This wave shape was pulsed in ramped bursts of 200 pulses per

second for 1.5 s with 0.75 s rest time before the next ramped burst. Peak-to-peak voltages of approximately 30 V were common, and current flow was on the order of 1 mA."

The diathermy frequency of 27.12 MHz used for PRF therapy has also been employed, with ELF amplitude modulation, to induce sleep. At least one double blind study [Reite et al., 1994] has shown the effectiveness for this purpose of "Low Energy Emission Therapy" (LEET). In this application, the 27.12 MHz carrier is periodically amplitude modulated with a 42.7 Hz sine wave for 3 s and then interrupted for 1 s during a 15-min treatment period. The signal is applied by means of an electrically conducting mouthpiece in direct contact with the oral mucosa. "The estimated local SAR ("specific absorption rate") is less than 10 W/kg in the oral mucosa and 0.1 to 100 mW/kg in brain tissue, with high SAR values only in the conduction area between tongue and mouth piece" [Reite et al., 1994]. The RF electrical field strength in the fluid surrounding neurons is calculated to be in the range of the ELF electrical fields generated by normal brain activity. No information is available on either positive or negative long term health effects, if any, due to treatment with this device.

Mechanisms and Dosimetry

Although clinical effects of electric currents, PEMF, and sinusoidal electric and magnetic fields are well documented, and although some specific biochemical results have been obtained *in vitro* or in animal experiments that suggest explanation of bone and soft tissue effects, the mechanism of field-to-cell or field-to-protein transduction is presently not understood. As a consequence, optimum "dose" (what field magnitude for how long) and optimum waveshapes or frequencies for particular clinical applications are unknown, and dosimetry relies largely on trial and error methods.

It is known that electrokinetic or **streaming potentials** rather than piezoelectricity make the principal contribution to electric potentials generated by mechanically stressed bone [Gross and Williams, 1982; Lavine and Grodzinsky, 1987]. Thus potential differences appear when mechanical loading displaces fluid that contains "counterions" which normally reside opposite ions fixed to cell or intercellular matrix surfaces. These potentials are likely to play a role in intercellular signaling and in bone as well as cartilage and soft tissue development. While the original intent of electrical bone therapy was to simply mimic endogenously generated fields, a much wider range of signals was found to be clinically useful. Furthermore it was found later that some weak ionic currents ($\approx 5 \times 10^{-2}$ A/m^2) [Levine and Grodzinsky, 1987] appear endogenously without mechanical stress and that extremely weak sinusoidal electric fields can produce profound effects on cells *in vitro*. For example, both cell number and phosphotase activity in monolayers of osteoblasts (bone forming cells) was significantly affected by a magnetically induced (1.8 mT) 30 Hz electrical field estimated to be 0.6 mV/m [McLeod et al., 1993]. Sinusoidally alternating fields between 50 and 300 Hz, as small as 0.6 mV/m, were shown to affect ATP splitting activity of the membrane enzyme Na, K-ATPase [Blank and Soo, 1992]. A 60-Hz magnetic field of 1.2 NT was reported to inhibit the oncostatic action of the hormone melatonin on estrogen positive human breast cancer cells ("MCF-7") *in vitro* [Liburdy et al.]. Calcium metabolism was affected significantly in mitogen-activated lymphocytes by a 4-μT, 16-Hz magnetic field (in the presence of a 23.4-μT dc magnetic field) that induced an average electric field of about (2) 10^{-5} V/m [Yost and Liburdy, 1992]. Sinusoidal 15-Hz magnetic fields at the 0.5-mT level, giving an estimated mean electric field in the affected tissue of less than 10^{-3} V/m, significantly affected cartilage development in immature rats [Ciombor et al., 1991].

It is likely that the mechanism involved when direct currents are directly applied to injured bone (or other tissue) differs from the transduction sequence which must be acting when low-intensity alternating fields are employed. Even the 5-μA continuous current of the implantable devices (F, G in Table 113.2) when distributed over an (estimated) 5-cm^2 area corresponds to a steady electric field of 1 V/m in bone tissue of 0.01-S/m conductivity. This value is large compared with the average (but not the peak) fields induced by some PEMF devices and very large compared with the average fields induced in tissue by other PEMF devices [Rubin et al., 1989] or the mV/m ELF sinusoidal fields that affect cartilage and bone development [Ciomber et al., 1991; McLeod and Rubin, 1990]. For example, if one assumes a mean radius of 4 cm for a particular human bone fracture, one obtains from Eq. (113.17) the electric field values between 0.18 and 1.74 V/m shown on Table 113.3. When electrodes are implanted, as for the dc signals, chemical reactions at the electrodes may play a role in bone and cartilage formation. For example, the reaction at a stainless steel cathode involves consumption of dissolved oxygen and increase in local pH [Lavine and Grodzinsky, 1987].

TABLE 113.3 Electric Field (V/m) Induced by PEMF Signals at Radius of 4 cm into Electrically Uniform Medium; B Perpendicular to Plane in which Radius is Defined

	Positive Peak	Negative Peak	Average of Rectified Signal
PEMF device A (Table 113.2, Figs. 113.25, 113.26)	0.2	1.74	0.024
PEMF device B (Table 113.2)	0.18	0.015	$(1.4)\ 10^{-4}$

If the mechanism of interaction were to involve simple charge transfer by the applied electric field, it would be useful to compare the magnitude of the charge transferred by a single pulse, within a specified volume, with the random charge fluctuation due to thermal excitation during the pulse. An equation due to Einstein [1956] gives the mean square value of the charge fluctuation δq in terms of Boltzmann's constant k, the absolute temperature T, the conductance G of the current-carrying region, and the observation time t:

$$<\delta q^2> = 2GkTt \tag{113.18}$$

It is then easy to find the electric field required to transfer during time t a charge at least equal to δq over the length of a conductance of volume v (v = length × cross-sectional area) and uniform conductivity σ. One obtains

$$E \geq \left(\frac{2kT}{v\sigma t}\right)^{1/2} \tag{113.19}$$

Assuming bone tissue with conductivity 10^{-2} S/m, a physiological temperature of 37°C, an interaction volume of 10^{-14} m^3 (about equal to the volume of a cell with 10-μm radius), and an observation time of 200 μs equal to the duration of the positive pulse of device A, one obtains from Eq. (113.19) a minimum value of 6.6 V/m. Unless v and σ can be assumed to be much larger, this would indicate that the values given in Table 113.3 should be below thermal noise. If one considers, instead of charge transfer over some as yet unknown path, the voltage induced by the applied field across the membrane of the idealized spherical cell with 10-μm radius, and compares the energy of the repetitive pulse below 100 Hz—where biological action apparently occurs [McLead and Rubin, 1990]—with the thermal noise voltage given by

$$<V_n^2> = 4kTR\ (\Delta f) \tag{113.20}$$

where (Δf) is the bandwidth and R the transmembrane resistance, one finds again that the electric field due to the PEMF devices would be at best only marginally above thermal noise. The induced electric fields of the *in vitro* experiments mentioned above are also clearly below thermal noise.

It is possible to obtain somewhat better signal-to-noise ratios if one considers either larger interaction volumes (assuming electrical phenomena involving the intercell volume), elongated cells, or cells connected by gap junctions [Polk, 1992]. It is also important to note that in the extremely inhomogeneous biological system, the actual electric field at a particular point can be considerably larger or smaller than its spatial average. Nevertheless very substantial improvement of signal-to-noise ratios would require signal averaging and limitation of bandwidth by resonance phenomena [Weaver and Astumian, 1990; Adair, 1991]. Weak steady and time-varying magnetic fields could also be detected above thermal equilibrium noise by ferrimagnetic single-domain particles that have recently been detected in the human brain [Kirschvink et al., 1992]. In addition, the applicability of Eqs. (113.18) and (113.20) is questionable, because living systems are often far from thermal equilibrium. For example, only mitogen-stimulated—and not quiescent—lymphocytes are affected by weak electric and magnetic fields [Yost and Liburdy, 1992]. Also, some molecules inside cells may at times be involved in systematic and guided rather than random thermal motion [Hoffman, 1992].

Several attempts have been made to construct theoretical models that would explain narrowband resonances in biological systems. Experiments to confirm or reject these hypotheses have thus far given ambiguous results.

One theory assumes that ion transfer through cell membranes is affected by cyclotron resonance [Liboff and McLeod, 1988]. It is based on the fact that the cyclotron resonance frequency $\omega_c = 2\pi f_c$ of several physiologically important ions of charge Q and mass m, in the steady magnetic field B_0 of the earth, falls into the ELF range:

$$\omega_c = \frac{QB_0}{m} \tag{113.21}$$

For example, the Ca^{2+} ion has a resonance frequency of 16 Hz in a dc field of 21.0 μT. However, it has been pointed out that the collision frequency in the physiological environment would be very much larger than the cyclotron frequency and would therefore wipe out any resonance motion, that the usually hydrated ions would have a total mass larger than m, and that the energy gain caused by an alternating field of frequency ω_c (as in a cyclotron) would require an orbital radius larger by many orders of magnitude than a typical cell radius [Polk, 1986]. Another theory postulates that the binding of Ca^{2+} to the protein calmodulin (ubiquitous in all vertebrates) should be affected by magnetic fields at frequencies ω_c and ω_c/n (where n is an integer) [Lednev, 1991; Adair, 1992]. This mechanism involves Zeeman splitting at ELF, due to B_0, of infrared vibrational modes that are chemically or thermally excited.

Some experiments showing a resonant cell response at the frequency given by (113.21) could not be replicated, while others were performed only over a very narrow frequency range. Nevertheless, apparently successful attempts have been made to stimulate bone cell proliferation at these field combinations [Fitzsimmons *et al.,* 1991]. The "Orthologic" device (H in Table 113.2) employs an experimentally determined and apparently clinically useful combination of a 20 μT static field with a 40 μT 76.6 Hz alternating field. This frequency lies near the fifth harmonic (76.2 Hz) of the calcium ion resonance frequency given by Eq. (103.21); however, since the device is always used in the presence of the geomagnetic field of the order of 50 μT, the total DC magnetic field parallel to the 76.6 Hz field can have any value between 0 and 70 μT which would essentially eliminate unique determination of a 76.6 Hz "resonance".

To clarify the sequence of biological events that occurs when PEMF signals are applied to developing bone, the following *in vivo* experiment was performed [Aaron et al., 1989]. Twenty-five milligrams of demineralized rat bone matrix in powdered form was implanted along the thoracic musculature of immature rats. This powder recruits cells from the surrounding tissue leading to formation of cartilage within 6 to 10 days; thereafter progressive calcification occurs, leading to formation of fibrous particles by days 12 to 14 and formation of a small bone ("ossicle"). These developments were compared in a large number of paired rats, with equal numbers unexposed and exposed (8 h/day) to the PEMF signal illustrated on Figs. 113.32 and 113.33 (A in Table 113.2). Estimates of the mean electric fields in the exposed tissues give values equal to about one-fourth of those listed on line 1 in Table 113.3. Chemical and histological analysis of ossicles harvested from animals, sacrificed on every second day, showed that exposure to this PEMF signal at the applied level significantly increased both rate and quantity of cartilage formation and enhanced maturation of the subsequent bone. The experimenters concluded that field exposure either enhanced recruitment or proliferation of cartilage precursor cells, increased differentiation of precursor to cartilage cells, or accelerated maturation of cartilage cells.

Getting even closer to fundamental events at the cellular level, both signals A and B (Table 113.2) were used to expose cultured mouse bone cells and mouse skin fibroblasts, as well as explanted mouse pineal cells in organ culture [Luben, 1993]. In all three cases various chemical procedures were employed to examine "beta-adrenergic **receptors**." These are cell surface protein strands that span the cell membrane and emerge from it; they mediate cell response to agents such as epinephrine (= adrenaline) or norepinephrine through so-called **G proteins** which act essentially as molecular amplifiers at the interior surface of the cell. Other G proteins are involved in the response to growth factors. The arrangement of the exposure coils and culture plates was such as to give mean electric fields equal to about one-tenth of the values shown in Table 113.3. Exposures were of 4 h duration. Specific types of G proteins were stimulated by the A signal, others by the B signal. The total number of binding sites on the cells was not affected, but the affinity of the receptors for specific hormones was changed, suggesting a change in receptor conformation. It is interesting to compare this work, which employed peak values of the order of 10^{-2} V/m and time average values not greater than 2 (10^{-3}) V/m, with

other *in vitro* experiments showing effects on enzyme activity at the cell surface by ELF sinusoidal fields between 5 (10^{-4}) V/m and 30 V/m [Blank, 1992]. Related experiments at higher field intensities and the theory of field effects on catalysis [Robertson and Astumian, 1991] also show that electric field action at the exterior of the cell surface can be translated via enzyme-catalyzed chemical reactions to the cell interior.

Defining Terms

Angiogenesis: Formation of blood vessels.

ATPase: An enzyme that converts adenosine triphosphate (ATP) to adenosine diphosphate (ADP); energy released thereby spontaneously in hydrolysis is used to drive an energy requiring reaction such as one producing muscle movement.

Capacitively coupled fields or currents: Fields applied to the affected limb by electrodes touching the skin (the current from the electrodes has both displacement and conduction components).

Chondrogenesis: Formation of cartilage.

Exponential decay time t_d: *Defined by* $B = B_0 \exp(-t/t_d)$

Exponential rise time t_r: *Defined by* $B = B_0[1 - \exp(-t/t_r)]$

G protein: Guanine nucleotide-binding protein, serves to couple receptors to cell membrane-associated enzymes for purposes of signal transduction.

LEET: Low energy emission therapy.

Nonunion: Bone fracture that fails to heal within normally expected period with conventional management.

Osteoblast: A bone-forming cell.

Osteonecrosis: Death of bone within a living vertebrate.

PEMF: Pulsed electromagnetic field.

PRF: Pulsed radio frequency.

Pseudoarthrosis: Formation of a pseudojoint in a broken or not completely formed bone (usually of congenital origin).

Receptor: Large protein molecule that usually protrudes from, and is embedded in, the membrane of a eucaryotic (nucleus-containing) cell. The part of the receptor outside the cell binds only to selected molecules that then cause chemical activity of proteins bound to the end at the cell interior. Activity continues as long as a single molecule (for example, of a hormone, the first "messenger") is bound to the exterior part, and many molecules of a "second messenger" are released on the cell interior.

Streaming potential: Potential difference produced when liquid pressure displaces "counterions" that are normally held by electrostatic forces near ions of the opposite sign embedded in the surface of a stationary material.

Related Topic

35.1 Maxwell Equations

References

R.K. Aaron, D.M. Ciombor, and J. Grant, "Stimulation of experimental endochondral ossification by low-energy pulsing electromagnetic fields," *J. Bone and Mineral Research,* vol. 4, no. 2, pp. 227–233, 1989.

R.K. Aaron and E. Steinberg, "Electrical stimulation of the femoral head," *Seminars in Arthroplasty,* vol. 2, no. 3, pp. 214–224, 1991.

K.R. Adair, "Constraints on biological effects of weak extremely low frequency electromagnetic fields," *Physical Review A,* pp. 1039–1048, 1991.

R.K. Adair, "Criticism of Lednev's mechanism for the influence of weak magnetic fields on biological systems," *Bioelectromagnetics,* vol. 13, pp. 231–235, 1992.

C.A.L. Bassett and R.O. Becker, "Generation of electric potentials by bone in response to mechanical stress," *Science,* vol. 13, p. 1063, 1963.

C.A.L. Bassett, R.J. Pawluk, and R.O. Becker, "Effects of electric currents on bone *in vivo*," *Nature,* vol. 204, p. 652, 1964.

C.A.L. Bassett, R.J. Pawluck, and A.A. Pilla, "Augmentation of bone repair by inductively coupled electromagnetic fields," *Science,* vol. 184, pp. 575–577, 1974.

C.A.L. Bassett, "Biology of fracture repair, nonunion and pseudoarthrosis," in *Compilations of Fracture Management,* H.R. Gossling and S.L. Pillsbury (eds.), New York: J. B. Lippincott, 1984, pp. 1–8.

R.H.C. Bentall, "Low level pulsed radiofrequency fields and the treatment of soft-tissue injuries," *Bioelectrochemistry and Bioenergetics,* vol. 16, pp. 531–548, 1986.

A. Binder, G. Parr, B. Hazelman, and S. Fitton-Jackson, "Pulsed electromagnetic field therapy of persistent rotator cuff tendinitis. A double-blind controlled clinical assessment," *Lancet,* pp. 695–698, March 31, 1984.

M. Blank, "Na, K-ATPase function in alternating electric fields," *FASEB Journal,* vol. 6, pp. 2434–2438, April 1992.

M. Blank and L. Soo, "Threshold for inhibiting of Na,K-ATPase by ELF alternating currents", *Bioelectromagnetics,* 13, 329–333, 1992.

C.T. Brighton, Z.B. Friedenberg, and J. Black, "Evaluation of the use of constant direct current in the treatment of nonunion," in *Electrical Properties of Bone and Cartilage,* C.T. Brighton, J. Black, and S.R. Pollack (eds.), New York: Grune and Stratton, 1979, pp. 519–545.

D.J. Canaday and R.C. Lee, "Scientific basis for clinical applications of electric fields in soft tissue repair," in *Electromagnetics in Medicine and Biology,* C.T. Brighton and S.R. Pollack (eds.), San Francisco: San Francisco Press, 1991, pp. 275–280.

D.M. Ciombor, R. Aaron, H. Fisher, C. Polk, D. Gautreau, and D. Cherlin, "Effect of 15 Hz Sinusoidal Magnetic Field on Cartilage Development *In Vivo* Depends Non-Linearly on Duration of Daily Stimulation," *Project Resumes, The Annual Review of Research on Biological Effects of 50 and 60 Hz Electric and Magnetic Field,* U.S. Dept. of Energy, Office of Energy Management, 1991, p. P-8.

A. Einstein, in *Investigation on the Theory of Brownian Movement,* M.R. Furth and A.D. Cowper, (eds.), New York: Dover Publications, 1956 (originally published 1905 in German), p. 33.

W. Ellis, "Pulsed subcutaneous electrical stimulation in spinal cord injury," *Bioelectromagnetics,* vol. 8, pp. 159–164, 1987.

R.J. Fitzsimmons, J. Farley, W.R. Adey, and D.J. Baylink, "Embrionic bone matrix formation is increased after exposure to a low amplitude capacitively coupled electric field *in vitro*," *Biochimica et Biophysica Acta,* vol. 882, pp. 51–56, 1986.

R. Fitzsimmons, D. Baylink, F.P. Magee, and A.M. Weinstein, "Electromagnetic field stimulated bone cell proliferation" (Abstract), *J. Bone and Mineral Research,* vol. 6 (Supplement 1), August 1991.

E. Fukada and I. Yasuda, "On the piezoelectric effect in bone," *J. Phys. Soc. Japan,* vol. 12, p. 1158, 1957.

A.J. Ginsberg, "Ultrashort radiowaves as a therapeutic agent", *Med. Record,* 140, 651–653, 1934.

H.R. Gossling, R.A. Bernstein, and J. Abbott, "Treatment of ununited tibial fractures, a comparison of surgery and pulsed electromagnetic fields (PEMF)," *Orthopaedics,* vol. 15, no. 6, pp. 711–719, 1992.

D. Gross and W.S. Williams, "Streaming potential and the electromechanical response of physiologically moist bone," *J. Biomechanics,* vol. 15, pp. 227–295, 1982.

M. Hoffman, "Motor molecules on the move," *Science,* vol. 256, pp. 1758–1760, June 26, 1992.

M. Ieran, S. Zaffuto, M. Bagnacani, M. Annovi, A. Moratti, and R. Cadossi, "Effect of low frequency pulsing electromagnetic fields on skin ulcers of venous origin in humans: A double-blind study," *J. Orthop. Research,* vol. 8, no. 2, pp. 276–282, 1990.

H. Ito and C.A.L. Bassett, "Effect of weak, pulsing electromagnetic fields on neural regeneration in the rat," *Clinical Orthopaedics,* vol. 181, pp. 283–290, 1983.

J.L. Kirschvink, A. Kobayashi-Kirschvink, and B.J. Woodford, "Magnetic biomineralization in the human brain," *Proc. Natl. Acad. Sci. USA,* vol. 89, 1992.

L.C. Kloth and M.C. Ziskin, "Diathermy and pulsed radio frequency radiation", chapt. 8 in *Thermal Agents in Rehabilitation,* 3rd ed., S.L. Michlovitz, Ed. Philadelphia: F.A. Davis Co., 1996.

J. Kort, H. Ito, and C.A.L. Bassett, "Effects of pulsing electromagnetic fields on peripheral nerve regeneration," *J. Bone Joint. Surg. Orthop. Trans.,* vol. 4, p. 238, 1980.

W. Kraus, "Magnetfeld Therapie und magnetisch induzierte Elektrostimulation in der Orthopädie," *Orthopäde,* vol. 13, pp. 78–92, 1984.

L.S. Lavine and A.J. Grodzinsky, "Electrical stimulation of repair of bone," *J. Bone Joint Surgery,* vol. 69, no. 4, pp. 626–630, 1987.

V.V. Lednev, "Possible mechanism for the influence of weak magnetic fields on biological systems," *Bioelectromagnetics,* vol. 12, pp. 71–75, 1991.

A.R. Liboff and B.R. McLeod, "Kinetics of channelized membrane ions in magnetic fields," *Bioelectromagnetics,* vol. 9, pp. 39–51, 1988.

R.P. Liburdy, T.R. Sloma, R. Sokolic, and P. Yaswen, "ELF magnetic fields, breast cancer, and melatonin: 60 Hz fields block melatonin's oncostatic action on ER+ breast cancer cell proliferation", *J. Pineal Research,* 14, 89–97.

R.A. Luben, "Effects of low energy electromagnetic fields on signal transduction by G protein linked receptors," in *Electricity and Magnetism in Biology and Medicine,* San Francisco: San Francisco Press, pp. 57–62, 1993.

E.A. Luce and G.C. Bryant, "Dose-response of electromagnetic field current in rat skin flap survival," *Trans. Bioelectrical Repair and Growth Society,* vol. 6, p. 72, 1986.

M.S. Markov, "Electric current and electromagnetic field effects on soft tissues: Implications for skin and wound healing", *Wounds,* 15(3), 94–110, 1995.

M.S. Markov and A.A. Pilla, "Electromagnetic field stimulation of soft tissues: Pulsed radio frequency treatment of post-operative pain and edema", *Wounds,* 7(4), 143–151, 1995.

K.J. McLeod and C.T. Rubin, "Frequency specific modulation of bone adaptation by induced electric fields," *J. Theor. Biol.,* vol. 145, pp. 385–396, 1990.

K.J. McLeod, H.J. Donahue, P.E. Levin, M.-A. Fontaine, and C.T. Rubin, "Electric fields modulate bone cell function in a density dependent manner", *J. Bone Mineral Res.,* 8(8), 977–984, 1993.

J. Nerubay, B. Marganit, J.J. Bubis, A. Tadmar, and A. Katznelson, "Stimulation of bone formation by electrical current on spinal fusion," *Spine,* vol. 11, p. 167, 1986.

C. Polk, "Physical mechanisms by which low-frequency magnetic fields can affect the distribution of counterions on cylindrical biological cell surfaces," *J. Biol. Phys,* vol. 14, pp. 3–8, 1986.

C. Polk and J.H. Song, "Electric fields induced by low frequency magnetic fields in inhomogeneous biological structures that are surrounded by an electric insulator," *Bioelectromagnetics,* vol. 11, pp. 235–249, 1990.

C. Polk, "Dosimetric extrapolations across biological systems: dosimetry of ELF magnetic fields," *Bioelectromagnetics,* vol. 13 (S1), 1992.

C. Polk, "Introduction", Fig. 11, p 5 in *Biological Effects of Electromagnetic Fields,* 2nd ed., C. Polk and E. Postow, Eds., Boca Raton, FL: CRC Press, 1995.

S.R. Pollack and C.T. Brighton, "Dosimetry in electrical stimulation," *Trans. Bioelectric Repair and Growth Society,* vol. IX, p. 40, 1989.

M. Reite, L. Higgs, J.-P. Lebet, A. Barbault, C. Rossel, N. Kuster, U. Dafni, D. Amato, and B. Pasche, "Sleep inducing effect of low energy emission therapy", *Bioelectromagnetics,* 15, 67–75, 1994.

B. Robertson and R.D. Astumian, "Frequency dependence of catalyzed reactions in a weak oscillating field," *J. Chem. Phys.,* vol. 94, pp. 7414–7419, 1991.

C.T. Rubin, K.J. McLeod, and L.E. Lanyon, "Prevention of osteoporosis by pulsed electromagnetic fields," *J. Bone and Joint Surgery,* vol. 71A, no. 3, pp. 411–418, 1989.

B.F. Sisken and E. Herbst, "Wound healing: electrical and electromagnetic fields," *Proc. 12th Ann. Intern. Conf. IEEE Engineering in Medicine and Biology Society,* vol. 4, no. 5, p. 1533, 1990.

B.F. Sisken, M. Kanje, G. Lundborg, and W. Kurtz, "Pulsed electromagnetic fields stimulate nerve regeneration *in vitro* and *in vivo*," *Restorative Neurology and Neuroscience,* vol. 1, pp. 303–309, 1990.

A.M.J. Van Amelsfort, *An Analytical Algorithm for Solving Inhomogeneous Electromagnetic Boundary-Value Problems for a Set of Coaxial Circular Cylinders,* Eindhoven, The Netherlands: James Clerk Maxwell Foundation, 1990.

J.C. Weaver and R.D. Astumian, "The response of living cells to very weak electric fields: the thermal noise limit," *Science,* vol. 247, pp. 459–562, January 26, 1990.

M.G. Yost and R.P. Liburdy, "Time-varying and static magnetic fields act in combination to alter calcium signal transduction in the lymphocyte," *FEBS,* vol. 296, no. 2, pp. 117–122, 1992.

Further Information

W.R. Adey, "Electromagnetic fields, cell membrane amplification, and cancer promotion," in *Extremely Low Frequency Electromagnetic Fields: The Question of Cancer,* B.W. Wilson, R.G. Stevens, and L. E. Anderson (eds.), Columbus, Ohio: Battelle Press, 1990, pp. 211–249.

C.A.L. Bassett, "Bioelectromagnetics in service of medicine," *Bioelectromagnetics,* 13:7–17, 1992.

Bioelectromagnetics, the bi-monthly (formerly quarterly) journal of the Bioelectromagnetics Society. Published by Wiley-Liss, New York, since 1980.

J. Black, *Electrical Stimulation,* New York: Praeger, 1987.

M. Blank (ed.), *Electricity and Magnetism in Biology and Medicine,* San Francisco: San Francisco Press, 1993.

C. Branden and J. Tooze, *Introduction to Protein Structure,* New York: Garland Publishing, 1991.

C.T. Brighton, J. Black, and S. Pollack (eds.), *Electrical Properties of Bone and Cartilage,* New York: Grune and Stratton, 1979.

C.T. Brighton and S.R. Pollack (eds.), *Electromagnetics in Medicine and Biology,* San Francisco: San Francisco Press, 1991.

C. Polk and E. Postow, *CRC Handbook of Biological Effects of Electromagnetic Fields,* (second edition) Boca Raton, Fla.: CRC Press, 1996.

Transactions of the Bioelectric Repair and Growth Society. Vols. I through XI. Published annually since 1980 by the Bioelectric Repair and Growth Society, Dresher, Pennsylvania.

Neuman, M.R. "Biomedical Sensors"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

114

Biomedical Sensors

114.1 Introduction
114.2 Physical Sensors
114.3 Chemical Sensors
114.4 Bioanalytical Sensors
114.5 Applications
114.6 Summary

Michael R. Neuman
Case Western Reserve University

114.1 Introduction

Any instrumentation system can be described as having three fundamental components: a sensor, a signal processor, and a display and/or storage device. Although all these components of the instrumentation system are important, the sensor serves a special function in that it interfaces the instrument with the system being measured. In the case of biomedical instrumentation a **biomedical sensor** (which in some cases may be referred to as a biosensor) is the interface between the electronic instrument and the biologic system. There are some general concerns that are very important for any sensor in an instrumentation system regarding its ability to effectively carry out the interface function. These concerns are especially important for biomedical sensors, since the sensor can affect the system being measured and the system can affect the sensor performance. Sensors must be designed so that they minimize their interaction with the biologic host. It is important that the presence of the sensor does not affect the variable being measured in the vicinity of the sensor as a result of the interaction between the sensor and the biologic system. If the sensor is placed in a living organism, that organism will probably recognize the sensor as a foreign body and react to it. This may in fact change the quantity being sensed in the vicinity of the sensor so that the measurement reflects the foreign body reaction rather than a central characteristic of the host.

Similarly, the biological system can affect the performance of the sensor. The foreign body reaction might cause the host to attempt to break down the materials of the sensor as a way to remove it. This may, in fact, degrade the sensor package so that the sensor can no longer perform in an adequate manner. Even if the foreign body reaction is not strong enough to affect the measurement, just the fact that the sensor is placed in a warm, aqueous environment may cause water to eventually invade the package and degrade the function of the sensor.

Finally, as will be described below, sensors that are implanted in the body are not accessible for calibration. Thus, such sensors must be extremely stable so that frequent calibrations are not necessary.

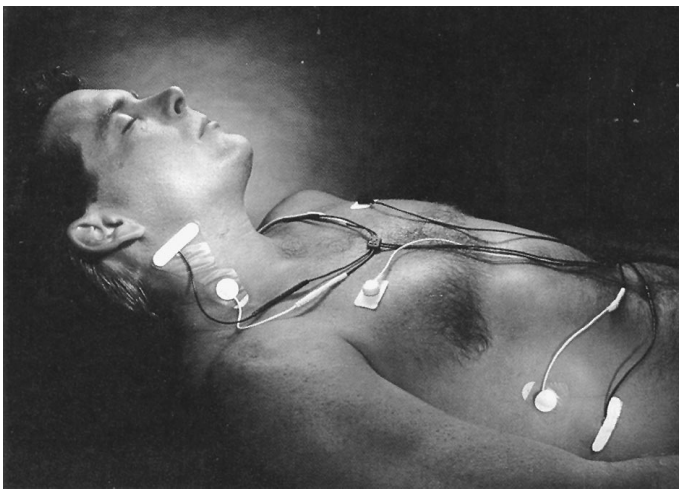
Biomedical sensors can be classified according to how they are used with respect to the biologic system. Table 114.1 shows that sensors can range from noninvasive to invasive as far as the biologic host is concerned. The most noninvasive of biomedical sensors do not even contact the biological system being measured. Sensors of radiant heat or sound energy coming from an organism are examples of noncontacting sensors. **Noninvasive sensors** can also be placed on the body surface. Skin surface thermometers, biopotential electrodes, and strain gauges placed on the skin are examples of noninvasive sensors. Indwelling sensors are those which can be placed into a natural body cavity that communicates with the outside. These are sometimes referred to as minimally invasive sensors and include such familiar sensors as oral-rectal thermometers, intrauterine pressure transducers, and stomach pH sensors. The most invasive sensors are those that need to be surgically placed and that

Cardiac Monitor

The basic method of assessing heart function is thermodilution, a procedure that involves insertion of a catheter into the pulmonary artery and is demanding in terms of cost, equipment, and skilled personnel time. For monitoring astronauts in flight, NASA needed a system that was non-invasive and considerably less complex.

In 1965, Johnson Space Center contracted with the University of Minnesota to explore the then-known but little-developed concept of impedance cardiography (ICG) as a means of astronaut monitoring. A five-year program led to the development of the Minnesota Impedance Cardiograph (MIC), an electronic system for measuring impedance changes across the thorax that would be reflective of cardiac function and blood flow from the heart's left ventricle into the aorta. ICG clearly had broad potential for hospital applications but further development and refinement was needed. A number of research institutions and medical equipment companies launched development of their own ICGs, using MIC technology as a departure point. Among them were Renaissance Technologies, Inc., Newtown, Pennsylvania, and Drexel University of Philadelphia, who jointly developed the IQ System. The system provides a simple, repeatable, non-invasive way of assessing cardiac function at dramatically reduced cost. The IQ System is in wide use in hospital intensive care units, emergency rooms, operating rooms, and laboratories in the U.S. and abroad.

IQ has two basic elements: the non-invasive, disposable patient interface known as IQ-Connect and the touch screen monitor, which calculates and displays cardiac output values and trends. The hardware design of the original MIC was retained but IQ has advanced automated software that features the signal processing technology known as TFD (Time Frequency Distribution). TFD provides three-dimensional distribution of the hemodynamic signals being measured, enabling visualization of the changes in power, frequency, and time. This clinically proven capability allows IQ to measure all cardiac events without using estimation techniques required in some earlier systems. (Courtesy of National Aeronautics and Space Administration.)



The IQ-Connect interface electronically measures impedance changes across the thorax to reflect heart function. (Photo courtesy of National Aeronautics and Space Administration.)

TABLE 114.1 Classification of Biomedical Sensors According to Their Interface with the Biologic Host

Noninvasive	Noncontacting
	Body surface
Invasive	Indwelling
	Implanted

TABLE 114.2 Physical Variables Sensed by Biomedical Sensors

Displacement, velocity, acceleration (linear and angular)
Temperature
Force (weight and mass)
Pressure
Flow
Radiant energy (optical)

involve some tissue damage associated with their installation. For example, a needle electrode for picking up electromyographic signals directly from muscles; a blood pressure sensor placed in an artery, vein, or the heart itself; or a blood flow transducer positioned on a major artery are all examples of invasive sensors.

We can also classify sensors in terms of the quantities that they measure. **Physical sensors** are used in measuring physical quantities such as displacement, pressure, and flow, while **chemical sensors** are used to determine the concentration of chemical substances within the host. A subgroup of the chemical sensors that are concerned with sensing the presence and the concentration of biochemical materials in the host are known as **bioanalytical sensors**, or sometimes they are referred to as biosensors.

In the following paragraphs we will look at each type of sensor and present some examples as well as describe some of the important issues surrounding these types of sensors.

114.2 Physical Sensors

Physical variables associated with biomedical systems are measured by a group of sensors known as physical sensors. A list of typical variables that are frequently measured by these devices is given in Table 114.2. These quantities are similar to physical quantities measured by sensors for nonbiomedical applications, and the devices used for biomedical and nonbiomedical sensing are, therefore, quite similar. There are, however, two principal exceptions: pressure and flow sensors.

The measurement of blood pressure and blood flow in humans and other animals remains a difficult problem in biomedical sensing. Direct blood pressure measurement refers to evaluation of the blood pressure using a sensor that is in contact with the blood being measured or contacts it through an intermediate fluid such as a physiologic saline solution. Direct blood pressure sensors are invasive. Indirect blood pressure measurement involves a sensor that does not actually contact the blood. The most familiar indirect blood pressure measurement is the sphygmomanometer cuff that is usually used in most medical examinations. It is a noninvasive instrument. Until recently, the primary sensor used for direct blood pressure measurement was the unbonded strain gauge pressure transducer shown in Fig. 114.1. The basic principle of this device is that a differential pressure seen across a diaphragm will cause that diaphragm to deflect. This deflection is then measured by a displacement transducer. In the unbonded strain gauge sensor a closed chamber is covered by a flexible diaphragm. This diaphragm is attached to a structure that has four fine gauge wires drawn between it and the chamber walls. A dome with the appropriate hardware for coupling to a pressure source covers the diaphragm on the side opposite the chamber such that when the pressure in the dome exceeds the pressure in the chamber, the diaphragm is deflected into the chamber. This causes two of the fine wires to stretch by a small amount while the other two wires contract by the same amount. The electrical resistance of the wires that are stretched increases while that of the wires that contract decreases. By connecting these wires, or more correctly these unbonded strain gauges, into a Wheatstone bridge circuit, a voltage proportional to the deflection of the diaphragm can be obtained.

In recent years semiconductor technology has been applied to the design of pressure transducers. Silicon strain gauges that are much more sensitive than their wire counterparts are formed on a silicon chip, and micromachining technology is used to form this portion of the chip into a diaphragm with the strain gauges integrated into its surface. This structure is then incorporated into a plastic housing and dome assembly. The entire sensor can be fabricated and sold inexpensively so that disposable, single-use devices can be made. These

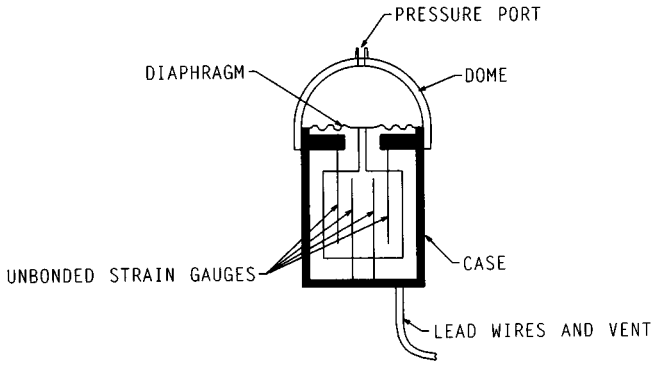


FIGURE 114.1 An unbonded strain gauge pressure transducer.

have the advantage that they are only used on one patient and they do not have to be cleaned and sterilized between patients. By using them on only one patient, the risk of transmitting blood-borne infections is eliminated.

In biomedical applications pressure is generally referenced to atmospheric pressure. Therefore, the pressure in the chamber of the pressure transducer must be maintained at atmospheric pressure. This is done by means of a vent in the chamber wall or a fine bore, flexible capillary tube that couples the chamber to the atmosphere. This tube is usually included in the electrical cable connecting the pressure transducer to the external instrumentation such that the tube is open to the atmosphere at the cable connecter.

In using this sensor to measure blood pressure the dome is coupled to a flexible plastic tube, and the dome and tube are filled with a physiological saline solution.[1] As described by Pascal's Law, the pressure in the dome, and hence against the diaphragm, will be the same as that at the tip of the tube provided the tip of the tube is at the same horizontal level as the dome. Thus by threading the tube into a blood vessel, an invasive procedure, the blood pressure in that vessel can be transmitted to the dome and hence the diaphragm of the pressure transducer. The pressure transducer will, therefore, sense the pressure in the vessel. This technique is known as external direct blood pressure measurement, and the flexible plastic tube that enters the blood vessel is known as a catheter. It is important to remember that the horizontal level of the blood pressure transducer dome must be the same as that of the tip of the catheter in the blood vessel to accurately measure the pressure in that vessel without adding an error due to the hydrostatic pressure in the catheter.

In addition to problems due to hydrostatic pressure differences between the chamber and the dome, catheters introduce pressure errors as a result of the dynamic properties of the catheter, fluid, dome, and diaphragm. These properties as well as air bubbles in the catheter, or obstructions due to clotted blood or other materials, introduce resonances and damping. These problems can be minimized by utilizing miniature pressure transducers fabricated using microelectronic semiconductor technology that are located at the tip of a catheter rather than at the end that is external to the body. A general arrangement for such a pressure transducer is shown in Fig. 114.2. As with the disposable sensors, strain gauges are integrated into the diaphragm of the transducer such that they detect very small deflections of this diaphragm. Because of the small size, small diaphragm displacement, and lack of a catheter with a fluid column, these sensors have a much broader frequency response, give a clearer signal, and do not have any hydrostatic pressure error.

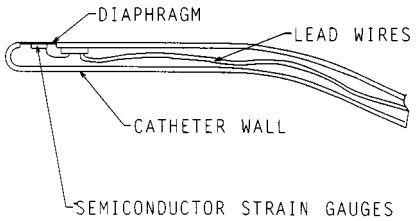


FIGURE 114.2 A catheter tip pressure transducer.

[1]It must be pointed out that the use of such a sensor is not limited to blood pressure measurement. The strain gauge pressure sensor can be used to measure the pressure of any fluid to which it is appropriately coupled.

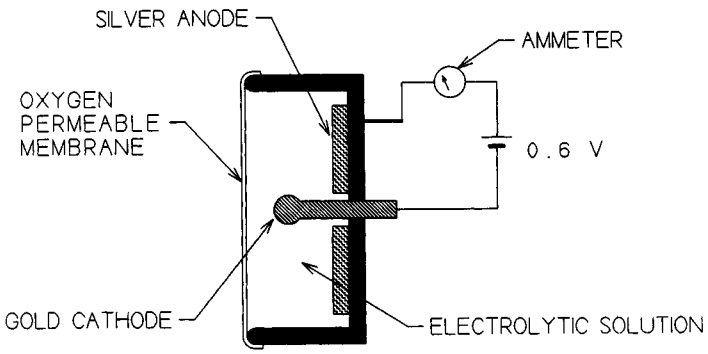


FIGURE 114.3 The Clark electrode, an amperometric electrochemical sensor of oxygen.

Although the indwelling catheter tip pressure transducer appears to solve many of the problems associated with the external pressure transducer, there are still important problems in pressure transducer design that need to be addressed. Long-term stability of pressure transducers is not very good. This is especially problematic for venous pressure measurements which are carried out at relatively low pressure. Long-term changes in baseline pressure require pressure transducers to be frequently adjusted to be certain of zero pressure. While this can be done relatively easily for external and indwelling pressure transducers, there is no way to carry out this procedure for implanted transducers, since there is not a way to establish zero pressure at the sensor. Thus devices that have very low long-term baseline drift are essential for implantable applications.

The packaging of the pressure transducer also represents a problem that needs to be addressed. Packaging must both protect the transducer and be biocompatible. It also must allow the appropriate pressure to be transmitted from the biologic fluid to the diaphragm. The amount of packaging material required should be kept at a minimum so as not to substantially increase the size of implantable or indwelling sensors. Furthermore, the material must be mechanically stable so that it does not swell or contract, since this will most likely change the baseline pressure seen by the sensor. These problems need to be overcome before miniature pressure transducers can be used reliably in implantable applications.

114.3 Chemical Sensors

There are many biomedical problems where it is necessary to know the concentration of a particular substance in a biological sample. Chemical sensors provide the interface between an instrument and the specimen to allow one to determine this concentration. These sensors can be used on a biological specimen taken from the host and tested in a laboratory, or they can be used for *in vivo* measurements either as noninvasive or invasive sensors, the latter being the most frequently used.

TABLE 114.3 Classifications of Chemical Biomedical Sensors

1. Electrochemical
   a. Amperometric
   b. Potentiometric
   c. Coulometric
2. Optical
   a. Colorimetric
   b. Emission and absorption spectroscopy
   c. Fluorescence
   d. Chemiluminescence
3. Thermal methods
   a. Calorimetry
   b. Thermoconductivity
4. Nuclear magnetic resonance

There are many types of chemical sensors used in biomedical instrumentation. Table 114.3 lists some general categories of sensors. Electrochemical and optical sensors are most frequently used for biomedical measurements both *in vivo* and *in vitro*. An example of an electrochemical sensor is the Clark electrode illustrated in Fig. 114.3. This consists of an electrochemical cell separated from the specimen being measured by an oxygen-permeable membrane. The cell is driven at a fixed potential of 600 mV, and under these conditions the following reaction occurs at the noble metal cathode:

$$O_2 + 4e^- + H_2O \rightarrow 4OH^-$$

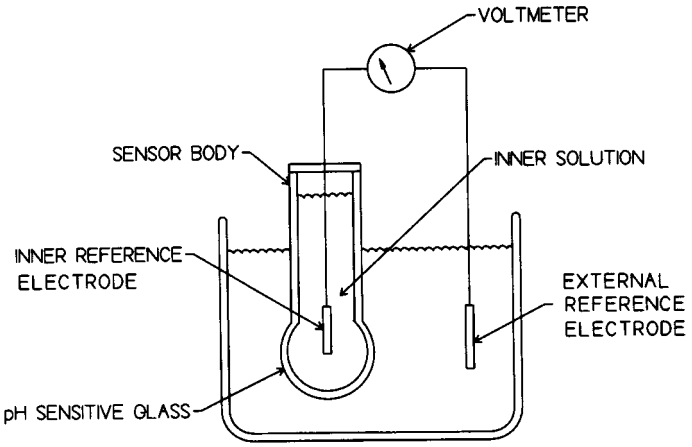


FIGURE 114.4 A glass electrode pH sensor.

This reaction involves the reduction of molecular oxygen that diffuses into the cell through the oxygen-permeable membrane. Since the other components of the reaction are in abundance, the rate of the reaction is limited by the amount of oxygen available. Thus, the rate of electrons used at the cathode is directly related to the available oxygen. In other words, the cathode current is proportional to the partial pressure of oxygen in the specimen being measured.

The electrochemical cell is completed by the silver anode. The reaction at the anode involves forming the low-solubility salt, silver-chloride, from the anode material itself and the chloride ion contained in the electrolyte. The cell is designed so that these materials are also in abundance so that their concentration does not affect the sensor performance. This type of sensor is an example of an **amperometric** electrochemical sensor.

Another type of electrochemical sensor that is frequently used in biomedical laboratories is the glass pH electrode illustrated in Fig. 114.4. The acidity or alkalinity of a solution is characterized by its pH. This quantity is defined as

$$\text{pH} = -\log_{10}[\text{H}^+]$$

where $[\text{H}^+]$ is the activity of the hydrogen ions in solution, a quantity that is related to the concentration of the hydrogen ions. This sensor only works in an aqueous environment. It consists of an inner chamber containing an electrolytic solution of known pH and an outer solution with an unknown pH that is to be measured. The membrane consists of a specially formulated glass that will in essence allow hydrogen ions to pass in either direction but will not pass other chemical species. If the concentration of hydrogen ions in the external solution is greater than that in the internal solution, there will be a gradient forcing hydrogen ions to diffuse through the membrane into the internal solution. This will cause the internal solution to have a greater positive charge than the external solution so that an electrical potential and, hence, an electric field will exist across the membrane. This field will counteract the diffusion of hydrogen ions due to the concentration difference and so an equilibrium will be eventually established. The potential across the membrane at this equilibrium condition will be related to the hydrogen ion concentration difference (or more accurately the activity difference) between the inner and outer solutions. This potential is given by the Nernst equation

$$E = -\frac{RT}{nF}\ln\left(\frac{a_1}{a_2}\right)$$

where E is the potential measured, R is the universal gas constant, T is the absolute temperature, n is the valence of the ion, and a_1 and a_2 are the activities of the ions on each side of the membrane. Thus the potential measured across the glass membrane will be proportional to the pH of the solution being studied. At room temperature

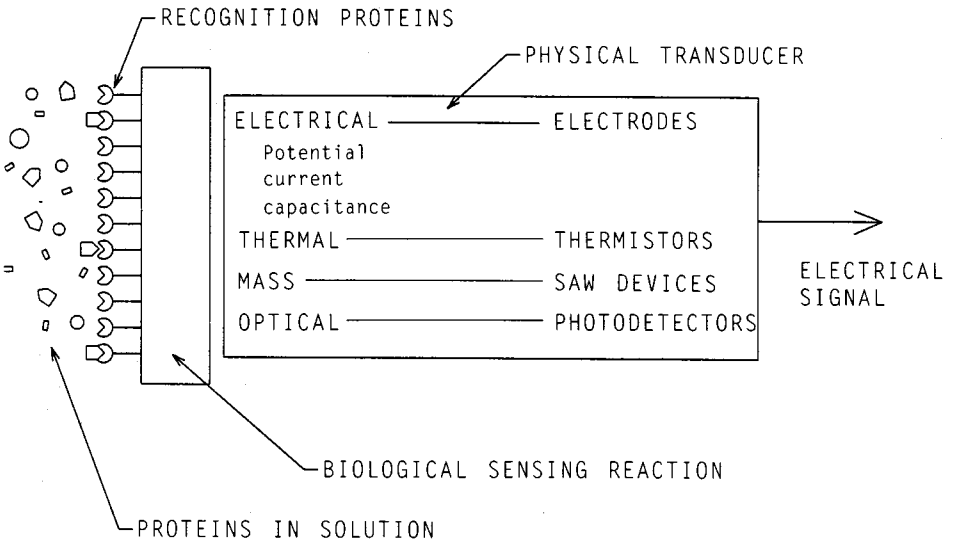


FIGURE 114.5 A generalized bioanalytical sensor.

the theoretical sensitivity of the electrode is approximately 60 mV/pH. It is not practical to measure the potential across the membrane directly and so reference electrodes, sensors that can be used to measure electrical potential of an electrolytic solution, are used to contact the solution on either side of the membrane to measure the potential difference across it. The reference electrodes and the glass membrane are incorporated into the structure shown in Fig. 114.4 known as a glass pH electrode. This is an example of a **potentiometric** measurement made using an ion-selective membrane.

There are other types of ion-selective membrane potentiometric chemical sensors that are used for biomedical applications. The membranes of these sensors determine the ion being sensed. The membrane can be based upon glass or a polymeric material such as polyvinyl chloride, but the key component is the substance that is added to the membrane that allows it to selectively pass a single ion.

Important problems in the development of chemical biomedical sensors are similar to those discussed above for the pressure sensor. Issues of long-term stability and packaging are critical to the success of a chemical sensor. The package is even more critical in chemical sensors than it was in pressure sensors in that the package must protect portions of the sensor that require isolation from the solutions being measured while it provides direct contact of the chemically sensitive portions of the sensor to the solution. The maintenance of a window through the package for this contact represents a critical aspect of sensor development. Frequent calibration is also necessary for chemical sensors. Just about every type of chemical sensor requires some sort of calibration using a standard solution with known concentration of the **analyte** being sensed. The best calibration method is a two-point procedure where two standards are used to establish the slope and the intercept of the calibration line. Some chemical sensors have stable slopes but need to be calibrated in terms of the baseline or intercept. In this case a single-point calibration can be used.

114.4 Bioanalytical Sensors

A special class of sensors of biological molecules has evolved in recent years. These bioanalytical sensors take advantage of one of the following biochemical reactions: (1) enzyme-substrate, (2) antigen-antibody, or (3) ligand-receptor. The advantage of using these reactions in a sensor is that they are highly specific for a particular biological molecule, and sensors with high sensitivity can be developed based upon these reactions. The basic structure of a bioanalytical sensor is shown in Fig. 114.5. There are two principal portions of the sensor. The first contains one component of the biological sensing reaction such as the enzyme or the antibody, and the second component involves a means of detecting whether the biological reaction has taken place. This second portion of a bioanalytical sensor is made up of either a physical or chemical sensor that serves as the detector

of the biological reaction. As illustrated in Fig. 114.5, this detector can consist of an electrical sensor such as used in electrochemical sensors, a thermal sensor, a sensor of changes in capacitance, a sensor of changes in mass, or a sensor of optical properties.

An example of a bioanalytical sensor is a glucose sensor. The first portion of the sensor contains the enzyme glucose oxidase. This enzyme promotes the oxidation of glucose to glucuronic acid and consumes oxygen in the process. Thus, by placing an oxygen sensor along with the glucose oxidase in the bioanalytical sensor, one can determine the amount of glucose oxidized by measuring the amount of oxygen consumed. An even better approach is to have two identical sensor structures in the same package. The only difference is that only one of the sensors contains the enzyme. When there is no glucose present, both sensors will measure the same oxygen partial pressure. The presence of glucose, however, will cause the sensor with the glucose oxidase to have a reduced partial pressure of oxygen due to the oxygen consumption of the reaction. By making a differential measurement of oxygen partial pressure with both sensors, other factors that can cause an apparent change in oxygen partial pressure such as temperature will have a much lower effect than if a single sensor was used.

Stability problems are important for bioanalytical sensors, especially those that are used for long-term measurements. Not only are the stability issues the same as for the physical and chemical sensors, but they are also related to preservation of the biological molecules used in the first stage of the sensor. These molecules can often be degraded or destroyed by heat or exposure to light. Even aging can degrade some of these molecules. Thus, an important issue in dealing with bioanalytical sensors is the preservation of the biochemical components of the sensor. Not all biochemical reactions are entirely reversible, and so the bioanalytical sensors based on them will not be reversible as well. This may be acceptable for some applications but not for others and must be taken into consideration in choosing a bioanalytical sensor.

114.5 Applications

Biomedical sensors and instrumentation are used in biomedical research and patient care applications. In terms of patient care, sensors are used as a part of instruments that carry out patient screening by making measurements such as blood pressure using automated apparatus. Specimen analysis is another important application of biomedical sensors in patient care. This can include analyses that can be carried out by the patients themselves in their homes such as is done with home blood glucose analyzers. Instrumentation based upon biomedical sensors can be used in the physician's office for carrying out some chemical analyses of patient specimens such as urinalysis or elementary blood chemistries such as serum glucose and electrolytes. Sensors also are a part of large multicomponent automatic blood analyzers used in the central clinical laboratory of major medical centers.

Another application for biomedical sensors is in patient monitoring. Sensors represent the front end of critical care monitors used in the intensive care unit and in the operating and recovery rooms. Measurements cover a wide range of biomedical variables such as continuous recordings of blood pressure and transcutaneous measurement of the partial pressure of carbon dioxide in the blood. The performance of these instruments is strongly dependent on biomedical sensors. Patient monitoring can also be carried out in the various clinical units of the hospital. Devices such as ambulatory cardiac monitors that allow patients to be observed while they are free to move around if they desire are becoming important in clinical care in "step-down" units for patients who have completed their stay in the intensive care unit. Patient monitoring has even made its way into the home. Home cardiorespiratory monitors are thought to have some potential value in identifying infants at risk of sudden infant death.

114.6 Summary

Sensors serve an important function in biomedical instrumentation systems in that they provide the interface between the electronic instrument and the biologic system being measured. Very often the quality of the instrument is based upon the quality of the sensor at the instrument's front end. Although electronic signal processing has been developed to a high level, the signals are no better than the quality of the sensors that provide them. Although there have been many advances in biomedical sensor technology, many problems remain. Biomedical sensors will continue to be an import area for research and development in biomedical engineering.

Defining Terms

Amperometric sensor: An electrochemical sensor that determines the amount of a substance by means of an oxidation-reduction reaction involving that substance. Electrons are transferred as a part of the reaction, so that the electrical current through the sensor is related to the amount of the substance seen by the sensor.

Analyte: The substance being measured by a chemical or bioanalytical sensor and instrumentation system.

Bioanalytical sensor: A special case of a chemical sensor for determining the amount of a biochemical substance. This type of sensor usually makes use of one of the following types of biochemical reactions: enzyme-substrate, antigen-antibody, or ligand-receptor.

Biomedical sensor: A device for interfacing an instrumentation system with a biological system such as a biological specimen or an entire organism. The device serves the function of detecting and measuring in a quantitative fashion a physiological property of the biologic system.

Chemical sensor: The interface device for an instrumentation system that determines the concentration of a chemical substance.

Noninvasive sensor: The interface device of an instrumentation system that measures a physiologic variable from an organism without interrupting the integrity of that organism. This device can be in direct contact with the surface of the organism or it can measure the physiologic quantity while remaining remote from the organism.

Physical sensor: An interface device at the input of an instrumentation system that quantitatively measures a physical quantity such as pressure or temperature.

Potentiometric sensor: A chemical sensor that measures the concentration of a substance by determining the electrical potential between a specially prepared surface and a solution containing the substance being measured.

Related Topics

56.1 Introduction • 56.2 Physical Sensors • 56.3 Chemical Sensors • 56.4 Biosensors • 56.5 Microsensors

References

R.S.C. Cobbold, *Transducers for Biomedical Measurements: Principles and Applications,* New York: John Wiley, 1974.

B.R. Eggins, *Biosensors: An Introduction,* Chichester; New York: John Wiley, 1996.

D.G. Fleming, W.H. Ko, and M.R. Neuman, Eds., *Indwelling and Implantable Pressure Transducers,* Cleveland: CRC Press, 1977.

L.A. Geddes, *The Direct and Indirect Measurement of Blood Pressure,* Chicago: Year Book Medical Publishers, 1970.

L.A. Geddes, *Electrodes and the Measurement of Bioelectric Events,* New York: John Wiley, 1972.

W. Göpel, J. Hesse and J.N. Zemel, *Sensors; A Comprehensive Survey,* Weinheim, Germany: VCH Verlagsgesellschaft, 1989.

A.H. Hall, *Biosensors,* Englewood Cliffs, N.J.: Prentice Hall, 1991.

J. Janata, *Principles of Chemical Sensors,* New York: Plenum Press, 1989.

M.R. Neuman, R.P. Buck, V.V. Cosofret, E. Lindner, and C.C. Liu, "Fabricating biomedical sensors with thin-film technoogy, *IEEE Engineering in Medicine and Biology Magazine,* 13, 409–419, 1994.

R. Pallas-Areny and J.G. Webster, *Sensors and Signal Conditioning,* New York: John Wiley, 1991.

J.I. Peterson and G.G. Vurek, "Fiber-optic sensors for biomedical applications," *Science,* vol. 224, pp. 123–127, 1984.

P. Rolfe, "Review of chemical sensors for physiological measurement," *J. Biomed. Eng.,* vol. 10, pp. 138–145, 1988.

J.G. Webster, Ed., *Encyclopedia of Medical Devices and Instrumentation,* New York: John Wiley, 1988.

O.S. Wolfbeis, Ed., *Fiber Optic Chemical Sensors and Biosensors,* Boca Raton, Fla: CRC Press, 1991.

Further Information

Research reports on biomedical sensors appear in many different journals ranging from those that are concerned with clinical medicine through those that are engineering and chemistry oriented. Three journals, however, represent major sources of biomedical sensor papers. These are listed as follows:

The *IEEE Transactions on Biomedical Engineering* is a monthly journal devoted to research papers on biomedical engineering. Papers on biomedical sensors frequently appear, and the February 1986 issue was devoted entirely to the topic of biomedical sensors. For more information or subscriptions, contact IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

The international journal *Medical and Biological Engineering and Computing* is published bi-monthly by the International Federation for Medical and Biological Engineering. This journal also contains frequent reports on biomedical sensors and related topics. Subscription information can be obtained from Peter Peregrinus Ltd., P.O. Box 96, Stevenage, Herts SG12SD, United Kingdom.

The journal *Biomedical Instrumentation and Technology* is published by the Association for the Advancement of Medical Instrumentation. This bimonthly journal has reports on biomedical instrumentation for clinical applications, and these include papers on biomedical sensors. Subscription information can be obtained from Hanley and Belfus, 210 S. 13th Street, Philadelphia, PA 19107.

There are also several scientific meetings that include biomedical sensors. The major meeting in the area is the international conference of the IEEE Engineering in Medicine and Biology Society. An extensive book or CD ROM of extended abstracts for this meeting is published each year by the IEEE. Further information can be obtained by contacting the IEEE at the address listed above.

Bronzino, J.D., Berbari, E.J., Johnson, P.L., Smith, W.M. "Bioelectronics and Instruments"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

115

Bioelectronics and Instruments

Joseph D. Bronzino
Trinity College/Biomedical Allience for Central Connecticut (BOACON)

Edward J. Berbari
Purdue University

Philip L. Johnson
University of Alabama at Birmingham

William M. Smith
University of Alabama at Birmingham

115.1 The Electroencephalogram
The Language of the Brain • Historical Perspective • EEG Recording Techniques • Frequency Analysis of the EEG • Nonlinear Analysis of the EEG • Topographic Mapping

115.2 The Electrocardiograph
Physiology • Instrumentation • Conclusions

115.3 Pacemakers/Implantable Defibrillators
Pacemakers • Implantable Cardioverter Defibrillators

115.1 The Electroencephalogram

Joseph D. Bronzino

Electroencephalograms (EEGs) are recordings of the minute (generally less than 300 μV) electrical potentials produced by the brain. Since 1924, when Hans Berger reported the measurements of rhythmic electrical activity on the human scalp, it has been suggested that these patterns of bioelectrical origin may provide clues regarding the neuronal bases for specific behaviors and has offered great promise to reveal correlations between pathological processes and the electrical activity of specific regions of the brain.

Over the years, EEG analyses have been conducted primarily in clinical settings, to detect gross organic pathologies and the epilepsies, and in research facilities to quantify the central effect of new pharmacological agents. As a result of these efforts, cortical EEG patterns have been shown to be modified by a wide variety of variables including biochemical, metabolic, circulatory, hormonal, neuroelectric, and behavioral factors. In the past, interpretation of the EEG was limited to visual inspection by a trained electroencephalographer capable of distinguishing normal activity from localized or generalized abnormalities of particular types from relatively long EEG records. This approach has left clinicians and researchers alike lost in a sea of EEG paper records. Computer technology has permitted the application of a host of methods to quantify EEG changes. With this in mind, this section provides an introduction to some of the basic concepts underlying the generation of the EEG, a review of the basic approaches used in quantifying alterations in the EEG, and some insights regarding quantitative electrophysiology techniques.

The Language of the Brain

The mass of brain tissue is composed of bundles of nerve cells (neurons) which constitute the fundamental building blocks of the nervous system. Figure 115.1 is a schematic drawing of just such a cell. It consists of three major components: the cell body (or soma), the receptor zone (or dendrites), and the axon, which carries electrical signals from the soma to target sites such as muscles, glands, or other neurons. Numbering approximately 20 billion in each human being, these tiny cells come in a variety of sizes and shapes. Although neurons are anatomically distinct units having no physical continuity between their processes, the axon ends on the soma and the dendrites of other cells in what is called a synapse. Under the microscope this often stands out as a spherical enlargement at the end of the axon to which various names have been given, for example, boutons,

end-plate, or synaptic terminals. This ending does not actually make physical contact with the soma or dendrite but is separated by a narrow cleft (gap) of approximately 100 to 200 Å (10^{-9} m) wide. This is known as the synaptic cleft. Each of these synaptic endings contains a large number of submicroscopic spherical structures (synaptic vesicles) that can be detected only under an electron microscope. These synaptic vesicles, in turn, are essentially "chemical carriers" containing transmitter substance that is released into the synaptic cleft on excitation.

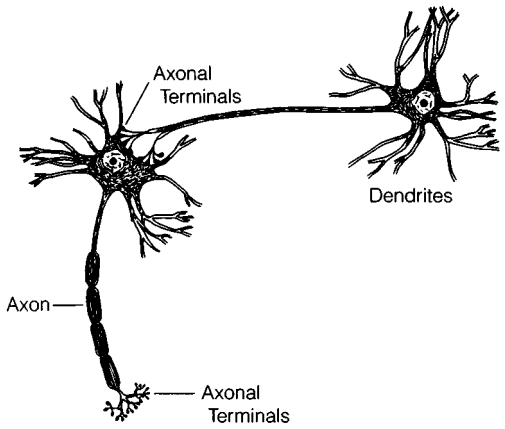


FIGURE 115.1 Basic structure of the neuron.

When an individual neuron is excited, an electrical signal is transmitted along its axon to many tiny branching, diverging fibers near its far end. These axonal terminals end as synapse on a large number of other neurons. When an electrical pulse arrives at the synapse, it triggers the release of a tiny amount of transmitter substance which crosses the synaptic cleft thereby altering the membrane potential of the receiving neuron. If the change is above a certain threshold value, the neuron is activated and generates an action potential of its own which is propagated along its axon, and the process is repeated.

Neurons are involved in every conceivable action taken by the body, whether it is to control its own internal environment or to respond to changes in the external world. As a result, they are responsible for such essential functions as:

- Accepting and converting sensory information into a form that can be processed within the nervous system by other neurons.
- Processing and analyzing this information so that an "integrated portrait" of the incoming data can be obtained.
- Translating the final outcome or "decision" of this analysis process into appropriate electrical or chemical form needed to stimulate glands or activate muscles.

Evolution has played a role in the development of these unique neurons and in the arrangement and development of interconnections between nerve cells in the various parts of the brain. Since the brain is a most complex organ, it contains numerous regions designed for specific tasks. One might, in fact, consider it to be a collection of organs arranged together to act in the harmony of activity we recognize as the individual's state of consciousness or as life itself. Over the years, anatomists and physiologists have identified and named most pathways (tracts), most groups of neurons (nuclei), and most of the major parts of the human brain. Such attention to detail is certainly not necessary here. It will serve our purpose to simply provide a broad overview of the organization of the brain and speak of three general regions: the brainstem, cerebellum, and the cerebral cortex.

The brainstem, or old brain, is really an extension and elaboration of the spinal chord. This section of the brain evolved first and is the location of all the centers that control the regulatory systems, such as respiration, necessary for physical survival of the organism. In addition, all sensory pathways find their way into the brainstem, thereby permitting the integration of complex input patterns to take place within its domain.

Above the brainstem is a spherical mass of neuronal tissue called the cerebellum. This remarkable structure is a complex monitor and modifier of body movements. The cerebellum does not initiate movements, but only modifies motor control activated in other areas. Cerebellar operation is not only dependent on evolutionary development, but relies heavily on actual use and patterns of learned motor behavior acquired throughout life. It is for this reason that the movements of a gymnast are smooth and seemingly effortless.

The most conspicuous part of all in the human brain is the cerebral cortex. Compared to most mammals, it is so large in man that it becomes a covering that surrounds and hides most of the other regions of the brain. Wrinkled and folded, the cerebral tissue is literally pressed into the limited space allocated to it. Although it has been possible to ascertain that certain cortical areas such as visual cortex, the sensory projection area, and the motor strip are associated with specific functions, the overall operation of this complex structure is still

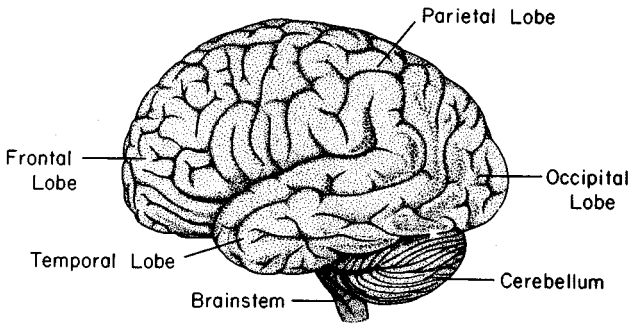


FIGURE 115.2 Major divisions of the cerebral cortex.

not completely understood. However, for the sake of convenience, it has been arbitrarily divided (based primarily on anatomical considerations) into the following areas: frontal lobe, parietal lobe, temporal lobe, and occipital lobe (Fig. 115.2). Each of these segments of the cortex, which is the source of intellectual and imaginative capacities, includes millions of neurons and a host of interconnections.

It is generally agreed that brain function is based on the organization of the activity of large numbers of neurons into coherent patterns. Since the primary mode of activity of these nerve cells is electrical in nature, it is not surprising that a composite of this activity can be detected in the form of electrical signals. Of extreme interest, then, are the actual oscillations, rhythms, and patterns seen in the cryptic flow of electrical energy coming from the brain itself, i.e., in the EEG.

Historical Perspective

In 1875, Caton published the initial account of the recording of the spontaneous electrical activity of the brain from the cerebral cortex of an experimental animal. The amplitude of these electrical oscillations was so low, that is, on the order of microvolts, that Caton's discovery is all the more amazing because it was made 50 years before suitable electronic amplifiers became available. In 1924, Hans Berger, of the University of Jena in Austria, carried out the first human EEG recordings using electrical metal strips pasted to the scalps of his subjects as electrodes and a sensitive galvanometer as the recording instrument. Berger was able to measure the irregular, relatively small electrical potentials (i.e., 50 to 100 μV) coming from the brain. By studying the successive positions of the moving element of the galvanometer recorded on a continuous roll of paper, he was able to observe the resultant patterns in these brain waves as they varied with time. From 1924 to 1938, Berger laid the foundation for many of the present applications of electroencephalography. He was the first to use the word electroencephalogram in describing these brain potentials in man. Berger noted that these brain waves were not entirely random, but instead displayed certain periodicities and regularities. For example, he observed that although these brain waves were slow (i.e., exhibited a synchronized patter of high amplitude and low frequency, <3 Hz) in sleep and states of depressed function, they were faster (i.e., exhibited a desynchronized pattern of low amplitude and high frequency, 15–25 Hz) during waking behavior. He suggested, quite correctly, that the brain's activity changed in a consistent and recognizable fashion when the general status of the subject changed, as from relaxation to alertness. Berger also concluded that these brain waves could be greatly affected by certain pathological conditions after noting the marked increase in the amplitude of these brain waves brought about by convulsive seizures. However, in spite of the insights provided by these studies, Berger's original paper published in 1929 did not excite much attention. In essence, the efforts of this most remarkable pioneer were largely ignored until similar investigations were carried out and verified by British investigators.

It was not until 1934 when Adrian and Matthews published their classic paper verifying Berger's findings that the reality of human brain waves was accepted and EEG studies were put on a firmly established basis. One of their primary contributions was the identification of certain rhythms in the EEG, regular oscillations at approximately 10–12 Hz in the occipital lobes of the cerebral cortex. They found that this alpha rhythm in the EEG would disappear when the brain displayed any type of attention or alertness or focused on objects in the visual field. The physiological basis for these results, the "arousing influence" of external stimuli on the

cortex, was not formulated until 1949 when Moruzzi and Magoun demonstrated the existence of widely spread pathways through the central reticular core of the brainstem capable of exerting a diffuse activating influence on the cerebral cortex. This reticular activating system has been called the brain's response selector because it alerts the cortex to focus on certain incoming information while ignoring other. It is for this reason that a sleeping mother will immediately be awakened by her crying baby or the smell of smoke, and yet ignore the traffic outside her window or the television still playing in the next room. An in-depth discussion of these early studies is beyond the scope of this presentation; however, for the interested reader an excellent historical review of this early era in brain research has been recorded in a fascinating text by Brazier [1968].

EEG Recording Techniques

Scalp recordings of spontaneous neuronal activity in the brain, identified as the EEG, allow measurement of potential changes over time between a signal electrode and a reference electrode [Kondraski, 1986]. Compared to other biopotentials, such as the electrocardiogram, the EEG is extremely difficult for an untrained observer to interpret. As might be expected, partially as a result of the spatial mapping of functions onto different regions of the brain, correspondingly different waveforms are visible, depending on electrode placement. Recognizing that some standardization was necessary for comparison of research as well as clinical EEG records, the International Federation in Electroencephalography and Clinical Neurophysiology adopted the 10–20 electrode placement system, [Jasper, 1958]. Additional electrodes to monitor extracerebral contaminants of the EEG such as eye movement, EKG, and muscle activity are essential. The acquisition of EEG for quantitative analysis should also require the ability to view the EEG during collection on a polygraph or high-resolution video display.

Since amplification, filtering, and digitization determine the frequency characteristics of the EEG and the source of potential artifacts, the acquisition parameters must be chosen with an understanding of their effects on signal acquisition and subsequent analysis. Amplification, for example, increases the amplitude range (volts) of the analog-to-digital (A/D) converter. The resolution of the A/D converter is determined by the smallest amplitude of steps that can be sampled. This is calculated by dividing the voltage range of the A/D converter by 2 to the power of the number of bits of the A/D converter. For example, an A/D converter with a range of ±5 V with 12-bit resolution can resolve samples as small as ±2.4 mV. Appropriate matching of amplification and A/D converter sensitivity permits resolution of the smallest signal while preventing clipping of the largest signal amplitudes.

The bandwidth of the filters and the rate of digitization determine the frequency components of interest that are passed, while other frequencies outside the band of interest that may represent potential artifacts, such as aliasing, are rejected. A filter's characteristics are determined by the rate of the amplitude decrease at the bandwidth's upper and lower edges. Proper digital representation of the analog signal depends on the rate of data sampling, which is governed by the Nyquist theorem that states that data sampling should be at least twice the highest frequency of interest.

In addition to the information available from spontaneous electrical activity of the EEG, the brain's electrical response to sensory stimulation can contribute data as to the status of cortical and subcortical regions activated by sensory input. Due to the relatively small amplitude of a stimulus-evoked potential as compared to the spontaneous EEG potentials, the technique of signal averaging is used to enhance the stimulus-evoked response. Stimulus averaging takes advantage of the fact that the brain's electrical response is time-locked to the onset of the stimulus and the nonevoked background potentials are randomly distributed in time. Consequently, the average of multiple stimulus responses will result in the enhancement of the time-locked activity, while the averaged random background activity will approach zero. The result is an evoked response that consists of a number of discrete and replicable peaks that occur, depending upon the stimulus and the recording parameters, at predicted latencies from the onset of stimulation. The spatial localization of maximum peak amplitudes has been associated with cortical generators in primary sensory cortex.

Instrumentation required for EEG recordings can be simple or elaborate [Kondraski, 1986]. (Note: Although the discussion presented in this section is for a single-channel system it can be extended to simultaneous multichannel recordings simply by multiplying the hardware by the number of channels required. In cases that do not require true simultaneous recordings, special electrode selector panels can minimize hardware requirements.) Any EEG system consists of electrodes, amplifiers (with appropriate filters) and a recording device.

Commonly used scalp electrodes consist of Ag-AgCl disks, 1 to 3 mm in diameter, with a very flexible long lead that can be plugged into an amplifier. Although it is desirable to obtain a low-impedance contact at the electrode ski interface (less than 10 kΩ), this objective is confounded by hair and the difficulty of mechanically stabilizing the electrodes. Conductive electrode paste helps obtain low impedance and keep the electrodes in place. A type of cement (collodion) is used to fix small patches of gauze over electrodes for mechanical stability, and leads are usually taped to the subject to provide some strain relief. Slight abrasion of the skin is sometimes used to obtain better electrode impedances, but this can cause irritation and sometimes infection (as well as pain in sensitive subjects).

For long-term recordings, as in seizure monitoring, electrodes present major problems. Needle electrodes, which must be inserted into the tissue between the surface of the scalp and skull, are sometimes useful. However, the danger of infection increases significantly. Electrodes with self-contained miniature amplifiers are somewhat more tolerant because they provide a low-impedance source to interconnecting leads, but they are expensive. Despite numerous attempts to simplify the electrode application process and to guarantee long-term stability, none has been widely accepted.

Instruments are available for measuring impedance between electrode pairs. The procedure is recommended strongly as good practice, since high impedance leads to distortions that may be difficult to separate from actual EEG signals. In fact, electrode impedance monitors are built into some commercial devices for recording EEGs. Standard dc ohmmeters should not be used, since they apply a polarizing current that causes build-up of noisy electrode potential at the skin-electrode interface. Commercial devices apply a known-amplitude sinusoidal voltage (typically 1 kHz) to an electrode pair circuit and measure root mean square (rms) current, which is directly related to the magnitude of the impedance.

From carefully applied electrodes, signal amplitudes of 1 to 10 μV can be obtained. Considerable amplification (gain = 10^6) is required to bring these levels up to an acceptable level for input to recording devices. Because of long electrode leads and the common electrically noisy environment where recordings take place, differential amplifiers with inherently high input impedance and high common mode rejection ratios are essential for high-quality EEG recordings.

In some facilities, special electrically shielded rooms minimize environmental electrical noise, particularly 60-Hz alternating current (ac) line noise. Since much of the information of interest in the EEG lies in the frequency bands less than 40 Hz, low-pass filters in the amplifier can be switched into attenuate 60-Hz noise sharply.

For attenuating ac noise when the low-pass cutoff is greater than 60 Hz, many EEG amplifiers have notch filters that attenuate only frequencies in a narrow band centered around 60 Hz. Since important signal information may also be attenuated, notch filtering should be used as a last resort; one should try to identify and eliminate the source of interference instead.

In trying to identify 60-Hz sources to eliminate or minimize their effect, it is sometimes useful to use a dummy source, such as a fixed 100-kΩ resistor attached to the electrodes. An amplifier output represents only contributions from interfering sources. If noise can be reduced to an acceptable level (at least by a factor of 10 less than EEG signals) under this condition, one is likely to obtain uncontaminated EEG records.

Different types of recording instruments obtain a temporary or permanent record of the EEG. The most common recording device is a pen or chart recorder (usually multichannel) that is an integral part of most commercially available EEG instruments. The bandwidth of clinical EEGs is relatively low (less than 40 Hz) and therefore within the frequency response capabilities of these devices. Recordings are on a long sheet of continuous paper (from a folded stack), fed past the moving pen at one of several selectable constant speeds. The paper speed translates into distance per unit time or cycles per unit time, to allow EEG interpreters to identify different frequency components or patterns within the EEG. Paper speed is selected according to the monitoring situation at hand: slow speeds (10 mm/s) for observing the spiking characteristically associated with seizures and faster speeds (up to 120 mm/s) for the presence of individual frequency bands in the EEG.

In addition to (or instead of) a pen recorder, the EEG may be recorded on a multichannel frequency modulated (FM) analog tape recorder. During such recordings, a visual output device such as an oscilloscope or video display is necessary to allow visual monitoring of signals, so that corrective action (reapplying the electrodes and so on) can take place immediately if necessary.

Sophisticated FM cassette recording and playback systems allow clinicians to review long EEG recordings over a greatly reduced time, compared to that required to flip through stacks of paper or observe recordings as they occur in real time. Such systems take advantage of time compensation schemes, whereby a signal recorded at one speed (speed of the tape moving past the recording head of the cassette drive) is played back at a different, faster speed. The ratio of playback to recording speed is known, so the appropriate correction factor can be applied to played-back data to generate a properly scaled video display. A standard ratio of 60:1 is often used. Thus, a trained clinician can review each minute of real-time EEG in 1 s. The display appears to be scrolled at a high rate horizontally across the display screen. Features of these instruments allow the clinician to freeze a segment of EEG on the display and to slow down or accelerate tape speed from the standard playback as needed. A time mark channel is usually displayed as one of the traces as a convenient reference (vertical "tick" mark displayed at periodic intervals across the screen).

Computers can also be recording devices, digitizing (converting to digital form) one or several amplified EEG channels at a fixed rate. In such sampled data systems, each channel is repeatedly sampled at a fixed time interval (sample interval) and this sample is converted into a binary number representation by an A/D converter. The A/D converter is interfaced to a computer system so that each sample can be saved in the computer's memory. A set of such samples, acquired at a sufficient sampling rate (at least two times the highest frequency component in the sampled signal), is sufficient to represent all the information in the waveform. To ensure that the signal is band-limited, a low-pass filter with a cutoff frequency equal to the highest frequency of interest is used. Since physically realizable filters do not have the ideal characteristics, the sampling rate is usually greater than two times the filter's cutoff frequency. Furthermore, once converted to a digital format, digital filtering techniques can be used.

On-line computer recordings are only practical for short-term recordings or for situations in which the EEG is immediately processed. This limitation is primarily due to storage requirements. For example, a typical sampling rate of 128 Hz yields 128 new samples per second that require storage. For an 8-channel recording, 1,024 samples are acquired per second. A 10-minute recording period yields 614,400 data points. Assuming 8-bit resolution per sample, over 0.5 megabyte (MB) of storage is required to save the 10-minute recording.

Processing can consist of compression for more efficient storage (with associated loss of total information content), as in data record or epoch averaging associated with evoked responses, or feature extraction and subsequent pattern recognition, as in automated spike detection in seizure monitoring.

Frequency Analysis of the EEG

In general, the EEG contains information regarding changes in the electrical potential of the brain obtained from a given set of recording electrodes. These data include the characteristic waveform with its variation in amplitude, frequency, phase, etc. and the occurrence of brief electrical patterns, such as spindles. Any analysis procedure cannot simultaneously provide information regarding all of these variables. Consequently, the selection of any analytic technique will emphasize changes in one particular variable at the expense of the others. This observation is extremely important if one is to properly interpret the results obtained by any analytic technique. In this chapter, special attention is given to frequency analysis of the EEG.

In early attempts to correlate the EEG with behavior, analog frequency analyzers were used to examine single channels of EEG data. Although disappointing, these initial efforts did introduce the utilization of frequency analysis to study gross brain wave activity. Although, **power spectral analysis**, i.e., the magnitude square of Fourier transform, provides a quantitative measure of the frequency distribution of the EEG, it does so as mentioned above, at the expense of other details in the EEG such as the amplitude distribution, as well as the presence of specific patterns in the EEG.

The first systematic application of power spectral analysis by general-purpose computers was reported in 1963 by Walter; however, it was not until the introduction of the **fast Fourier transform (FFT)** by Cooley and Tukey in the early 1970s that machine computation of the EEG became commonplace. Although an individual FFT is ordinarily calculated for a short section of EEG data (e.g., from 1 to 8 s epoch), such segmentation of a signal with subsequent averaging over individual modified periodograms has been shown to provide a consistent estimator of the power spectrum, and an extension of this technique, the compressed spectral array, has been particularly useful for computing EEG spectra over long periods of time. A detailed review of the

development and use of various methods to analyze the EEG is provided by Givens and Redmond [1987].

Figure 115.3 provides an overview of the computational processes involved in performing spectral analysis of the EEG, i.e., including computation of auto and **cross spectra** [Bronzino, 1984]. It is to be noted that the power spectrum is the autocorrellogram, i.e., the correlation of the signal with itself. As a result, the power spectrum provides only magnitude information in the frequency domain; it does not provide any data regarding phase. The power spectrum is computed by:

$$P(f) = \mathrm{Re}^2[X(f)] + \mathrm{Im}^2[X(f)] \tag{115.1}$$

where $X(f)$ is the Fourier transform of the EEG.

Power spectral analysis not only provides a summary of the EEG in a convenient graphic form, but also facilitates statistical analysis of EEG changes which may not be evident on simple inspection of the records. In addition to absolute power derived directly from the power spectrum, other measures calculated from absolute power have been demonstrated to be of value in quantifying various aspects of the EEG. Relative power expresses the percent contribution of each frequency band to the total power and is calculated by dividing the power within a band by the total power across all bands. Relative power has the benefit of reducing the intersubject variance associated with absolute power that arises from intersubject differences in skull and scalp conductance. The disadvantage of relative power is that an increase in one frequency band will be reflected in the calculation by a decrease in other bands; for example, it has been reported that directional shifts between high and low frequencies are associated with changes in cerebral blood flow and metabolism. Power ratios between low (0–7 Hz) and high (10–20 Hz) frequency bands have been demonstrated to be an accurate estimator of changes in cerebral activity during these metabolic changes.

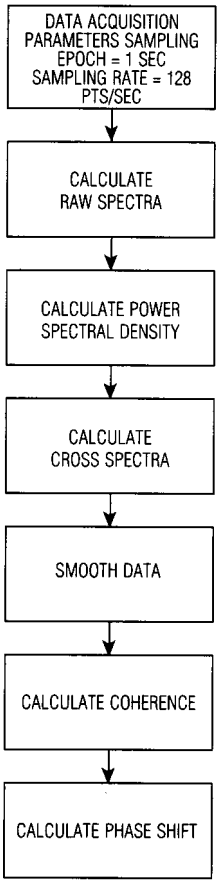


FIGURE 115.3 Block diagram of measures determined from spectral analysis.

Although the power spectrum quantifies activity at each electrode, other variables derivable from FFT offer a measure of the relationship between activity recorded at distinct electrode sites. Coherence (which is a complex number), calculated from the cross-spectrum analysis of two signals, is similar to cross-correlation in the time domain. The **magnitude squared coherence (MSC)** values range from 1 to 0, indicating maximum or no synchrony, respectively, and are independent of power. The temporal relationship between two signals is expressed by phase, which is a measure of the lag between two signals for common frequency components or bands. Phase is expressed in units of degrees, 0° indicating no time lag between signals or 180° if the signals are of opposite polarity. Phase can also be transformed into the time domain, giving a measure of the time difference between two frequencies.

Cross spectrum is computed by:

$$\text{Cross spectrum} = X(f)\ Y^*(f) \tag{115.2}$$

where $X(f)$, $Y(f)$ are Fourier transforms and * indicates complex conjugates and coherence is calculated by

$$\text{Coherence} = \frac{\text{Cross spectrum}}{\sqrt{PX(f) - PY(f)}} \tag{115.3}$$

Since coherence is a complex number, the phase is simply the angle associated with the polar expression of that number. MSC and phase represent measures that can be employed to investigate the cortical interactions of cerebral activity. For example, short (intracortical) and long (cortico-cortical) pathways have been proposed

as the anatomic substrate underlying the spatial frequency and patterns of coherence. Therefore, discrete cortical regions linked by such fiber systems should demonstrate a relatively high degree of synchrony, whereas the time lag between signals, as represented by phase, quantifies the extent to which one signal leads another.

Nonlinear Analysis of the EEG

As mentioned earlier, the EEG has been studied extensively using signal-processing schemes, most of which are based on the assumption that the EEG is a linear, gaussian process. Although linear analysis schemes are computationally efficient and useful, they only utilize information retained in the autocorrelation function (i.e., the second-order cumulant). Additional information stored in higher-order cumulants is therefore ignored by linear analysis of the EEG. Thus, while the power spectrum provides the energy distribution of a stationary process in the frequency domain, it cannot distinguish nonlinearly coupled frequencies from spontaneously generated signals with the same resonance condition [Nikias and Raghvveer, 1987].

There is evidence showing that the amplitude distribution of the EEG often deviates from gaussian behavior. It has been reported, for example, that the EEG of humans involved in the performance of mental arithmetic task exhibits significant nongaussian behavior. In addition, the degree of deviation from gaussian behavior of the EEG has been shown to depend to the behavioral state, with the state of slow-wave sleep showing less gaussian behavior than quiet waking, which is less gaussian than rapid eye movement (REM) sleep [Ning and Bronzino, 1989a,b]. Nonlinear signal-processing algorithms such as bispectral analysis are therefore necessary to address nongaussian and nonlinear behavior of the EEG in order to better describe it in the frequency domain.

But what exactly is the bispectrum? For a zero-mean, stationary process $\{X(k)\}$, the bispectrum, by definition, is the Fourier transform of its third-order cumulant (TOC) sequence:

$$B(\omega_1, \omega_2) = \sum_{m=-\alpha}^{a} \sum_{n=-\alpha}^{a} C(m, n) e^{-j(w_1 m + w_2 n)} \tag{115.4}$$

The TOC sequence $\{C(m, n)\}$ is defined as the expected value of the triple product

$$C(m, n) = E\{X(k)X(k + m)X(k + n)\} \tag{115.5}$$

If process $X(k)$ is purely gaussian, then its third-order cumulant $C(m, n)$ is zero for each (m, n), and consequently, its Fourier transform, the bispectrum, $B(\omega_1, \omega_2)$ is also zero. This property makes the estimated bispectrum an immediate measure describing the degree of deviation from gaussian behavior. In our studies [Ning and Bronzino, 1989a,b], the sum of magnitude of the estimated bispectrum was used as a measure to describe the EEG's deviation from gaussian behavior, that is,

$$D = \sum_{(\omega_1 \omega_2)} |B(\omega_1, \omega_2)| \tag{115.6}$$

Using bispectral analysis, the existence of significant quadratic phase coupling (QPC) in the hippocampal EEG obtained during REM sleep in the adult rat was demonstrated [Ning and Bronzino, 1989a,b, 1990]. The result of this nonlinear coupling is the appearance, in the frequency spectrum, of a small peck centered at approximately 13 to 14 Hz (beta range) that reflects the summation of the two theta frequency (i.e., in the 6- to 7-Hz range) waves. Conventional power spectral (linear) approaches are incapable of distinguishing the fact that this peak results from the interaction of these two generators and is not intrinsic to either.

To examine the phase relationship between nonlinear signals collected at different sites, the *cross-bispectrum* is also a useful tool. For example, given three zero-mean, stationary processes "$x_j(n) j = 1, 2, 3\}$, there are two conventional methods for determining the cross-bispectral relationship, *direct* and *indirect*. Both methods first divide these three processes into M segments of shorter but equal length. The direct method computes the

Fourier transform of each segment for all three processes and then estimates the cross-bispectrum by taking the average of triple products of Fourier coefficients over M segments, that is,

$$B_{x_1x_2x_3}(\omega_1, \omega_2) = \frac{1}{M}\sum_{m=1}^{M} X_1^m(\omega_1)X_2^m(\omega_2)X_3^{m*}(\omega_1 + \omega_2) \tag{115.7}$$

where $X_j^m(\omega)$ is the Fourier transform of the mth segment of $\{x_j(n)\}$, and * indicates the complex conjugate.

The indirect method computes the third-order cross-cumulant sequence for all segments:

$$C_{x_1x_2x_3}^m(k, l) = \sum_{n\epsilon\tau} x_1^m(n)x_2^m(n + k)x_3^m(n + l) \tag{115.8}$$

where τ is the admissible set for argument n. The cross-cumulant sequences of all segments will be averaged to give a resultant estimate:

$$C_{x_1x_2x_3}(k, l) = \frac{1}{M}\sum_{m=1}^{M} C_{x_1x_2x_3}^m(k, l) \tag{115.9}$$

The cross-bispectrum is then estimated by taking the Fourier transform of the third-order cross-cumulant sequence:

$$B_{x_1x_2x_3}(\omega_1, \omega_2) = \sum_{k=-\alpha}^{\alpha}\sum_{l=-\alpha}^{\alpha} C_{x_1x_2x_3}(k, l)^{-j(\omega_1 k+\omega_2 l)} \tag{115.10}$$

Since the variance of the estimated cross-bispectrum is inversely proportional to the length of each segment, computation of the cross-bispectrum for processes of finite data length requires careful consideration of both the length of individual segments and the total number of segments to be used.

The cross-bispectrum can be applied to determine the level of cross-QPC occurring between $\{x_1(n)\}$ and $\{x_2(n)\}$ and its effects on $\{x_3(n)\}$. For example, a peak at $Bx_1x_2x_3(\omega_1, \omega_2)$ suggests that the energy component at frequency $\omega_1 + \omega_2$ of $\{x_3(n)\}$ is generated due to the QPC between frequency ω_1 of $\{x_1(n)\}$ and frequency ω_2 of $\{x_2(n)\}$. In theory, the absence of QPC will generate a flat cross-bispectrum. However, due to the finite data length encountered in practice, peaks may appear in the cross-bispectrum at locations where there is no significant cross-QPC. To avoid improper interpretation, the cross-bicoherence index, which indicates the significance level of cross-QPC, can be computed as follows:

$$bic_{x_1x_2x_3}(\omega_1, \omega_2) = \frac{B_{x_1x_2x_3}(\omega_1, \omega_2)}{\sqrt{P_{x_1}(\omega_1)P_{x_2}(\omega_2)P_{x_3}(\omega_1 + \omega_2)}} \tag{115.11}$$

where $P_{xj}(\omega)$ is the power spectrum of process $\{x_j(n)\}$. The theoretical value of the bicoherence index ranges between 0 and 1, i.e., from nonsignificant to highly significant.

In situations where the interest is the presence of QPC and its effects on $\{x(n)\}$, the cross-bispectrum equations can be modified by replacing $\{x_1(n)\}$ and $\{x_3(n)\}$ with $\{x(n)\}$ and $\{x_2(n)\}$ with $\{y(n)\}$, that is,

$$B_{xyz}(\omega_1, \omega_2) = \frac{1}{M}\sum_{m=1}^{M} X^m(\omega_1)\ Y^m(\omega_2)\ X^{m*}(\omega_1 + \omega_2) \tag{115.12}$$

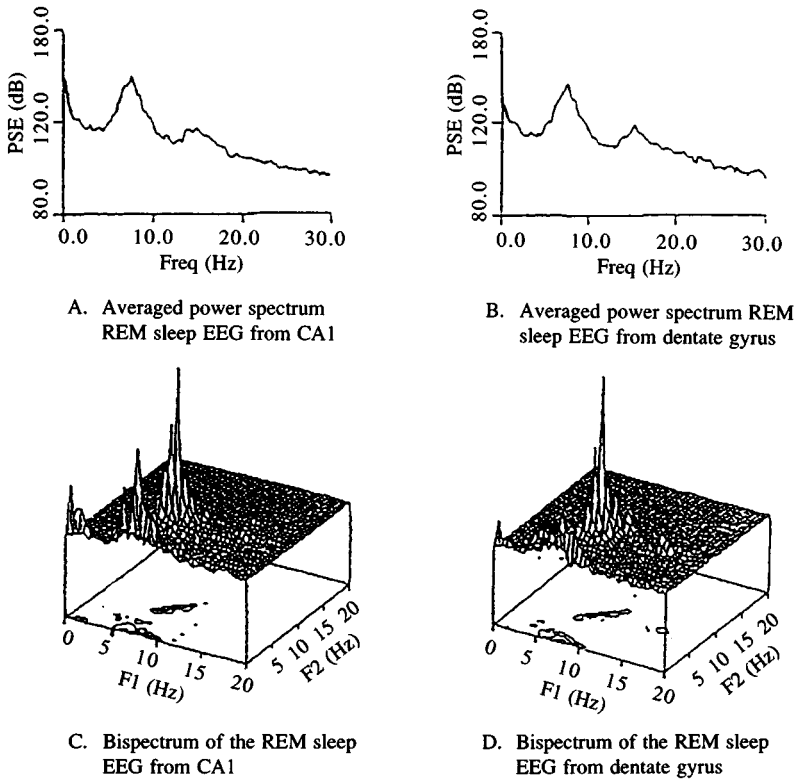


FIGURE 115.4 A and B represent the averaged power spectra of 80 4-s epochs of REM sleep (sampling rate = 128 Hz) obtained from hippocampal CA1 and the dentate gyrus, respectively. Note that both spectra exhibit clear power peaks at 7-Hz (theta) and 14-Hz (beta) frequencies. C and D represent the bispectrum of these same epochs from CA1 and the dentate gyrus, respectively. Computation of the bicoherence index at 7 Hz shows significant quadratic phase coupling at this frequency, indicating that the 14-Hz peak is not spontaneously generated, but results from quadratic phase coupling.

In theory, both methods will lead to the same cross-bispectrum when data length is infinite. However, with finite data records, direct and indirect methods generally lead to cross-bispectrum estimates with different shapes (Fig. 115.4). Therefore, like power spectrum estimation, users have to choose an appropriate method to extract the information desired.

Topographic Mapping

Computerized tomography (CT) and magnetic resonance imaging (MRI) have demonstrated the impact of spatial displays on data interpretation and analysis. Similarly, mapping techniques have been applied to electrophysiologic data to depict the spatial information available from multielectrode recordings. This effort has been assisted by the development and implementation of low-cost, high-resolution graphic displays on microcomputer systems. The data are frequently presented as two-dimensional topographic color maps [Zappulla, 1991]. In the time domain, color values depict the changes in potential across the scalp at each time point. This is exemplified by mapping peaks of an evoked potential or the spatial distribution of an epileptic spike. Temporal changes in the spatial distribution of voltage can be presented graphically as a series of maps constructed at adjacent time points or by cartooning the topographic maps over the time interval of interest. In the frequency domain, color coding can be used to spatially map power, covariance, and phase values. These maps may be constructed for the broadband activity or for selective frequency components.

Unlike CT and MRI displays where each picture element or pixel value represents real data, most of the pixels comprising an EEG and ER topographic map consist of interpolated values. This is because the activity

from a finite number of electrodes represents a sampling of the spatial activity over the scalp. Consequently, the remaining values of the map located outside the electrode positions must be estimated from this sampled activity. One technique for deriving these values is linear interpolation. In the case of a four-point interpolation, the map is divided into boxes whose corners are defined by real data. The interpolated points within the boxes are calculated by the weighted sum of the four real data points, based on their distance from the interpolated point. Although linear interpolation is the most popular technique, polynomial regression and surface spline interpolation have been employed as alternative procedures. These methods reduce the discontinuities inherent in linear interpolation and offer better estimates of extreme values. Polynomial regression has the additional advantage of permitting quantitative comparisons between maps by taking into account the topographic information represented in the map.

Maps can be presented in any of several projections to assist in interpretation [Zappulla, 1991]. The most common projection is the top view which presents the spatial distribution of variables from all leads simultaneously. Lateral, posterior, and anterior projections highlight focal areas of interest. Although mapping presents a method by which spatial information can be efficiently communicated, it is important to be alert to the artifacts that can arise from map construction and manipulation. Topographic spatial artifacts that can lead to misinterpretation include ring enhancement around a spike using source-derivation references, spatial aliasing arising from linear interpolation which causes maximal activity to be mapped at electrode sites, the enhancement of activity away from the midline, and the attenuation of midline activity on amplitude asymmetry maps (centrifugal effect).

The quality of the spatial information derivable from EEG recordings depends upon the number of recording electrodes, the choice of the reference electrode, and the conductive properties of intracranial and extracranial structures. The localization of cortical activity from scalp recordings assumes that the potentials recorded from the scalp reflect cortical activity generated in proximity to the recording electrode. Therefore, the greater the density of recording electrodes, the more accurate the estimate of the spatial distribution of scalp potentials and the localization of cortical generators. However, since the distance between the cortical source and recording electrode, as well as the low conductivity of the skull, results in a selective attenuation of small dipole fields, most available EEG information can be obtained with an average scalp-electrode spacing of 2 cm.

Topographic maps are constructed from monopolar electrodes referenced to a common cephalic (linked ears or mandible, chin and nose) or noncephalic (linked clavicles or a balanced sternum-vertebra) electrode. Although the reference electrode should be free of any EEG activity, in practice most cephalic electrodes contain some EEG activity, while noncephalic electrodes are a potential source of EKG or muscle activity. Differential amplification of an EEG-contaminated reference electrode can decrease or cancel similar activity in neighboring electrodes, while at electrodes distant from the reference, the injected activity will be present as a potential of opposite polarity. Similarly, noncerebral potentials can be injected into scalp electrodes and misinterpreted as cerebral activity. Therefore, a nonneutral reference electrode can result in misleading map configurations. Several techniques have been applied to circumvent this problem. The construction of multiple maps using several different references can sometimes assist in differentiating active and reference electrode activity. This can be accomplished by acquiring serial EEG records using different references. Alternatively, various references can be acquired simultaneously during acquisition, and various montages can be digitally reconstructed, post hoc.

A more computationally intensive method for localizing a source at an electrode involves calculating the local source activity at any one electrode based on the average activity of its neighbors, weighted by their distance from the source. The technique has the advantage of suppressing potentials that originate outside the measurement area and weighing factors for implementing source deviation techniques for each of the electrodes in the 10–20 system are available.

Another reference technique, the average head reference, uses the average activity of all active electrodes as the common reference. In this approach, the activity at any one electrode will vary depending upon the activity at the site of the reference electrode, which can be anywhere on the recording montage. Therefore, for N number of recording electrodes, each being a potential reference, there are $N - 1$ possible voltage measurements at each instant of time for each electrode. Maps constructed using the average head reference represent a unique solution to the problem of active reference electrodes in that the average reference produces an amplitude-weighted reference-free map of maximal and minimal field potentials. Power maps constructed from the average reference

best depict the spatial orientation of the generating field, and the areas with extreme values are closest to the generating processes [Zappulla, 1991].

Topographical maps represent an efficient format for displaying the extensive amount of data generated by quantitative analysis. However, for reasons discussed above, the researcher and clinician must be cautious in deriving spatial and functional conclusions from mapped data. Although the replicability of map configurations across subjects or experimental conditions may represent a useful basis for experimental and diagnostic classification, judgments concerning the localization of cortical generators or functional localization of cerebral activity are less certain and more controversial. Research continues on defining models and validating assumptions that relate scalp potentials to cortical generators in an attempt to arrive at accurate mathematical solutions that can be applied to mapping functions.

Defining Terms

Bispectra: Computation of the frequency distribution of the EEG exhibiting nonlinear behavior.
Cross spectra: Computation of the energy in the frequency distribution of two different electrical signals.
Electroencephalogram (EEG): Recordings of the electrical potentials produced by the brain.
Fast Fourier transform (FFT): Algorithms that permit rapid computation of the Fourier transform of an electrical signal, thereby representing it in the frequency domain.
Magnitude squared coherence (MSC): A measure of the degree of synchrony between two electrical signals at specific frequencies.
Power spectral analysis: Computation of the energy in the frequency distribution of an electrical signal.
Quadratic phase coupling: A measure of the degree to which specific frequencies interact to produce a third frequency.

Related Topic

108.1 Introduction

References

M. Brazier, *Electrical Activity of the Nervous System,* 3rd ed., Baltimore: Williams and Wilkins, 1968.

J.D. Bronzino, M. Kelly, C. Cordova, "Utilization of amplitude histograms to quantify the EEG: Effects of systemic administration of morphine in the chronically implanted rat," *IEEE Trans. Biomed. Eng.,* 28(10), 673, 1981.

J.D. Bronzino, "Quantitative analysis of the EEG: General concepts and animal studies," *IEEE Trans. Biomed. Eng.,* 31(12), 850, 1984.

J.W. Cooley and J.S. Tukey, "An algorithm for the machine calculatio of complex Fourier series," *Math Comput.,* 19, 267, 1965.

A.S. Givens and A. Remond, Eds., "Methods of analysis of brain electrical and magnetic signals," in *EEG Handbook,* vol. 1, Amsterdam: Elsevier, 1987.

S.M. Kay and S.L. Maple, "Spectrum analysis—A modern perspective," *Proc. IEEE.* 69, 1380, 1981.

G.V. Kondraski, "Neurophysiological measurements," in *Biomedical Engineering and Instrumentation,* J.D. Bronzino, Ed., Boston: PWS Publishing, pp. 138–179, 1986.

C.L. Nikias and M.R. Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proc. IEEE,* 75, 869, 1987.

T. Ning and J.D. Bronzino, "Bispectral analysis of the rat EEG during different vigilance states," *IEEE Trans. Biomed. Eng.,* 36(4), 497, 1989a.

T. Ning and J.D. Bronzino, "Bispectral analysis of the EEG in developing rats," in *Proc. Workshop Higher-Order Spectral Anal.,* Vail, Colo.: 1989b, pp. 235–238.

T. Ning and J.D. Bronzino, "Autoregressive and bispectral analysis techniques: EEG applications," *Special Issue on Biomedical Signal Processing, IEEE Eng. Med. Biol. Mag.,* 9, 47, 1990.

J.R. Smith, "Automated analysis of sleep EEG data," in *Clinical Applications of Computer Analysis of EEG and Other Neurophysiological Signals, EEG Handbook,* revised series, vol. 2, Amsterdam: Elsevier, 1986, pp. 93–130.

Further Information

The Biomedical Engineering Handbook, J.D. Bronzino, Ed., Boca Raton, Fla.: CRC Press, 1995.

The Electroencephalogram: Its Patterns and Origins, by J.S. Barlow (Cambridge, Mass., MIT Press, 1993). See also the journals, *IEEE Transactions in Biomedical Engineering and Electroencephalography and Clinical Neurophysiology.*

115.2 The Electrocardiograph

Edward J. Berbari

The electrocardiogram (**ECG**) is the recording on the body surface of the electrical activity generated by the heart. It was originally observed by Waller in 1889 using his pet bulldog as the signal source and the capillary electrometer as the recording device. In 1903 Einthoven enhanced the technology by using the string galvanometer as the recording device and using human subjects with a variety of cardiac abnormalities. Einthoven is chiefly responsible for introducing some concepts still in use today including the labeling of the various waves, defining some of the standard recording sites using the arms and legs, and developing the first theoretical construct whereby the heart is modeled as a single time varying dipole. We also owe the "EKG" acronym to Einthoven's native Dutch language where the root word "cardio" is spelled with a "k".

In order to record an ECG waveform, a differential recording between two points on the body is made. Traditionally each differential recording is referred to as a lead. Einthoven defined three leads numbered with the Roman numerals I, II, and III. They are defined as:

$$\mathrm{I} = V_{LA} - V_{RA} \tag{115.13}$$

$$\mathrm{II} = V_{LL} - V_{RA} \tag{115.14}$$

$$\mathrm{III} = V_{LL} - V_{LA} \tag{115.15}$$

where *RA* = right arm, *LA* = left arm, and *LL* = left leg. Because the body is assumed to be purely resistive, at ECG frequencies, the four limbs can be thought of as wires attached to the torso. Hence lead I could be recorded from the respective shoulders without a loss of cardiac information. Note that these are not independent and the following relationship holds: II = I + III.

For 30 years the evolution of the ECG proceeded when F. N. Wilson [1934] added concepts of a "unipolar" recording. He created a reference point by tying the three limbs together and averaging their potentials so that individual recording sites on the limbs or chest surface would be differentially recorded with the same reference point. Wilson extended the biophysical models to include the concept of the cardiac source enclosed within the volume conductor of the body. He erroneously thought that the central terminal was a true zero potential. However, from the mid-1930s until today the 12 leads composed of the three limb leads, three leads in which the limb potentials are referenced to a modified **Wilson terminal** (the augmented leads [Goldberger, 1942]), and six leads placed across the front of the chest and referenced to the Wilson terminal form the basis of the standard **12-lead ECG**. Figure 115.5 summarizes the 12-lead set. These sites are historically based, have a built in redundancy, and are not optimal for all cardiac events. The voltage difference from any two sites will record an ECG, but it is these standardized sites with the massive 90-year collection of empirical observations that has firmly established their role as the standard. Figure 115.6 is a typical or stylized ECG recording from lead II. Einthoven chose the letters of the alphabet from P to U to label the waves and to avoid conflict with other physiologic waves being studied at the turn of the century. The ECG signals are typically in the range of ±2 mV and require a recording bandwidth of 0.05–150 Hz. Full technical specification for ECG equipment has been proposed by both the American Heart Association [1984] and the Association for the Advancement of Medical Instrumentation [Bailey et al., 1990].

There have been several attempts to change the approach for recording the ECG. The vectorcardiogram used a weighted set of recording sites to form an orthogonal *XYZ* lead set. The advantage here was minimum lead set but in practice it gained only a moderate degree of enthusiasm among physicians. Body surface mapping

Infrared Camera

NASA, teaming with an industry partner, has developed a revolutionary infrared camera that offers important applications not only in aerospace research but in other areas such as air transportation, environment monitoring, and medicine.

An innovative feature of the infrared camera shown is its use of highly sensitive quantum-well photodetectors of QWIPS. The greater sensitivity of long wavelength QWIPS could allow physicians to detect tumors using thermographic techniques; improve pilots' night vision to allow better landings; and enable environmental scientists to monitor pollution and weather patterns with enhanced measurement accuracy. Other possible applications include law enforcement, industrial process control, search and rescue, and military antimissile surveillance.

The camera weighs only 9.9 pounds and measures 4.4 inches wide, 10.3 inches deep, and 7.2 inches long. The prototype plugs into a wall socket for power but the camera can be converted readily to battery power for portability.

Because infrared light detectors must operate at extremely low temperatures, the camera contains a Stirling cryocooler, a closed-cycle refrigerator about the size of a fist that cools the camera from room temperature to about 343 degrees below zero Fahrenheit in about 10 minutes.

The camera was developed by the Center for Space Microelectronics Technology at the Jet Propulsion Laboratory in cooperation with Amber, a Raytheon company. (Courtesy of National Aeronautics and Space Administration.)

This revolutionary infrared camera, developed by an industry/government team, has broad applications in medicine, environment monitoring, industrial processing, and law enforcement, as well as in aerospace research. (Photo courtesy of National Aeronautics and Space Administration.)

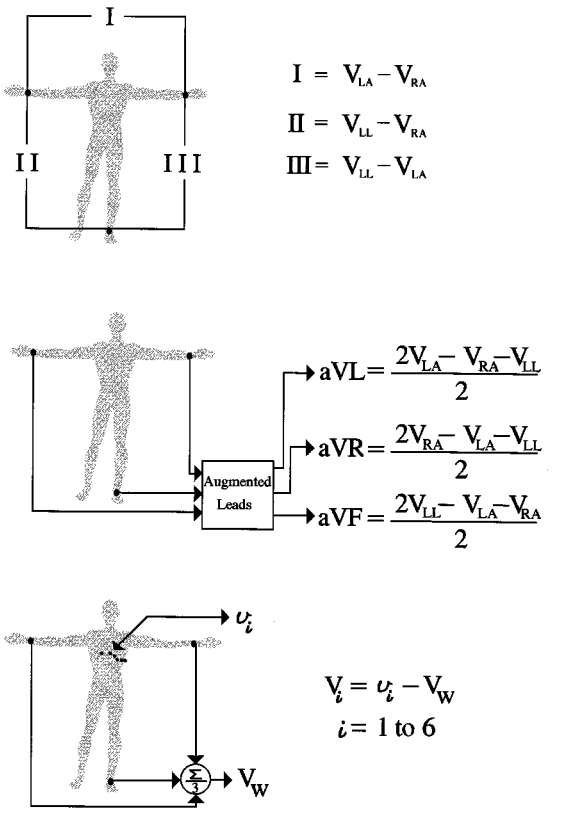


FIGURE 115.5 The 12-lead ECG is formed by the three bipolar surface leads: I, II, and III; the augmented Wilson terminal referenced limb leads: *aVR, aVL, aVF;* and the Wilson terminal referenced chest leads: V_1, V_2, V_3, V_4, V_5, and V_6.

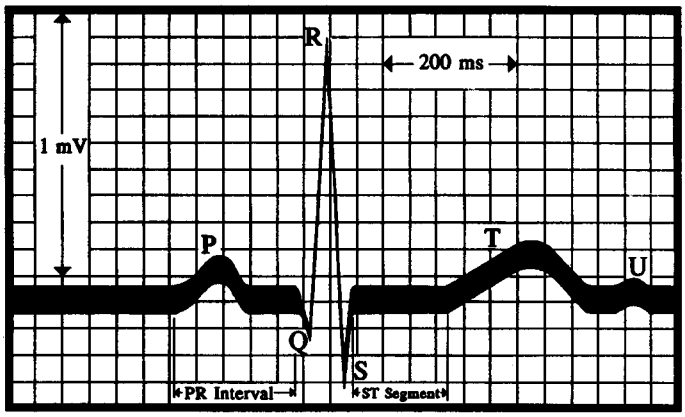


FIGURE 115.6 Stylized version of a normal lead II recording showing the P wave, QRS complex, and the T and U waves. The PR interval and the ST segment are significant time windows. The peak amplitude of the QRS is about 1 mV. The vertical scale is usually 1 mV/cm. The time scale is usually based on mm/s scales with 25 mm/s being the standard form. The small boxes of the ECG are 1×1 mm.

refers to the use of many recording sites (>64) arranged on the body so that isopotential surfaces could be computed and analyzed over time. This approach still has a role in research investigations. Other subsets of the 12-lead ECG are used in limited mode recording situations such as the tape recorded ambulatory ECG

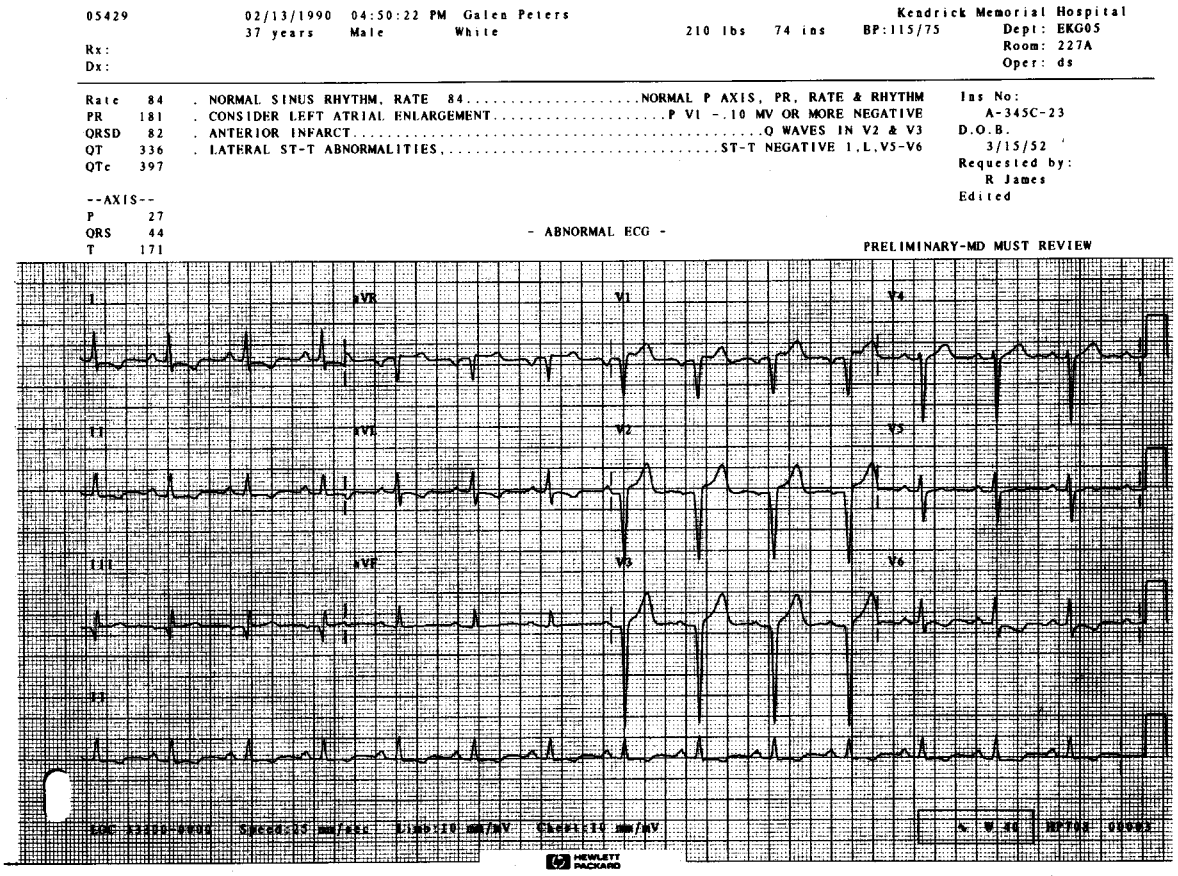


FIGURE 115.7 Example of an interpreted 12-lead ECG. A 2 1⁄2-second recording is shown for each of the 12 leads. The bottom trace is a continuous 10-second rhythm strip of lead II. Patient information is given in the top area, below which is printed the computerized interpretive statements. (Tracing is courtesy of the Hewlett-Packard Co., Palo Alto, CA.)

(usually two leads) or in intensive care monitoring at the bedside (usually one or two leads) or telemetered within regions of the hospital from patients who are not confined to bed (one lead). The recording electronics of these ECG systems have followed the typical evolution of modern instrumentation, e.g., vacuum tubes, transistors, ICs, and microprocessors.

Application of computers to the ECG for machine interpretation was one of the earliest uses of computers in medicine [Jenkins, 1981]. Of primary interest in the computer-based systems was the replacement of the human reader and the elucidation of the standard waves and intervals. Originally this was performed by linking the ECG machine to a centralized computer via phone lines. The modern ECG machine is completely integrated with an analog front end, a 12- to 16-bit A/D converter, a computational microprocessor, and dedicated I/O processors. These systems compute a measurement matrix derived from the 12 lead signals and analyze this matrix with a set of rules to obtain the final set of interpretive statements. The depiction of the 12 analog signals and this set of interpretive statements form the final output, with an example shown in Fig. 115.7. The physician will over-read each ECG and modify or correct those statements which are deemed inappropriate. The larger hospital-based system will record these corrections and maintain a large database of all ECGs accessible by any combination of parameters, e.g., all males, older than 50, with an inferior myocardial infarction.

More recently the high-resolution ECG (**HRECG**) as been developed whereby the digitized ECG is signal averaged to reduce random noise (Berbari et al., 1973, 1977). This approach, coupled with post averaging high-pass filtering, is used to detect and quantify low-level signals (≈1.0 μV) not detectable with standard approaches. This computer-based approach has enabled the recording of events which are predictive of future life-threatening cardiac events [Berbari et al., 1978; Simson, 1981].

Physiology

The heart has four chambers; the upper two chambers are called the atria and the lower two chambers are called the ventricles. The atria are thin-walled, low-pressure pumps which receive blood from venous circulation. Located in the top right atrium are a group of cells which act as the primary pacemaker of the heart. Through a complex change of ionic concentration across the cell membranes (the current source) an extracellular potential field is established which then excites neighboring cells and a cell-to-cell propagation of electrical events occurs. Because the body acts as a purely resistive medium, these potential fields extend to the body surface [Geselowitz, 1989]. The character of the body surface waves depends upon the amount of tissue activating at one time and the relative speed and direction of the activation wavefront. Therefore the pacemaker potentials which are generated by a small tissue mass are not seen on the ECG. As the activation wavefront encounters the increased mass of atrial muscle, the initiation of electrical activity is observed on the body surface and the first ECG wave of the cardiac cycle is seen. This is the P wave and it represents activation of the atria. Conduction of the cardiac impulse proceeds from the atria through a series of specialized cardiac cells (the A-V node and the His-Purkinje system) which again are too small in total mass to generate a signal large enough to be seen on the standard ECG. There is a short relatively isoelectric segment following the P wave. Once the large muscle mass of the ventricles is excited, a rapid and large deflection is seen on the body surface. The excitation of the ventricles causes them to contract and provides the main force for circulating blood to the organs of the body. This large wave appears to have several components. The initial downward deflection is called the Q wave, the initial upward deflection is the R wave, and the terminal downward deflection is the S wave. The polarity and actual presence of these three components depends upon the position of the leads on the body as well as a multitude of abnormalities that may exist. In general, the large ventricular waveform is generically called the QRS complex regardless of its makeup. Following the QRS complex is another short relatively isoelectric segment. After this short segment the ventricles return to their electrical resting state and a wave of repolarization is seen as a low-frequency signal called the T wave. In some individuals a small peak occurs at the end or after the T wave and is called the U wave. Its origin has never been fully established but is believed to be a repolarization potential.

Instrumentation

The general instrumentation requirements for the ECG have been addressed by professional societies through the years [American Heart Association, 1984; Bailey et al., 1990]. Briefly, they recommend a system bandwidth 0.05–150 Hz. Of great importance in ECG diagnosis is the low-frequency response of the system because shifts in some of the low-frequency regions, e.g., the ST segment, have critical diagnostic value. While the heart rate may only have a 1-Hz fundamental frequency, the phase response of typical analog high-pass filters is such that the system corner frequency must be much smaller than the 3-dB corner frequency where only the amplitude response is considered. The system gain depends upon the total system design. The typical ECG amplitude is ± 2mV and if A/D conversion is used in a digital system, then enough gain to span the full range of the A/D converter is appropriate.

To first obtain an ECG the patient must be physically connected to the amplifier front end. The patient/amplifier interface is formed by a special bioelectrode which converts the ionic current flow of the body to the electron flow of the metallic wire. These electrodes typically rely on a chemical paste or gel with a high ionic concentration. This acts as the transducer at the tissue-electrode interface. For short-term applications silver-coated suction electrodes or "sticky" metallic foil electrodes are used. Long-term recordings, such as the case for the monitored patient, require a stable electrode/tissue interface and special adhesive tape material surrounds the gel and a Ag^+/Ag^+Cl electrode.

At any given time, the patient may be connected to a variety of devices, e.g., respirator, blood pressure monitor, temporary pacemaker, etc., some of which will invade the body and provide a low resistance pathway to the heart. It is essential that the device not act as a current source and inject the patient with enough current to stimulate the heart and cause it to fibrillate. Some bias currents are unavoidable for the system input stage and recommendations are that these leakage currents be less than 10 μA per device. This not only applies to the normal setting but if a fault condition arises whereby the patient comes in contact with the high voltage side of the ac power lines, then the isolation must be adequate to prevent 10 μA of fault current as well. This mandates that the ECG reference ground not be connected physically to the low side of the ac power line or

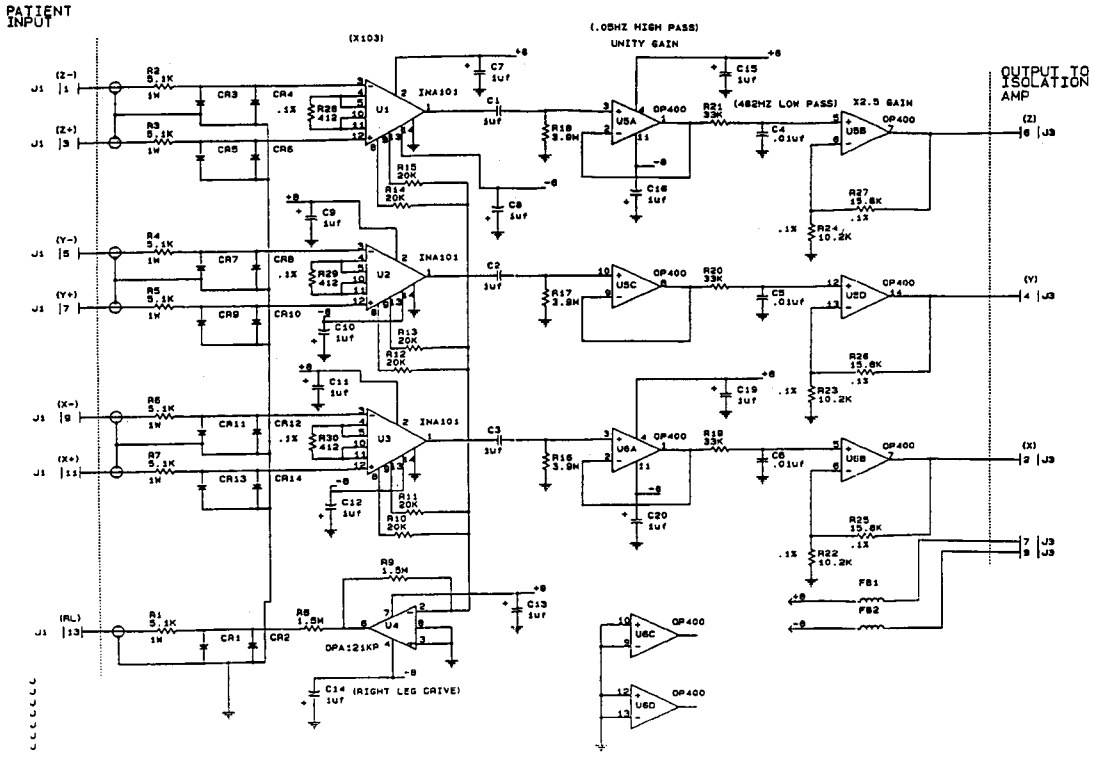


FIGURE 115.8 This schematic represents a typical three-lead *XYZ* amplifier set used in a high-resolution ECG. The instrumentation amplifier (INA101) and bandpass filter (OP400) for each channel are on the isolated side of the power supply. The diode pairs and 4.1 kΩ resistors on each lead wire provide high-voltage defibrillation protection. The outputs of each differential amplifier are averaged through an amplifier (OPA121) and provide the right leg drive. (Schematic is courtesy of Corazonix Corp., Oklahoma City.)

its third wire ground. For ECG machines the solution has typically been to AM modulate a medium-frequency carrier signal (≈400 kHz) and use an isolation transformer with subsequent demodulation. Other methods of signal isolation can be used but the primary reason for the isolation is to keep the patient from being part of the ac circuit in the case of a patient to power line fault. In addition, with many devices connected in a patient monitoring situation it is possible that ground loop currents will be generated. To obviate this potential hazard a low-impedance ground buss is often installed in these rooms and each device chassis will have an external ground wire connected to the buss. Another unique feature of these amplifiers is that they must be able to withstand the high-energy discharge of a cardiac defibrillator.

Figure 115.8 shows a three-channel ECG amplifier schematic used in a high-resolution ECG system. The patient is dc coupled to the front end differential, instrumentation amplifier. The first stage of gain is relatively low (≈100) because there can be a significant signal drift due to a high static charge on the body or low-frequency offset potentials generated by the electrolyte in the tissue/electrode interface. In this particular amplifier the signal is bandpass filtered prior to the isolation stage. To further limit the high floating potential of the patient and to improve the system common mode rejection a driven ground is usually used. This ground is simply an average of the limb potentials inverted by a single amplifier and connected to the right leg.

Older style ECG machines recorded one lead at a time, then evolved to three simultaneous leads. This necessitated the use of switching circuits as well as analog weighting circuits to generate the various 12 leads. This is usually eliminated in modern digital systems by using an individual single-ended amplifier for each electrode on the body. Each potential signal is then digitally converted and all of the ECG leads can be formed mathematically in software. This would necessitate a nine-amplifier system. By performing some of the lead calculations with the analog differential amplifiers this can be reduced to an eight-channel system. Thus only

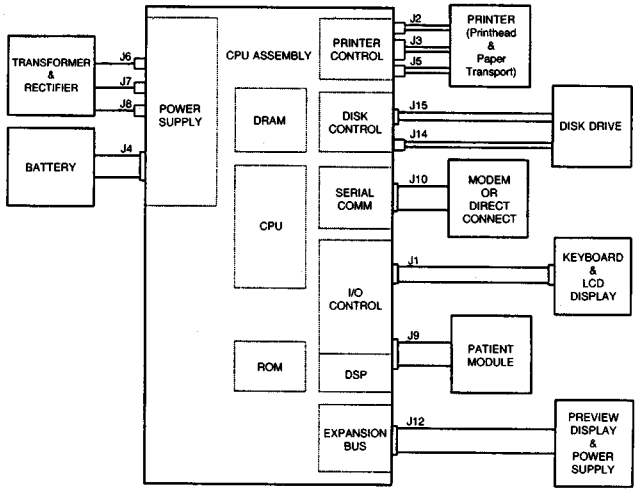


FIGURE 115.9 Block diagram of a microprocessor-based ECG system. It includes all of the elements of a personal computer class system, e.g., 80386 processor, 2 Mbytes of RAM, disk drive, 640 × 480 pixel LCD display, and is battery operable. In addition, it includes a DSP56001 chip and multiple controllers which are managed with a real-time, multitasking operating system. (Diagram is courtesy of the Hewlett-Packard Co., Palo Alto, Calif.)

the individual chest leads V_1 through V_6 and any 2 of the limb leads, e.g., I and III, are needed to calculate the full 12-lead ECG. Figure 115.9 is a block diagram of a modern digital-based ECG system. This system uses up to 13 single-ended amplifiers and a 16-bit A/D converter, all within a small lead wire manifold or amplifier lead stage. The digital signals are optically isolated and sent via a high-speed serial link to the main ECG instrument. Here the 32-bit CPU and DSP chip perform all of the calculations and a hard copy report is generated (Fig. 115.7). Notice that each functional block has its own controller and the system requires a real-time, multitasking operating system to coordinate all system functions. Concomitant with the data acquisition is the automatic interpretation of the ECG. These programs are quite sophisticated and are continually evolving. It is still a medical/legal requirement that these ECGs be over-read by the physician.

High-resolution capability is now a standard feature on most digitally based ECG systems or as a stand-alone microprocessor-based unit [Berbari, 1988]. The most common application of the HRECG is to record very low-level (≈1.0 μV) signals which occur after the QRS complex but are not evident on the standard ECG. These "late potentials" are generated from abnormal regions of the ventricles and have been strongly associated with the substrate responsible for a life-threatening rapid heart rate (ventricular tachycardia). The typical HRECG is derived from three bipolar leads configured in an anatomic *XYZ* coordinate system. These three ECG signals are then digitized at a rate of 1000–2000 Hz/channel, time aligned via a real-time QRS correlator, and summated in the form of a signal average. Signal averaging will theoretically improve the signal-to-noise ratio by the square root of the number of beats averaged. The underlying assumptions are that the signals of interest do not vary, on a beat-to-beat basis, and that the noise is random. Figure 115.10 has four panels depicting the most common sequence for processing the HRECG to measure the late potentials. Panel A depicts a three-second recording of the *XYZ* leads close to normal resolution. Panel B was obtained after averaging 200 beats and with a sampling frequency of 10 times that shown in panel A. The gain is also five times greater. Panel C is the high-pass filtered signal using a partially time reversed digital filter having a second-order Butterworth response and a 3-dB corner frequency of 40 Hz [Simson, 1981]. Note the appearance of the signals at the terminal portion of the QRS complex. A common method of analysis, but necessarily optimal, is to combine the filtered *XYZ* leads into a vector magnitude $(X^2 + Y^2 + Z^2)^{1/2}$. This waveform is shown in panel D. From this waveform several parameters have been derived such as total QRS duration, including late potentials, the rms voltage value of the terminal 40 ms, and the low-amplitude signal (LAS) duration from the 40-μV level to the end of the late potentials. Abnormal values for these parameters are used to identify patients at high risk of ventricular tachycardia following a heart attack.

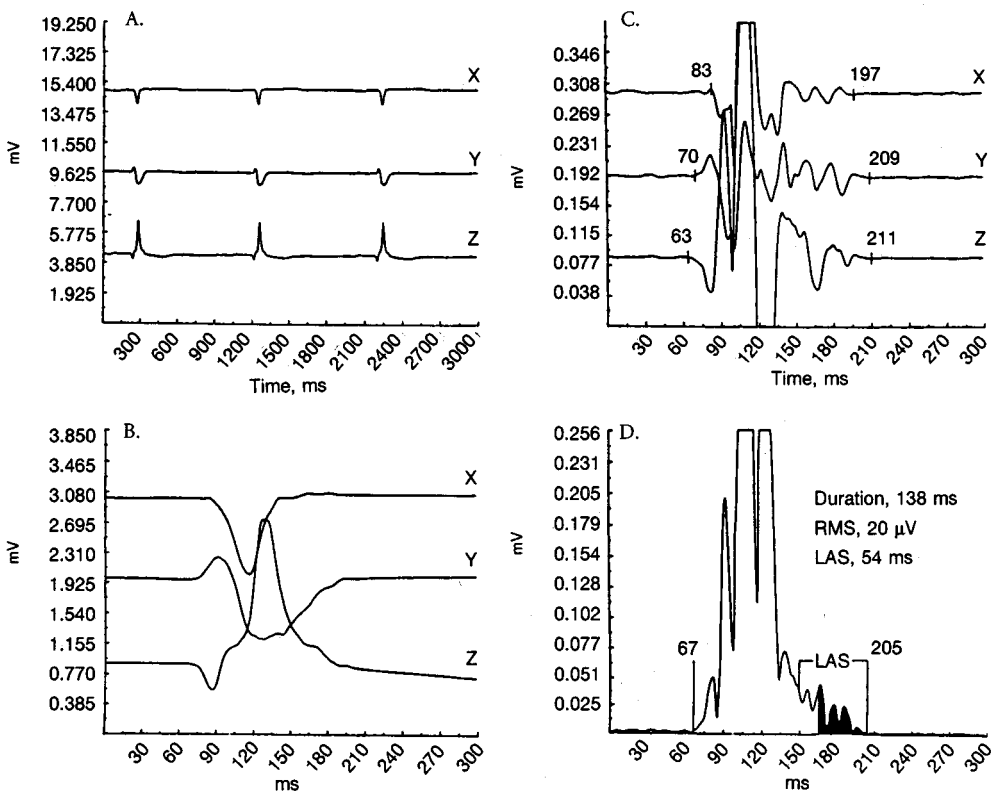


FIGURE 115.10 The signal processing steps typically performed to obtain a high-resolution ECG are shown in panels A–D. See text for a full description.

Conclusions

The ECG is one of the oldest instrument-bound measurements in medicine. It has faithfully followed the progression of instrumentation technology. Its most recent evolutionary step, to the microprocessor-based system, has allowed for an enhanced, high-resolution ECG which has opened new vistas of ECG analysis and interpretation.

Defining Terms

12-lead ECG: Twelve traditional ECG leads comprising the standard set.

ECG: Abbreviation for the device (electrocardiograph) or the output (electrocardiogram) depicting the body surface recording of the electrical activity of the heart.

ECG lead: Differential signal depicting one channel of the ECG record.

HRECG: High-resolution ECG used to detect microvolt-level cardiac potentials most commonly by signal averaging.

Wilson central terminal: Reference point for forming most of the standard ECG leads. It is the average of the right arm, the left arm, and the left potentials. It is a time-varying reference.

Related Topic

108.5 Instrumentation System

References

J.J. Bailey, A.S. Berson, A. Garson, L.G. Horan, P.W. Macfarlane, D.W. Mortara, and C. Zywietz, "Recommendations for standardization and specifications in automated electrocardiography: bandwidth and digital signal processing," A report for health professionals by an ad hoc writing group of the Committee on Electrocardiography and Cardiac Electrophysiology of the Council on Clinical Cardiology, American Heart Association, *Circulation,* vol. 81, no. 2, pp. 730–739, 1990.

E.J. Berbari, "High resolution electrocardiography," *CRC Crit. Rev. Bioeng.,* vol. 16, p. 67, 1988.

E.J. Berbari, R. Lazzara, P. Samet, and B.J. Scherlag, "Noninvasive technique for detection of electrical activity during the PR segment," *Circulation,* vol. 48, p. 1006, 1973.

E.J. Berbari, R. Lazzara, and B.J. Scherlag, "A computerized technique to record new components of the electrocardiogram," *Proc. IEEE,* vol. 65, p. 799, 1977.

E.J. Berbari, B.J. Scherlag, R.R. Hope, and R. Lazzara, "Recording from the body surface of arrhythmogenic ventricular activity during the ST segment," *Am. J. Cardiol.,* vol. 41, p. 697, 1978.

W. Einthoven, "Die galvanometrische Registrirung des menschlichen Elektrokardiogramms, zugleich eine Beurtheilung der Anwendung des Capillar-Elecktrometers in der Physiologie," *Pflugers Arch. Ges. Physiol.,* vol. 99, p. 472, 1903.

D.B. Geselowitz, "On the theory of the electrocardiogram," *Proc. IEEE,* vol. 77, p. 857, 1989.

E. Goldberger, "A simple, indifferent, electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar, extremity leads," *Amer. Heart J.,* vol. 23, p. 483, 1942.

J.M. Jenkins, "Computerized electrocardiography," *CRC Crit. Rev. Bioeng.,* vol. 6, p. 307, 1981.

M.B. Simson, "Use of signals in the terminal QRS complex to identify patients with ventricular tachycardia after myocardial infarction," *Circulation,* vol. 64, p. 235, 1981.

"Voluntary standard for diagnostic electrocardiographic devices," ANSI/AAMI EC11a, Arlington, Va.: Association for the Advancement of Medical Instrumentation, 1984.

A.D. Waller, "On the electromotive changes connected with the beat of the mammalian heart, and the human heart in particular," *Phil. Trans. B.,* vol. 180, p. 169, 1889.

F.N. Wilson, F.S. Johnston, and I.G.W. Hill, "The interpretation of the galvanometric curves obtained when one electrode is distant from the heart and the other near or in contact with the ventricular surface," *Amer. Heart J.,* vol. 10, p. 176, 1934.

Further Information

Comprehensive Electrocardiology: Theory and Practice in Health and Disease, Volumes 1–3, P. W. Macfarlane and T.D. Veitch Lawrie, Eds., England: Pergamon Press, 1989.

High-Resolution Electrocardiography, M.D. Nabil El-Sherif and M.D. Gioia Turitto, Eds., Mount Kisco, N.Y.: Futura Publishing Company, 1992.

Medical Instrumentation: Application and Design, 2nd ed., J. G. Webster, Ed., Boston: Houghton Mifflin, 1992.

115.3 Pacemakers/Implantable Defibrillators

Philip L. Johnson and William M. Smith

The heart is an amazing machine. Throughout an average lifetime, it contracts over 2.5 billion times to pump blood throughout the body. Without its proper function, an individual will die within minutes. The heart consists of four chambers. The upper two chambers, the **atria,** are used as primers for the lower two chambers, the **ventricles,** which serve as the main pump. Blood delivery will be inefficient if the atria and ventricles do not pump in mechanical synchrony (**AV synchrony**). Optimum efficiency occurs when the atria contract slightly before the ventricles. Electrical depolarization waves are responsible for controlling the contractions of the heart and thus maintaining AV synchrony. The depolarization waves originate from a specialized set of cells, known as the **sinus node,** that are modulated by neural input and are located in the top of the right atrium. The sinus

node is the heart's natural pacemaker. It is part of the atrioventricular (AV) conduction system, which serves to distribute the wavefronts throughout the heart and to connect the otherwise electrically isolated atria and ventricles. A normal depolarization wave spreads across the **atria** causing them to contract first, and then, after a brief delay while traversing the AV conduction system, across the ventricles causing them to contract shortly thereafter. With such a demanding, complex organ, it is no wonder there are multiple ways by which it can fail. Failures in the electrical system, known as **arrhythmias**, may impair the contraction sequence and compromise blood flow. These failures are often the result of some underlying heart disease, but may also have a genetic etiology. While antiarrhythmic drugs have been available for some time, contemporary treatment of arrhythmias relies heavily on two types of implantable medical devices: pacemakers and implantable cardioverter defibrillators.

Bradyarrhythmias

Bradyarrhythmias are defined as heart rates that are abnormally slow (<60 b.p.m.) [Katz, 1992]. They are generally caused by either **sinus node** disease or AV conduction disorders. In the former, disease of the body's natural pacemaker cells often results in an unnaturally slow heart rate and significant patient discomfort. Also, the heart rate may not increase in response to exercise due to a loss of neural control of the sinus node, which will inhibit the patient from performing strenuous activities; this is known as chronotropic incompetence. AV conduction disease results from a pathology of the cells that electrically connect the atria and the ventricles. This can result in inefficient blood delivery due to a loss of **AV synchrony.** Pacemakers are commonly used to attempt to restore a natural heart rate, AV synchrony, and chronotropic competence in patients with these and other diseases. Approximately 115,000 pacemakers are implanted in the U.S. every year [Ellenbogen, 1996].

Tachyarrhythmias

Tachyarrhythmias are generally defined as heart rates that are abnormally or inappropriately fast (>100 b.p.m.) [Katz, 1992]. There are many different types of tachyarrhythmias (tachycardias). They are caused by "runaway" depolarization wavefronts that may continue to rapidly activate the same tissue over and over again by a process known as reentry. This can result from a number of underlying physiological problems, such as dying tissue with altered conduction properties due to a blocked coronary artery, around which the wavefront can propagate. The atria and the ventricles can both experience tachycardia. Although somewhat debilitating, atrial tachycardias are not immediately life threatening; however, most ventricular tachycardias are life threatening. The most serious ventricular tachycardia, ventricular **fibrillation,** has been defined as the rapid, disorganized, and asynchronous contraction of ventricular muscle [Epstein and Ideker, 1995] during which the heart's ability to distribute blood to the body is completely compromised. If not immediately treated with a defibrillating electrical shock, loss of life will follow in only minutes. Ventricular fibrillation is the most common cause of sudden cardiac death, of which nearly 400,000 people die annually in the U.S. alone [Gillum, 1989]. The Implantable Cardioverter Defibrillator (ICD) was developed in an attempt to terminate ventricular fibrillation and prevent sudden death from occurring.

Pacemakers

The pacemaker is a medical device capable of controlling the heart rate through a set of implanted electrode leads (Fig. 115.11). The first devices in 1958 were simple, fixed rate oscillators controlled by two transistors [Elmquist and Senning, 1960]. Weighing more than 180 grams, they had a lifetime of around 3 years and paced only the ventricles [Sanders and Lee, 1996]. More sophisticated dual-chamber pacemakers, which sense and pace both the atria and the ventricles independently, were introduced in the late 1970s [Funke, 1982]. Modern pacemakers have shrunk to less than 15 grams (6 ccs) and evolved into sophisticated, implantable computers capable of complex pacing algorithms, telemetry, extensive diagnostics, data storage, and a lifetime greater than 5 years [Sanders and Lee, 1996].

Clinical Indications

Pacemakers are generally indicated for three major disorders, including sinus node disease, AV conduction system disorders, and certain atrial tachyarrhythmias, as well as other less-common pathologies. There are varying degrees of each of these disorders, and the severity of the symptoms and age of the patient may suggest

Figure 115.11 The Guidant/CPI Discovery family of pacemakers. (Courtesy of Guidant/CPI, St. Paul, MN.)

if a pacemaker is warranted. Symptoms may include syncope, dizziness, seizures, heart failure, depression, and dementia. Although bradycardia accounts for the majority of implantations, pacemakers can be used to treat some tachycardias. Antitachycardia pacing is a special type of pacing that may be indicated for atrial or ventricular tachycardias; however, when managing ventricular tachycardias, an ICD is preferable to a pacemaker so that, if the tachycardia degenerates into ventricular **fibrillation,** it can be halted with a **defibrillation** shock. Pacemaker implants may be permanent or temporary.

Surgery

A pacemaker implant is a fairly standard procedure in which the pacing leads are inserted intravenously into the heart and the pacemaker is implanted subcutaneously in a pocket on the chest just below the clavicle. A number of veins have been used to implant the leads, but the cephalic and the subclavian veins are the most common. Access to either of these veins is obtained from the same incision that is used to form the pocket on the chest. A number of tools, such as guidewires, dilator sheaths, and fluoroscopy (real time X-ray), are used to work the lead down the veins into position in the right side of the heart. Depending on the type of pacemaker used, one or two leads may be implanted. In dual-chamber pacemakers, a lead is required in both the **ventricles** and the **atria**. Figure 115.12 shows a diagram of a typical dual-chamber system in the body. The right ventricular lead is usually implanted first and is advanced to the right ventricular apex, where it is fixed. The atrial lead is then advanced to the right atrial appendage, where it is also fixed. A pacing system analyzer is used to test for sufficient electrode contact by measuring impedance, pacing thresholds, and sensing thresholds. The pacing threshold is the smallest charge necessary to stimulate cardiac tissue and initiate a depolarizing wavefront. The

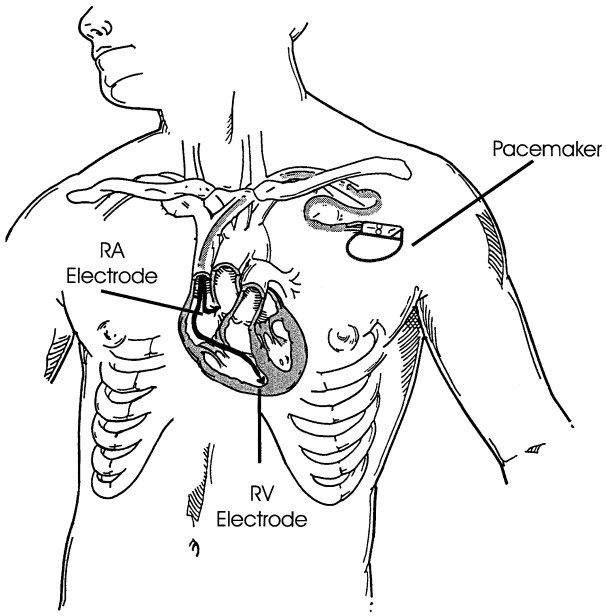


Figure 115.12 Placement of a dual-chamber pacemaker in the body and the leads in heart. The right atrial (RA) electrode is placed in the right atrial appendage and the right ventricular (RV) electrode is placed in the right ventricular apex. Both electrodes are capable of pacing and sensing.

sensing threshold is the smallest acceptable amplitude and slew rate of a sensed cardiac signal. Electrode repositioning may be required if efficient pacing and sensing are not realized. The leads are then connected to the pacemaker, sutured in place to prevent significant pacemaker movement, and the pocket is closed. Pacemaker programming can be accomplished through a wireless telemetry system.

Design

The pacemaker must be able to reliably detect the wavefronts on the **atria** and the ventricles, determine their rate, determine if the chambers are synchronized, and intervene in the appropriate chamber if these conditions are not met. Some patients may receive continuous pacing, while others may rarely need pacing at all. The pacemaker must perform over a period of greater than 5 years, in all types of environments, with maximum reliability, and using the smallest, most comfortable design possible. In addition, it must be adaptable to various types of patients through remotely programmable parameters and informative diagnostics. A pacemaker system is made up of three major components, including the leads, the pulse generator, and the programmer. The pulse generator (i.e., the pacemaker itself) houses all of the controlling and pacing electronics in a biologically compatible titanium shell. The leads provide the electrical link from the pulse generator to the heart. The programmer allows the physician to remotely program the pacemaker parameters and assess pacemaker function through a telemetry system.

Programmer

The programmer provides a bi-directional wireless link to the implanted pacemaker through a telemetry wand. The physician can program parameters that customize functions, check battery capacity, send functional commands to the pacemaker, check for past events, and monitor realtime or stored electrograms measured from the implanted electrodes. Diagnostics provide an invaluable tool for the physician to assess patient welfare, and a well-designed programmer will effectively display this information. Because of the overwhelming availability of parameters and functions built into modern pacemakers, the programmer is an essential part of the pacemaker system. The major problem associated with current telemetry systems is the speed of data transfer. Transfer speeds will need to be significantly increased in the future to keep pace with the increasing number of available diagnostics.

Leads

The purpose of the pacemaker lead(s) is to provide a link between the pulse generator and the cardiac tissue in order to efficiently sense and stimulate the heart. The majority of leads are inserted intravenously and attached to the inside of the heart. Modern leads are composed of five major parts, including the connector, conductor(s), insulation, electrode(s), and a fixation mechanism [Kay, 1996]. The electrode design is critical to minimizing current drain during pacing while ensuring reliable sensing. It has been shown that the pacing threshold is a function of the current density at the electrode [Stokes and Bornzin, 1985]. Minimizing the radius of the electrode will maximize current density and therefore reduce the pacing threshold and current drain. In addition, a small radius will increase the electrode resistance, which also helps to reduce current drain. Conversely, it has been shown that a large electrode surface area decreases sensing impedance and electrode polarization [Kay, 1996]. Electrode polarization is caused by a buildup of charge on the cardiac tissue after a stimulation pulse and can affect the electrode's ability to sense properly. The ideal electrode would, therefore, minimize the radius while maximizing the surface area [Sinnaeve et al., 1987]. This has been accomplished by building electrodes with a small radius but with complex microscopic mesh or porous structure to maximize surface area [Kay, 1996; Bornzin et al., 1983]. Electrode material varies among manufacturers, but is often comprised of a platinum alloy. The conductor serves as the electrical pathway between the pulse generator and the electrodes. It must have low impedance and be able to withstand repeated flexing due to heart motion. This is generally accomplished using a nickel alloy and coiling the wire to resist stress. There will be one conductor in the lead for unipolar configurations and two conductors for bipolar. Unipolar leads use a single electrode at the tip of the lead with the reference being the metal shell of the pacemaker. Bipolar leads use two closely spaced electrodes near the tip of the lead, which is helpful in rejecting external noise when sensing. The conductor must be insulated by a material that can withstand flexing and is resistant to harsh biological conditions. Silicone rubber and polyurethane are two commonly used insulating materials. The connector pin provides a physical link between the lead and the pulse generator and was standardized by an international meeting of manufacturers to avoid confusion when mixing brands of leads and pacemakers [Calfee and Saulson, 1986]. Finally, the fixation mechanism is responsible for holding the lead in place on the heart. The two major types of fixation are known as active and passive fixation. The most common active mechanism involves a helical screw that is advanced into the tissue [Markewitz et al., 1988]. The most common passive mechanism uses flexible tines that become entrapped in the cardiac tissue [Furman et al., 1979]. Fibrous tissue often grows around the fixation mechanism due to tissue injury. This further stabilizes the lead but can cause pacing thresholds to increase over time. A good fixation mechanism will minimize tissue injury while ensuring a stable anchor. Some designs use a steroid-eluting electrode to minimize and stabilize fibrous tissue growth [Timmis et al., 1983].

Pulse Generator

The primary function of the pulse generator is to interpret the information gained from the atrial and ventricular electrodes and other sensors to determine if the patient requires pacing and to deliver the pacing pulses, if necessary. A number of sophisticated algorithms are applied by the pacemaker to determine when pacing is necessary. The pulse generator also performs several secondary functions, such as telemetry and diagnostics. A hermetically sealed titanium shell is used to house the electronics of the pulse generator because titanium is a strong, lightweight metal that is biocompatible with human tissue and does not corrode. Figure 115.13 shows a block diagram of a typical pacemaker pulse generator. A battery supplies the power for the electronics as well as for the pacing pulses. The sensing circuitry is used to amplify the electrical signals that are present on the heart in order to determine the heart rate. If rate-adaptive sensors are used (discussed below), additional sensing circuitry is needed. The output circuitry generates the pacing pulses by storing a charge on a capacitor so it can be delivered to the heart on demand. The backup pacing circuit is capable of asynchronously pacing either the **atria** or the **ventricles** at a preprogrammed rate in the event of excess noise. It also serves as a rate-limiting protection circuit to prevent the heart from being paced at an excessive rate due to a main system failure. The pacing control consists of timing circuitry and logic sections; it is responsible for interpreting sensed data and reacting with an appropriate pacing response. A microprocessor can be integrated with the pacing control and is often utilized as an overall system control to allow for flexibility of design. Memory in the form of RAM and ROM is required to store the microprocessor program, pacing parameters, and diagnostics. The telemetry

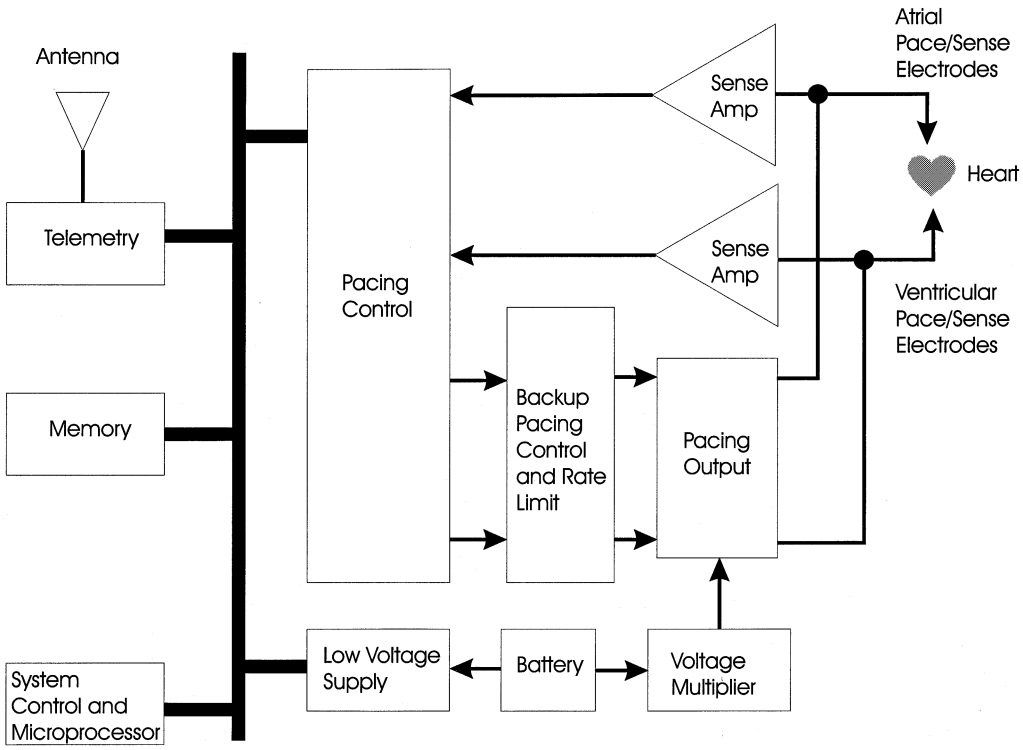


Figure 115.13 Block diagram of typical dual-chamber pacemaker components.

circuit is capable of swapping information with the programmer and is activated by a magnetic reed switch. The above circuits are currently built using CMOS integrated circuits, VLSI design, and hybrid technology. Continuing advances in the electronics industry will allow for further size reductions.

Battery

The battery for a pacemaker must be a safe and reliable energy source with a high energy density capable of supplying several microamperes for longer than 5 years. In addition, it must be possible to reliably predict the end of its life so that the pulse generator can be replaced before pacing fails. Because the battery accounts for the majority of pacemaker volume, these requirements must be met while ensuring that the battery is as small as possible. Nearly all modern pacemakers use lithium-iodine technology [Sanders and Lee, 1996]. This battery has a high energy density and a low internal self-discharge, which combine to give a longer lifetime than past batteries. Lithium serves as the anode, iodine combined with poly-2-vinyl pyridine serves as the cathode, and a semisolid layer of lithium iodide serves as the electrolyte. The cell is hermetically sealed to prevent corrosion. A new battery produces 2.8 V and declines linearly to 2.4 V near the end of its life. Either multiple batteries in series or voltage multipliers can be used to achieve voltages greater than 2.8 V. The status of the battery can be determined by several methods and can be telemetered out to the physician. The current drain of the device ultimately determines the lifetime of the battery and is dependent on many factors, such as circuit operating current, electrode impedance, and frequency, duration, and amplitude of the output pulses. Significant improvements in these areas have enabled the pacemaker battery to shrink in size.

Amplifier Sense System

The purpose of the amplifier sense system is to reliably detect the rate of electrical activity of the heart so that the pacemaker can determine if there is a need for therapy. The amplifier must have a high input impedance to ensure adequate signal amplitude. The system must reject all forms of external noise so that no inappropriate counting occurs. The electrical activity on the heart is manifested as deflections in an electrogram measured

by the sense electrodes. There are multiple deflections in an electrogram during one heartbeat due to near-field and far-field cardiac activity. In a ventricular electrogram, the far-field activity is associated with the depolarization of the **atria** (P waves), and the near-field activity is associated with the depolarization (R waves) and repolarization (T waves) of the ventricles. The R waves, which are the largest and fastest deflections, indicate a depolarization wavefront located directly below the sensing electrode and are, therefore, used as a means to count the heart rate. Circuitry used to sense R waves usually involves voltage comparators and slew rate detectors. Since the R wave deflection normally has a larger amplitude and faster slew rate than the other deflections, the circuitry can use reference thresholds to detect the R waves. The thresholds may be dynamic and determined by a complicated algorithm that has been termed autosensing [Jacobson and Kroiss, 1996; Castro et al., 1996; Kim, 1998]. In addition, there may be a brief amplifier blanking period after an R wave is detected during which all further deflections are ignored (so as to avoid potential inappropriate sensing of following T waves). In dual-chamber pacemakers, it is necessary to measure an atrial electrogram in addition to the ventricular electrogram. Similar detection strategies are employed in the atria; however, due to the large mass of the ventricles compared to the atria, the far-field ventricular activity in an atrial electrogram is more difficult to reject. A sensed cardiac event is marked by a digital pulse that is input into the timing circuitry. The timing circuitry determines if the heart rate is too slow and also controls all amplifier blanking periods and pacing rates.

In addition to far-field effects, many other sources of external noise exist in an electrogram, such as motion artifact, electrode polarization, noise from the skeletal muscles, and environmental noise (e.g., 60-Hz noise and cellular phones). Using bandpass filters and closely spaced bipolar electrodes minimizes external noise. In the event of extreme interference, backup pacing circuits can assume control and asynchronously pace the heart until the noise is gone.

Output Circuitry

The output circuitry is responsible for delivering the pacing pulses through the electrodes to the cardiac tissue in order to artificially control (capture) the heart. Because the amount of energy needed for capture can vary over the lifetime of a patient (e.g., due to changing electrode impedance or position) and between different patients, it is necessary to be able to deliver a controlled amount of energy per pacing pulse. Many pacemakers use an output voltage much higher than the pacing threshold to ensure capture for every pulse; however, this may unnecessarily waste energy. The minimum reliable pulse voltage necessary for capture is desired in order to minimize the current drain on the battery. Autocapture algorithms are capable of monitoring every pulse for capture and adjusting the output voltage to the minimum necessary value on demand [Jones et al., 1999]. This feature will become more prevalent in future pacemakers. Timing circuitry and output amplifiers are used to control the frequency, pulse width, and amplitude of the stimuli. Capacitors controlled by electronic switches physically deliver the energy. The capacitors are charged by the battery up to the desired voltage in between pacing pulses and then discharged into the heart by the switches at the proper timing. Voltage multipliers can be used to double or triple the battery voltage if necessary.

Rate Adaptive Pacing

An important feature of modern pacemakers is their ability to modulate the heart rate based on the metabolic needs of the body. There are many conditions that call for heart rate modulation, such as exercise, fever, stress, or sleep. Because it is under neural control, the ideal rate modulator is a normally functioning **sinus node.** In the case that the patient has atrioventricular conduction problems, the sinus node and the atria may still be functional; therefore, atrial sensing may be used by the pacemaker to regulate the ventricular rate. Patients with sinus node disease or atrial **arrhythmias,** however, require an artificial means of regulating heart rate. Many attempts have been made to design metabolic sensors that can be used to naturally modulate the heart rate. Control variables tested include blood pH [Cammilli, 1977], blood temperature [Alt et al., 1986], venous blood oxygen saturation [Eityzgrlf et al., 1982], respiratory rate [Rossi et al., 1983], minute ventilation [Alt et al., 1987], vibration [Anderson et al., 1983], acceleration [Matula et al., 1992], right ventricular pressure [Yee and Bennett, 1995], and QT interval [Donaldson and Rickards, 1983]. The ideal sensor would be reliable, have low current drain, require no additional surgery, and accurately reflect metabolic needs. All of the above sensors could potentially be used to reflect metabolic needs; however, only a few are currently practical and in use. The most common sensors in pacemakers today are the activity sensors, which attempt to indicate activity level by

transducing vibration and acceleration. Piezoelectric crystals functioning as strain gauges can be mounted to the inside of the pacemaker to detect mechanical vibrations [Anderson and Moore, 1986]. In addition, accelerometers can be mounted directly to the hybrid circuit to detect acceleration [Kay, 1996]. Although both of these sensors are subject to motion that may not be due to exercise, accelerometers have generally proven to be more proportional to exercise than strain gages [Kay, 1996]. Another sensor that has been successfully integrated into pacemakers is the minute ventilation sensor. Minute ventilation is representative of the amount of air a person breathes and can be estimated by measuring the transthoracic impedance over time. This can be accomplished by emitting a train of very small pulses from one pole of a bipolar pacing electrode and measuring the voltage between the other pole and the pacemaker can [Nappholz et al., 1986]. The impedance calculated from the measured voltage rises when a person breathes in and falls when the person breathes out. Combining activity sensors with minute ventilation sensors has proven to be a clinically successful approach to estimating metabolic needs during exercise [Alt et al., 1995].

Future

There are many exciting advancements currently under development for pacemakers. Efforts are being concentrated on making devices more sophisticated but less complicated [Jones et al., 1999]. One approach to realizing these goals is to expand the number and ability of the automatic features, such as autosense and autocapture algorithms. Pacemakers are trending toward automatic self-optimization abilities that are based on the individual patient's needs. Improvements in rate adaptive sensors and algorithms are key to realizing this goal and represent an area of significant research. In the future, integrated circuit technology will allow for much smaller and more efficient designs. This, in addition to improved battery technology, will allow for increased memory, advanced signal processing, and faster telemetry. More memory will permit more extensive patient diagnostics. These diagnostics will be displayed on advanced programmer interfaces that are more clinically relevant. Pacemaker lead research may ultimately yield single-pass leads, capable of pacing and sensing in both the atria and the **ventricles.** Finally nontraditional pacemaker uses are currently being explored. A major area of interest is in the treatment of congestive heart failure (CHF). CHF is a debilitating and deadly disease characterized by an enlarged heart that is incapable of pumping adequate blood to the body. It is possible that the size of the heart facilitates an electromechanical asynchrony between the ventricles. Recent studies have shown that appropriately positioned and timed pacing stimuli can help to improve cardiac output by synchronizing the ventricles [Foster et al., 1994; Bakker et al., 1994].

Implantable Cardioverter Defibrillators

The Implantable Cardioverter Defibrillator (ICD) was first conceived of by Dr. Michael Mirowski in the mid-1960s [Mirowski et al., 1970]. He imagined a device that would continuously monitor the hearts of high-risk individuals for life-threatening **arrhythmias** and intervene by electrical shock to restore normal sinus rhythm.Just over a decade later, the first patient was successfully implanted with an ICD [Mirowski et al., 1980]. From there, rapid development ensued, and FDA approval was obtained in 1985, at which time Cardiac Pacemakers Incorporated (CPI) took over the marketing and development of the ICD. Since then a number of competitors have arisen, and the ICD has evolved into a remarkably sophisticated medical device capable of bradycardia and antitachycardia pacing, low-energy **cardioversion**, high-energy **defibrillation** shocks, and extensive diagnostics (Fig.115.14).

Clinical Indications

Initially, ICDs were only indicated for certain patients that had survived an episode of cardiac arrest. Time of intervention is critical to survival of cardiac arrest; only 25% of people that have an episode are successfully resuscitated by first responders [Shuster and Keller, 1993]. ICDs have decreased the first-year mortality rate of these survivors from 30 to 2% [Winkle et al., 1991]. More recently, in addition to cardiac arrest survivors, patients deemed at risk for a first arrest due to sustained ventricular **tachyarrhythmias** are receiving ICDs as well [Saksena et al., 1996]. In the future, patients with more subtle predictors of sudden death may be indicated for ICDs.

Surgery

The modern ICD is now implanted similarly to a pacemaker. In the past, it was necessary to open the chest (i.e., thoracotomy) in order to suture large patch electrodes directly onto the ventricles and situate the ICD abdominally. Patch electrodes were required to achieve a low defibrillation threshold (DFT), which is a measure

Figure 115.14 The Guidant/CPI Ventak Mini IV ICD with the Endotak lead system attached (a single-pass lead). The Mini IV is a single chamber defibrillator. (Courtesy of Guidant/CPI, St. Paul, MN.)

of the amount of energy needed to reliably terminate **fibrillation** (defibrillate). Due to significant advances in the size and energy efficiency of ICD systems, it is now possible to implant the ICD pectorally and insert specially developed electrode leads into the heart intravenously without the need for a thoracotomy. This ensures greater patient comfort and significantly reduces the risk associated with surgery. There are several different lead configurations available, depending on the manufacturer, each with its own advantages. All of these systems place one electrode in the right ventricular apex, while the position of the return electrode varies between systems. Figure 115.15 shows a typical system layout in the body. The entire surgery can be performed through a single incision, using only local anesthetics and heavy sedation, and is similar to that of the pacemaker. Lead positioning is critical to obtaining low DFTs [Lang et al., 1995; Usui et al., 1995]. Once the lead(s) are in place, the ICD is tested by inducing fibrillation by artificial means and then giving a shock of known energy to halt the arrhythmia. The DFT can be determined by a number of different methods, such as by decreasing the energy of each successive shock until the **defibrillation** attempt is unsuccessful. The DFT must be well below the maximum output energy of the device before a successful implant is declared; a 10-joule safety margin is typically used [Moss et al., 1996]. If an adequate safety margin cannot be obtained through optimal electrode placement, additional electrodes may be required in order to obtain an acceptable DFT. In extreme cases, a thoracotomy may still be required. It is also necessary to thoroughly test the pacing/sensing characteristics as with pacemakers.

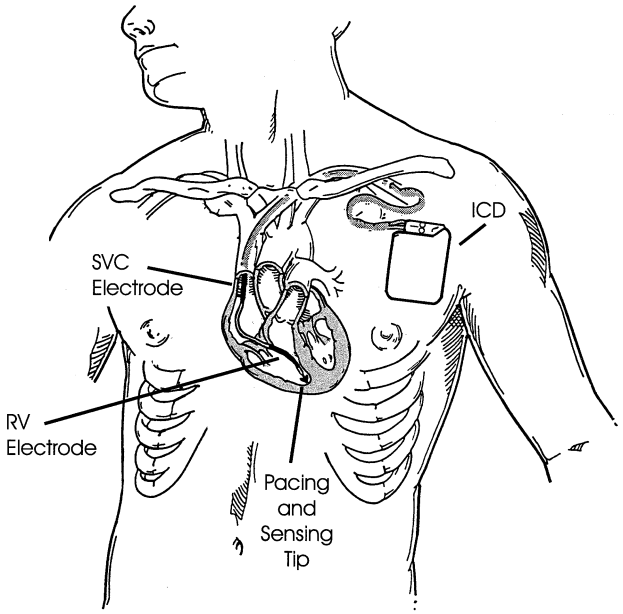


Figure 115.15 Placement of an ICD in the body and a single-pass lead system in the heart. The right ventricular (RV) and the superior vena cava (SVC) electrodes serve as the defibrillation electrodes, and there are pace/sense electrodes at the tip of the lead in the right ventricle.

Design

An ICD system consists of three main components: the programmer, the leads, and the pulse generator. The programmer provides a link to the ICD after it has been implanted and is similar to the pacemaker programmer. The leads deliver the energy from the ICD to the heart for defibrillation, as well as provide for pacing and sensing capabilities. The ICD pulse generator is an ever-evolving technology capable of delivering sophisticated cardiac rhythm management to the patient and diagnostics to the physician. As technology advances, ICDs will become smaller, more reliable, and more versatile.

Leads

The leads provide the means by which to deliver the energy of the defibrillation shock from the pulse generator to the heart, as well as pacing and sensing capabilities. They are insulated with either medical-grade silicone rubber or polyurethane, except at the electrodes. A major disadvantage of these leads is that they deliver current in a largely nonuniform manner as compared to the older patch electrodes, which results in higher energy requirements for defibrillation. Fortunately, advances in battery, capacitor, circuit, and waveform technology have compensated for this greater energy requirement while still allowing the ICDs to become smaller. Because electrode position is important, many different electrode shocking configurations have been attempted. One of the most popular is to place one electrode in the right ventricle (RV) and the return electrode in the superior vena cava (SVC). In addition to defibrillation electrodes, there must be pacing and sensing electrodes as well. Currently, three major configurations exist to accommodate these requirements. The first, known as the single pass lead, integrates two defibrillation coil electrodes (RV, SVC) and ventricular pace/sense electrodes onto a single lead. The second configuration consists of one lead, which contains the RV defibrillation electrode and the ventricular pace/sense electrodes, and a second lead, which contains the return defibrillation electrode. The third configuration consists of one lead containing both defibrillation electrodes and a separate lead containing the ventricular sense/pace electrodes. Some systems use the titanium housing of the ICD as an additional return electrode in the "active can" configuration in an attempt to distribute the current more evenly throughout the **ventricles**. Also, the most modern ICD systems utilize atrial pace/sense electrodes in addition to ventricular electrodes to achieve dual-chamber pacing and sensing capabilities. Another requirement of the defibrillation

lead is that it have very low impedance due to the large **defibrillation** currents, but can still withstand repeated flexing due to millions of heartbeats. This is achieved using a combination of high-strength, low-impedance metals. The pace/sense electrodes are similar in design to those used by pacemakers. The leads are fixed in place by either a screw-in mechanism or flexible tines.

Pulse Generator

The pulse generator is the core of the ICD system. It consists of the batteries, capacitors, and accompanying electronics enclosed in a hermetically sealed titanium can. The can may be used as a return electrode. A header, typically made of epoxy, is attached to the can and provides the link from the electronics to the leads via silicone-sealed ceramic feedthroughs. The size of the pulse generator has steadily declined since the introduction of the ICD and is currently around 40 cc. Significant efforts are underway to further reduce the size of the pulse generator in order to increase patient comfort. As research in defibrillation progresses, more efficient defibrillation strategies will no doubt be developed and allow the size to be decreased further. Currently, the major barriers to size reduction are the battery and capacitor size required to create waveforms capable of ventricular defibrillation. In addition, designing the electronics for a system that can measure cardiac signals on the order of 100 μV and produce high-energy waveforms on the order of 750 V and 40 A, all within the same small space, presents a significant engineering challenge. A problem associated with this includes high-voltage arcing among internal components. To prevent this, nitrogen gas is sealed inside the can because of its high breakdown voltage barrier.

Figure 115.16 shows a block diagram of the key components of a typical ICD pulse generator. The brain of the ICD is the microprocessor. Most ICD manufacturers use industry-standard microprocessors, such as the Z80, 6502, or 8852, to control the ICD [Warren et al., 1996]. In order to conserve energy, it is desirable to put the microprocessor into a sleep mode as often as possible and to wake it only when necessary, such as when an **arrhythmia** is suspected. To accomplish this, many of the monitoring and pacing functions are implemented using analog and digital circuits. Modern ICD designs have reduced the size of the circuitry to a small number of integrated circuits on a hybrid chip [Warren et al., 1996].

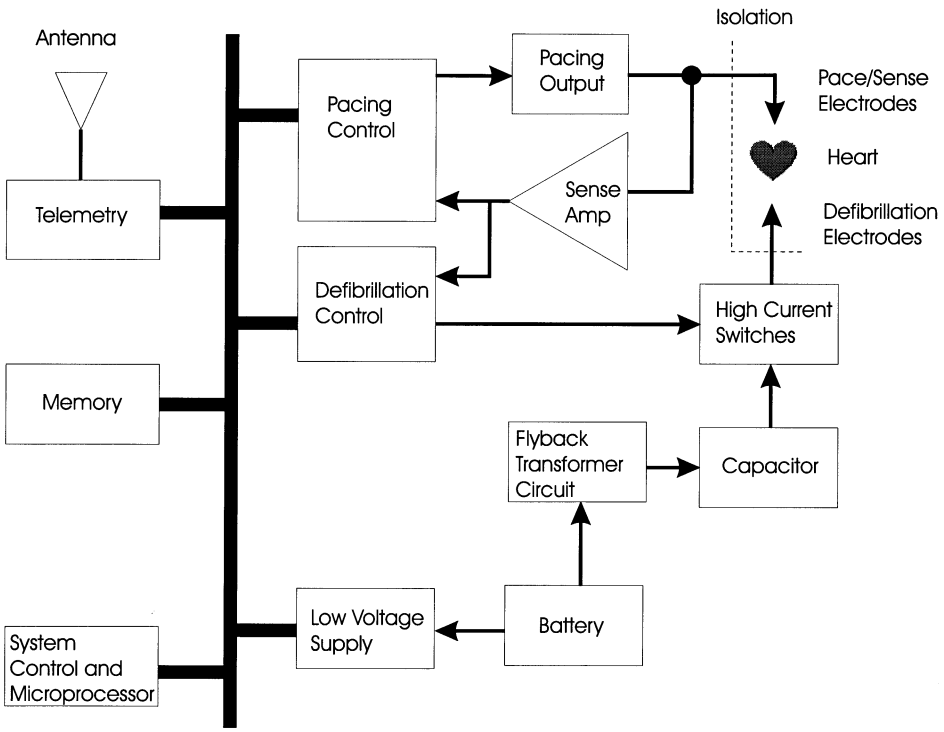


Figure 115.16 Block diagram of typical ICD components.

Memory is required to store the program and individual patient parameters for the operation of the ICD. Startup code and some of the main program is often stored in ROM, while the remaining program, parameters, patient diagnostics (e.g., electrograms), and event markers are stored in RAM. Clinicians and investigators are increasingly interested in diagnostics obtained from the ICD, which warrants future memory increases.

Additional circuitry includes support for the microprocessor, timers, the telemetry interface, low-voltage power supplies, the high-voltage system, the pacing control, the **defibrillation** control, and isolation and external protection circuits. The telemetry interface is the link to the external programmer and consists of a coil, which serves as the antenna, support circuitry, and a magnetic reed activator switch. The low-voltage supplies power the analog and digital circuitry as well as the pacing pulses. The defibrillation control determines when defibrillation is necessary and controls the process of defibrillation. The high-voltage system is used to generate the defibrillation shocks and consists of high-current batteries, capacitors, a fly-back transformer, and output switching circuits. The pacing control includes the circuitry to deliver pacing pulses, to interpret the signals from the sense amplifiers, and timers that monitor the current heart rate and wake the microprocessor if necessary. The isolation and protection circuits provide protection against external defibrillation attempts and external noise.

Amplifier

The purpose of the amplifier sense system is to reliably detect the rate of electrical activity on the heart so that the ICD can determine if there is a need for intervention. The amplifier must be immune to noise and be able to quickly respond to a large range of heart rates (30 to 360 b.p.m.) [Warren et al., 1996]. In order to obtain an accurate heart rate, it is desirable to digitally count R waves, while rejecting all other electrical activity and noise. Since the R wave is much larger in amplitude than the other waves in the electrogram, a simple method to detect them would be to use a comparator with a set threshold. Unfortunately, the amplitude of R waves is not constant; therefore, a simple comparator circuit is not reliable. For example, during tachycardias and **fibrillation**, the amplitude of the signals decreases significantly. One technique commonly used to solve this is to use a dynamically adjusting comparator threshold, in which the threshold level exponentially decreases over time until the next R wave is detected, at which time, the threshold level is reset [Brumwell et al., 1996]. This ensures that low-amplitude signals will be detected. Another technique is to use an automatic gain control to slowly increase the gain between detected R waves, while keeping the threshold constant [Brumwell et al., 1996]. To avoid double counting caused by undesired T-wave detection, there is, typically, a brief period of time after each detected R wave during which sensed signals are ignored. Detection schemes are often implemented with carefully designed analog chips, known as Application Specific Integrated Circuits (ASICs), in order to keep the current drain on the order of 10 μA [Warren et al., 1996]. An advantage of these chips is that they provide near-perfect component matching, which is critical in engineering predictable gain control and frequency response.

Battery

The design for an ICD battery has many stringent requirements and presents a unique challenge to the engineer. While the pacemaker battery is optimized for high energy density, a defibrillator battery must sacrifice some energy density for high current capability. An average defibrillator battery must be able to supply a steady background current of 10 to 20 μA for at least 5 years for monitoring and pacing functions, as well as, provide around 200, 2-A pulses for 10 to 15 seconds each in order to charge the capacitors for multiple defibrillation shocks [Holmes, 1996]. All of these criteria must be met while minimizing the size of the battery and maximizing its safety and reliability. In addition, it is necessary to be able to reliably predict the end of life of the battery. Nearly all modern ICDs use lithium silver vanadium oxide battery technology to accomplish these requirements [Liang et al., 1982]. Lithium pressed into a nickel current collector serves as the anode, and silver vanadium oxide serves as the active cathode. The electrolyte is generally a lithium salt dissolved in a mixed organic solvent. This produces approximately 3.2 V, but two cells are often connected in series to give around 6 V. Some newer systems are using single battery technology. A large electrode surface area and low internal impedance are required to achieve the high current pulses [Holmes, 1996]. This is accomplished by folding the anode in an accordion-like fashion and placing cathode plates in the folds. A disadvantage of this battery is that it exhibits a phenomenon in its mid-life known as voltage delay, in which the voltage goes low in the first second or two during a charging pulse. This can cause a prolongation of the capacitor charging time, which can be dangerous to the patient. It is due to an initial high resistance caused by a chemical buildup on the cathode in the cell

and can be alleviated by periodic pulsing of the batteries into the capacitors and internally dumping the charge. This shortens the life of the battery slightly but is not a total waste because the capacitors need this type of reforming as well. Several methods can be used to predict the end of life of a battery. Common indicators are the battery's open-circuit voltage, voltage during charging, and the time it takes to charge the capacitor [Holmes, 1996]. Future advancements in battery technology are critical in reducing defibrillator size and increasing longevity.

Charging Circuit

In order for a defibrillation shock to occur, it is necessary to convert the 6 V from the battery to an output voltage of up to 750 V to be stored across a capacitor. This is generally done with a dc/dc converter, or inverter. Unique design considerations include the large size of the conversion, a demand for high efficiency, and a minimized transformer and circuit size. The circuit includes the battery and a low-voltage, high-current switch at the input, a fly-back transformer, and a rectifying diode and storage capacitor at the output. A controlling oscillator typically operates the switch between 30 and 60 kHz for high efficiency [Warren et al., 1996]. This creates a simulated ac current, which is converted by the transformer to the higher voltage. The current in the input stage is, typically, around 2 A. The rectifier diode prevents current from flowing back into the secondary winding of the transformer. The voltage on the capacitor increases as a function of the square root of the time that the oscillator is on and typically reaches full capacity in 10 to 15 seconds [Warren et al., 1996]. It is possible for the transformer to be small because of the high-speed switching. In addition, the diameter of the core and the wire windings in the fly-back transformer can be small because the converter is only used intermittently to charge the capacitor, which allows for ease of heat dissipation compared to continuous conversion [Bach and Monroe, 1996]. There is a trade-off between the size of the converter and the efficiency. The high clock rates allow the transformer to be small, but also introduce losses in energy due to hysteresis in the coils. A typical ICD charging circuit achieves about 75% efficiency [Bach and Monroe, 1996]. This plays a major role in the charging time and, therefore, the delay before shocking therapy can be delivered.

Capacitor

The function of the capacitor is to store the energy generated from the high-voltage charging circuitry and to deliver that energy on demand to the heart over a few milliseconds. The commercially available aluminum electrolytic photoflash capacitor is currently used in the ICD. This is because of its high energy density of 1.7 J/cm^3 made possible due to special etching techniques that maximize surface area in the aluminum foil. Because the capacitors are commercially available and are not custom-designed for defibrillators, there are several shortcomings that ICD manufacturers must deal with. Since the highest energy density aluminum electrolytic capacitors are designed to operate at around 375 V, two of these capacitors must be used in series to attain the necessary 750 V used in most designs; therefore, the capacitors play a major role in determining the size of the defibrillator. Their geometry is not ideal for efficient packaging. The round shape results in wasted space in the ICD. Perhaps the greatest shortcoming is the requirement that the capacitor be reformed after periods of no use to ensure there is no leakage current during charging. This involves automatic application of the rated voltage to the capacitor for a few minutes every few weeks in order to repair damage that has occurred due to aging. The latest generation capacitors have minimized the need for reforming. As mentioned before, because the battery requires reforming as well, the energy is not completely wasted. Current research in capacitors is centered on reducing the size by increasing the energy density and eliminating the need for reforming.

Waveform and Output Switching

Optimizing the defibrillation waveform has been the subject of much investigation. Because a capacitor is used to deliver the energy, the waveform is some form of a decaying exponential. Still, there are many parameters that can be varied when creating a waveform, such as the pulse width, amplitude, decay, and the polarity and number of phases. In the past, monophasic, truncated, exponential waveforms were common. More recently, biphasic waveforms, in which the direction of current is reversed at some point during the waveform by switching the electrode polarity, have become more popular. This is due to the work of Schuder and others, who have shown that biphasic waveforms defibrillate with less energy than monophasic waveforms [Schuder et al., 1984; Feeser et al., 1990].

Output switching circuitry is needed in order to time and create these waveforms. Since the load impedance of the defibrillation system ranges from 20 to 70 Ω, it is possible to have peak currents of 40 A in the output circuit. There is, therefore, a need for high-power electronic switches to carry the current, such as Silicon Controlled Rectifiers (SCRs), Metal Oxide Semiconductor Field Effect Transistors (MOSFETs), and Insulated Gate Bipolar Transistors (IGBTs). In addition, these switches must be mounted so that there are very low junctional resistances to minimize power loss. MOSFETs and IGBTs are often used in bridge circuits to facilitate switching of biphasic waveforms. SCRs were used mostly in generating monophasic waveforms, but are still used in some biphasic waveform circuits. Because these types of switches require around 15 V for the control, it is necessary to use a low-power dc/dc converter to boost the 6 V from the battery. Timing of the switching is either controlled by timing circuitry, or by voltage monitoring circuitry that causes a polarity reversal to occur when the voltage falls to a certain threshold.

Future

In the future, ICDs will continue to evolve into increasingly sophisticated cardiac rhythm management devices. The ICD is no longer simply a safeguard against ventricular fibrillation; it is, moreover, being called to better manage bradycardias and tachycardias, and may eventually be used to predict and prevent **arrhythmias** from ever occurring. Management of atrial arrhythmias is also an important emerging frontier. To facilitate these and other demands, a number of advancements are currently being explored [Morris et al., 1999]. New lead systems will be smaller, more reliable, and better designed to manage atrial arrhythmias. Additional sensors may be incorporated, such as pressure transducers to measure hemodynamic stability. More complex rhythm discrimination algorithms will be developed to ensure that appropriate therapy is given at the appropriate time. Advances in electronics, battery, and capacitor technology will allow the ICD to continue to shrink in size. Diagnostic capabilities will be expanded due to increases in memory, and efforts will be made to dramatically decrease the complexity of programming the ICD. Finally, basic research will produce important advances in arrhythmia management, the significance of which cannot yet be imagined.

Defining Terms

Arrhythmia: A general term referring to a disorder in the electrical system of the heart.
Atria: The upper two chambers in the heart that act as primer pumps for the ventricles.
AV synchrony: The timing that must be maintained between the atria and the ventricles in order to pump blood efficiently.
Bradyarrhythmia: A class of arrhythmia that results in an abnormally slow heart rate.
Cardioversion: Termination of a tachyarrythmia, other than ventricular fibrillation, by a low-energy electrical shock.
Defibrillation: Termination of fibrillation by an electrical shock.
Fibrillation: A type of tachyarrhythmia characterized by a disorganized rhythm that can occur in either the atria or the ventricles and completely compromises their ability to pump blood.
Sinus node: Specialized cells in the top of the right atrium, which act as the heart's natural pacemaker.
Tachyarrhythmia: A class of arrhythmia that results in an abnormally fast heart rate.
Ventricles: The lower two chambers of the heart, which are responsible for pumping the blood to the body.

References

Alt, E., Heinz, M., Hirgstetter, C., Emslander, H. P., Daum, S. and Blomer, H., Control of pacemaker rate by impedance-based respiratory minute ventilation, *Chest,* 92, 247–252, 1987.

Alt, E., Hirgstetter, C., Heinz, M. and Blomer, H., Rate control of physiologic pacemakers by central venous blood temperature, *Circulation,* 73, 1206–1212, 1986.

Alt, E., Millerhagen, J. O. and Heemels, J.-P., Accelerometers. *Clinical Cardiac Pacing,* Ellenbogen, K. A., Kay, G. N. and Wilkoff, B. L., Eds., W.B. Saunders, Philadelphia, PA, 1995, 275.

Anderson, K. and et al., A rate-variable pacemaker which automatically adjusts for physical activity, *Pacing and Clin. Electrophys.,* 6, A-12, 1983. Abstract.

Anderson, K. M. and Moore, A. A., Sensors in pacing, *Pacing Clin. Electrophysiol.,* 9, 954–9, 1986.
Bach, S. M. and Monroe, P., High power circuitry, *Implantable Cardioverter Defibrillator Therapy. The Engineering-Clinical Interface,* Kroll, M. W. and Lermann, M. H., Eds., Kluwer Academic Publishers, Norwell, MA, 1996, 257–274.
Bakker, P. J., Meijburg, H., de Jonge, N., van Mechelen, R., Wittkampf, F. H., Mower, M., et al., Beneficial effects of biventricular pacing in congestive heart failure, *Pacing and Clin. Electrophys.,* 17, 820, 1994. Abstract.
Bernstein, A. D., Camm, A. J., Fletcher, R. D., Gold, R. D., Rickards, A. F., Smyth, N. P., et al., The NASPE/BPEG generic pacemaker code for antibradyarrhythmia and adaptive-rate pacing and antitachyarrhythmia devices, *Pacing Clin Electrophysiol.,* 10, 794–799, 1987.
Bornzin, G. A., Stokes, K. B. and Wiebush, W. A., A low threshold, low polarization, platonized endocardial electrode, *Pacing and Clin. Electrophys.,* 6, A70, 1983.
Brinker, J. and Midel, M., Techniques of pacemaker implantation, *Cardiac Pacing,* Ellenbogen, K. A., Ed., Blackwell Science, Cambridge, MA, 1996, 216–269.
Brumwell, D. A., Kroll, K. and Lehmann, M. H., The amplifier: sensing the depolarization, *Implantable Cardioverter Defibrillator Therapy. The Engineering-Clinical Interface,* Kroll, M. W. and Lermann, M. H., Eds., Kluwer Academic Publishers, Norwell, MA, 1996, 275–302.
Calfee, R. V. and Saulson, S. H., A voluntary standard for 3.2 mm unipolar and bipolar pacemaker leads and connectors, *Pacing Clin Electrophysiol.,* 9, 1181–5, 1986.
Cammilli, L., *A New Pacemaker Autoregulating the Rate of Pacing in Relation to Metabolic Needs,* Excerpta Medica, Amsterdam,1977.
Castro, A., Liebold, A., Vincente, J., Dungan, T. and Allen, J. C., Jr., Evaluation of autosensing as an automatic means of maintaining a 2:1 sensing safety margin in an implanted pacemaker. Autosensing Investigation Team, *Pacing Clin. Electrophysiol.,* 19, 1708–13, 1996.
Donaldson, R. M. and Rickards, A. F., Rate responsive pacing using the evoked QT principle. A physiological alternative to atrial synchronous pacemakers, *Pacing Clin. Electrophysiol.,* 6, 1344–1349, 1983.
Eityzgrlf, A., Goedel-Meinem, L., Bock, T. and et al., Central venous saturation for the control of automatic rate-responsive pacing, *Pacing and Clin. Electrophys.,* 5, 829, 1982. Abstract.
Ellenbogen, K. A., *Cardiac Pacing,* Blackwell Science, 1996.
Elmquist, R. and Senning, A. *An implantable pacemaker for the heart, Proceedings of the 2nd International Conference on Medical Electronics,* London, Iliffe and Sons, 1960.
Epstein, A. E. and Ideker, R. E., Ventricular fibrillation, *Cardiac Electrophysiology: From Cell to Bedside,* Zipes, D. P. and Jalife, J., Eds., W. B. Saunders Company, Philadelphia, PA, 1995, 927–934.
Feeser, S. A., Tang, A. S. L., Kavanagh, K. M., Rollins, D. L., Smith, W. M., Wolf, P. D., et al., Strength-duration and probability of success curves for defibrillation with biphasic waveforms, *Circulation,* 82, 2128–2141, 1990.
Foster, A. H., McLaughlin, J. S. and Fisher, M. L., Improved hemodynamics with biventricular pacing, *J. Am. Coll. Cardiol.,* 23, 156A, 1994. Abstract.
Funke, H. D., Cardiac pacing with the universal DDD pulse generator: Technological and electrophysioological considerations, *The Third Decade of Cardiac Pacing: Advances in Technology and Clinical Applications,* Barold, S. and Mugica, J., Eds., Futura Publ. Co., Mt. Kisco, NY, 1982, 191–223.
Furman, S., Pannizzo, F. and Campo, I., Comparison of active and passive adhering leads for endocardial pacing, *Pacing Clin. Electrophysiol.,* 2, 417–27, 1979.
Gillum, R. F., Sudden coronary death in the United States, *Circulation,* 79, 756–765, 1989.
Holmes, C. F., The Battery. *Implantable Cardioverter Defibrillator Therapy. The Engineering-Clinical Interface,* Kroll, M. W. and Lermann, M. H., Eds., Kluwer Academic Publishers, Norwell, MA, 1996, 205–222.
Jacobson, P. and Kroiss, D. Automatic control of the sensing threshold for monitoring cardiac rhythm in an implantable device**:** United States Patent 5564430, 1996.
Jones, B. R., Kim, J., Zhu, Q., Nelson, J. P., KenKnight, B. H., Lang, D. J., Warren, J.A., Future of bradyarrhythmia therapy systems: Automaticity, *Am. J. Cardiol.,* 83(5B), 192D–201D, 1999 Mar 11.
Katz, A. M., The Arrhythmias. I. Introduction and mechanisms, *Physiology of the Heart,* Raven Press,1992, 515–545.
Kay, G. N., Basic Concepts of Pacing, *Cardiac Pacing,* Ellenbogen, K. A., Ed., Blackwell Science, Cambridge, MA, 1996, 37–106.

Kim, J., Zhu, Q. and Lang, D., An autosense algorithm with dual thresholds, *Arch. Mal Cœur,* 91, 77, 1998. Abstract.

Lang, D. J., Heil, J. E., Hahn, S. J., Lindstrom, C. C. and Derfus, D. L., Implantable cardioverter defibrillator lead technology: improved performance and lower defibrillation thresholds, *Pacing and Clin. Electrophys.,* 18, 548–559, 1995.

Liang, C. C., Bolster, M. E., Murphy, R. M. and inventors (1982 Jan 12.). Metal oxide composite cathode material for high energy density batteries. U.S. Patent No. 4,310,609, Wilson Greatbatch Ltd., assignees.

Markewitz, A., Wenke, K. and Weinhold, C., Reliability of atrial screw-in leads, *Pacing Clin. Electrophysiol.,* 11, 1777–83, 1988.

Matula, M., Alt, E., Fotuhi, P. and et al., Influence of varied types of exercise on the rate adaptation of activity pacemakers, *Pacing and Clin. Electrophys.,* 15, 578, 1992. Abstract.

Mirowski, M., Mower, M. M., Staewen, W. S., Tabatznik, B. and Mendeloff, A. I., Standby automatic defibrillator. An approach to prevention of sudden coronary death, *Arch. Intern. Med.,* 126, 158–61, 1970.

Mirowski, M., Reid, P. R., Mower, M. M., Watkins, L., Gott, V. L., Schauble, J. F., et al., Termination of malignant ventricular arrhythmias with an implanted automatic defibrillator in human beings, *N. Engl. J. Med.,* 303, 322–4, 1980.

Morris, M. M., KenKnight, B. H., Warren, J. A. and Lang, D. J., A preview of implantable cordioverter defibrillator systems in the next millennium: An integrative cardiac rhythm management approach, *Am. J. Cardiol.,* 83(5B), 48D–54D, 1999 Mar 11..

Moss, A. J., Hill, W. J., Cannom, D. S., Daubert, J. P., Higgins, S. L., Klein, H., et al., Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia, *N. Engl. J. Med.,* 335, 1933–1940, 1996.

Nappholz, T., Valenta, H., Maloney, J. and Simmons, T., Electrode configurations for a respiratory impedance measurement suitable for rate responsive pacing, *Pacing and Clin. Electrophys.,* 9, 960, 1986.

Rossi, P., Plicchi, G., Canducci, G., Rognoni, G. and Aina, F., Respiratory rate as a determinant of optimal pacing rate, *Pacing Clin. Electrophysiol.,* 6, 502–10, 1983.

Saksena, S., Prakash, A., Hill, M., Krol, R. B., Munsif, A. N., Mathew, P. P., et al., Prevention of recurrent atrial fibrillation with chronic dual-site right atrial pacing, *J. Am. Coll. Cardiol.,* 28, 687–694, 1996.

Sanders, R. S. and Lee, M. T., Implantable pacemakers, *Proceedings of the IEEE,* 84, 480–6, 1996.

Schuder, J. C., McDaniel, W. C. and Stoeckle, H., Transthoracic defibrillation of 100 kg calves with bidirectional truncated exponential shocks, *Trans. Am. Soc. Artif. Intern. Organs,* 30, 520–525, 1984.

Shuster, M. and Keller, J. L., Effect of fire department first-responder automated defibrillation, *Ann. Emerg. Med.,* 22, 721–7, 1993.

Sinnaeve, A., Willems, R., Backers, J., Holovoet, G. and Stroobandt, R., Pacing and sensing: how can one electrode fulfill both requirements?, *Pacing Clin. Electrophysiol.,* 10, 546–54, 1987.

Stokes, K. and Bornzin, G., The electrode-biointerface: Stimulation, *Modern Cardiac Pacing,* Barold, S. S., Ed., Futura Publ. Co., Mount Kisco, NY, 1985, 37–77.

Timmis, G. C., Gordon, S., Westveer, D. C. and et al. (1983). A new steroid-eluting low threshold lead, *Proceedings of the Seventh World Symposium on Cardiac Pacing,* Darmstadt, Vienna, Steinkopff-Verlag.

Usui, M., Walcott, G. P., KenKnight, B. H., Walker, R. G., Rollins, D. L., Smith, W. M., et al., Influence of malpositioned transvenous leads on defibrillation efficacy with and without a subcutaneous array electrode, *Pacing and Clin. Electrophys.,* 18, 2008–2016, 1995.

Warren, J. A., Dreher, R. D., Jaworski, R. V., Putzke, J. J. and Russie, R. J., Implantable cardioverter defibrillators, *Proceedings of the IEEE,* 84, 468–79, 1996.

Winkle, R., Mead, H., Ruder, M. and et al., Ten year experience with implantable defibrillators, *Circulation,* 84, II-426, 1991. Abstract.

Yee, R. and Bennett, T. D., Rate-adaptive pacing controlled by dynamic right ventricular pressure (dP/dtmax), *Clinical Cardiac Pacing,* Ellenbogen, K. A., Kay, G. N. and Wilkoff, B. L., Eds., W.B. Saunders Company, Philadelphia, PA, 1995, 212–218.

Further Information

Cardiac Pacing, second edition, Kenneth A. Ellenbogen, Ed., Blackwell Science, 1996.

Clinical Cardiac Pacing, Kenneth A. Ellenbogen, G. Neal Kay, and Bruce L. Wilkof, Eds., W.B. Saunders Company, 1995.

Implantable Cardioverter Defibrillator Therapy: The Engineering-Clinical Interface, Mark W. Kroll and Michael H. Lehmann, Eds., Kluwer Academic Publishers, 1996.

Implantable Cardioverter-Defibrillators: A Comprehensive Textbook, N.A. Mark Estes III, Antonis S. Manolis, and Paul J. Wang, Eds., Marcel Dekker, 1994.

Fox, M..D., Frizzell, L.A., Franks, L.A., Darken, L.S., James, R.B. "Medical Imaging"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

116

Medical Imaging

M. D. Fox
University of Connecticut

Leon A. Frizzell
University of Illinois

Larry A. Franks
Sandia National Laboratories

Larry S. Darken
Oxford Instruments

Ralph B. James
Sandia National Laboratories

116.1 Tomography
Computerized Tomography • Positron Emission Tomography • Single Photon Emission Computed Tomography • Magnetic Resonance Imaging • Imaging

116.2 Ultrasound
Fundamentals of Acoustics • Principles of Pulse-Echo Ultrasound • Future Developments

116.3 Semiconductor Detectors for Radiation Measurements
Cryogenic Detectors • True Room-Temperature Detectors • Silicon Detectors • Prices and Availability

116.1 Tomography

M. D. Fox

The term **tomography** derives from the Greek *tomos* (cutting) and *grapho* (to write). Originally the term was applied to sectional radiography achieved by a synchronous motion of the x-ray source and detector in order to blur undesired data while creating a sharp image of the selected plane. The term *tomography* was used to distinguish between such slices and the more conventional plain film radiograph, which represents a two-dimensional shadowgraphic superposition of all x-ray absorbing structures within a volumetric body.

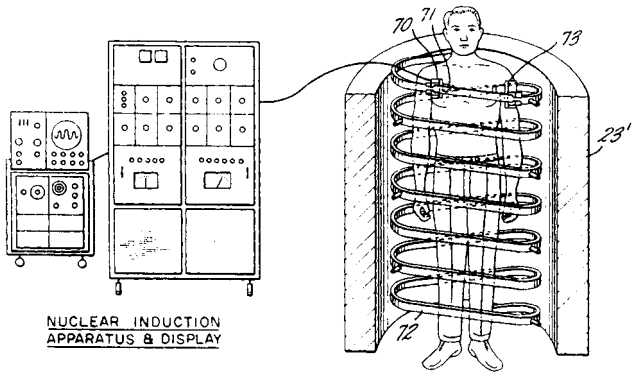
Computerized tomography, also known as **computerized axial tomography,** was introduced by EMI, Ltd. in 1973 and transformed medical imaging by obviating the superposition of intervening structures present in conventional radiographic images. Initially, the clinical application was for imaging the head, but soon the technique found wide application in body imaging.

As medical imaging has evolved into a multimodality field, the meaning of tomography has broadened to include any images of thin cross-sectional slices, regardless of the modality utilized to produce them. Thus, tomographic images can be generated by **magnetic resonance imaging** (MRI), ultrasound (US), computerized tomography (CT), or such nuclear medicine techniques as **positron emission tomography** (PET) or **single photon emission computerized tomography** (SPECT). For the purposes of this discussion we will cover all of the foregoing modalities with the exception of ultrasound, which will be treated separately.

Since the power of such computerized techniques was recognized, the practice of radiology has been revolutionized by making possible much more precise diagnosis of a wide range of conditions. In this necessarily brief discussion we will describe the basic physical principles of the major tomographic modalities as well as their key clinical applications.

Computerized Tomography

The basic concept of computerized tomography can be described by consideration of Fig. 116.1. An x-ray source is passed through an aperture to produce a fan-shaped beam that passes through the body of interest with absorption along approximately parallel lines. The natural logarithm of the detected intensity will be the integral of the linear attenuation coefficient of the object along the ray directed from the source to the detector element. If the source and the detector array are synchronously rotated about a point within the object, a number of



Apparatus and Method for Detecting Cancer in Tissue

Raymond V. Damadian
Patented February 5, 1974
#3,789,832

Excerpts from Raymond Damadian's patent application:

...It has now been found that, by measuring the degree of organization of these selected molecules in cells being studied and comparing this with the degree of organization in a known cancerous cell, cancer cells can be detected. Furthermore, it has now been found that the less the organization the greater the malignancy, therefore a scale can be made to provide a standard for basing a decision on the degree of malignancy...

...Further apparatus is provided for scanning throughout the entire body during which time the relaxation times are measured for selected nuclei and compared with standards. In this way a determination can be made of the existence of cancer together with the location and degree of malignancy of the cancerous cells present....

This patent describes a device that uses very powerful magnetic fields to resonate the nuclei in cells in a body. Collapsing the field and measuring the relaxation times gave a comparison to healthy cells. Later advances in digital signal processing have resulted in magnetic resonance imaging (MRI) equipment with color-coded image viewing of living tissue and its chemical composition.

lines of data can be collected, each representing the projected density of the object as a function of lateral position and angle.

A number of mathematical techniques can and have been used to recover the two-dimensional distribution of the linear attenuation coefficient from this array of measurements. These include iterative solution of a set of simultaneous linear equations, Fourier transform approaches, and techniques utilizing back-projection followed by deconvolution [Macovski, 1983]. Conceptually, the Fourier transform approach is perhaps the most straightforward, so we will describe it in some detail.

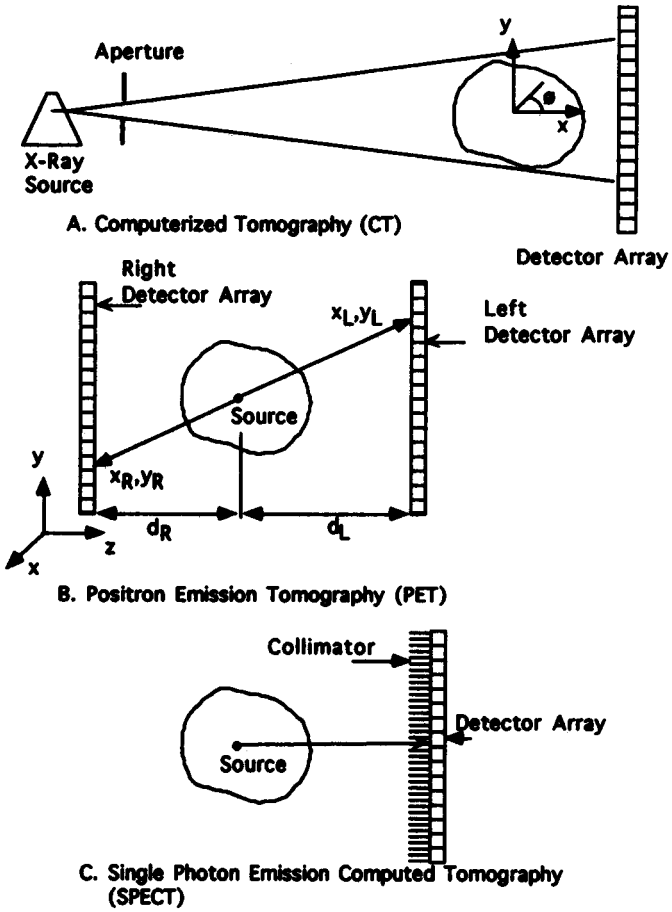


FIGURE 116.1 Comparison of three photon-based tomographic imaging modalities.

Using the coordinate system of Fig. 116.1(A) and assuming parallel rays, the intensity picked up by the detector array can be expressed as

$$I_d(y) = I_0 \exp[-\int a(x,y)\,dx]$$

where $a(x,y)$ represents the linear attenuation coefficient to x-ray photons within the body as a function of x,y position, and I_0 is the source intensity. Rearranging, we see that

$$a_p(y) = \int_{-\infty}^{\infty} a(x, y)dx = \ln[I_d(y)/I_0]$$

where $a_p(y)$ is the projected attenuation function. Taking a one-dimensional Fourier transform of this projected density function we see that

$$F[a_p(y)] = A_p(f_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(x,y)dx\,e^{-j2\pi f_y y}dy$$

where $A_p(f_y)$ is the Fourier transform of a single line of detected data. But this can also be written

$$A_p(0, f_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(x, y)dx\, e^{-j2\pi(0x+f_y y)}dy$$

Thus, the one-dimensional Fourier transform of the projection of the linear attenuation function, $a_p(y)$, is equal to the two-dimensional Fourier transform of the original attenuation function evaluated along a line in the frequency domain (in this case the $f_x = 0$ line).

It can readily be demonstrated that if we rotate a function $a(x,y)$ through an angle ϕ in the x,y plane, its transform will be similarly rotated through an angle ϕ [Castleman, 1979]. Thus as we rotate the source and detector around the object, each projected density function detected $a_p(\rho,\phi_i)$ can be Fourier transformed to provide one radial line of the two-dimensional Fourier transform of the desired reconstructed image, $A(\rho,\phi_i)$, where ρ is a radial spatial frequency. The set of all $A(\rho,\phi_i)$ for small angular displacements ϕ_i form a set of spokes in the transform domain which can be interpolated to estimate $A(f_x,f_y)$, the two-dimensional Fourier transform of the image in rectangular coordinates. The image can then be recovered by inverse transformation of $A(f_x,f_y)$, which can readily be carried out digitally using fast Fourier transform algorithms, i.e,

$$a(x,y) = F^{-1}[A(f_x,f_y)]$$

While the Fourier transform approach is mathematically straightforward, many commercial scanners utilize the equivalent but more easily implemented back-projection/deconvolution approach, where each ray is traced back along its propagation axis. When all rays have been back-projected and the result summed, one obtains an approximate (blurred) image of that plane. This image can then be sharpened (deblurred) through the use of an appropriate filter, which is usually implemented by convolving with an appropriate two-dimensional deblurring function. Refer to Macovski [1983] for the details of this process.

Clinically, the impact of computerized tomography was dramatic due to the vastly increased density resolution, coupled with the elimination of the superposition of overlying structures, allowing enhanced differentiation of tissues with similar x-ray transmittance, such as blood, muscle, and organ parenchyma. CT scans of the head are useful for evaluation of head injury and detection of tumor, stroke, or infection. In the body, CT is also excellent in detecting and characterizing focal lesions, such as tumors and abscesses, and for the evaluation of the skeletal system. [Axel et al., 1983]. In recent years the advent of magnetic resonance systems has provided even greater soft tissue contrast, and thus the role of CT has been constrained by this at times competing modality.

Positron Emission Tomography

Unlike computerized tomography, which relies on photons produced by an external source, in the modalities of positron emission tomography (PET) and single photon emission computed tomography (SPECT), the source of radiation is a radioisotope that is distributed within the body, and thus these modalities are sometimes referred to as forms of emission computed tomography (ECT). While conventional CT can produce images based upon anatomy of organs, emission CT techniques can quantitate the distribution of tracer materials that can potentially elucidate physiologic function.

The positron or positive electron is a positively charged particle that can be emitted from the nucleus of a radionuclide. The positron travels at most a few millimeters before being annihilated by interaction with a negative electron from the surrounding tissue. The product of this event is the emission of 511-keV gamma ray photons which travel in almost exactly opposite directions. The detectors themselves can be either discrete detectors or a modified Anger camera like those used in conventional nuclear imaging. A coincidence detector is employed to limit recorded outputs to cases in which events are detected simultaneously in both detector arrays, thus reducing the pickup of noise or scattering.

A possible detection scheme is illustrated in Fig. 116.1(B). The detector arrays shown can be made energy selective to eliminate lower energy scattered gamma rays. While the distribution of radioactivity can be reconstructed using the reconstruction from projection techniques described in the section on CT [Hurculak, 1987],

the *x,y* source position of an event can be determined directly from the detection geometry as follows [Macovski, 1983]:

$$x \approx x_L\, d_R/(d_R + d_L) + x_R\, d_L/(d_R + d_L)$$

$$y \approx y_L\, d_R/(d_R + d_L) + y_R\, d_L/(d_R + d_L)$$

Typically a single plane is studied, and no collimators are required. A drawback of PET has been that because of the short half-lives of positron-producing radioisotopes, the use of this modality has required the presence of an expensive cyclotron facility located near the hospital.

One important radionuclide commonly used in PET is oxygen 15 with a half-life of 2.07 minutes, which can be bonded to water for measurement of cerebral blood flow or to O_2/CO_2 to assess cerebral oxygen utilization. Another is carbon 11 with a half-life of 20.4 minutes, which can be bonded to glucose to trace glucose utilization. F-18 fluorodeoxyglucose (FDG) has been used to demonstrate the degree of malignancy of primary brain tumors, to distinguish necrosis from tumor, and to predict outcome [Coleman, 1991]. Perhaps the most unusual feature of this modality is the ability to quantitate the regional metabolism of the human heart [Schelbert, 1990].

Single Photon Emission Computed Tomography

In contrast to PET, SPECT can be utilized with any radioisotope that emits gamma rays, including such common radioisotopes as Tc-99m, I-125, and I-131 which have been utilized in conventional nuclear imaging for the last 30–35 years and which due to their relatively long half-lives are available at reasonable cost at nearly every modern hospital. Due to the need for direction sensitivity of the detector, a collimator must be used to eliminate gamma rays from other than the prescribed direction, thus resulting in a 1–2 order of magnitude decrease in quantum efficiency as compared with PET scanning [Knoll, 1983].

The basic concept of SPECT is illustrated in Fig. 116.1(C). A gamma ray photon from a radionuclide with energy above 100 keV will typically escape from the body without further interaction, and thus the body can be regarded as a transparent object with luminosity proportional to the concentration of the radionuclide at each point. The reconstruction mathematics are similar to those derived for absorption CT, with the exception that the variable reconstructed is a source distribution rather than an attenuation coefficient. Some errors can be introduced in the reconstruction because of the inevitable interaction of gamma rays with overlying tissue, even at energies above 100 keV, although this can be compensated for to some extent. Detection of scattered radiation can be reduced through the use of an energy acceptance window in the detector.

Technetium 99m can be used to tag red blood cells for blood pool measurements, human serum albumin for blood pool and protein distribution, or monoclonal antibodies for potential detection of individual tumors or blood cells. Emission computed tomography techniques such as PET and SPECT follow the recent trend toward imaging techniques that image physiologic processes as opposed to anatomic imaging of organ systems. The relatively low cost of SPECT systems has led to a recent resurgence of interest in this modality.

Magnetic Resonance Imaging

The basic magnetic resonance concept has been used as a tool in chemistry and physics since its discovery by Bloch in 1946, but its use expanded tremendously in the 1980s with the development of means to represent magnetic resonance signals in the form of tomographic images. Magnetic resonance imaging is based on the magnetic properties of atomic nuclei with odd numbers of protons or neutrons, which exhibit magnetic properties because of their spin. The predominant source of magnetic resonance signals in the human body is hydrogen nuclei or protons. In the presence of an external magnetic field, these hydrogen nuclei align along the axis of the field and can precess or wobble around that field direction at a definite frequency known as the Larmour frequency. This can be expressed:

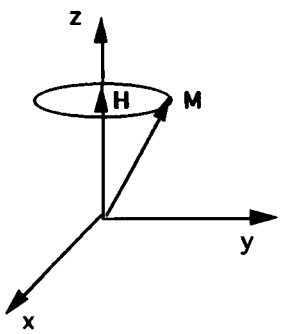


FIGURE 116.2 Geometry of precessing proton in a static magnetic field oriented in the *z* direction.

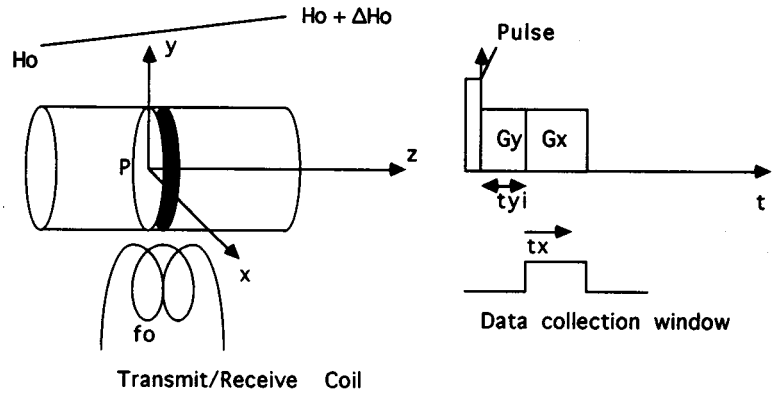


FIGURE 116.3 Concept of magnetic resonance imaging. The static magnetic field H_0 has a gradient such that excitation at frequency f_0 excites only the plane *P*. Gradient G_y in the *y* direction is applied for time t_{yi}, causing a phase shift along the *y* direction. Gradient G_x in the *x* direction is applied for time t_x, causing a frequency shift along the *x* direction. Repetition of this process for different t_{yi} allows the receive coil to pick up a signal which is the two-dimensional Fourier transform of the magnetic resonance effect within the slice.

$$f_0 = \gamma H$$

where f_0 is the Larmour frequency, γ is the gyromagnetic ratio which is a property of the atomic element, and H is the magnitude of the external magnetic field. For example, given a gyromagnetic ratio of 42.7 MHz/tesla for hydrogen and a field strength of 1 tesla (10 kilogauss), the Larmour frequency would be 42.7 MHz, which falls into the radio frequency range.

The magnetic resonance effect occurs when nuclei in a static magnetic field H are excited by a rotating magnetic field H_1 in the x,y plane, resulting in a total vector field **M** given by

$$\mathbf{M} = H\,\mathbf{z} + H_1(\mathbf{x}\cos\omega_0 t + \mathbf{y}\sin\omega_0 t)$$

Upon cessation of excitation, the magnetic field decays back to its original alignment with the static field H, emitting electromagnetic radiation at the Larmour frequency, which can be detected by the same coil that produced the excitation [Macovski, 1983].

Imaging

As shown in Fig. 116.3, one method for imaging utilizes a transmit/receive coil to emit a magnetic field at frequency f_0 which is the Larmour frequency of plane *P*. Subsequently, magnetic gradients are applied in the *y* and *x* directions. The detected signal during the data collection window can be expressed as

$$S(t_x, t_{yi}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(x, y)\exp[-i\gamma(G_x x t_x + G_y y t_{yi})]\,dx\,dy$$

where $s(x,y)$ represents the magnetic resonance signal at position (x,y) (G_x,G_y) are the *x* and *y* gradients, t_x is time within the data collection window, t_{yi} is the *y* direction gradient application times, and γ is the gyromagnetic ratio. The two-dimensional spatial integration is obtained by appropriate geometry of the detection coil. Collecting a number of such signals for a range of t_{yi}, we can obtain the two-dimensional function $S(t_x,t_y)$. Comparing this to the two-dimensional Fourier transform relation

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\exp[-i2\pi(ux + vy)]\,dx\,dy$$

we see that the detected signal $S(t_x,t_y)$ is the two-dimensional Fourier transform of the magnetic resonance signal $s(x,y)$ with $u = \gamma G_x t_x/2\pi$, $v = \gamma G_y t_y/2\pi$. The magnetic resonance signal $s(x,y)$ depends on the precise sequence of pulses of magnetic energy used to perturb the nuclei. For a typical sequence known as spin-echo consisting of a 90-degree pulse followed by a 180-degree pulse spaced at time τ with the data collection at $t_e = 2\tau$, and t_r being the repetition time between 90-degree pulses, the detected magnetic resonance signal can be expressed

$$s(x,y) = \rho(1 - e^{-tr}/T_1)(e^{-te}/T_2)$$

where ρ is the proton density, and T_1 (the spin-lattice decay time) and T_2 (the spin-spin decay time) are constants of the material related to the bonding of water in cells [Wolf and Popp, 1984]. Typically T_1 ranges from 0.2 to 1.2 seconds, while T_2 ranges from 0.05 to 0.15 seconds.

By modification of the repetition and orientation of excitation pulses, an image can be made T_1, T_2, or proton density dominated. A proton density image shows static blood and fat as white and bone as black, while a T_1 weighted image shows fat as white, blood as gray, and cerebrospinal fluid as black. T_2 weighted images tend to highlight pathology since pathologic tissue tends to have longer T_2 than normal.

In general, magnetic resonance imaging has greater intrinsic ability to distinguish between soft tissues than computerized tomography. It also has some ability to visualize moving blood. As the preceding discussion indicates, magnetic resonance is a richer and more complex modality than CT. Typically MRI has been more expensive than CT. Both MRI and CT have been used primarily for anatomic imaging, but MRI has the potential through spectroscopy (visualization of other nuclei than hydrogen) to become a factor in physiologic imaging. Thus, it can be anticipated that magnetic resonance imaging will continue to increase and become an even more important modality in the next decade.

Defining Terms

Computerized axial tomography (CATscan, CT): A form of medical imaging based upon the linear attenuation coefficient of x-rays in which a tomographic image is reconstructed from computer-based analysis of a multiplicity of x-ray projections taken at different angles around the body.

Magnetic resonance imaging (MRI, NMR): Aform of medical imaging with tomographic display which represents the density and bonding of protons (primarily in water) in the tissues of the body, based upon the ability of certain atomic nuclei in a magnetic field to absorb and reemit electromagnetic radiation at specific frequencies.

Positron emission tomography (PET scan): A form of tomographic medical imaging based upon the density of positron-emitting radionuclides in an object.

Single photon emission computed tomography (SPECT): A form of tomographic medical imaging based upon the density of gamma ray-emitting radionuclides in the body.

Tomography: A method of image presentation in which the data is displayed in the form of individual slices that represent planar sections of the object.

Related Topic

35.1 Maxwell Equations

References

L. Axel, P.H. Arger, and R. Zimmerman, "Applications of computerized tomography to diagnostic radiology," *Proceedings of the IEEE*, vol. 71, no. 3, p. 293, March 1983.

K.R. Castleman, *Digital Image Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1979.

R.E. Coleman, "Single photon emission computed tomography and positron emission tomography," *Cancer,* vol. 67 (4 Suppl.), pp. 1261–1270, Feb. 1991.

P.M. Hurculak, "Positron emission tomography," *Canadian Journal of Medical Radiation Technology,* vol. 18, no. 1, March 1987.
G.F. Knoll, "Single-photon emission computed tomography," *Proceedings of the IEEE*, vol. 71, no. 3, p. 320, March 1983.
A. Macovski, *Medical Imaging Systems*, Englewood Cliffs, N.J.: Prentice-Hall, 1983.
H.R. Schelbert, "Future perspectives: Diagnostic possibilities with positron emission tomography," *Roentgen Blaetter,* vol. 43, no. 9, pp. 384–390, Sept. 1990.
G.L. Wolf and C. Popp, *NMR, A Primer for Medical Imaging*, Thorofare, N.J.: Slack, Inc., 1984.

Further Information

The journal *IEEE Transactions on Medical Imaging* describes advances in imaging techniques and image processing. *Investigative Radiology,* published by the Association of University Radiologists, emphasizes research carried out by hospital-based physicists and engineers. *Radiology,* published by the North American Society of Radiologists, contains articles which emphasize clinical applications of imaging technology. *Diagnostic Imaging,* publishing by Miller Freeman, Inc., is a good source of review articles and information on the imaging marketplace.

116.2 Ultrasound

Leon A. Frizzell

Ultrasound, acoustic waves at frequencies higher than those audible by humans, has developed over the past 35 years into an indispensable clinical diagnostic tool. Currently, ultrasound is used to image most parts of the body. More than half of all pregnant women in the United States are examined with ultrasound. This widespread utilization has resulted from ultrasound's proven clinical utility for imaging soft tissues compared to more expensive imaging techniques. The development of ultrasound, particularly for fetal examinations, has also been fostered by its safety record; no case of an adverse biological effect induced by diagnostic ultrasound has ever been reported in humans [AIUM, 1988].

Diagnostic ultrasound systems are used primarily for soft tissue imaging, motion detection, and flow measurement. Except for some Doppler instruments, these systems operate in a **pulse-echo** mode. A brief summary of some of the fundamentals of acoustic wave propagation and the principles of ultrasound imaging follows.

Fundamentals of Acoustics

Unlike electromagnetic waves, acoustic waves require a medium for propagation. The acoustic wave phenomenon causes displacement of particles (consisting of many molecules), which results in pressure and density changes within the medium. For a traveling sinusoidal wave, the variation in acoustic pressure (the difference between the total and ambient pressure), excess density, particle displacement, particle velocity, and particle acceleration can all be represented by the form

$$p = P e^{-\alpha x} \cos(\omega t - kx) \tag{116.1}$$

for a wave propagating in the positive x direction, where p is the pressure (or one of the other parameters listed above), P is its amplitude, ω is the angular frequency, and $\omega = 2\pi f$ where f is the frequency in hertz, k is the propagation constant and $k = \omega/c$ where c is the propagation speed, α is the attenuation coefficient, and t is the time. The wave can experience significant attenuation, as represented by the exponential decay of amplitude with distance, during propagation in tissues. The attenuation coefficient varies greatly among tissues [Goss et al., 1978, 1980; Haney and O'Brien, 1986] but is low for most body fluids, much higher for solid tissues, and very high for bone and lung (see Table 116.1). The skin depth is the distance that the wave can propagate before being attenuated to e^{-1} of its original amplitude and is thus simply the inverse of the attenuation coefficient.

TABLE 116.1 Approximate Ultrasonic Attenuation Coefficient, Speed, and Characteristic Impedance for Water and Selected Tissues at 3.5 MHz

Tissue	Attenuation Coefficient (m^{-1})	Speed (m/s)	Characteristic Impedance (10^6 Pa s/m)
Water	0.2	1520	1.50
Amniotic fluid	0.7	1510	1.51
Blood	7	1550	1.60
Liver	35	1580	1.74
Muscle	50	1560	1.72
Bone	800	3360	5.70
Lung	1000	340	0.25

Ultrasound is typically used to image soft body tissues such as liver, but the sound beam often travels through fluids, for example, through amniotic fluid when imaging the fetus. Generally, bone and lung are not imaged with ultrasound. The attenuation processes include absorption, which is the conversion of acoustic energy to heat, and scattering, which will be addressed later. The attenuation increases roughly linearly with frequency in the 2- to 10-MHz range typically used for medical imaging. This range represents a compromise between increased penetration at lower frequencies (because of decreased attenuation) and improved resolution associated with higher frequencies as discussed below. Thus, the lower frequencies are used when greater penetration is required, such as for fetal imaging in the obese patient, and higher frequencies for lesser penetration, such as the examination of peripheral vascular flow.

When an acoustic wave impinges on an interface between two media of different specific acoustic impedance, a portion of the incident energy is reflected. For normal incidence on an infinite plane interface, the pressure reflection coefficient is given by [Kinsler et al., 1982]

$$R = \frac{z_2 - z_1}{z_2 + z_1} \tag{116.2}$$

where z_1 and z_2 are the specific acoustic impedance of the incident and transmitting media, respectively. For a plane wave the specific acoustic impedance is equal to the characteristic impedance which is the product of the density and acoustic speed in the medium (see Table 116.1). The speed is dependent upon the density and the elastic properties of the medium. Thus, at an interface between media exhibiting different densities or elastic properties, i.e., compressibility, some acoustic energy will be reflected. Although the reflection coefficient at an interface between muscle and bone is large (approximately 0.54) the reflection coefficient between two soft tissues such as liver and muscle is quite small (approximately 0.006). Reflection at oblique incidence obeys Snell's law in the same way it applies to electromagnetic waves.

In addition to the specular reflection that occurs at an interface between two media of different specific acoustic impedance as described above (where any curvature along the interface is negligible over distances comparable to a wavelength), energy may also be scattered in all directions by inhomogeneities in the medium. An acoustic image is formed by using this scattered energy as well as specular reflections. The fraction of the incident energy reflected or scattered is very small for soft tissues.

Although it is convenient to consider plane waves of infinite lateral extent, as was done above, real sources generate finite beams of ultrasound. These sources may be unfocused, but for the typical diagnostic system they are focused. Figure 116.4 shows the acoustic field from a typical focused source. The source consists of a piezoelectric transducer which converts electrical to acoustic energy and vice versa. Most transducers for medical applications are made from ceramic materials such as a lead zirconate titanate (PZT) mixture. For a circular aperture these may be circular disks with a plano-concave lens mounted in front to produce spherical focusing. Alternatively, the transducer itself may be a spherical segment that produces a focused field without a lens. Some probes utilize electronic focusing methods. Such a phased array probe consists of many individual elements which can be excited with signals having a controlled delay with respect to one another such that the

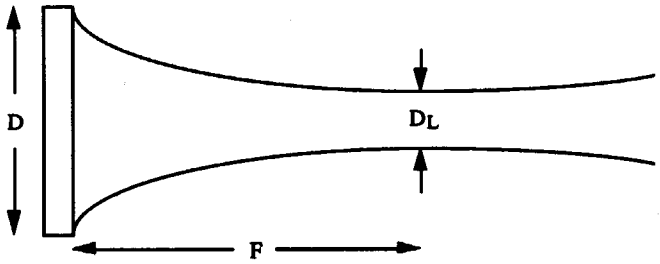


FIGURE 116.4 Cross section of a typical focused circular ultrasound source of aperture diameter D and focal length F, showing the focused beam of lateral beam width at the focus D_L.

signals constructively interfere at the desired focal region. At a receiver the signals are combined with delays associated with various elements to provide reinforcement of the signals from a receiving focal region.

The 3 db lateral beam width D_L is directly dependent upon the wavelength λ and focal length F and inversely related to the aperture diameter (diameter of the transducer) D [Kino, 1987]:

$$D_L = 1.02 \frac{F\lambda}{D} \tag{116.3}$$

Because $f\lambda = c$, the higher the frequency the smaller is λ and D_L. The smaller D_L, the better the lateral resolution near the focus, but the beam spread is greater with distance from the focus. Thus, the strength of the focusing varies among transducers so that the user may choose very good resolution over a short region or somewhat poorer resolution that is maintained over a greater depth. With phased array transducers, the focal region can be varied dynamically to optimize lateral resolution at all distances.

Principles of Pulse-Echo Ultrasound

Ultrasound imaging usually employs frequencies in the 2- to 10-MHz range, though some of the new intravascular probes use higher frequencies. Images are formed by using a transducer within a probe to generate a short pulse (typically on the order of 1 μs in duration) of ultrasound which is propagated through the tissue. A portion of the energy in this pulse is reflected back toward the transducer from specular reflectors and from scatterers in the tissue. These acoustic echoes, with amplitudes much lower than the transmitted pulse, are converted by the transducer to electrical signals which are converted to a (rectified) video signal, amplified by a time gain controlled amplifier, and displayed. The **A-mode display** is rarely used but simply involves display of the received echoes as amplitude versus time of arrival. The time of arrival is related by the wave speed to the tissue depth from which the echo returns, i.e., $d = ct/2$. Figure 116.5 provides a very simple representation of this process where the A-mode display associated with specular reflection from three different interfaces is illustrated. For clinical imaging the interfaces would not necessarily be perpendicular to the axis of the sound beam, and there would be a continuum of echoes, a continuous received signal, due to energy backscattered from within the tissues. Since the ultrasound pulse is attenuated as it propagates, all ultrasonic imaging systems use a logarithmic variation of amplifier gain with time to compensate the exponential attenuation of the tissue. Thus, echoes from structures reflecting or backscattering the same fraction of the incident signal will have the same amplitude after passing through the time gain controlled amplifier.

A **B-mode display** is typically used for ultrasound imaging. It involves display of the echoes at various brightness or gray levels corresponding to their amplitude. A **two-dimensional B-mode display** involves movement of the transducer (manually or automatically), movement of a mirror to change the direction of the field (automatically), or movement of the ultrasound beam directly (electrically) such that it scans a plane through the body. Figure 116.6 provides a simplified representation (again, echoes are shown as arising from interfaces only) of the formation of a B-mode image. The direction of the beam is monitored so that the received signals along each path are placed in their correct location on the display. Typically, the orientation information and echoes are processed by a digital scan converter for appropriate display of the two-dimensional image on

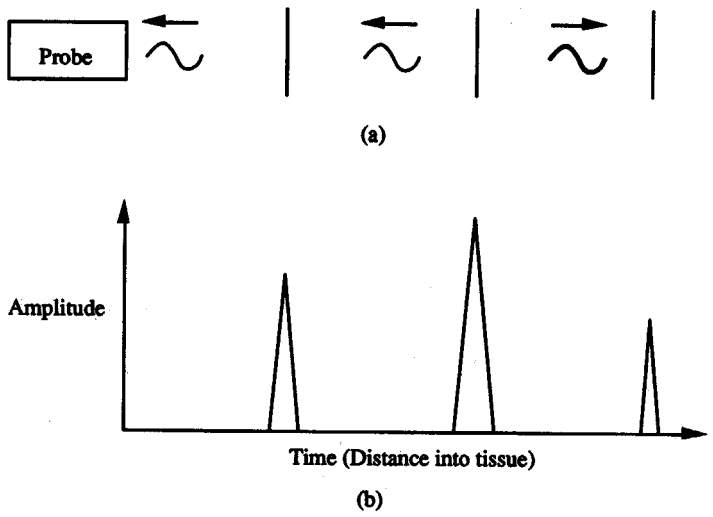


FIGURE 116.5 (a) The transmitted pulse (heavy wave) and echoes from reflecting structures; (b) the resulting A-mode display.

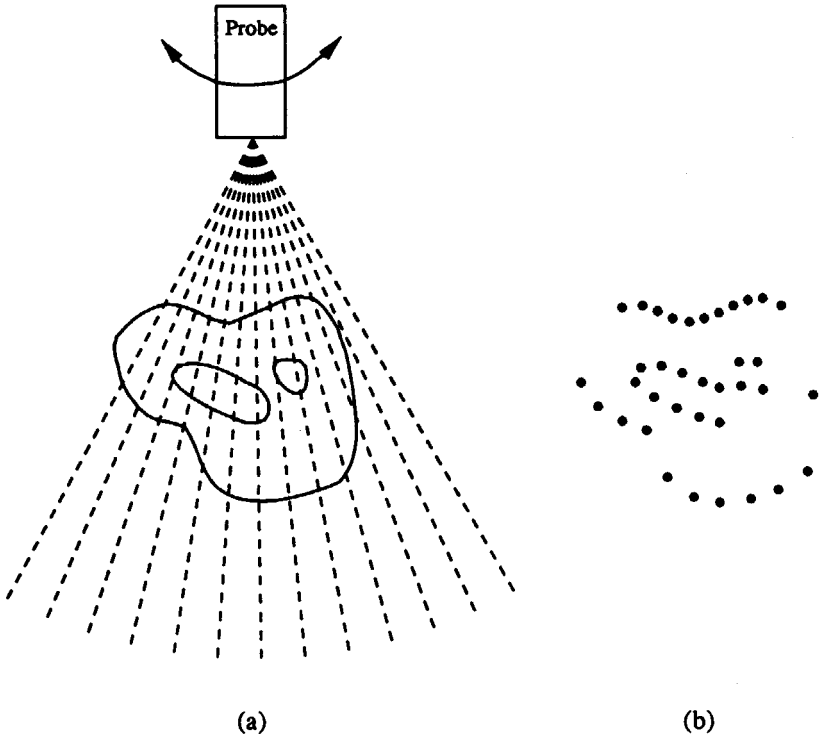


FIGURE 116.6 (a) The transmitted pulse paths for a rotating transducer probe; (b) the resulting two-dimensional B-mode display of echoes from the interfaces only.

a cathode ray tube in the standard format used for television picture display. Most B-mode systems in use today create an image in 0.1 s or less, so that the image is displayed in real-time for viewing of moving structures, such as structures in the heart or the fetus moving within the womb. This is not possible with the typical magnetic resonance or computed tomography system.

Many systems now use digital processing to enhance portions of the image. For example, it is possible to emphasize the large amplitude, small amplitude, or midrange signals. It is also possible to perform a more sophisticated analysis to enhance edges.

Many specialty probes have been designed for intracavitary examination. Examples include examination of the fetus with a vaginal probe, the prostate with a rectal probe, and blood vessel walls with intravascular probes. The intracavitary probe offers the advantage of decreasing the distance from the transducer to the tissue of interest and thus decreasing attenuation such that higher frequencies can be used for greater resolution. The lateral resolution of a focused probe is improved with frequency as discussed in the preceding subsection, but the axial (along the ultrasound beam) resolution also improves with frequency. The shorter the transmitted pulse, the better the axial resolution. Shorter pulses are generated by sources with a larger bandwidth, which corresponds to a higher center frequency when the sources have bandwidths which are approximately the same fraction of the center frequency.

The use of ultrasound for motion detection and measurement has increased tremendously in recent years. Most of these systems use the Doppler principle, but some use time domain detection. In Doppler detection, if the ultrasound is reflected from a target moving at some speed v_t toward (away from) the source at an angle θ with respect to the beam axis, the frequency of the transmitted signal f is shifted up (down) by an amount f_D, the Doppler shift, according to the following relation:

$$f_D = \frac{2 f v_t \cos\theta}{c} \tag{116.4}$$

In principle a measurement of f_D, when f, c, and θ are known, will yield the speed of the target v_t. However, it is often difficult to determine θ because the angle the transducer axis makes with a blood vessel, for example, is often unknown. Even when that angle is known, the flow is not necessarily along the direction of the vessel at every location and for all times. Time domain detection of motion, by measuring the movement of specific echoes from one pulse to another, is a recently developed alternative to Doppler detection that is not currently widely used.

For many years **duplex** systems, which provide both a two-dimensional image and a Doppler signal, showing the change of target speed with time, from a particular selected target area, have been in wide use. More recently, **color flow imaging** has been employed, which provides a two-dimensional color (typically red or blue) image of flow toward and away from the transducer superimposed on the gray scale image of stationary tissue structures. For these systems the speed, whether from Doppler or time domain detection schemes, is indicated by color saturation, hue, or luminance. These systems have proven very valuable for detecting the existence of flow in a region, detecting obstructions to flow and the turbulence associated with this, detecting reduced flow, and so on. Some other systems add to the color flow image a display of speed versus time for a region that is defined by a user-movable box.

Future Developments

It seems clear that the continuing development of intracavitary transducers, particularly for intravascular imaging, and the use of ultrasound intraoperatively will lead to more high-frequency commercially available probes that will produce better resolution images for these applications. The development of useful three-dimensional ultrasonic imaging is progressing rapidly and should immediately improve the measurement of tissue volumes [Gilja et al., 1995].

Defining Terms

A-mode display: Returned ultrasound echoes displayed as amplitude versus depth into the body.

B-mode display: Returned ultrasound echoes displayed as brightness or gray scale levels corresponding to the amplitude versus depth into the body.

Color flow imaging or color Doppler: Two-dimensional image showing color-coded flow toward and away from the transducer displayed with the two-dimensional gray scale image of stationary targets.

Duplex ultrasound: Simultaneous display of speed versus time for a chosen region and the two-dimensional B-mode image.

Pulse-echo ultrasound: Using a probe containing a transducer to generate a short ultrasound pulse and receive echoes of that pulse, associated with specular reflection from interfaces between tissues or scattering from inhomogeneities within the tissue, to form a display of the tissue backscatter properties.

Two-dimensional B-mode display: Echoes from a transducer, or beam, scanned in one plane displayed as brightness (or gray scale) versus location for the returned echo to produce a two-dimensional image.

Related Topic

48.1 Introduction

References

American Institute of Ultrasound in Medicine, "Bioeffects considerations for the safety of diagnostic ultrasound," *J. Ultrasound Med.*, vol. 7, no. 9 (supplement), 1988.

O.H. Gilja, A.I. Smievoll, N. Thune, K. Matre, T. Hausken, S. Odegaard, and A. Berstad, "In vivo comparison of 3D ultrasonography and magnetic resonance imaging in volume estimation of human kidneys," *Ultrasound Med. Biol.,* vol. 21, pp. 25–32, 1995.

S.A. Goss, R.L. Johnston, and F. Dunn, "Comprehensive compilation of empirical ultrasonic properties of mammalian tissues," *J. Acoust. Soc. Am.,* vol. 64, pp. 423–457, 1978.

S.A. Goss, R.L. Johnston, and F. Dunn, "Compilation of empirical ultrasonic properties of mammalian tissues. II," *J. Acoust. Soc. Am.*, vol. 68, pp. 93–108, 1980.

M.J. Haney, and W.D. O'Brien, Jr., "Temperature dependency of ultrasonic propagation properties in biological materials," in *Tissue Characterization with Ultrasound*, vol. 1, J. Greenleaf, Ed., Boca Raton, Fla.: CRC Press, 1986, pp. 15–55.

G. Kino, *Acoustic Waves,* Englewood Cliffs, N.J.: Prentice-Hall, 1987, p. 185.

L.E. Kinsler, A.R. Frey, A.B. Coppens, and J.V. Sanders, *Fundamentals of Acoustics,* 3rd ed., New York: John Wiley, 1982.

Further Information

The American Institute of Ultrasound in Medicine publishes monthly the *Journal of Ultrasound in Medicine,* which contains largely clinically oriented articles, and many clinically oriented ultrasound texts are available covering almost any medical discipline. However, there are only a few books that provide more than a cursory treatment of the basic physics and instrumentation of ultrasound imaging.

One text that has been regularly updated since the first volume appeared in 1980 is *Diagnostic Sonography: Principles and Instruments,* 4th edition, by F.W. Kremkau (W.B. Saunders, Philadelphia, 1993). Though this text is designed primarily to train sonographers who do not have a technical background, it provides the fundamentals of ultrasound imaging in a format that is very easy to read and understand.

Other texts that provide a more technical background (though some are a bit dated) include:

Biomedical Ultrasonics, by P. N. T. Wells (Academic Press, New York, 1977).

New Techniques and Instrumentation in Ultrasonography, edited by P.N.T. Wells and M.C. Ziskin (Churchill Livingstone, New York, 1980).

Medical Physics of CT and Ultrasound, edited by G.D. Fullerton and J.A. Zagzebski (American Institute of Physics, New York, 1980).

Physical Principles of Medical Ultrasonics, edited by C. R. Hill (John Wiley, New York, 1986).

Ultrasonic Bioinstrumentation, by D.A. Christensen (John Wiley, New York, 1988).

The Physics of Medical Imaging, edited by S. Webb (IOP Publishing, New York, 1988).

Principles of Medical Imaging, by B. Tsui, M. Smith, and K.K. Shung (Academic Press, New York, 1992).

116.3 Semiconductor Detectors For Radiation Measurements

Larry A. Franks, Larry S. Darken, and Ralph B. James

Since their introduction in the early 1960s, **semiconductor radiation** detectors have become the devices of choice for numerous X-ray, gamma-ray, and charged particle measurements. They are essential in applications where maximum **energy resolution** (i.e., the ability to record the energy of the incident photon or particle) is required. In this characteristic, their performance greatly exceeds that of gaseous sensors (proportional counters, for example) or scintillator/photocell-based spectrometers. Their superior energy resolution stems, in the main, from the greater number of information carriers generated in semiconductors per unit of absorbed energy than in, for example, scintillator/photocell combinations, which are widely used as low-resolution spectrometers. In scintillator-based spectrometers, approximately 100 eV of absorbed energy is required to generate a single information carrier. In a typical semiconductor, only 3 to 5 eV are required to create one electron-hole pair — the information carrier in semiconductor detectors. The statistical variation of the number of carriers produced per ionizing event is thus substantially greater in the case of the scintillator and is reflected in reduced energy resolution. Similar arguments apply to gaseous detectors, where the relatively small number of information carriers is due to the combination of the low density of the absorbing gas and the significant energy required to produce ion-pairs (≈30 eV per ion-pair, [1]), the information carrier in the gas detector.

It is convenient to divide semiconductor detectors into two groups: those requiring cooling (normally to 77K) and those capable of room temperature (or near room temperature) operation. The former group is dominated by high-purity germanium (HPGe) and lithium drifted silicon (Si:Li) — a group characterized by particularly high energy resolution. The latter group includes detectors based on cadmium telluride, cadmium zinc telluride, and mercuric iodide, as well as a number of silicon-based devices. This group finds application in portable X-ray and gamma-ray spectrometers and counters and imaging systems where the freedom from the weight and maintenance of a cryogenic cooler is particularly valued. This chapter section is divided into sections on cryogenic detectors and room-temperature devices. Details of the physics of semiconductor detectors as well as their performance characteristics can be found in several texts [1–7].

Cryogenic Detectors

Detectors in this group are limited principally to high-purity germanium (HPGe) and lithium drifted silicon (Si:Li). HPGe detectors are available in a number of configurations for operation in the X-ray and gamma-ray regions. Si:Li detectors are primarily for the X-ray region. Cryogenic detectors provide the ultimate in energy resolution, as well as good detection efficiency. Common to the group is the requirement that they be operated in the region of 77K. In most cases, the detector temperature is maintained by liquid nitrogen (LN_2) contained in an attached cryogenic vessel (dewar). Dewars are available in capacities ranging from a few liters to several tens of liters, depending on the service interval that is acceptable and the degree of portability required. Alternatively, electromechanical coolers are also available.

Germanium Detectors

High-purity germanium (HPGe) detectors are widely used for gamma-ray spectroscopy due to their combination of efficient absorption and high energy resolution. Figure 116.7 shows the cross sections for photoelectric absorption, Compton scattering, and absorption by electron-positron pair production in several materials used for solid-state nuclear radiation detectors. Attenuation is significantly stronger in germanium than in silicon. Over much of the gamma spectrum, the dominant interaction is Compton scattering. However, it is principally the stronger photoelectric absorption in germanium that makes it more suitable than silicon for gamma-ray spectroscopy. In the typical germanium detector, a gamma-ray can be scattered several times before it is photoelectrically absorbed. Thus, the energy of the gamma-ray is primarily transmitted directly to a small number of electrons. These energetic electrons, in turn, interact with electrons in the valence bands to create mobile pairs of electrons and holes. The average number of electron-hole pairs N produced by a completely absorbed gamma-ray of energy E becomes independent of the details of the initial reaction path and varies linearly with E:

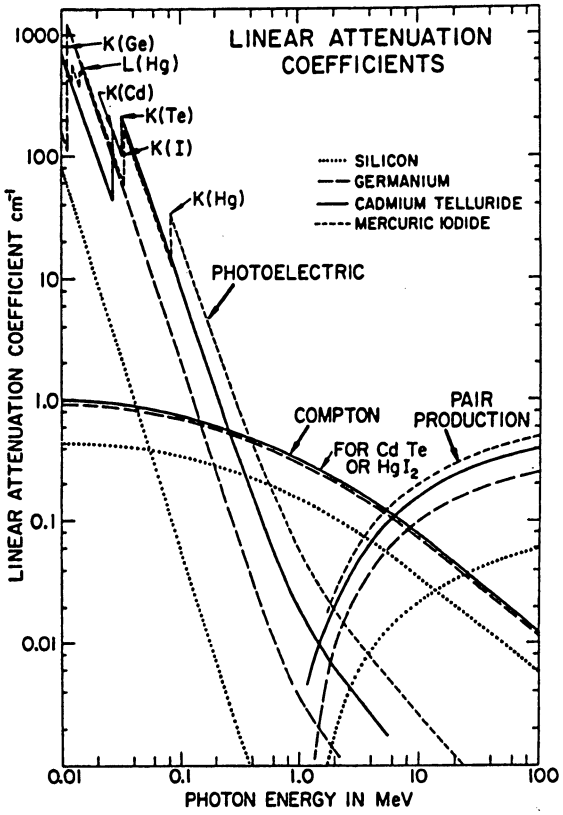


FIGURE 116.7 Attenuation coefficients vs. energy common semiconductor materials. (T. E. Schlesinger and R. B. James, *Semiconductors for Room Temperature Detector Applications.*)

$$N = E/\varepsilon \tag{116.5}$$

This relationship is more broadly valid and is the foundation of energy spectroscopy of gamma-rays using semiconductors, gases, and cryogenic liquids (ε depending on the material). While ε is independent of the gamma-ray energy (and is also virtually the same for energy deposited by charged particles), ε in germanium does increase slightly with decreasing temperature, as does the energy gap. At 77K, ε is 2.96 eV and the energy gap is 0.72 eV.

Practical exploitation of Eq. (116.5) depends on electronically detecting the motion of the ionized charge in an electric field. The signal-to-noise ratio is improved by reducing current flow in the detector from other mechanisms. In germanium, this is achieved by producing a rectifying and a blocking contact and by cooling to about 100K. For a planar detector, a slice of high-purity germanium is diffused with the donor lithium on one side, forming a strongly *n*-type layer. The opposite side is implanted with the acceptor boron, forming a p^+ layer. When voltage is properly applied, the electric field direction prevents the majority carriers in the contact regions from being injected across the device. As the voltage is applied, a region depleted of holes will advance into the slice from the n^+ contact if the slice is *p*-type. If the slice is *n*-type, the a region depleted of electrons will advance from the p^+ contact. At the depletion voltage V_d, the depletion region reaches the opposite contact. For germanium:

$$V_d = 565\ V * \left|N_A - N_D\right| \Big/ 10^{10}\ dm^{-3} * d^2 \Big/ cm^2 \tag{116.6}$$

Here, $N_A - N_D$ is the net charge density in the depleted or active region of the detector and d is the thickness of this region. This is a key relationship in high-purity germanium technology as it quantifies the effect of the residual impurity concentration on device size and depletion voltage. Techniques to grow germanium pure enough for gamma detectors were pioneered by Hall [8] and the detector group at Lawrence Berkeley Laboratory [Haller, Hansen, and Goulding [9], based on purification methods of Pfann [10] and crystal-growing techniques developed by Teal and Little [11] to produce crystals for germanium transistors.

Leakage Current

Germanium detectors need to be cooled to reduce **leakage current.** There are several potential sources of leakage current, including diffusion of minority carriers from either doped contact into the depletion region, thermal generation of carriers at either bulk or surface defects in the depletion region, and electrical breakdown at points where the electrical field is concentrated due to irregularities in the contact geometry, large-scale inhomogeneities in the bulk, or surface states. Current will also be generated if the detector is not shielded from room-temperature infrared radiation. Background nuclear radiation from materials near the detector and cosmic radiation also generate current.

Germanium detectors are typically liquid-nitrogen cooled and operated between 85K and 100K. In this temperature range, leakage current is typically less than 40 pA in "good" detectors, and is not a significant contributor to system noise (400–900 eV). Leakage current increases with temperature and eventually becomes the predominate noise component. Pehl, Haller, and Cordi [12] reported a leakage current-driven system noise of 2 KeV at 150K and 7 KeV at 170K for an 8 cm^3 planar detector. These authors also reported that above about 120K, the leakage current had an activation energy of approximately one-half the bandgap, and attributed this to generation at mid-gap surface states. Below 120K, the temperature dependence was milder.

A typical detector/cryostat configuration is shown in Fig. 116.8. The detector resides in an evacuated cryostat and is cooled by means of a copper rod inserted into a liquid nitrogen dewar. The first stage of amplification is an FET, also cooled, positioned nearby the detector. Mechanical fixturing is designed to stabilize the detector and the mechanisms for contacting it, to provide a cooling path from the detector to liquid nitrogen, and to electrically insulate the high-voltage contact.

A variety of detector geometries is shown in Fig. 116.9. These different electrode configurations allow the detectors' efficiency and **energy resolution** to be optimized for different gamma-ray energies and applications. For example, the detector in Fig. 116.9(c) minimizes noise by the lower capacitance of its electrode configuration at the expense of the reduced stopping power. Thus, this detector would be more suitable for lower-energy gamma-rays.

Coaxial Detectors

The detector type shown in Fig. 116.8 and in Fig. 116.9(e) has a closed-end coaxial geometry. Nearly all the largest volume (active volumes of 100 cm^3 to 800 cm^3) HPGe detectors are of this type. This electrode geometry reduces both capacitance and depletion voltage with respect to a planar detector of the same volume. This latter benefit relaxes the constraint on material purity. In addition, charge collection distances are shortened, and the uncontacted surface area, frequently troublesome in processing, is reduced. Also, the HPGe is grown by the Czochralski technique, and is therefore nearly cylindrical even before machining. It is important, however, to note that the reduction in depletion voltage is realized only when the device is contacted so that it depletes from the outer contact to the inner contact. Thus, *p*-type HPGe to be fabricated into a coaxial detector is lithium diffused on the outer diameter; and in the case of *n*-type HPGe, the outer diameter is boron implanted.

The boron-implanted contact (depth approximately 0.2 μ) is thinner than the lithium-diffused contact (depth approximately 750 μ), so the *n*-type coaxial detector can detect lower-energy radiation and is usually built with a beryllium window in the aluminum end-cap to take full advantage of this feature. The difference in the range of use is illustrated in Fig. 116.10. The geometric asymmetry of the contacting electrodes in the coaxial detector makes charge collection more dependent on the carriers (electrons or holes) traversing to the inner contact. As more gamma-rays are absorbed near the outer contact, the carriers traversing to the inner contact must travel on average a longer distance. Also, charge traversal near the inner contact is particularly effective in inducing current in the external circuit [13]. Thus, the *p*-type coaxial detector with positive bias on the outer electrode is more sensitive to hole collection, and the *n*-type coaxial detector with negative bias on the outer electrode is more sensitive to electron collection. This is a crucial consideration in applications where hole collection is going to be degraded during use by exposure to fast neutrons or other damaging radiation. The

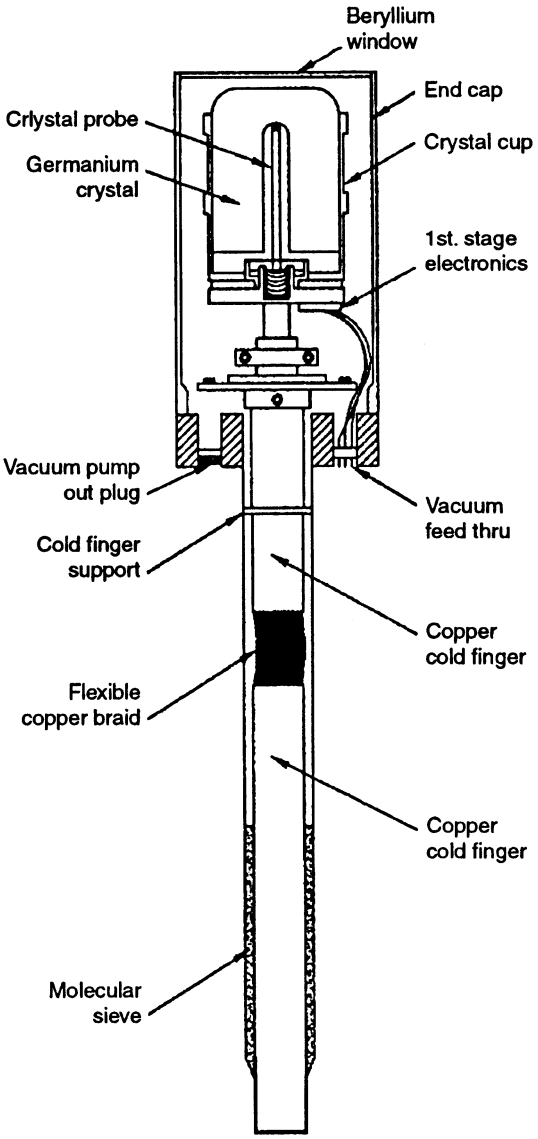


FIGURE 116.8 Schematic cross section of a dipstick cryostat. (Darken and Cox, 1993; reprinted with permission of Oxford Instruments, Inc.)

superior neutron damage resistance of the electrode biasing polarity on *n*-type coaxial detectors was demonstrated by Pehl et al. [14].

A typical gamma-ray spectrum of a ^{60}Co source taken with a coaxial HPGe detector is shown in Fig. 116.11. The salient features are the full-energy peaks at 1.17 MeV and 1.33 MeV, and the lower-energy plateaus due to incomplete energy absorption of Compton scattered gamma-rays. The peak-to-Compton ratio [15] is generally 40 to 100, depending on the size and quality of the detector. The 1.33-MeV **peak** is shown separately in Fig. 116.12. The **energy resolution** measured as the full width at half the peak maximum (FWHM) for typical coaxial germanium detectors is between 1.6 KeV and 2.1 KeV for 1.33-MeV gamma-rays, again depending on the size and quality of the detector. The variance in the peak L^2 (FWHM = 2.35 × L, L being the standard deviation for a Gaussian distribution) can be divided into three additive components: the electronic noise component L_N^2, a component reflecting the variance in the number of electron-hole pairs created L_F^2, and a component due to incomplete charge collection L_T^2:

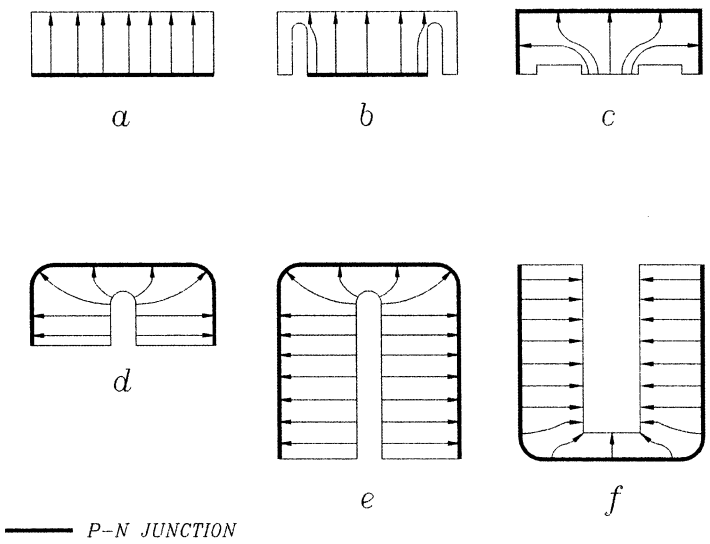


FIGURE 116.9 Schematic cross section and electrostatic field distribution in high-purity germanium detectors. The dark line represents the *p-n* **junction:** (a) true planar, (b) grooved planar, (c) low-capacity planar, (d) truncated coaxial, (e) closed end coaxial, (f) well geometry.

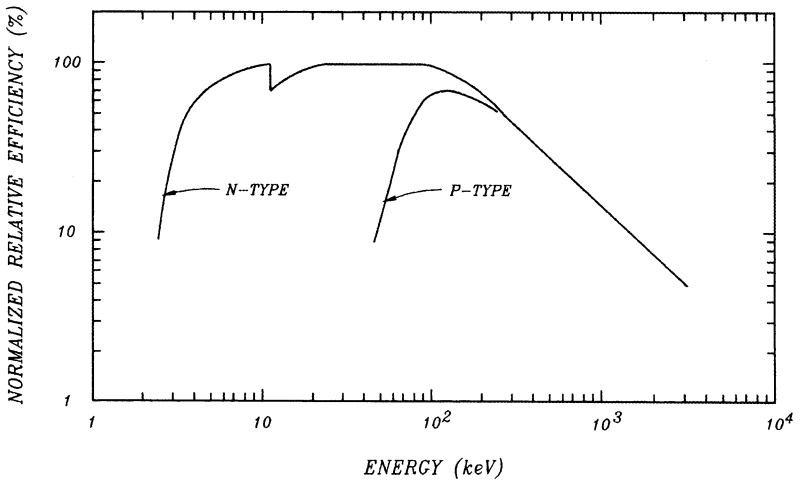


FIGURE 116.10 Relative absorption efficiencies for typical *n*- and *p*-type detectors. (Darken and Cox, 1993; reprinted with permission of Oxford Instruments, Inc.)

$$L^2 = L_N^2 + L_F^2 + L_T^2$$
$$L_F^2 = \varepsilon E F \tag{116.7}$$

F is called the Fano factor and has been experimentally determined to be no greater than 0.08 for germanium [16]. $F < 1$ implies that electron-hole pair creation events are not uncorrelated. L_T^2 is usually dominated by the trapping of electrons and holes at defect sites. However, shorter electronic shaping times, lower electric fields, and larger detectors accentuate **ballistic deficit** (loss of collected charge in the external electronics due to the finite traversal time of the electrons and holes across the detector). L_N^2 is independent of gamma-ray E and is

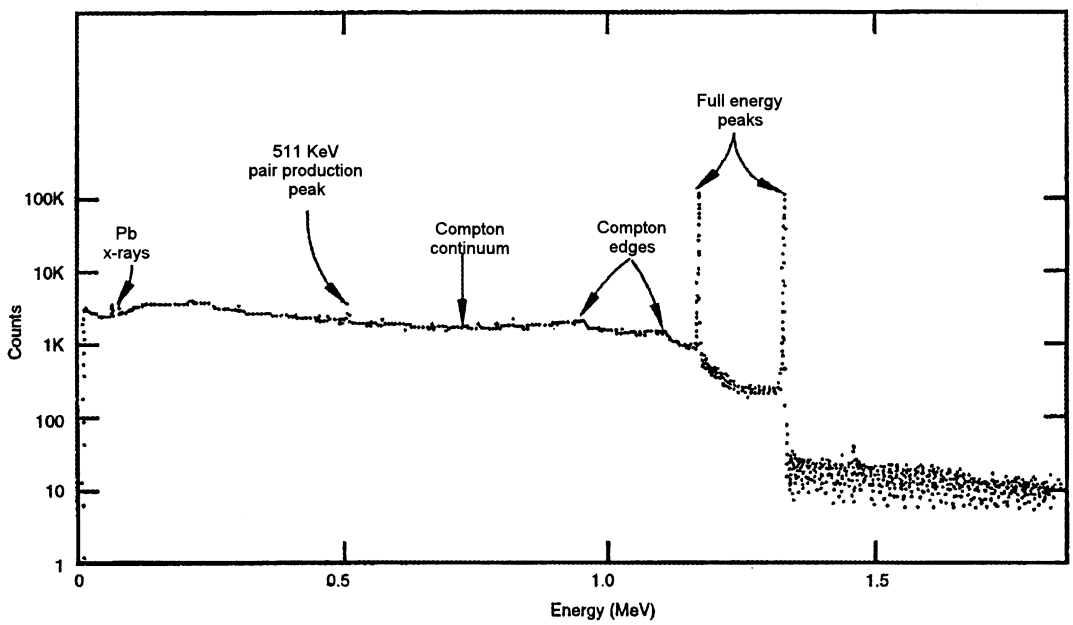


FIGURE 116.11 A ^{60}Co spectrum collected with a 15% *p*-type detector showing typical features of the germanium detector spectrum.

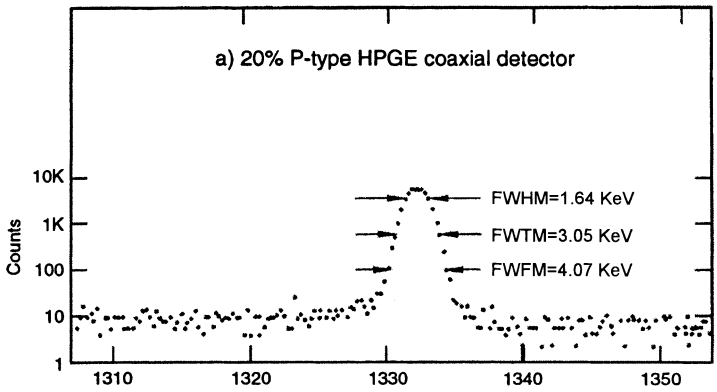


FIGURE 116.12 A ^{60}Co spectrum collected with a 22% relative efficiency *p*-type detector. (Darken and Cox, 1993; reprinted with permission of Oxford Instruments, Inc.)

the dominant resolution limiting factor at low energies. L_F^2 depends linearly on E, and for a coaxial detector usually dominates L_N^2 for E over a few hundred KeV. The energy dependence of L_T^2 is not given simply from first principles for an arbitrary trap distribution, but an E^2 dependence seems to fit under many circumstances. Thus, at high enough E, L_T^2 is expected to be the largest component. For "good" detectors at 1.33 MeV though, L_T^2 is always smaller than L_F^2. However, the magnitude of L_T^2 is variable enough between detectors that it distinguishes between acceptable, very good, and excellent detectors. L_T^2 is usually also the only component of resolution drawn from a nongaussian distribution and is thus responsible for any low-energy tailing of the peak.

X-ray Detection

Both silicon and germanium detectors are used in low noise systems for the detection of fluorescent X-rays produced by electron beams (usually in an electron microscope) or X-rays (XRF). For both materials, the detector is liquid-nitrogen cooled to reduce **leakage current,** and small volume devices (Fig. 116.9, typically

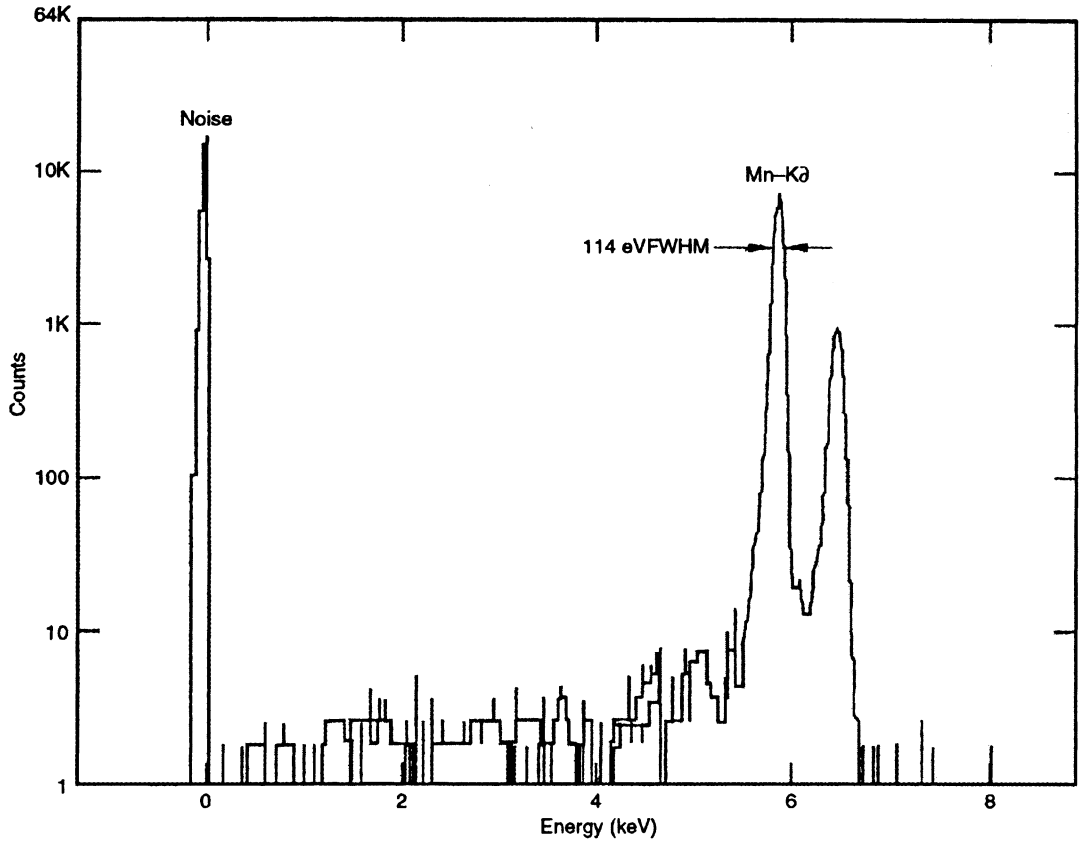


FIGURE 116.13 Manganese X-ray spectrum from an ^{55}Fe source collected with an HPGe detector. (Darken and Cox, 1993; reprinted with permission of Oxford Instruments, Inc.)

10 mm^2 active area, and 3 mm depth) are used in order to decrease capacitance and therefore further reduce electronic noise. Lithium-drifted silicon (Si:Li) detectors were used first for these applications. Early germanium detectors displayed poor peak shape for X-ray energies just above the L absorption edges (attributed to diffusion against the field to the front contact by some electrons and their resulting loss to the photopeak [17]). However, as was first demonstrated by Cox et al. [18], this is not a fundamental problem, but can be solved by the contacting technology. An X-ray spectrum taken with an HPGe detector is shown in Fig. 116.13. Germanium has the advantages with respect to silicon of a smaller ε (2.96 eV/pair vs. 3.96 eV/pair at 77K) for better **energy resolution,** and a higher Z (32 vs. 14) for better photoelectric absorption of higher-energy X-rays.

Current Status of HPGe Detector Technology

High-purity germanium detectors are a mature commercial technology. Process development in crystal growing and diode fabrication have been conducted in private industry where significant advances are proprietary. However, the results of technological advances in these areas are quite evident in the continual improvement in the size, performance, and availability of HPGe detectors. Maximum photopeak efficiency for HPGe gamma-ray detectors is doubling every 6 to 8 years. Concurrently, energy resolutions are moving toward the theoretical limits of Eq. (116.7) as the concentrations of trapping centers are reduced.

The reliability as well as the performance of germanium gamma-ray detectors has also continued to improve, although this is harder to quantify. Cryostats have been redesigned to reduce virtual and direct leaks, reduce microphonics, implement modular design, and improve ruggedness. Detector makers are also making more serious attempts to offer models with reduced backgrounds by judicious design changes and careful selection of materials.

New applications for gamma-ray spectroscopy have emerged. The HPGe detector industry has recently supplied over 100 detectors each to two different experimental facilities (GAMMASPHERE in the U.S., and EUROBALL in Europe), where they were arranged spherically in a modular fashion around the target of an ion accelerator to study the decay of nuclei from excited states of high angular momentum.

For users of single detector systems, developments in the pulse processing electronics necessary for data acquisition and in the hardware and software for data analysis have resulted in both more compact and more flexible systems. Plug-in cards for a personal computer are available now that not only contain the functions of the ADC and multichannel analyzer, but also the high voltage power supply and amplifier as well. Software developments also allow for control of many pulse processing parameters that were previously set manually.

Si:Li Type Silicon Detectors

As with germanium for gamma-ray spectroscopy, the impurity requirements on silicon for nuclear radiation detectors are also stringently low and difficult to obtain. Such silicon must be grown by the float zone technique to eliminate contamination from a crucible. Unlike germanium, little dedicated effort has been expended trying to improve silicon growth techniques to achieve superior detector characteristics. Most progress in material quality has come from technology improvements aimed at other applications. The purest silicon commercially available typically has a net electrically active impurity concentration of a few times 10^{11} cm^{-3} (compared to 10^{10} cm^{-3} for HPGe), which usually limits device thicknesses to less than a millimeter. When thicker silicon devices are required, as in X-ray spectroscopy, silicon of higher net purity can be obtained by lithium drifting [19], but such material cannot be subsequently processed above room temperature.

In X-ray spectroscopy for microanalysis, liquid-nitrogen cooled Si:Li detectors are being challenged by similarly sized HPGe detectors, but Si:Li detectors are still more widely used. Recent developments in specialized low-noise silicon drift detectors and CCD-based detector (see Room Temperature Semiconductor section) designs that have yielded promising results at room temperature may find future application in liquid nitrogen-cooled systems for microanalysis using X-ray spectroscopy.

Room-Temperature Semiconductors

Applications arise that require **energy resolution** beyond the capability of scintillator systems and where cryogenically cooled semiconductors are not suitable. Examples include detector probes for monitoring restricted areas, monitoring at remote sites where replenishing the coolant is impractical, spectral imaging, and many portable instrument applications. There is available a class of semiconductor detectors that satisfy many such needs by providing energy resolution substantially better than the best scintillators (although inferior to cryogenic semiconductors) while operating at ambient temperature. In addition to spectroscopy, these devices are also useful for counting applications where high detection efficiency per unit volume is required. In these applications, the devices are operated in **pulse mode** wherein the charge associated with single photon absorption events is recorded. They also can be operated in a **current mode** in the manner of a solid-state ion chamber. In their current stage of development, room-temperature detectors are limited in size and best suited for the energy region below 1 MeV.

There are numerous detector types in the general category of room-temperature detectors: those capable of ambient temperature operation (true room-temperature devices) and those requiring cooling to the region of –30°C. The former group includes wide bandgap materials such as cadmium zinc telluride, mercuric iodide, cadmium telluride, and a number of silicon structures. The latter group consist primarily of *p*-intrinsic-*n* (PIN) silicon devices. As a group, room-temperature detectors are employed mainly in X-ray and gamma-ray spectroscopy, although silicon surface barrier and ion-implanted devices are highly valued for charge particle spectroscopy.

True Room-Temperature Detectors

True room-temperature detectors are distinguished from cryogenic semiconductors by the magnitude of the energy gap that separates the normally vacant conduction band from the highest filled band. If this energy gap is small, as in the case of germanium (0.67 eV), electrons can be thermally stimulated across the bandgap at room temperature. The resultant current competes with the gamma-ray generated signal, precluding room-temperature operation of germanium (and many high-resolution applications of silicon). Thermally induced

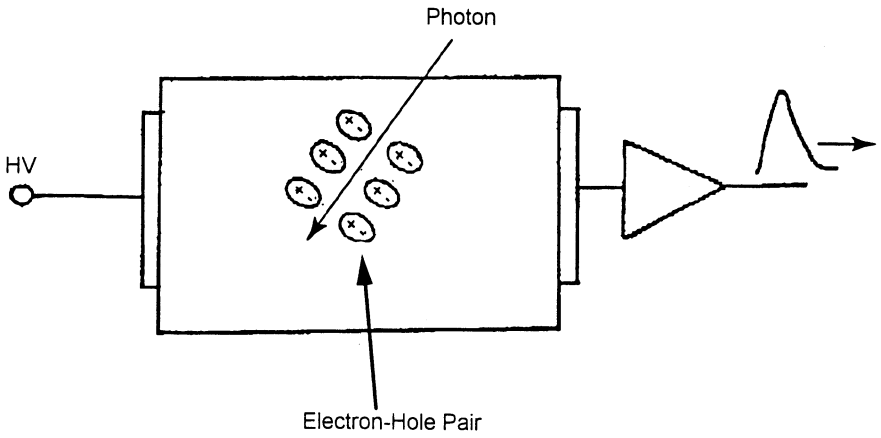


FIGURE 116.14 Schematic illustration of charge generation in a planar detector.

current is reduced to acceptable levels at bandgap energies of about 1.4 eV and above. This phenomenon has been successfully exploited in the development of room-temperature detector materials, including cadmium zinc telluride (acronym CZT), cadmium telluride (CdTe), and mercuric iodide (HgI_2).

Theory of Operation

Operating principles of room-temperature detectors are similar to those governing the more familiar cryogenic semiconductor devices. Gamma-radiation is absorbed in the material and generates electron-hole pairs that move under the influence of an applied electric field to contacts and external electronics for processing and production of the familiar pulse height spectrum. The process is shown schematically in Fig. 116.14. Fundamental to the charge-transfer process is the carrier mobility (μ) and the carrier life time (τ). The product $\mu\tau E$ defines a drift length (λ), which should be long compared to the inter-contact dimensions. Owing to the substantially higher average atomic number of the room-temperature detector materials, the probability of gamma-ray absorption is much higher than in silicon or germanium (see Fig. 116.7). As a result, room-temperature detectors provide greater detection efficiency per unit volume.

The energy required to produce an electron-hole pair (ε) is typically a few times the **energy bandgap** of the material. In silicon where the bandgap is 1.14 eV, the energy to produce an electron-hole pair (ε) is 3.62 eV. In HgI_2, where the bandgap is 2.13 eV, about 4.2 eV is required to create an electron-hole pair. The absorption of a 1 MeV photon thus produces about 276,000 e-h pairs in silicon and 240,000 e-h pairs in HgI_2. Values of ε for room-temperature materials are in the region of 4.2 to 5.0 eV/e-h pair (see Table 116.2) and consequently fewer electron-hole pairs are generated per unit of absorbed energy. Complete collection of the charge is desired although charge trapping, which may not affect the two carrier types equally, prevents this in most cases. The drift length for holes (λ_h) in these materials is often less than the inter-contact dimensions and creates a condition where the collection efficiency depends on the photon interaction depth. This phenomenon is illustrated in Fig. 116.15 where induced charge from single gamma absorption events originating at various

TABLE 116.2 Physical Parameters of Common Room-Temperature Semiconductor Materials

Material	E_g, eV	Z	ε, eV	ρ, Ω	$(\mu\tau)_e$ cm^2/V	$(\mu\tau)_h$ cm^2/V
Cadmium zinc telluride	1.65	48	5.0	10^{11}	1×10^{-3}	6×10^{-6}
Cadmium telluride	1.5	50	4.4	10^{9}	3.5×10^{-3}	2.3×10^{-4}
Mercuric iodide	2.13	62	4.2	10^{13}	1×10^{-4}	4×10^{-5}

Note: E_g = bandgap energy; Z = average atomic number; ε = energy to create an electron-hole pair; ρ = resistivity.

Source: Semiconductors for Room-Temperature Radiation Detector Applications, R. B. James, T. E. Schlesinger, P. Siffert, and L. A. Franks (Eds.), Materials Research Society, Vol. 32, Pittsburgh, PA, 1993.

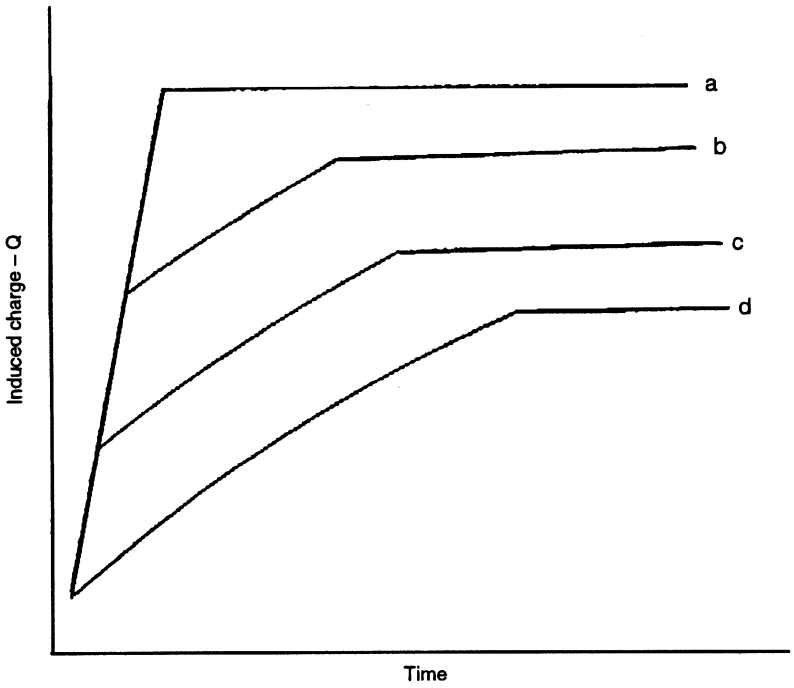


FIGURE 116.15 Charge collection in a planar detector for single-photon interaction. Curves a to d depict the charge from photon interactions at increasing depths below the cathode.

depths in the material is plotted as a function of time. The initial fast-rising segment is due to the more mobile electrons; the slower component is due to holes. In this example, hole trapping is assumed and manifests itself in the curvature of the hole segment. The charge collection efficiency (η) can be derived from the Hecht relation [20]. For a photon absorbed at a distance X from the cathode of a planar detector of thickness L operated with a uniform electric, the relationship becomes

$$\eta = \frac{\lambda_e}{L}\left[1-\exp\frac{\left(L-\mathrm{X}\right)}{\lambda_e}\right]+\frac{\lambda_h}{L}\left[1-\exp\frac{-\mathrm{X}}{\lambda_h}\right] \qquad (116.8)$$

The dependence of the collection efficiency on interaction depth reduces **energy resolution** and without mitigation would limit high resolution to thin devices. Fortunately, methods have been developed that permit high-energy resolution to be achieved in relatively thick samples. As with cooled semiconductor detectors, the energy resolution of the combined detector-electronics system is normally specified by the full width of a monoenergetic spectral peak at its half amplitude points (ΔE). The FWHM is, in turn, related to the variance in the peak L^2 (see Eq. 116.7). It is useful to note that the energy resolution is inversely related to the product $\mu\tau$.

Operational Considerations

Important physical parameters for the leading room-temperature detectors are summarized in Table 116.2. Detectors are available with surface areas of a few square centimeters and thicknesses up to about 1 cm. The performance of detectors based on the different materials varies considerably, as can the performance for detectors of the same material. The choice of specific detector material is normally dictated by the application. The exceptionally high resistivity and high photoelectric cross section in mercuric iodide permit good resolution and high efficiency in the X-ray region, particularly below 10 KeV. For example, ΔE of 5% at 5.9 KeV has been reported [21] with 1 cm^2 devices, with typical values in the region of 10%. For mercuric iodide applications in the region of 0.5 MeV, trade-offs between efficient gamma absorption and resolution may be required.

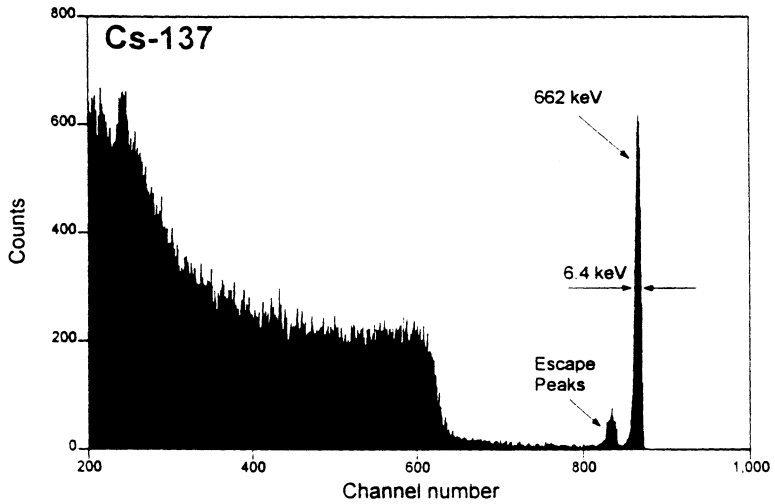


FIGURE 116.16 Energy spectrum of a ^{137}Cs source obtained with a multi-electrode CZT spectrometer. (Courtesy AMPTEK Inc. Bedford, MA.)

In gamma-ray applications where **energy resolution** is the primary concern, thinner devices that minimize charge trapping are often utilized, although procedures have been developed to mitigate many of the thickness and area limitations. Two general approaches have been developed to mitigate the effects of incomplete hole collection. Those approaches are based on electronic pulse processing via ancillary circuitry or novel electrode structures that achieve monopolar (electron transport only) operation. These methods have been particularly successful with CZT and CT devices. CZT- and CT-based devices incorporating such technology are available in commercial products. In CZT, for example, $1.5 \times 1.5 \times 0.8$ cm^3 detectors (electron only collection) are available with 3% resolution at 662 KeV. Somewhat better resolution can be obtained with smaller devices. For example, $1 \times 1 \times 0.5$ cm^3 detectors, operated in the electron-only mode, provide 2% at 662 KeV. Similarly, cylindrical detectors 6 mm in diameter and 5 mm high provide about 1% resolution at 662 KeV. An **energy spectrum** obtained with a typical CZT spectrometer obtained with a ^{137}Cs source is shown in Fig. 116.16.

Improvements in material quality can be expected to further improve the performance of thick detectors. While the area of single crystal detectors is limited currently to a few square centimeters, electronics have been developed to facilitate the operation of large arrays of single-crystal detectors and thus achieve high detection efficiency. Multiple CZT detector arrays 20×20 cm have been produced using such technology.

Further information concerning the performance of single charge collection and array structures as well as electronic processing and design details are available in the literature [2, 3, 22, 23].

Silicon Detectors

In contrast to germanium detectors, silicon detectors can be operated at room temperature in applications where some current noise can be tolerated. Compared to gas and scintillation detectors, silicon detectors have good energy resolution and are reasonably compact. They are fabricated from slices of a silicon single crystal and are available in a variety of areas (25 mm^2 to 3000 mm^2), and the active thickness is usually a few hundred microns. Specialized detectors have been developed for a wide variety of applications.

Charged Particle Detectors

Energetic heavy-charged particles lose kinetic energy continuously along a linear path in an absorbing material. Energy is transferred primarily to the electrons in the absorbing material, but to a lesser extent to the nuclei also via Rutherford scattering. Although only energy transferred directly to the electronic system generates electron-hole pairs, Eq. (116.5) (with $\varepsilon = 3.62$/pair for silicon at 300K) is still a good approximation. Energy loss is characterized by two parameters: specific ionization loss dE/dx, which depends on the incident particle, its energy, and the absorbing material, and the range R (i.e., the penetration depth of the particle), which

determines the detector thickness required for complete energy absorption. The continuous nature of energy loss leads to substantial window effects.

Diffused Junction Detector

Silicon detectors can be generically categorized by the type of rectifying contact employed. The diffused junction detector is fabricated by diffusing phosphorus from the gas phase into *p*-type silicon. This is a high-temperature (900°C to 1200°C) operation that is prone to introducing faster diffusing metals into the bulk that can act either as generation centers increasing **leakage current,** or as trapping centers degrading charge collection. The thickness of the diffused region, from 0.1 to 2.0 μ, also presents a **dead layer** to incident particles that is reduced in alternate technologies. Nonetheless, these detectors find use due to their ruggedness and economy.

Surface Barrier Detector

Surface barrier junctions are fabricated by either evaporating gold onto *n*-type silicon or aluminum onto *p*-type silicon. A typical entrance window is equivalent to 80 nm of silicon. The rectification properties depend on the charge density of surface states of the silicon and of the thin oxide layer over the silicon, as well as on the evaporated metal. The wafer is epoxied in an insulating ring before metallization. The finished detector is encapsulated in a can that has a front window for particle entry and a single contact in the back for the combined function of applying bias and for extracting the signal pulse. Devices can be operated either in the partially depleted or totally depleted mode. As fabrication is entirely at room temperature, there is no opportunity for metal contamination by diffusion. Generally, surface barrier detectors have lower leakage current, and less system noise than a diffused junction detector of comparable area and depth. However, detectors currently fabricated by ion implantation have still lower leakage current and electronic noise, together with a thinner and more rugged front contact. On the other hand, implanted detectors are not available in the same range of active thicknesses as surface barrier detectors. Below 100 μ and above 500 μ, only surface barrier detectors are currently available. Surface barrier detectors can be made in small quantities with rather simple equipment.

Ion-Implanted Detector

A simplified representation of ion-implanted detector fabrication is shown in Fig. 116.17. The first successful implementation of silicon planar processing to silicon detectors was reported by Kemmer [24]. The procedure starts with the thermal growth of an oxide film on a high-purity, *n*-type silicon wafer. Windows are then opened in the oxide by photolithographic techniques. The front contact area is implanted with boron to form the rectifying contact, and arsenic is implanted into the backside. The wafer is then annealed to activate the implant, and aluminum is evaporated on both sides to reduce sheet resistivity. Typical entrance windows are 50 nm silicon equivalent. Electrical connections are made by wire bonding to the aluminum layers. Finished detectors are canned in a manner similar to surface barrier detectors. More than one detector can be fabricated on the same wafer using the appropriate masks during photolithography. In fact, quite elaborate detector geometries can be achieved via photolithography. The detector in Fig. 116.17 is actually a strip type.

The ion implantation planar process technology is well suited for mass production of wafer sizes compatible with the rest of the silicon industry. Minimum wafer diameters are now 4 in. or 5 in. At this diameter, breakage during fabrication is an issue for thicknesses less than 150 μ. For thicknesses over 500 μ, the availability of enough sufficiently pure material to justify the cost of photolithographic masks is an issue. Ion-implanted detectors can be baked at 200°C to reduce outgassing. This is a significant improvement over surface barrier detectors, which irreversibly degrade by device processing above room temperature. This is a useful feature as most heavy charged particle spectroscopy is done in a vacuum.

Leakage currents are, at room temperature, typically 1 to 10 nA cm^{-2} active area and per 100-μ depletion depth. These values represent an order of magnitude reduction in leakage current with respect to surface barrier detectors. Two factors are relevant. Passivation of silicon surfaces by thermal oxidation is extremely effective in reducing leakage current around the rectifying contact. Also, the bulk generation current is reduced by the gettering of metal impurities during the high-temperature oxidation. Float zone silicon for radiation detectors usually has a minority carrier lifetime longer than 1 ms, and this can be increased an order of magnitude during detector fabrication [25]. Thus, not only is leakage current reduced, but potential charge collection problems are also eliminated.

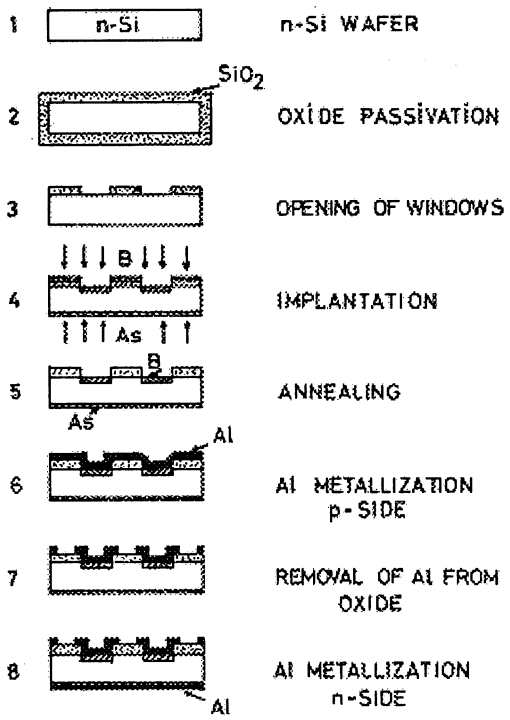


FIGURE 116.17 Steps in the fabrication of passivated planar silicon diode detectors. (Kemmer et al., 1982).

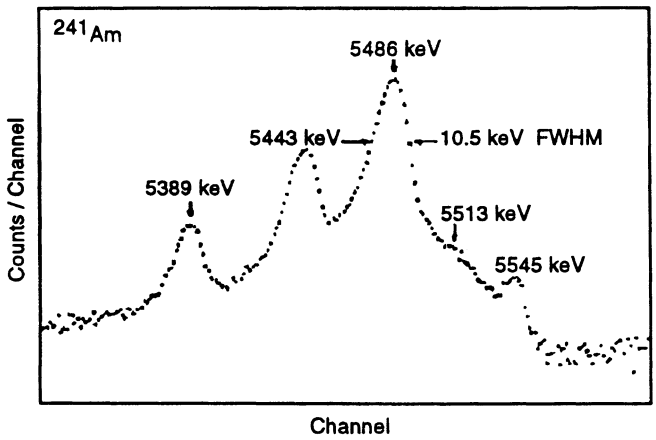


FIGURE 116.18 Spectrum of a ^{241}Am alpha-particle source (log scale) measured with an IP detector (25 mm^2 area, 300 μm thick) at room temperature. Resolution at 5.486 MeV is 10.6 KeV (FWHM). (Kemmer et al., 1982.)

Energy Resolution

A typical spectrum of an ^{241}Am alpha-particle source taken with an ion-implanted detector is shown in Fig. 116.18. While the factors considered in Eq. (116.7) for germanium gamma-ray spectrometers are still valid, additional considerations also apply. In particular, if the source is moved closer to the detector to improve collection efficiency, larger differences in the angle of incidence will produce peak broadening due to larger variation in effective window thickness. Even when the source is sufficiently distanced from the detector, there

will still be spatial variations in window thickness, as well as some variation in energy lost escaping from the source and traversing to the detector.

Another source of peak broadening is the variation in the small amount of particle energy lost during Rutherford scattering. This energy is transmitted directly to the scattering nuclei and does not generate electron-hole pairs, and a small pulse deficit results. These events are relatively few but large, and therefore contribute disproportionately to peak variance. The FWHM contribution of this effect on a 6-MeV alpha particle peak has been estimated to be 3.5 KeV [26].

Silicon Detectors for Spatial Resolution

The uninterrupted progress of the semiconductor silicon industry in achieving both larger wafers and smaller device features has allowed the development of larger and more complex silicon detectors that can provide position information in addition to (or instead of) energy information. Spatial detection can be obtained by fabricating detectors as pixels (two-dimensional) or strips (one-dimensional) on the same wafer. For penetrating radiation, two strip detectors, one behind the other but with the strip pattern rotated 90°, provide two-dimensional positioning. Frequently, such detectors are individually designed and fabricated for a particular application. Strip detectors and CCD (charge-coupled device) detectors will be discussed here.

Strip Detectors

Silicon strip detectors are currently fabricated on silicon wafers (typically approximately 300 μ thick) using photolithographic masking to implant the rectifying contact in strips [27]. The strips usually have a pitch on the order of 100 μ and a width less than half of this to minimize strip-to-strip capacitance and hence electronic noise [28]. The device is biased past depletion, and the back blocking contact is continuous. Each strip requires, in principle, its own signal processing electronics; however, charge division readout (capacitive or resistive) can reduce the number of amplifiers by a factor of 10. Detectors are fabricated in rectangular segments from a single wafer, and can be ganged together if a larger area is needed.

Strip detectors are well established in high-energy physics experiments for reconstruction on the micron scale of the tracks of ionizing particles. The particles being tracked result from the collision of accelerated particles with a target and are highly energetic ($>10^{10}$ eV). Frequently, experimental interest is focused on short-lived particles created in the collision but which decay before they can be directly detected. Spatial resolution of the decay vertex from the original collision is necessary to detect such a particle and to determine its lifetime.

The requirements of new high-energy experiments and advances in silicon technology have produced much evolution and innovation in the strip detector concept. For example, a double-sided microstrip detector with an oxide-nitride-oxide capacitor dielectric film has been reported [29]. The use of intermediate strips to improve spatial resolution has become common [30], and the biasing network has been integrated onto the detector [31].

Silicon Drift Detectors

Silicon drift detectors were first proposed by Gatti and Rehak [32] as an alternative to silicon strip detectors in high-energy physics experiments. The primary motivation was to significantly reduce the number of readout channels. Drift detectors have subsequently been adapted for X-ray spectroscopy. These detectors are usually fabricated on *n*-type silicon wafers with holes collected to either a p^+ contact on the back side of the detector, or to concentric annular p^+ contacts on the front side. The detector is depleted from both sides. The reverse bias applied to the p^+ annular rings is varied in such a way that electrons are collected radially in a potential energy trough to an n^+ anode at the center of the detector on the front side.

A cross section through a circular drift detector is shown in Fig. 116.19. The electron collecting anode ring surrounds the integrated FET used for the first stage of signal amplification. Enough negative bias is applied to the back contact (actually the entrance window) to deplete the wafer to the anode, which is near ground potential. At the same time, negative bias progressively increasing in magnitude is applied from the ring next to the anode (near ground potential) to the outermost ring, which is maintained at about two times the bias of the back contact. These applied biases deplete the detector in such a way that there is an electrostatic potential minimum for electrons that varies in depth across the detector from right under the front surface at the anode to near the back contact at the last ring. Ionized electrons will drift first to this minimum, then drift radially to the anode as shown in Fig. 116.19. A feature of this contacting arrangement is that the anode capacitance,

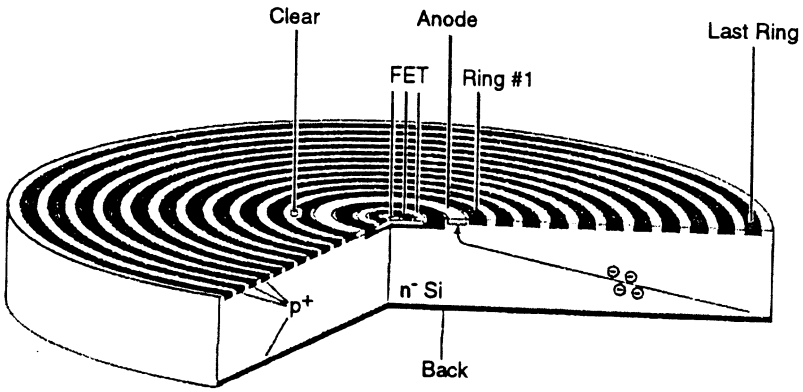


FIGURE 116.19 Cross section of a cylindrical silicon drift detector with integrated *n*-channel JFET. The gate of the transistor is connected to the collecting mode. The radiation entrance window for the ionizing radiation is the non-structured backside of the device. (P. Lechner, S. Eckbauer, R. Hartmann, S. Krisch, D. Hauff, R. Richter, H. Soltau, L. Struder, C. Fiorini, E. Gatti, A. Longoni, M. Sampietro, Nuclear Instruments and Methods in Physics Research, A377, 346–351, 1996.)

and hence amplifier series noise, is low and nearly independent of the active area of the detector. Radial position is deduced from the signal risetime.

CCD Detectors

The design of CCD (charge-coupled device) detectors has similarities to the silicon drift detector [32]. The CCD detector is normally fabricated on an *n*-type silicon wafer depleted both from the backside with a continuous p^+ contact on the back, and from p^+ CCD registers on the front. Reverse bias voltages are such that the wafer is totally depleted and the electron potential minimum is about 10 μ below the CCD registers. After an ionizing event, holes are collected to the p^+ contacts, and electrons are trapped under a nearby register, then transported down a channel of registers by properly clocked voltage pulses to the registers. Each channel has its own readout anode, which can be made small to minimize capacitance — a prerequisite for minimizing noise. The first stage of amplification is frequently integrated onto the same wafer. Spatial resolution is limited to the register (pixel) size. Brauniger et al. [33] described initial results on a 6×6 cm CCD array of 150×150 μ pixels intended for satellite X-ray imaging. The system also had an **energy resolution** of 200 eV FWHM for 5.9 KeV X-rays at room temperature.

Silicon pixel detectors have also been designed using other highly integrated device structures to optimize particular performance aspects such as timing resolution. Pixel detectors using MOS transistors [34] and using reverse biased diodes with individual readout circuitry [35] have been described.

PIN Silicon X-ray Spectrometers

Detectors based on PIN silicon structures have found considerable success as X-ray spectrometers in the region below 10 KeV. The devices consist of *n* and *p* layers on opposite sides of a high purity (10–20 K ohm-cm) silicon wafer. Particularly good performance is obtained if the devices are operated in the region of –30°C as can be provided by a thermoelectric cooler. On the order of 1 W is required to maintain a typical device at –30°C. Energy resolution (FWHM) of 186 eV at 5.9 KeV can be obtained in 7-mm^2 devices (with 20 μs shaping time). Spectrometers with areas up to about 25 mm^2 are also available, but with reduced resolution. The energy spectrum of ^{55}Fe obtained with a 7-mm^2 detector is shown in Fig. 116.20(a), together with a schematic showing construction details (Fig. 116.20(b)).

Status of Silicon Detector Technology

The simply structured silicon detectors fabricated with parallel contacts on a silicon wafer continue to serve a well established need for charged particle spectroscopy. Where economies of scale can be applied, ion implanted detectors have replaced surface barrier detectors. In projects of sufficient size to support their development, specialized low-noise silicon drift detectors and CCD-based detectors have been designed and fabricated with promising room-temperature energy resolution: 200 eV FWHM at 5.9 KeV. These highly structured detector

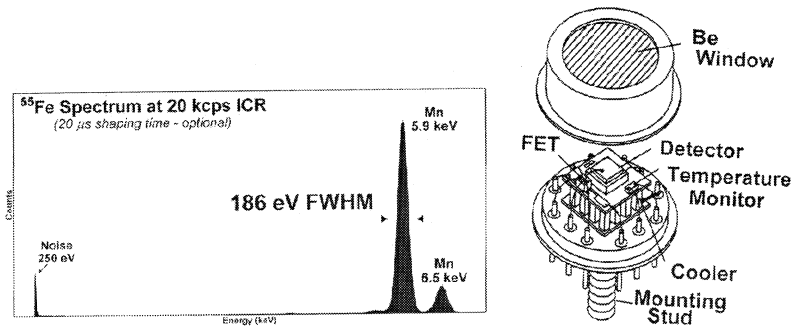


FIGURE 116.20 (a) Energy spectrum of ^{55}Fe source obtained with a 7-mm^2 PIN silicon detector using a shaping time of 20 μs. (b) Schematic diagram showing construction details of typical PIN-silicon detector with thermoelectric cooler. (Courtesy AMTEK Inc. Bedford, MA.)

technologies may find future application in liquid nitrogen-cooled or room-temperature systems for microanalysis using X-ray spectroscopy.

In high-energy physics, the use of various strip, drift, and pixellated detectors for tracking and vertex determination has flourished. These efforts will intensify as experimental requirements for spatial resolution increase. However, radiation damage to the detector is already an issue in this application, and higher luminosity beams will only increase the problem. Nevertheless, it appears that the continuing need of the high-energy physics community for a higher number and density of signal paths forecasts continued reliance on the ever-improving integration technology of the semiconductor silicon industry.

Prices and Availability

The detectors described in this section are available commercially. Their prices vary widely, depending on type, size, and performance. Room-temperature semiconductor detectors range from a few hundred dollars for small, low-resolution devices to over $1000 for large, high-resolution devices. Because of the dynamic nature of wide bandgap detector technology, buyer guides should be consulted for the latest price and availability information.

Pricing of coaxial HPGe detectors is based largely on their gamma-ray efficiency, which is specified relative to a 3 in. × 3 in. sodium iodide scintillator at 1.33 MeV. Coaxial detectors are available with relative efficiencies up to about 150% with cost in the area of several hundred dollars per percent efficiency. Planar HPGe detectors are normally less expensive than coaxial designs. In either case, the price includes the cryostat, dewar, and preamplifier. Cryogenic silicon detectors are available in areas up to several tens of square millimeters. Cost ranges to over $10,000 depending on size, performance, and complexity of design.

Defining Terms

Ballistic deficit: The loss of signal amplitude that occurs when the charge collection time in a detector is a significant fraction of the amplifier's time constant.

Current mode: Measurement mode in which the detector signal provides information on the flux X-rays, gamma-rays, or charged particles.

Dead layer: A layer (frequently associated with a contact region) in which no significant part of the energy lost by photons or particles can contribute to the resulting signal.

Electrical junction: The metallurgical transition boundary between the semiconductor regions of different electrical properties (for example, PIN or between a metal and a semiconductor).

Energy bandgap: The energy difference between the bottom of the conduction band and the top of the valence band.

Energy resolution: The full width of the half maximum of a peak in the energy spectrum, after subtraction of the background under the peak; expressed in units of energy, usually KeV or as a percentage of the energy of the peak.

Energy spectrum: A differential distribution of the intensity of the radiation as a function of the energy.

Leakage current: In the absence of external ionizing radiation and at the operating bias, the total current flowing through or across the surface of the detector element.

Line of peak (in a spectrum): A sharply peaked portion of the spectrum that represents a specific feature of the incident radiation, usually the full energy of a monoenergetic X-ray, gamma-ray, or charged particle.

Pulse mode: Measurement mode in which the detector signal provides information on a single X-ray, gamma-ray, or charged particle.

Semiconductor: Material in which the conductivity is due to charge carriers of both signs (electrons and holes) and is normally in the range between metals and insulators, and in which the charge carrier density can be changed by external means.

Semiconductor radiation detector: A semiconductor device in which the production and motion of excess free carriers is used for the detection and measurement of incident particles or photons.

References

1. G. F. Knoll, *Radiation Detection and Measurement,* 2nd edition, New York: John Wiley & Sons, 1989.
2. R. B. Rossi and H. H. Staub, *Ionization Chambers and Counters,* New York: McGraw-Hill, 1949.
3. M. J. Weber, P. Lecoq, R. C. Ruchti, C. Woody, W. M. Yen, and R. Y. Zhu (Eds.), *Scintillator and Phosphor Materials, Vol. 348,* Materials Resarch Society, Pittsburgh, PA, 1994.
4. T. E. Schlesinger and R. B. James (Eds.), Semiconductors for room-temperature nuclear detector applications, Vol. 43, *Semiconductors and Semimetals,* San Diego: Academic, 1995.
5. R. B. James, T. E. Schlesinger, P. Siffert, and L. Franks (Eds.), *Semiconductors for Room-Temperature Radiation Detector Applications, Vol. 302,* Materials Research Society, Pittsburgh, PA, 1993.
6. J. Fraden, *AIP Handbook of Modern Sensors,* New York: American Institute of Physics, 1993.
7. M.Cuzin, R.B.James, P.F. Manfredi, and P. Siffert (Eds.), *Proceedings of the 9th International Workshop on Room Temperature Semiconductor X- and Gamma-Ray Detectors,* Grenoble, France, Sept. 18–22 1995, *Nucl. Instru. and Meth.,* 380, 1996.
8. R. N. Hall, *IEEE Trans. Nucl. Sci.,* NS-21, 260, 1974.
9. E. E. Haller, W. L. Hansen, and F. S. Goulding, *Adv. in Physics,* 93, 30, No. 1, 1981.
10. W. G. Pfann, *Zone Melting,* New York: John Wiley & Sons, 1966.
11. G. K. Teal and J. B. Little, *Phys. Rev.,* 78, 647, 1950.
12. R. H. Pehl, E. E. Haller, and R. C. Cordi, *IEEE Trans. Nucl. Sci.,* NS-20, 494, 1973.
13. L. S. Darken and C. E. Cox, *Semiconductors for Room-Temperature Detectors,* T. E. Schlesinger and R. B. James (Eds.), Academic Press, 43, 1995.
14. R. H. Pehl, N. W. Madden, J. H. Elliott, T. W. Raudorf, R. C. Trammell, and L. S. Darken, Jr., *IEEE Trans. Nucl. Sci.,* NS-26, 321, 1979.
15. IEEE Test Procedures for Germanium Detectors for Ionizing Radiation, ANSI/IEEE Standard 325-1989.
16. R. H. Pehl and F. S. Goulding, *Nucl. Instru. Meth.,* 81, 329, 1970.
17. J. Llacer, E. E. Haller, and R. C. Cordi, *IEEE Trans. Nucl. Sci.,* NS-24 53.
18. C. E. Cox, B. G. Lowe, and R. Sareen, *IEEE Trans. Nucl. Sci.,* 35, 28, 1988.
19. E. M. Pell, *J. Appl. Phys.,* 31, 291, 1960.
20. H. K. Hecht, *Z. Phys.,* 77, 235, 1932.
21. L. van den Berg, Constellation Technology Corporation, St. Petersburg, FL, private communication.
22. B. E. Patt, J. S. Iwanczyk, G. Vikelis, and Y. J. Wang, *Nuclear Instruments and Methods in Physics Research A,* 380, 276–281, 1996.
23. P. N. Luke, *Nuclear Instruments and Methods in Physics Research A,* 380, 232-237, 1996.
24. J. Kemmer, *Nucl. Instru. Meth.,* 499, 169, 1980.
25. J. Kemmer and G. Lutz, *Nucl. Instru. Meth.,* 365 A235, 1987.
26. G. D. Alkhazov, A. P. Komar, and A. Vorob'ev, *Nucl. Instru. Meth.,* 48, 1, 1967.

27. J. Kemmer, P. Burger, R. Henck, and E. Heijne, *IEEE Trans. Nucl. Sci.,* NS-29, 733, 1982.
28. T. Dubbs, S. Kashigin, M. Kratzer, W. Kroeger, T. Pulliam, H. F.-W. Sadrozinski, E. Spencer, R. Wichmann, M. Wilder, W. Bialas, W. Daabrowski, Y. Unno, and T. Oshugi, *IEEE Trans. Nucl. Sci.,* 42, 1119, 1996.
29. Y. Saitoh, T. Akamine, M. Inoue, J. Yamanaka, K. Kadoi, R. Takano, Y. Kojima, S. Miyahara, M. Kamiaya, H. Ikeda, T. Matsuda, T. Tsuboyama, H. Ozaki, M. Tanaka, H. Iwasaki, J. Haba, Y. Higashi, Y. Yamada, S. Okuno, S. Avrillon, T. Nemota, I. Fulunishi, and Y. Asano, 1123, *IEEE Trans. Nucl. Sci.,* 1996.
30. P. Chochula, V. Cindro, R. Jeraj, S. Macek, D. Zontar, M. Krammer, H. Pernegger, M. Pernicka, and C. Mariotti, *Nucl. Instru. Meth.,* 409, 1996.
31. I. Westgaard, B. S. Avset, N. N. Ahmed, and L. Eversen, *Nucl. Instru. Meth.,* A377, 429, 1996.
32. E. Gatti and P. Rehak, *Nucl. Instru. Meth.,* 225, 608, 1984.
33. H. Brauninger, R. Danner, D. Hauff, P. Lechner, G. Lutz, N. Meidinger, E. Pinotti, C. Reppin, L. Struder, and J. Trumper, *Nucl. Instru. Meth.,* 129, 1993.
34. K. Misiakos and S. Kavadias, *IEEE Trans. Nucl. Sci.,* 43, 1102, 1996.
35. E. Beauville, C. Cork, T. Earnest, W. Mar, J. Millaud, D. Nygen, H. Padmore, B. Truko, G. Zizka, P. Datte, and N. Xuong, A 2D smart pixel detector for time resolved protein crystallogrpahy, *IEEE Trans. Nucl. Sci.,* 43(3),1243, 1996.

Kun, L., Baretich, M.F. "Biocomputing"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

117

Biocomputing

Luis Kun
Cedars-Sinai Medical Center

Matthew F. Baretich
University of Colorado

117.1 Clinical Information Systems
Computer-Based Record • Clinical Information Standards • Bedside Terminals/Point-of-Care Systems • Imaging and the CIS • Systems Integration • Smart/Optical Cards

117.2 Hospital Information Systems
The Clinical Environment • Healthcare Codes and Standards

117.1 Clinical Information Systems

Luis Kun

The main objective of this section is to provide the reader with a summary of areas that relate to clinical information systems. Since this field is so wide, the following topics will be covered mainly because of their importance within the field of medical informatics and the impact that these areas will have in healthcare delivery in the near future. At the end of this section there is a list of definitions that should help the reader not used to related acronyms and a list of suggested bibliographic references which should allow those interested to further increase their knowledge.

Computer-Based Record

Besides improvements in patient care, enhancing the productivity of physicians, nurses, and all healthcare-related personnel is very high on the agenda of all hospitals. Hospitals, clinics, HMOs, doctors' offices, emergency care centers, group practices, laboratories, radiology clinics, and nursing homes among others have a need to share patients' records. Aside from the direction that all of these medical-related centers will have with a required connection to the insurance companies/agencies to speed up payments and their accuracy, the growing need is to have the ability to transfer patients' medical files electronically anywhere in the world. As medical centers become more competitive, they will become worldwide centers of excellence for their given specialties. In turn then, their services will be marketed to the entire world population, becoming true global resources.

The trend of converting hospitals into "paperless hospitals" is becoming one of the most important topics of the 1990s. In 1970, chartered by the National Academy of Sciences, the Institute of Medicine working under the Policy Matters for Public Health has actively pursued the creation of a computer-based record (CBR). In July of 1991 a book was published by the Institute of Medicine in regards to the CBR. The requirements to compile an all-digital medical record (ADMR)will require ways to combine data, graphics, voice, signals, and images, both clinical and document. The architecture that will accommodate all these forms of information for capturing, storing, communicating, and displaying is extremely complex. Some of the technologies involved include optical fibers, LANs, compact/optical disks, bedside terminals, medical image display stations, image diagnostic workstations, and picture archival and communications systems to name a few.

The High Performance Computing and Communications Initiative (HPCCI) was signed into law in December of 1991. Although most of the emphasis for this initiative was from a research and academic sense, some

of the true practical values of these highways of information will occur at the clinical level. While advances are taking place in different parts of the world in fighting diseases such as cancer, AIDS, heart disease, cystic fibrosis, Alzheimer's, Parkinson's, Gaucher's, and malignant hyperthermia, not sharing the knowledge learned by all the groups would be a terrible underutilization of extremely costly resources, causing duplication of effort and enormous waste of time and resources.

The four technologies that have been considered critical by the National Institutes of Health for the coming years are molecular medicine, vaccine development, structural biology, and biotechnology. The four will greatly be affected by the HPCCI. Finally, the integration of all medical-related information will be the most complex task that the healthcare arena will face this decade.

Clinical Information Standards

One of the most demanding and key areas for successfully integrating the hospital information system (HIS) with the clinical information systems (CIS) from multiple clinical departments and/or clinical areas deals with clinical information standards. Two of the driving forces behind the automation of the patient record deal with national concerns related to healthcare costs and quality of healthcare. These concerns have generated demand for managed care. The automated patient record could then be one of the vehicles to achieve managed care.

Clinical information standards are constantly evolving. They were developed (some are still in the process of development; e.g., IEEE/MIB P1073) by very different sets of requirements. What follows is a brief description and structure of most of these standards.

Communications/Storage (e.g., HL/7, IEEE/MEDIX P1157, ANSI ASC X12, ACR/NEMA, IEEE/MIB P1073)

The HL/7 standards group aimed to define vendor-independent communications standards among components of hospital information systems. The IEEE, ANSI, ACR/NEMA, and ASTM have been very active in creating standards through subcommittees from organizations within. As an example, ASTM has the following Healthcare Automation Committees (E31.XX):

- E31.10: Computer automation in the Hospital Pharmacy
- E31.11: Data exchange standards for Clinical Laboratory results
- E31.12: Medical Informatics
- E31.13: Clinical Laboratory Systems
- E31.14: Clinical Laboratory Instrument Interface
- E31.15: Health Knowledge Representation

The MEDIX mission was to establish a robust and flexible communications standard for the exchange of data between heterogeneous healthcare information systems. The MIB was created mainly to allow the exchange of data from medical instrumentation, e.g., monitoring devices and hospital information systems. Many of the manufacturers of these devices have proprietary hardware, e.g., buses and/or software, which complicates this exchange. Bedside terminals in the intensive care environment will benefit immensely from such a standard, since most hospitals' ICUs and CCUs have many vendors' equipment in their units. To effectively integrate and manage the data are major goals of the MIB.

Classification/Reimbursement (e.g., ICD, DRG, SNOMED, CPT, DSM, RCS, UMLS)

ICDs were originally used for public health morbidity statistics; now in the United States they are primarily used for reimbursement. Its structure is numbered classification of diseases grouped by anatomical areas. The DRGs facilitate the definition of case-mix for hospital reimbursement. Its structure is multi-axial: severity of illness, prognosis, treatment difficulty, need for intervention, and resource intensity. SNOMED provides description of pathological tests related to patient identification. It has four axes: function (primary symptoms), etiology (cause of disease), morphology (description of disease form), and topology (area of body). CPT is primarily used for reimbursement and utilization review. It derives codes from specialty nomenclatures divided into chapters: systemic (medicine, anesthesia, etc.), topological (cardiovascular, lymphatic, etc.), and technological

(radiology, laboratory, etc.). DSM provides consistent abbreviations for prescription and administrative use. It facilitates psychiatric education and research. Its structure is multi-axial: clinical syndromes, developmental and personality disorders, physical disorder, severity psychological stresses, and global assessment functioning. RCS is a comprehensive nomenclature and classification of medical terms for computerized records. UMLS facilitates the unification of clinical data classification systems into a single unified medical language system. It will also facilitate the creation of data into compatible automated patient record systems. Its structure reconciles clinical terminology, semantics, and formats of the major clinical coding and reference systems.

Knowledge (e.g., ARDEN SYNTAX)

The ARDEN SYNTAX is a standard for sharing medical knowledge bases in the form of medical logic modules (MLM). Its structure is derived from the HELP (LDS Hospital) and the CARE (Regenstrief MC) systems. The MLMs accommodate alerts, management critiques, therapy suggestions, diagnosis scoring, etc. Each MLM is limited to the knowledge to make a single decision.

HCFA (e.g., UCDS, WARP, UHDDS)

UCDS provides an electronic clinical data set that Medicare can use to perform clinical quality reviews. The quality evaluation is done by using algorithms related to surgical procedures, disease specific, organ specific, discharge status and disposition, etc. The UCDS permits the hospital to enter the data into a personal computer; then this information can be sent electronically to the HCFA. WARP provides an epidemiologic approach to quality assurance. It hopes to overcome about 50% of ICD miscoding and its initial focus is on ambulatory chart review rather than real-time patient care. It is not a diagnostic or procedural classification system. It basically provides a model for encoding clinical information. It is an object-oriented case tool. UHDDS was created for studies on quality of care and fraud. It is also used for auditing Medicare and Medicaid subsystems.

Bedside Terminals/Point-of-Care Systems

Patient information is generated on an ongoing basis, wherever the patient may be. Almost two decades ago with the creation of the first programmable calculators, a trend started in terms of calculating hemodynamic variables in the OR, etc. This approach was improved with the creation of personal computers, ending with the development of what are now called bedside terminals. Companies such as Clinicom, Emtek, Hewlett-Packard, Hospitronics, and Spacelabs offer systems that can go from doing simply patient monitoring, to a complete data acquisition, data management, and data analysis system that incorporates in some cases diagnosis and treatment therapy.

From the patients' point of view, it is critical to integrate their demographic information with their clinical data. Usually the HIS contains all the ADT, orders, laboratory, pharmacy, etc. while the CIS may be more of a departmental system such as ICU/CCU, which contains hemodynamic variables, i.e., blood pressure, stroke volume, heart rate, etc. Both systems need to coexist. Point-of-care systems, many times known as bedside terminals, include both general med/surgery and the ICU/CCU type. The general type include functions such as patient assessment, nursing diagnosis, patient care plans, kardex, discharge planning, discharge summary, medication administration record, I/O, vital signs, activities of daily living, patient classification/acuity, etc. The ICU/CCU systems in addition contain information regarding drug administration, fluid analysis, hemodynamic analysis (i.e., blood gas report, ECG, blood pressures, pulse oximeters, cardiac output), respiratory analysis (i.e., ventilator data, O_2/CO_2 analyzer), and real-time monitoring. Today's trends are incorporating imaging devices in both at the regular nursing stations, at the operating rooms, and at the recovery room/ICU/CCU. The motivation is to incorporate all patients' information and have it available wherever they may be. As a patient moves from a regular bed to the OR, back to an ICU, and later to a regular nursing station, the electronic record follows the patient. The one big difference with paper charts is that the electronic record can be shared simultaneously within and outside the institution.

Having the ability to look at electronic images in all of these locations not only opens the doors for consultation within the institution but also with outside institutions and/or expert individuals.

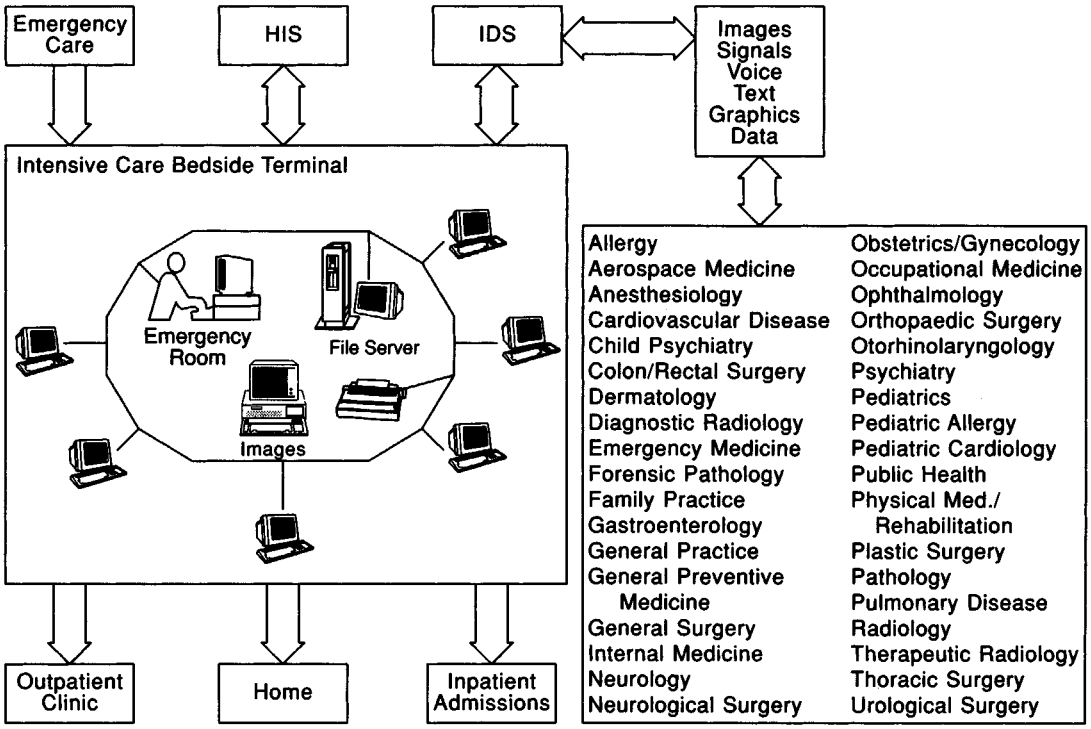


FIGURE 117.1 An example of medical systems integration.

Imaging and the CIS

Imaging plays two very important roles within the context of a computer-based record (CBR). Document imaging allows for all those records that exist today in storage for the medical records departments to be scanned and incorporated electronically with the rest of the patient's current records existing in the HIS and CIS. The second role is from the perspective of clinical images. Most imaging experts will call this PACS, which stands for picture archival and communications system and is mostly associated with the Radiology Department of the hospital. We can view clinical images as a form of data which can be generated in any department.

Some of these typical clinical departments utilizing clinical images are radiology, cardiology (e.g., echocardiography, fluoroscopic techniques, cine cameras, 3D modeling, gamma cameras), orthopedic surgery, plastic surgery, obstetrics/gynecology, laboratories (e.g., genetics, chromosome analysis, cytology, hematology, clinical chemistry, pathology, histology, electron microscope), maxillofacial clinics, sports medicine, and oncology (e.g., radiation therapy, chemotherapy), emergency rooms, intensive care units, etc.

There are five imaging modalities: x-ray, magnetic resonance imaging (MRI), computer tomography (CT), nuclear medicine (NM), and ultrasound (US). These modalities create images which are very different not only in medical terms but in their size and content. As a result, there are three main areas under PACS which are critical in succeeding with such systems: communications (i.e., network, transmission protocol, and image format), archiving (i.e., database and storage media), and image processing (i.e., display, user interface, and IP algorithms).

Systems Integration

As an example of systems integration in the emergency care environment (see Fig. 117.1), from an information-flow point of view we see the following:

1. Information coming and going to the HIS, e.g., laboratory, pharmacy, orders, etc.
2. Information going to outpatient clinics for referring services, admissions to the hospital, or even to the patient's physician at home.
3. In the emergency room, the utilization of an intensive care type of bedside terminal allowing data collection, analysis and management, and also the ability to view clinical images in the ER.
4. From a consulting point of view, the whole electronic patient record, under an integrated diagnostic system, allows for any (department) consulting physician within or outside the hospital to review the case.

Smart/Optical Cards

Smart/optical cards provide a wide range of applications in the medical field. The patient, the provider (e.g., physician, dentist, etc.), the hospital, and the insurer can all benefit from such a card. The card will eventually contain all data forms—voice, text, graphics, clinical images, document images, signals, and data values collected from medical instrumentation. Besides patient identification/demographics, medical history, medications, allergies, and insurance verification, the system could contain the patient's picture, fingerprint, digital signature, voice signature, and even genetic/blood information for security reasons.

The patient is admitted and treatment is provided more quickly, historical information is more accurate, and personal physicians and specialists can be consulted more quickly. Less testing may be a direct result, and faster diagnosis is accomplished. Since information needs to be entered only once, patients do not need to rely on their memory, particularly in emergency situations.

The hospital identifies the patient and accesses all the medical records information from multiple departments more quickly. It needs fewer staff to find records from the hospital/clinics (even from other institutions), and this could reduce the length of stay.

The provider is better informed for a quicker diagnosis by getting all the available history at admission and can consult with the patient's personal physician and specialist by having their respective phone numbers. All prior records from the same or a different set of institutions coexist in the card. It also can reduce exposure to malpractice.

The insurer reduces fraudulent claims, reduces costs for data entry, and has more complete and accurate claims data. Also, by eliminating redundant tests costs are reduced.

Most of the cards can be classified into five groups by the type of technologies used: microfilm, magnetic strip, softstrip, chip, and laser/optical. Microfilm is hard to change and can be damaged by both temperature and humidity. Magnetic strip contains little information, approximately 2K, and can be destroyed by electric and magnetic fields. The softstrip, because it is laser printed and optically read, is difficult to change information on. The chip card has only up to 10K of storage and is very expensive. Finally, the laser/optical card allows for approximately 1000 typed pages or approximately 4 Mb of memory and requires a read/write device.

Some of the complexities that are incorporated by using these types of technologies are associated with the access to the information. For someone to be able to either "read" and/or "write" in the card, it must possess technologies compatible with the ones where the information was created. It is a fundamental principle then that a set of international standards will be created so that any hospital that requires access to the card information can do so. Already the International Patient Cards Standards Council, the Health Industry Business Communications Council (HIBCC), and the Smart Card Applications and Technology (SCAT) have been created. These groups, among others, are working towards the goal of an international set of standards.

There is a large set of companies that are already marketing different types of card technologies. Some examples of projects and/or vendors include:

- Affiliated Healthcare in Princeton, New Jersey, which maintains a Health Summary Database with a Smart Card.
- CentraHealth, a Florida hospital network with about 12K users.
- Clinicard, a subscription service which provides a softstrip, 3K, PC DOS card that folds to a business card size.
- Drexler LaserCard from Mountain View, California, which has 4.1-Mb card being tested by both British Telecom with a hospital group specializing in obstetric patients and Baylor College of Medicine in Houston, Texas.

- Eltrax in St. Paul, Minnesota, which is associated with several HIS manufacturers (Spectrum, McDonald Douglas, SMS, and Meditech) and provides a magnetic strip card with about 900-character capacity.
- IMSG/INFODYNE from Englewood, Colorado, which has a medical information card on magnetic strip carrying up to 600 characters.
- IntelliScan from American Medical Data Corp. in Atlanta, Georgia, which has information stored as 350 characters of readable text printed on the top of the card and up to 850 characters of detailed medical information optically encoded at the bottom. It is being used by hospitals in both Texas and Mississippi. The cards are customized to each hospital's database.
- Lifecard, early pilot (1985) card for electronic claims provided by Blue Cross/Shield of Maryland.
- Medfirst credit card, combining both medical information and financial credit, in test by Humana and Discovery Card.
- Medi-Card, a chip card from MediData Systems in Allston, Massachusetts.
- MedKey, from Biloxi Regional Medical Center, Biloxi, Mississippi.
- Medical Information Systems, in St. Louis, Missouri, which has a microfilm card that can contain up to 18 pages of information, including signals, text, images, data, and color photos.
- Ulticard, a 64K RAM memory chip in a credit card sized pack being tested in Houston at Baylor and Methodist hospitals.

Some of the cards are being tested in different countries. Sweden, the leader for about 20 years, has been using a patient card which is issued at birth by the government together with an ID number. Sweden has a socialized medicine program and it has been in their best interest to develop uniform standards so that the information can be accessed by every institution in the country. Belgium, Canada, France, Great Britain, Spain, and Switzerland all have several systems on trial.

Acronyms

ACR/NEMA: American College of Radiology/National Equipment Manufacturers Association
ANSI ASC X12: American National Standards Institute Accredited Standards Committee
ARDEN SYNTAX: Syntax for Medical Logic Modules
ASTM: American Society for Testing and Materials
CIS: Clinical Information System
CPT: Current Procedural Terminology
DRG: Diagnostic Related Group
DSM: Diagnostic and Statistical Manual of Mental Disorders
EDI: Electronic Data Interchange
HCFA: Healthcare Financing Administration
HIS: Hospital Information System
HL/7: Health Level/7
ICD: International Classification of Diseases
IDS: Integrated Diagnostic System
IEEE: Institute of Electrical and Electronics Engineers
IEEE/MEDIX P1157: Medical Data Interchange
IEEE/MIB P1073: Medical Information Bus
OSI: Open Systems Interconnection
RCS: Read Classification System
SNOMED: Systemized Nomenclature of Medicine
UCDS: Uniform Clinical Data Set
UHDDS: Uniform Hospital Discharge Data Set
UMLS: Unified Medical Language System
WARP: Wisconsin Ambulatory Review Project

Related Topics

97.1 Introduction • 117.2 Hospital Information Systems

References

Bedside Terminals/Point-of-Care

W. Donovan and S. Corrales, *The Book on Bedside Computing,* Long Beach, Calif.: Inside Healthcare Computing, 1991.

L. Kun, "The use of a personal computer for patient-condition-treatment in a CUU/ICU environment," *IEEE Transactions on Biomedical Engineering,* vol. BME-30, no. 8, August 1983.

L. Kun, "Rapid assessment of hemodynamic cardiorespiratory function for the critically ill with a personal computer," *IEEE Transactions on Biomedical Engineering,* vol. BME-30, no. 8, August 1983.

D. O'Boyle, G. Feiherr, and R. Gough, *The Buyer's Guide to Bedside Computer Systems,* Rockville, Md.: National Report of Computers & Health, 1991.

M.M. Shabot et al., "Rapid bedside computation of cardiorespiratory variables with a programmable calculator," *Critical Care Med.,* vol. 5, p. 105, 1977.

Classification Systems Standards

C. Chute, "Tutorial 19: Clinical data representation," in *Proceedings of SCAMC 91,* November 1991.

B. Humphreys, *Building the Unified Medical Language System,* Bethseda, Md.: National Library of Medicine, 1989.

Communications Standards

J. Harrington, IEEE/EMBS P1158, "Medical Data Interchange (MEDIX) overview and status report," in *Proceedings of SCAMC 90,* November 1990.

National Electrical Manufacturers Association, "Digital Imaging and Communications," ACR-NEMA Standards Publication No. 300-1988, 1988.

R.E. Norden-Paul, IEEE Proposed Standard 1073, "Medical Information Bus: An Introduction and Progress Report," in *Proceedings of the 9th Annual Conference of the IEEE-EMBS,* vol. 2, MIB Symposium, Boston, pp. 1209–1211, 1987.

Knowledge Base Standards

"The ARDEN SYNTAX for medical logic modules," in *Proceedings of SCAMC 90,* November 1990.

"Emerging standards for medical logic," in *Proceedings of SCAMC 90,* November 1990.

Clinical Imaging/PACS

Y. Kim and F.A. Spelman, Eds., "Images of the twenty-first century," in *Proceedings of the Annual International IEEE-EMBS,* vol. 11, part 2, track 2, Imaging, pp. 345–630; Track 23, Picture Archiving and Communications Systems, pp. 775–793, 1989.

L. Kun, "Imaging and the clinical information system," in *Proceedings of '91 International Workshop on Medical Imaging,* Korea Institute of Science and Technology, Seoul, Korea.

Computerized Medical Record

M. Ball and M. Collin, Eds., *Aspects of the Computer-Based Patient Record,* New York: Springer-Verlag, 1992.

J. Blair, "Overview of clinical information representation and standard organization," in *Proceedings of the Fall 92 ECHO Meeting,* Palm Beach, Calif., 1992.

Institute of Medicine, *Computer-Based Patient Record,* Washington, D.C.: National Academy Press, 1991.
C.J. McDonald et al., "The benefits of automated medical record systems for ambulatory care," in *Proceedings of the Computer Applications in Medical Care Conference,* New York: IEEE Computer Society, pp. 157–171, October 1986.
W.W. Stead et al., "Practicing nephrology with a computerized medical record," *Kidney Int.,* vol. 24, pp. 446–454, 1983.
Q.E. Whiting-O'Keefe et al., "A computerized summary medical record system can produce more information than the standard medical record," in *Proceedings of MedInfo '86,* Washington, D.C., 1986.

High-Performance Computing and Communications, (HPCC)

D.A. Bromley, "The Federal High-Performance Computing Program," Washington, D.C.: Executive Office of the President, Office of Science and Technology Policy, 1989.
"National High-Performance Computer Technology Act," Congressional Record, U.S. Senate 101st Congress, First Session 5/18/89, Washington, D.C.

Smart/Optical Cards

Handbook of Optical Memory Systems. Bi-monthly updating service. Boston: Medical Records Institute.
Proceedings of the 13th Annual International Conference IEEE/EMBS, Track 21: Session 5, Medical Informatics V: Optical and Smart Cards, Orlando, Fla., pp. 1387–1392, October 1991.
1989 Smart Card Industry Directory, Palo Alto, Calif.: Palo Alto Management Inc., 1989.

117.2 Hospital Information Systems

Matthew F. Baretich

What does an electrical engineer need to know to be part of a team designing and implementing a hospital information system? For the most part, the necessary skills are those required to design and implement any comprehensive information system in a complex organization. Hospitals do, however, have unique characteristics that must be taken into account. These characteristics are described in the following pages.

The Clinical Environment

Hospitals are, indeed, complex organizations. They perform a vital function (patient care) but are subject to strict regulation and operate under severe financial constraints. Quality of patient care is the highest value, but a competitive marketplace demands efficient operation. Hospital information systems range from nonexistent to antique to state-of-the-art.

Hospitals are highly professionalized. Each professional group has a particular area of expertise and a unique perspective regarding the healthcare delivery system. Hospital administrators are much like administrators of other organizations. Recent graduates essentially have standard MBA (Master of Business Administration) degrees with some extent of healthcare specialization. However, many administrators in positions of authority received MHA (Master of Hospital Administration) degrees from programs more closely affiliated with medical schools than with business schools.

Hospitals also have large clinical staffs which include nurses and technologists (who are hospital employees) and medical doctors (who are usually not hospital employees). Clinicians are educated in the biological and medical sciences, and their preparation generally includes a large component of practical experience in the hospital as well as theoretical study in the classroom. As hospital employees, nurses and technologists (respiratory, laboratory, etc.) are part of the administrative structure of the hospital. Medical doctors (physicians and surgeons), on the other hand, are part of a separate medical staff structure that is largely independent of the hospital's administrative structure. However, medical doctors control the admission and discharge of the hospital's patients, and many hospital activities are the result of medical orders for patient services.

The number of hospital employees with an engineering background is limited. For the electrical engineer who is involved in the implementation of a hospital information system, hospital-based technical support may include an information systems department and a clinical engineering (or biomedical engineering) department.

The following aspects of the healthcare delivery system are worthy of study by an electrical engineer working in the clinical environment:

- The healthcare delivery system in the United States [Williams and Torrens, 1984]
- The organizational structure of hospitals [Goldberg and Buttaro, 1990]
- The characteristics of hospital information systems [Austin, 1988; Minard, 1991]

With this background information the electrical engineer will be better prepared to translate the concerns of hospital administrators and clinicians into the technical specifications of the hospital information system.

Healthcare Codes and Standards

The healthcare delivery system is a highly regulated industry. Numerous governmental and nongovernmental organizations have established codes and standards intended to promote safe and effective patient care. Although there can be significant differences in the regulatory environment from one hospital to another, the major codes and standards are relatively uniform.

The **National Electrical Code** (NFPA 70), promulgated by the National Fire Protection Association (NFPA: Quincy, Massachusetts) applies to hospitals. Specifically, Article 517 deals with "Health Care Facilities." A more focused document, however, is the *Standard for Health Care Facilities* (NFPA 99). The most accessible format for this information is the NFPA's *Health Care Facilities Handbook* [Klein, 1990] which includes the full text of NFPA 99 as well as interpretive and explanatory material.

Many of the healthcare-related provisions of the electrical code are based on two concerns. First, many patients in surgery and intensive care depend on electrical equipment for life support. Such equipment ranges from heart-lung bypass devices to mechanical ventilators. Therefore, much attention is devoted to ensuring the availability of electrical power in the event that the primary power distribution system fails. A hospital information system that provides life-support functions may be subject to these provisions.

Second, because of the use of invasive medical procedures, many patients are considered to be "electrically susceptible." Under certain conditions, electrical currents on the order of microamperes can cause ventricular fibrillation, a potentially fatal disruption of normal cardiac function. Therefore, the NFPA and other organizations have established strict standards for grounding, "leakage" current, and other electrical parameters. These standards apply to devices and cabling in patient-care locations of the hospital.

The **Joint Commission on Accreditation of Healthcare Organizations** (Chicago, Illinois) is another major source of standards affecting hospitals. The JCAHO's *Accreditation Manual for Hospitals* [JCAHO, 1993] covers the entire spectrum of hospital activities. Pursuit of JCAHO accreditation is voluntary but, in practice, essentially all hospitals seek accreditation to ensure eligibility for reimbursement under certain governmental programs. At present, JCAHO standards include little reference to information systems. However, this is expected to change and, therefore, familiarity with the latest edition of the *Accreditation Manual for Hospitals* is advisable.

Another standard unique to the healthcare system is **Health Level 7** (HL7) which is a data communications protocol intended to facilitate the interfacing of various components in a hospital information system [Walker, 1989]. These components range from accounting systems (financial data) to clinical laboratory information systems (laboratory test results) to medical records systems (documentation of patient care services) to patient data management systems (physiological data). In the recent past, each such component was independent and generally incompatible with other components. However, to achieve high quality in patient care at the lowest cost, both administrators and clinicians need integrated, comprehensive access to a wide variety of information.

HL7 is an attempt to specify the types of data (and their formats) to be shared within a hospital information system. For example, if all components of the system use a common format for a patient's name, then it is possible for a single database query to gather all data regarding that patient. This also allows automation of certain activities such as billing (through the accounting system) for laboratory tests ordered by clinicians (through the clinical laboratory system). Unfortunately, HL7 has not achieved its promise but it does represent a significant step away from the chaotic past [Bond et al., 1990].

Summary

The electrical engineer will be only one of many professionals involved in the implementation of a hospital information system. Successful participation in this team will depend on more than the electrical engineering skills that are applicable to any information system project. The critical success factor is an understanding of the hospital—the people (clinicians and administrators), their objectives (low cost and high quality), and the environment within which they work.

Defining Terms

HL7: A data communications protocol for interfacing components of a hospital information system.

JCAHO: The Joint Commission on Accreditation of Healthcare Organizations, an organization that promulgates standards affecting hospital operations.

NEC: The National Electrical Code, an NFPA standard that is commonly adopted by governmental units and, therefore, having the force of law.

NFPA: The National Fire Protection Association, an organization that promulgates standards affecting electrical systems in hospitals.

Related Topics

94.1 Databases • 117.1 Clinical Information Systems

References

C.J. Austin, *Information Systems for Health Services Administration*, 3rd ed., Ann Arbor, Mich.: Health Administration Press, 1988.

V. Bond, J. Lenahan, and W. Wagner, "HL7: A practical perspective," *Healthcare Informatics,* vol. 7, no. 10, p.46, 1990.

A.J. Goldberg and R.A. Buttaro, Eds., *Hospital Departmental Profiles,* 3rd ed., Chicago: American Hospital Publishing, 1990.

JCAHO, *Accreditation Manual for Hospitals,* 1993 ed., Chicago: Joint Commission on Accreditation of Healthcare Organizations, 1993.

B.R. Klein, Ed., *Health Care Facilities Handbook*, 3rd ed., Quincy, Mass.: National Fire Protection Association, 1990.

B. Minard, *Health Care Computer Systems for the 1990s*, Ann Arbor, Mich.: Health Administration Press, 1991.

J.M. Walker, "Integrating information systems with HL7," *Hospitals,* vol. 63, no. 13, p. FB60, 1989.

S.J. Williams and P.R. Torrens, *Introduction to Health Services,* 2nd ed., New York: John Wiley & Sons, 1984.

Further Information

Many of the major professional societies dealing with computer science and engineering have healthcare–related divisions. Further information can be obtained from each professional society.

The Healthcare Information and Management Systems Society is a division of the American Hospital Association that deals with information systems, telecommunications, and management engineering. For further information contact the American Hospital Association, Chicago, Illinois.

Major periodicals that focus on hospital information systems include *National Report on Computers and Health* and *Healthcare Informatics*. These publications, and other healthcare-related literature, can be found in the libraries of academic medical centers.

Luebbers, R. "Computer Design for Biomedical Applications"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

118

Computer Design for Biomedical Applications

Raymond Luebbers
Pennsylvania State University

The Finite Difference Time Domain (FDTD) [Yee, 1966; Kunz and Luebbers, 1993; Taflove, 1995] is a numerical method for the solution of electromagnetic field interaction problems. It utilizes a geometry mesh, usually of rectangular box-shaped cells. The constitutive parameters for each cell edge may be set independently, so that objects having irregular geometries and inhomogeneous dielectric composition can be analyzed.

The FDTD method solves Maxwell's differential equations at each cell edge at discrete time steps. Since no matrix solution is involved, electrically large geometries can be analyzed. FDTD solutions for three dimensional complex biological geometries involving millions of cells have become routine. FDTD may be used for both open region calculations, such as a human body in free space, or closed regions, such as within a TEM cell. Commercial FDTD software is available from several sources (CST, EMA, and Remcom), with some of these also offering FDTD meshes for human heads and bodies. These commercial packages provide a graphical user interface for viewing the FDTD mesh. Some provide interactive mesh editing (Remcom), while others allow for import of objects from CAD programs (CST and EMA).

The choice of cell size is critical in applying FDTD. It must be small enough to permit accurate results at the highest frequency of interest, and yet be large enough to keep resource requirements manageable. Cell size is directly affected by the materials present. The greater the permittivity and/or conductivity, the shorter the wavelength at a given frequency and the smaller the cell size required. Once the cell size is selected, the maximum time step is determined by the Courant stability condition. After the user determines the cell size, a problem space large enough to encompass the scattering object, plus space between the object and the absorbing outer boundary, is determined. From the number of Yee cells needed and the number of time steps required, resource requirements can be estimated.

The fundamental constraint is that the cell size must be much less than the smallest wavelength for which accurate results are desired. An often quoted constraint is "10 cells per wavelength", meaning that the side of each cell should be 1/10 of the wavelength at the highest frequency (shortest wavelength) of interest. Since FDTD is a volumetric computational method, if some portion of the computational space is filled with penetrable material, one must use the wavelength *in the material* to determine the maximum cell size. For problems containing biological materials, this results in cells in the material that are much smaller than if only free space and perfect conductors were being considered.

Another cell size consideration is that the important characteristics of the problem geometry must be accurately modeled. This will normally be met automatically by making the cells smaller than 1/10 λ unless some special geometry features smaller than this are factors in determining the response of interest.

In some situations there is a specific region of the object where smaller FDTD cells are needed, for example, a region of high dielectric material, or of fine geometry features such as eyes. But if uniform FDTD cells are used throughout the computation, then these small cells must be used even in regions where they are not needed. One approach to reduce the total number of FDTD cells for these situations is to mesh local regions with smaller cells than in the main mesh [Kim and Hoefer, 1990; Zivanovic et al., 1991]. All of the commercial FDTD software referenced above has this local grid capability.

The other basic constraint on FDTD calculations is the time step size. For a three-dimensional grid with cell edges of length Δx, Δy, Δz, with v the maximum velocity of propagation in any medium in the problem, usually the speed of light in free space, the time step size Δt is limited by

$$v\,\Delta t \;\le\; 1 \Big/ \sqrt{\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2} + \frac{1}{(\Delta z)^2}}$$

Now let us consider how to estimate the computer resources required. Given the shortest wavelength of interest, the cell dimensions are determined as 1/10 of this wavelength (or less if greater accuracy is required). From this and the physical size of the problem geometry the total number of cells in the problem space (here denoted as NC) can be determined. We assume that the material information for each cell edge is stored in 1 byte (INTEGER*1) arrays with only dielectric materials considered. Then, to estimate the computer storage in bytes required, and assuming single-precision FORTRAN field variables, we can use the relationship

$$storage \;=\; NC * \left(6\,\frac{components}{cell} \times 4\,\frac{bytes}{component} + 3\,\frac{edges}{cell} \times 1\,\frac{byte}{edge} \right)$$

where components indicate the vector electric and magnetic field components. If magnetic materials are included, then six edges must also be considered for the material arrays. In this equation, we have neglected the relatively small number of auxiliary variables needed for the computation process.

One can estimate the computational cost in terms of the number of floating point operations required using

$$Operations \;=\; NC \times 6\ components/cell \times 15\ operations/component \times N$$

where 15 operations is an approximation based on experience and where N is the total number of time steps. The number of time steps N is typically on the order of five to ten times the number of cells on one side of the problem space. It will be larger for resonant objects and smaller for lossy objects.

As an example, consider a human body meshed with 5-mm cubical cells. At 10 cells per free space wavelength, this would correspond to a maximum frequency of 6 GHZ. But, since the biological materials in the body have relatively high dielectric constants, the wavelength inside the body is reduced. If the maximum dielectric constant of body materials is 49, then the maximum frequency would be reduced by 7 to about 857 MHz. If results at higher frequencies are needed, then the cell size must be reduced.

For a human body that fits into a box of $63 \times 36 \times 183$ cm, with a 15-cell border around the body to separate it from the outer boundary, the problem space is about $160 \times 100 \times 400$ or 6.4 millions cells. Using the above formula, the computer RAM necessary to make this calculation is approximately 172 MBytes. Since this does not allow for storage of instructions and other arrays, and since the operating system will take some computer memory, a machine with about 256 MBytes of random access memory (RAM) should be sufficient to make this calculation.

A conservative estimate of the number of time steps needed is 10 times the longest dimension in cells, or 4000 time steps. Using the above equation, an estimate of 2.3×10^{12} operations results. Typical MFLOPS (Million Floating Point Operations per Second) ratings for computers are 15 for a Pentium PC or low end work station, 60 for a fast work station, and several hundred for a super computer. If we use 200 MFLOPS for the super computer, then the calculation times for the human body are 42 h for the PC or low end work station, 10.5 h for the fast work station, and 3.1 h for the super computer.

The preceding discussion primarily considers the high frequency limitations of FDTD calculations, which are based on the size of the object in wavelengths. The low frequency limitation is usually determined by a combination of the geometry features and time step. For example, consider applying FDTD for a 60-Hz calculation for a human body. Based on the wavelength, the FDTD cells could be huge, but then the body shape would be unrecognizable. Suppose that we pick FDTD cells of 10 cm to at least make a crude body shape. Then the maximum time step would be 19.2×10^{-10} s. If we further assume that we need to make FDTD calculations for at least one period of the sine wave in order to read some semblance of steady state, this would require about 86 million time steps, which is not feasible on current computers. This illustrates the difficulty of using FDTD for extremely low frequencies. For these very low frequencies other methods, such as finite elements, are preferred.

Depending on the application, human body models may be crude approximations or detailed meshes based on actual anatomy. A popular source of anatomical data suitable as the basis for an FDTD biological mesh is the Visible Human Project of the National Library of Medicine. Various types of data are available, with the most useful perhaps being the cross-sections. These are 1-mm slices for the male and 0.33-mm slices for the female. Both have a cross-sectional resolution of 0.33 mm. The FDTD meshing of this data still requires considerable effort, especially in assigning the colors of the slices to particular tissue types.

The actual FDTD calculations may be excited in different ways. Most commonly the electric fields on one or more mesh edges are determined by an analytical function of time, such as a Gaussian pulse or sine wave. This then acts as a driven voltage source. This may be used to excite an antenna. For example, a short monopole antenna on a rectangular box may approximate a portable telephone. This monopole antenna could be driven by a drive voltage source located on the mesh edge at the monopole base next to the top of the box. Both Kunz and Luebbers [1993] and Taflove [1995] describe methods for modeling RF sources. A variety of FDTD sources, including current sources, are described in Piket-May et al. [1994]. Alternatively a plane wave may be incident on the object as the excitation source.

The time variation of the excitation may be either pulsed or sine wave. The advantage of the pulse is that response for a wide frequency range can be obtained. But, for accurate results, the frequency-dependent behavior of biological materials must be included in the calculations. Methods for doing this are well known [Kunz and Luebbers, 1993; Taflove, 1995] so that transient electromagnetic field amplitudes for pulse excitation can be calculated using FDTD [Furse et al., 1994]. When results at a single frequency or at a few specific frequencies are desired, then sine wave excitation is preferred. This is especially true if results for the entire body, such as SAR, are needed, since storing the transient results for the entire body mesh and then applying fast Fourier transformation to calculate the SAR vs. frequency requires extremely large amounts of computer storage.

Related Topic

45.1 Introduction

References

C. M. Furse, J. Y. Chen, and O. P. Gandhi, "The use of the frequency-dependent finite-difference time-domain method for induced currents and SAR calculations for a heterogeneous model of the human body," *IEEE Trans. Electromagn. Comp.,* 36, 128–133, 1994.

L. S. Kim and W. J. R. Hoefer, "A local mesh refinement algorithm for the time-domain finite-difference method using Maxwell's equations," *IEEE Trans. Microwave Theory Techniques,* 38, 812–815, 1990.

K. S. Kunz and R. J. Luebbers, *The Finite Difference Time Domain Method for Electromagnetics,* Boca Raton, Fla.: CRC Press, 1993.

M. Piket-May, A. Taflove, and J. Baron, "FD-TD modeling of digital signal propagation in 3-D circuits with passive and active loads," *IEEE Trans. Microwave Theory Techniques,* 42, 1514–1523, 1994.

A. Taflove, *Computational Electrodynamics—The Finite-Difference Time-Domain Method,* Boston, Mass.: Artech House, 1995.

K. S. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas Propagation,* AP-17, 585–589, 1966.

S. S. Zivanovic, K. S. Yee, and K. K. Mei, "A subgridding method for the time-domain finite-difference method to solve Maxwell's equations," *IEEE Trans. Microwave Theory Techniques,* 39, 471–479, 1991.

Further Information

CST GmbH, Lauteschlägerstr, 38, D-64289 Darmstadt, Germany, +49(0)6151 717057, fax +49(0)6151 718057.

EMA Electromagnetic Applications, P.O. Box 260263, Denver, CO, 80226-2091, voice (303) 980-0070.

Remcom, Inc., Calder Square, Box 10023, State College, PA 16805-0023, voice (814) 353-2986, fax (814) 353-1420, URL http://www.remcominc.com, e-mail xfdtd@remcominc.com.

Visible Human Project, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894; fax (301) 402-4080; URL http://www.nlm.nih.gov/research/visible/visible-human.

Robinson, C.J. "Rehabilitation Engineering, Science, and Technology"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

119

Rehabilitation Engineering, Science, and Technology

119.1 Rehabilitation Concepts
119.2 Engineering Concepts in Sensory Rehabilitation
119.3 Engineering Concepts in Motor Rehabilitation
119.4 Engineering Concepts in Communications Disorders
119.5 Appropriate Technology
119.6 The Future of Engineering in Rehabilitation

Charles J. Robinson
Louisiana Tech University
Overton Brooks VA Medical Center

Rehabilitation engineering requires a multidisciplinary effort. To put rehabilitation engineering into its proper context, we need to review some of the other disciplines with which rehabilitation engineers must be familiar. Robinson [1993] has reviewed or put forth the following working definitions and discussions.

Rehabilitation: The (Re)integration of an individual with a **disability** into society. This can be done either by enhancing existing capabilities or by providing alternative means to perform various functions or to substitute for specific sensations.

Rehabilitation engineering: The *application* of science and technology to ameliorate the handicaps of individuals with disabilities [Reswick, 1982]. In actual practice, many individuals who say that they practice rehabilitation engineering are not engineers by training. While this leads to controversies from practitioners with traditional engineering degrees, it also has the *de facto* benefit of greatly widening the scope of what is encompassed by the term "rehabilitation engineering."

Rehabilitation medicine: A clinical *practice* that focuses on the physical aspects of **functional** recovery, but that also considers medical, neurological and psychological factors. Physical therapy, occupational therapy, and rehabilitation counseling are professions in their own right. On the sensory-motor side, other medical and therapeutical specialties practice rehabilitation in vision, audition, and speech.

Rehabilitation technology (or Assistive technology): Narrowly defined, the *selection*, *design*, or *manufacture* of *augmentative* or *assistive devices* that are appropriate for the individual with a disability. Such devices are selected based on the specific disability, the function to be augmented or restored, the user's wishes, the clinician's preferences, cost, and the environment in which the device will be used.

Rehabilitation science: The *development* of a body of knowledge, gleaned from rigorous basic and clinical research, that describes how a disability alters specific physiological functions or anatomical structures, and that details the underlying principles by which **residual function** or capacity can be measured and used to restore function of individuals with disabilities.

119.1 Rehabilitation Concepts

Effective rehabilitation engineers must be well versed in all of the areas described above because they generally work in a team setting, in collaboration with physical and occupational therapists, orthopedic surgeons, physical medicine specialists, and/or neurologists. Some rehabilitation engineers are interested in certain activities that we do in the course of a normal day that could be summarized as *activities of daily living* (**ADL**). These include eating, toileting, combing hair, brushing teeth, reading, etc. Other engineers focus on *mobility* and the limitations to mobility. Mobility can be personal (e.g., within a home or office) or public (automobile, public transportation, accessibility questions in buildings). Mobility also includes the ability to move functionally through the environment. Thus, the question of mobility is not limited to that of getting from place to place, but also includes such questions as whether one can reach an object in a particular setting or whether a paralyzed urinary bladder can be made functional again. Barriers that limit mobility are also studied. For example, an ill-fitted wheelchair cushion or support system will most assuredly limit mobility by reducing the time that an individual can spend in a wheelchair before he or she must vacate it to avoid serious and difficult-to-heal pressure sores. Other groups of rehabilitation engineers deal with *sensory disabilities*, such as sight or hearing, or with *communications disorders*, both on the production side (e.g., the nonvocal) or on the comprehension side. For any given client, a rehabilitation engineer might have all of these concerns to consider (i.e., ADLs, mobility, sensory, and communication dysfunctions).

A key concept in physical or sensory rehabilitation is that of *residual function* or *residual capacity*. Such a concept implies that the function or sense can be quantified, that the performance range of that function or sense is known in a nonimpaired population, and that the use of residual capacity by a disabled individual should be encouraged. These measures of human performance can be made subjectively by clinicians or objectively by some rather clever computerized test devices.

A rehabilitation engineer asks three key questions: Can a diminished function or sense be successfully augmented? Is there a substitute way to return the function or to restore a sense? And is the solution appropriate and cost-effective? These questions give rise to two important rehabilitation concepts: orthotics and prosthetics. An ***orthosis***is an appliance that aids an existing function. A ***prosthesis***provides a substitute.

An artificial limb is a *prosthesis*, as is a wheelchair. An ankle brace is an *orthosis*; so are eyeglasses. In fact, eyeglasses might well be the penultimate rehabilitation device. They are inexpensive, have little social stigma, and are almost completely unobtrusive to the user. They have let many millions of individuals with correctable vision problems lead productive lives. But in essence, a pair of eyeglasses is an optical device, governed by traditional equations of physical optics. Eyeglasses can be made out of simple glass (from a raw material as abundant as the sands of the earth) or complex plastics such as those that are ultraviolet sensitive. They can be ground by hand or by sophisticated computer-controlled optical grinders. Thus, crude technology can restore functional vision. Increasing the technical content of the eyeglasses (either by material or manufacturing method) in most cases will not increase the amount of function restored, but it might make the glasses cheaper, lighter, and more prone to be used.

119.2 Engineering Concepts in Sensory Rehabilitation

Of the five traditional senses, vision and hearing best define the interactions that permit us to be human. These two senses are the main input channels through which data with high information content can flow. We read; we listen to speech or music; we view art. A loss of one or the other of these senses (or both) can have a devastating impact on the individual affected. Rehabilitation engineers attempt to restore the functions of these senses, either through augmentation or via sensory substitution systems. Eyeglasses and hearing aids are examples of augmentative devices that can be used if some residual capacity remains. A major area of rehabilitation engineering research deals with *sensory substitution systems* [Kaczmarek et al., 1991].

The visual system has the capability to detect a single photon of light, yet also has a dynamic range that can respond to intensities many orders of magnitude greater. It can work with high contrast items and with those of almost no contrast, and across the visible spectrum of colors. Millions of parallel data channels form the optic nerve that comes from an eye; each channel transmits an asynchronous and quasi-random (in time) stream of binary pulses. While the temporal coding on any one of these channels is not fast (on the order of

200 bits per second or less), the capacity of the human brain to parallel process the entire image is faster than any supercomputer yet built.

If sight is lost, how can it be replaced? A simple pair of eyeglasses will not work, because either the sensor (the retina), the communication channel (the optic nerve and all of its relays to the brain), or one or more essential central processors (the occipital part of the cerebral cortex for initial processing; the parietal and other cortical areas for information extraction) has been damaged. For replacement within the system, one must determine where the visual system has failed and whether a stage of the system can be artificially bypassed. If one uses another sensory modality (e.g., touch or hearing) as an alternate input channel, one must determine whether there is sufficient bandwidth in that channel and whether the higher-order processing hierarchy is plastic enough to process information coming via a different route.

While the above discussion might seem just philosophical, it is more than that. We normally read printed text with our eyes. We recognize words from their (visual) letter combinations. We comprehend what we read via a mysterious processing in the parietal and temporal parts of the cerebral cortex. Could we perhaps read and comprehend this text or other forms of writing through our fingertips with an appropriate interface? The answer, surprisingly, is yes! And, the adaptation actually goes back to one of the earliest applications of coding theory — that of the development of Braille. Braille condenses all text characters to a raised matrix of 2 by 3 dots (2^6 combinations), with certain combinations reserved as indicators for the next character (such as a number indicator) or for special contractions. Trained readers of Braille can read over 250 words per minute of grade 2 Braille (as fast as most sighted readers can read printed text). Thus, the Braille code is in essence a rehabilitation engineering concept where an alternate sensory channel is used as a substitute and where a recoding scheme has been employed.

Rehabilitation engineers and their colleagues have designed other ways to read text. To replace the retina as a sensor element, a modern high-resolution, high-sensitivity, fast-imaging sensor (CCD, etc.) is employed to capture a visual image of the text. One method, used by various page scanning devices, converts the scanned image to text by using optical character recognition schemes, and then outputs the text as speech via text-to-speech algorithms. This machine essentially recites the text, much as a sighted helper might do when reading aloud to the blind individual. The user of the device is thus freed of the absolute need for a helper. Such *independence* is often the goal of rehabilitation.

Perhaps the most interesting method presents an image of the scanned data directly to the visual cortex or retina via an array of implantable electrodes that are used to electrically activate nearby cortical or retinal structures. The visual cortex and retina are laid out in topographic fashion such that there is an orderly mapping of the signal from different parts of the visual field to the retina, and from the retina to corresponding parts of the occipital cortex. The goal of stimulation is to mimic the neural activity that would have been evoked had the signal come through normal channels. And, such stimulation does produce the sensation of light. Since the "image" stays within the visual system, the rehabilitation solution is said to be ***modality specific***. However, substantial problems dealing with biocompatibility and image processing and reduction remain in the design of the electrode arrays and processors that serve to interface the electronics and neurological tissue.

Deafness is another manifestation of a loss of a communication channel, this time for the sense of hearing. Totally deaf individuals use vision as a substitute input channel when communicating via sign language (also a substitute code), and can sign at information rates that match or exceed that of verbal communication. Hearing aids are now commercially available that can adaptively filter out background noise (a predictable signal) while amplifying speech (unpredictable) using autoregressive, moving average (ARMA) signal processing. With the recent advent of powerful digital signal processing chips, true digital hearing aids are now available. Previous analog aids, or digitally programmable analog aids, provided a set of tunable filters and amplifiers to cover the low-, mid-, and high-frequency ranges of the hearing spectrum. But the digital aids can be specifically and easily tailored (i.e., programmed) to compensate for the specific losses of each individual client across the frequency continuum of hearing, and still provide automatic gain control and one or more user-selectable settings that have been adjusted to perform optimally in differing noise environments.

An exciting development is occurring outside the field of rehabilitation that will have a profound impact on the ability of the deaf to comprehend speech. Electronics companies are now beginning to market universal translation aids for travelers, where a phrase spoken in one language is captured, parsed, translated, and restated

(either spoken or displayed) in another language. The deaf would simply require that the visual display be in the language that they use for writing.

Deafness is often brought on (or occurs congenitally) by damage to the cochlea. The cochlea normally transduces variations in sound pressure intensity at a given frequency into patterns of neural discharge. This neural code is then carried by the auditory (eighth cranial) nerve to the brainstem, where it is preprocessed and relayed to the auditory cortex for initial processing and on to the parietal and other cortical areas for information extraction. Similar to the case for the visual system, the cochlea, auditory nerve, auditory cortex, and all relays in between maintain a topological map, this time based on tone frequency (tonotopic). If deafness is solely due to cochlear damage (as is often the case) and if the auditory nerve is still intact, a cochlear implant can often be substituted for the regular transducer array (the cochlea) while still sending the signal through the normal auditory channel (to maintain modality specificity).

At first glance, the design of a cochlear prosthesis to restore hearing appears daunting. The hearing range of a healthy young individual is 20 to 16,000 Hz. The transducing structure, the cochlea, has 3500 inner and 12,000 outer hair cells, each best activated by a specific frequency that causes a localized mechanical resonance in the basilar membrane of the cochlea. Deflection of a hair cell causes the cell to fire an all-or-none (i.e., pulsatile) neuronal discharge, whose rate of repetition depends to a first approximation on the amplitude of the stimulus. The outputs of these hair cells have an orderly convergence on the 30,000 to 40,000 fibers that make up the auditory portion of the eighth cranial nerve. These afferent fibers, in turn, go to brainstem neurons that process and relay the signals on to higher brain centers [Klinke, 1983]. For many causes of deafness, the hair cells are destroyed, but the eighth nerve remains intact. Thus, if one could elicit activity in a specific output fiber by means other than the hair cell motion, perhaps some sense of hearing could be restored. The geometry of the cochlea helps in this regard as different portions of the nerve are closer to different parts of the cochlea.

Electrical stimulation is now used in the cochlear implant to bypass hair cell transduction mechanisms [Loeb, 1985; Clark et al., 1990]. These sophisticated devices have required that complex signal processing, electronic, and packaging problems be solved. One current cochlear implant has 22 stimulus sites along the scala tympani of the cochlea. Those sites provide excitation to the peripheral processes of the cells of the eighth cranial nerve, which are splayed out along the length of the scala. The electrode assembly itself has 22 ring electrodes spaced along its length and some additional guard rings between the active electrodes and the receiver to aid in securing the very flexible electrode assembly after it is snaked into the cochlea's very small (a few millimeters) round window (a surgeon related to me that positioning the electrode was akin to pushing a piece of cooked spaghetti through a small hole at the end of a long tunnel). The electrode is attached to a receiver that is inlaid into a slot milled out of the temporal bone. The receiver contains circuitry that can select any electrode ring to be a source and any other electrode to be a sink for the stimulating current, and that can rapidly sequence between various pairs of electrodes. The receiver is powered and controlled by a radiofrequency link with an external transmitter, whose alignment is maintained by means of a permanent magnet imbedded in the receiver.

A digital signal processor stores information about a specific user and his or her optimal electrode locations for specific frequency bands. The object is to determine what pair of electrodes best produces the subjective perception of a certain pitch *in the implanted individual himself or herself*, and then to associate a particular filter with that pair via the controller. An enormous amount of compression occurs in taking the frequency range necessary for speech comprehension and reducing it to a few discrete channels. At present, the optimum compression algorithm is unknown, and much fundamental research is being carried out in speech processing, compression, and recognition. But, what is amazing is that a number of totally deaf individuals can relearn to comprehend speech exceptionally well without speech-reading through the use of these implants. Other individuals find that the implant aids in speech-reading. For some, only an awareness of environmental sounds is apparent; and for another group, the implant appears to have little effect. But, if you could (as I have been able to) finally converse in unaided speech with an individual who had been rendered totally blind and deaf by a traumatic brain injury, you would certainly begin to appreciate the power of rehabilitation engineering.

119.3 Engineering Concepts in Motor Rehabilitation

Limitations in mobility can severely restrict the quality of life of an individual so affected. A wheelchair is a prime example of a prosthesis that can restore personal mobility to those who cannot walk. Given the proper

environment (fairly level floors, roads, etc.), modern wheelchairs can be highly efficient. In fact, the fastest times in one of man's greatest tests of endurance, the Boston Marathon, are achieved by the wheelchair racers. Although they do gain the advantage of being able to roll, they still must climb the same hills, and do so with only one fifth of the muscle power available to an able-bodied marathoner.

While a wheelchair user could certainly go down a set of steps (not recommended), climbing steps in a normal manual or electric wheelchair is a virtual impossibility. Ramps or lifts are engineered to provide accessibility in these cases, or special climbing wheelchairs can be purchased. Wheelchairs also do not work well on surfaces with high rolling resistance or viscous coefficients (e.g., mud, rough terrain, etc.), so alternate mobility aids must be found if access to these areas is to be provided to the physically disabled. Hand-controlled cars, vans, tractors, and even airplanes are now driven by wheelchair users. The design of appropriate control modifications falls to the rehabilitation engineer.

Loss of a limb can greatly impair functional activity. The engineering aspects of artificial limb design increase in complexity as the amount of residual limb decreases, especially if one or more joints are lost. As an example, a person with a mid-calf amputation could use a simple wooden stump to extend the leg, and could ambulate reasonably well. But such a leg is not cosmetically appealing and completely ignores any substitution for ankle function.

Immediately following World War II, the U.S. government began the first concerted effort to foster better engineering design for artificial limbs. Dynamically lockable knee joints were designed for artificial limbs for above-knee amputees. In the ensuing years, energy-storing artificial ankles have been designed, some with prosthetic feet so realistic that beach thongs could be worn with them. Artificial hands, wrists, and elbows were designed for upper-limb amputees. Careful design of the actuating cable system also provided for a sense of hand grip force, so that the user had some feedback and did not need to rely on vision alone for guidance.

Perhaps the most transparent (to the user) artificial arms are the ones that use electrical activity generated by the muscles remaining in the stump to control the actions of the elbow, wrist, and hand [Stein et al., 1988]. This electrical activity is known as myoelectricity, and is produced as the muscle contraction spreads through the muscle. Note that these muscles, if intact, would have controlled at least one of these joints (e.g., the biceps and triceps for the elbow). Thus, a high level of modality specificity is maintained because the functional element is substituted only at the last stage. All of the batteries, sensor electrodes, amplifiers, motor actuators, and controllers (generally analog) reside entirely within these myoelectric arms. An individual trained in the use of a myoelectric arm can perform some impressive tasks with this arm. Current engineering research efforts involve the control of simultaneous multi-joint movements (rather than the single joint movement now available) and the provision for sensory feedback from the end effector of the artificial arm to the skin of the stump via electrical means.

119.4 Engineering Concepts in Communications Disorders

Speech is a uniquely human means of interpersonal communication. Problems that affect speech can occur at the initial transducer (the larynx) or at other areas of the vocal tract. They can be of neurological (due to cortical, brainstem, or peripheral nerve damage), structural, and/or cognitive origin. A person might only be able to make a halting attempt at talking, or might not have sufficient control of other motor skills to type or write.

If only the larynx is involved, an externally applied artificial larynx can be used to generate a resonant column of air that can be modulated by other elements in the vocal tract. If other motor skills are intact, typing can be used to generate text, which in turn can be spoken via text-to-speech devices described above. And the rate of typing (either whole words or via coding) might be fast enough so that reasonable speech rates could be achieved.

The rehabilitation engineer often becomes involved in the design or specification of *augmentative communication aids* for individuals who do not have good muscle control, either for speech or for limb movement. An entire industry has developed around the design of symbol or letter boards, where the user can point out (often painstakingly) letters, words, or concepts. Some of these boards now have speech output. Linguistics and information theory have been combined in the invention of acceleration techniques intended to speed up

the communication process. These include alternative language representation systems based on semantic (iconic), alphanumeric, or other codes; and prediction systems, which provide choices based on previously selected letters or words.

Some individuals can produce speech, but it is dysarthric and very difficult to understand. Yet the utterance does contain information. Can this limited information be used to figure out what the individual wanted to say, and then voice it by artificial means? Research labs are now employing neural network theory to determine which pauses in an utterance are due to content (i.e., between a word or sentence) and which are due to unwanted halts in speech production.

119.5 Appropriate Technology

Rehabilitation engineering lies at the interface of a wide variety of technical, biological, and other concerns. A user might (and often does) put aside a technically sophisticated rehabilitation device in favor of a simpler device that is cheaper and easier to use and maintain. The cosmetic appearance of the device (or cosmesis) sometimes becomes the overriding factor in acceptance or rejection of a device. A key design factor often lies in the use of the ***appropriate technology*** to accomplish the task adequately, given the extent of the resources available to solve the problem and the residual capacity of the client. Adequacy can be verified by determining that increasing the technical content of the solution results in disproportionately diminishing gains or escalating costs. Thus, a rehabilitation engineer must be able to distinguish applications where high technology is required from those where such technology results in an incremental gain in cost, durability, acceptance, and other factors. Further, appropriateness very much depends on location. What is appropriate to a client near a major medical center in a highly developed country might not be appropriate to one in a rural setting or in a developing country.

This is not to say that rehabilitation engineers should shun advances in technology. In fact, a fair proportion of rehabilitation engineers work in a research setting where state-of-the-art technology is being applied to the needs of the disabled. However, it is often difficult to transfer complex technology from a laboratory to disabled consumers not directly associated with that laboratory. Such devices are often designed for use only in a structured environment, are difficult to repair properly in the field, and often require a high level of user interaction or sophistication.

Technology transfer in the rehabilitation arena is difficult, due to the limited and fragmented market. Advances in rehabilitation engineering are often piggybacked onto advances in commercial electronics. For example, the exciting developments in text-to-speech and speech-to-text devices mentioned above are being driven by the commercial marketplace, and not by the rehabilitation arena. But such developments will be welcomed by rehabilitation engineers no less.

119.6 The Future of Engineering in Rehabilitation

The traditional engineering disciplines permeate many aspects of rehabilitation. Signal processing, control and information theory, materials design, and computers are all in widespread use from an electrical engineering perspective. Neural networks, microfabrication, fuzzy logic, virtual reality, image processing, and other emerging electrical and computer engineering tools are increasingly being applied. Mechanical engineering principles are used in biomechanical studies, gait and motion analysis, prosthetic fitting, seat cushion and back support design, and the design of artificial joints. Materials and metallurgical engineers provide input on newer biocompatible materials. Chemical engineers are developing implantable sensors. Industrial engineers are increasingly studying rehabilitative ergonomics.

The challenge to rehabilitation engineers is to find advances in *any* field — engineering or otherwise — that will aid their clients who have a disability.

Defining Terms

[*Note:* the first five terms below have been proposed by the National Center for Medical Rehabilitation and Research (NCMRR) of the U.S. National Institutes of Health (NIH).]

Activities of daily living (ADL): Personal activities that are done by almost everyone in the course of a normal day, including eating, toileting, combing hair, brushing teeth, reading, etc. ADLs are distinguished from hobbies and from work-related activities (e.g., typing).

Appropriate technology: The technology that will accomplish a task adequately, given the resources available. Adequacy can be verified by determining that increasing the technological content of the solution results in diminishing gains or increasing costs.

Disability: Inability or limitation in performing tasks, activities, and roles to levels expected within physical and social contexts.

Functional limitation: Restriction or lack of ability to perform an action in the manner or within the range consistent with the purpose of an organ or organ system.

Impairment: Loss or abnormality of cognitive, emotional, physiological, or anatomical structure or function, including all losses or abnormalities, not just those attributed to the initial **pathophysiology.**

Modality-specific: A task that is specific to a single sense or movement pattern.

Orthosis: A modality-specific appliance that aids the performance of a function or movement by augmenting or assisting the residual capabilities of that function or movement. An orthopedic brace is an orthosis.

Pathophysiology: Interruption or interference with normal physiological and developmental processes or structures.

Prosthesis: An appliance that substitutes for the loss of a particular function, generally by involving a different modality as an input and/or output channel. An artificial limb, a sensory substitution system, or an augmentative communication aid are prosthetic devices.

Residual function or **residual capacity:** *Residual function* is a measure of the ability to carry out one of more general tasks using the methods normally used. *Residual capacity* is a measure of the ability to to carry out these tasks using any means of performance. These residual measures are generally more subjective than other more quantifiable measures such as residual strength.

Societal limitation: Restriction, attributable to social policy or barriers (structural or attitudinal), that limits fulfillment of roles, or denies access to services or opportunities that are associated with full participation in society.

References

Much of this material also appeared in:

Clark, G.M., Y.C. Tong, and J.F. Patrick, 1990. *Cochlear Prostheses,* Churchill Livingstone, Edinburgh.

Goodenough-Trepagnier, C., 1994. Guest Editor of a special issue of *Assistive Technology,* 6(1), dealing with mental loads in augmentative communication.

Kaczmarek, K.A., J.G. Webster, P. Bach-y-Rita, and W.J. Tompkins, 1991. Electrotactile and vibrotactile displays for sensory substitution, *IEEE Trans. Biomed. Engr.,* 38:1–16.

Klinke, R., 1983. Physiology of the sense of equilibrium, hearing and speech. Chapter 12 in *Human Physiology* (eds: R.F. Schmidt and G. Thews), Springer-Verlag, Berlin.

Loeb, G.E., 1985. The Functional Replacement of the Ear, *Scientific American,* 252:104–111.

Reswick, J. 1982. What is a rehabiliation engineer? in *Annual Review of Rehabiltation,* Vol. 2 (eds. E.L. Pan, T.E. Backer, C.L. Vash), Springer-Verlag, New York.

Robinson, C.J. 1993. Rehabilitation Engineering — an editorial, *IEEE Transactions on Rehabilitation Engineering,* 1(1):1–2.

Robinson, C.J., 1995. Rehabilitation Engineering, Science, and Technology, *The Biomedical Engineering* (J.O. Bronzino, Editor), CRC Press LLC, Boca Raton, FL, pp. 2045–2054.

Stein, R.B., D. Charles, and K.B. James, 1988. Providing motor control for the handicapped: A fusion of modern neuroscience, bioengineering, and rehabilitation, *Advances in Neurology, Vol. 47: Functional Recovery in Neurological Disease,* (ed. S.G. Waxman), Raven Press, New York.

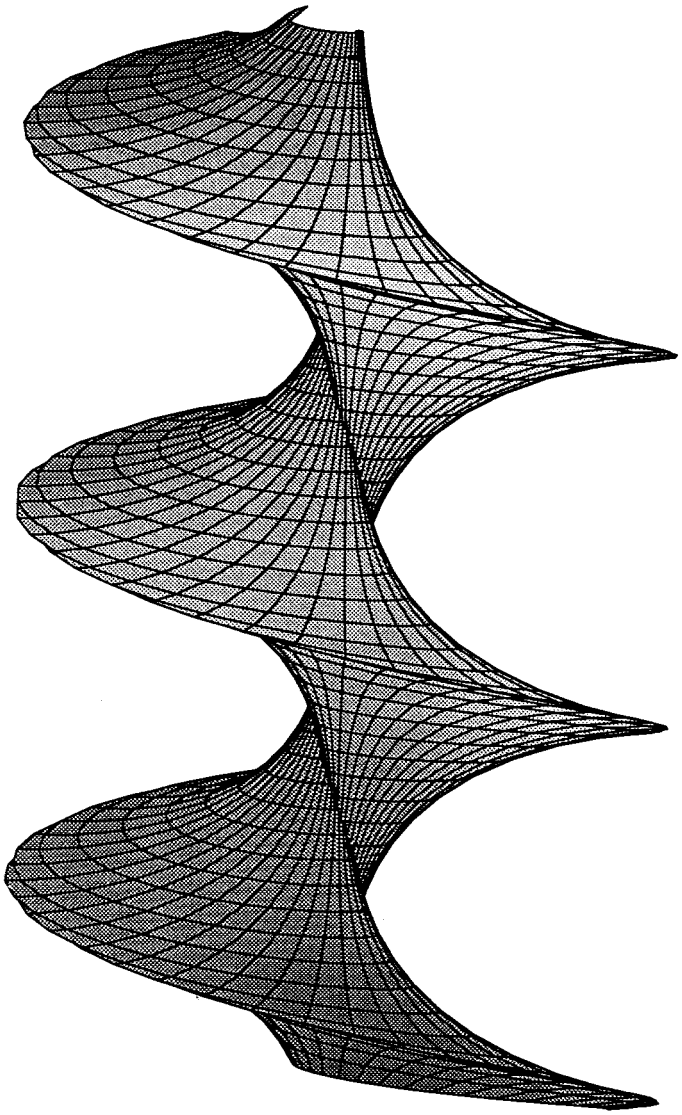
Futher Information

Readers interested in rehabilitation engineering can contact RESNA — an interdisciplinary association for the advancement of rehabilitation and assistive technologies — at 1101 Connecticut Ave., N.W., Suite 700, Washington, D.C. 20036. RESNA publishes a quarterly journal called *Assistive Technology*.

The U.S. Department of Veterans Affairs puts out a quarterly *Journal of Rehabilitation R&D*. The January issue each year contains an overview of most of the rehabilitation engineering efforts occurring in the U.S. and Canada, with over 500 listings.

The IEEE Engineering in Medicine and Biology Society publishes *IEEE Transactions on Rehabilitation Engineering*, a quarterly journal. The reader should contact the IEEE at P.O. Box 1331, 445 Hoes Lane, Piscataway, NJ 08855-1331 for further details.

Tallarida, R.J. "Section XII – Mathematics, Symbols, and Physical Constants"
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000



The mathematical equation used to generate this three-dimensional figure is worth a thousand words. It represents a single-solitron surface for the sine-Gordon equation $w_{uv} = \sin w$. Among the areas in which the sine-Gordon equation arises is that of wave propagation on nonlinear transmission lines and in semi-conductors. The equation is famous because it is known to have a nonlinear superposition principle obtainable by means of a Bäcklund Transformation. The sine-Gordon equation is an example of an evolution equation which has an infinite sequence of non-trivial conservation laws so important in the fields of engineering and physics. For further information on the Bäcklund Transformation see *Bäcklund Transformations and their Application,* Rogers and Shadwick, Academic Press, 1982.

This three-dimensional projection was generated using the MAPLE® software package. MAPLE® is one of three important mathematical computer packages that offer a variety of analytical and numerical software for use by scientists, engineers, and mathematicians.

This figure was developed by W.K. Schief and C. Rogers and the Center for Dynamical Systems and Nonlinear Studies at Georgia Institute of Technology and the University of New South Wales in Sydney, Australia. (Figure courtesy of Schief and Rogers.)

XII

Mathematics, Symbols, and Physical Constants

Greek Alphabet

International System of Units (SI)
Definitions of SI Base Units • Names and Symbols for the SI Base Units • SI Derived Units with Special Names and Symbols • Units in Use Together with the SI

Conversion Constants and Multipliers
Recommended Decimal Multiples and Submultiples • Conversion Factors—Metric to English • Conversion Factors—English to Metric • Conversion Factors—General • Temperature Factors • Conversion of Temperatures

Physical Constants
General • π Constants • Constants Involving e • Numerical Constants

Symbols and Terminology for Physical and Chemical Quantities
Classical Mechanics • Electricity and Magnetism • Electromagnetic Radiation • Solid State

Credits

Ronald J. Tallarida
Temple University

THE GREAT ACHIEVEMENTS in engineering deeply affect the lives of all of us and also serve to remind us of the importance of mathematics. Interest in mathematics has grown steadily with these engineering achievements and with concomitant advances in pure physical science. Whereas scholars in nonscientific fields, and even in such fields as botany, medicine, geology, etc., can communicate most of the problems and results in nonmathematical language, this is virtually impossible in present-day engineering and physics. Yet it is interesting to note that until the beginning of the twentieth century engineers regarded calculus as something of a mystery. Modern students of engineering now study calculus, as well as differential equations, complex variables, vector analysis, orthogonal functions, and a variety of other topics in applied analysis. The study of systems has ushered in matrix algebra and, indeed, most engineering students now take linear algebra as a core topic early in their mathematical education.

This section contains concise summaries of relevant topics in applied engineering mathematics and certain key formulas, that is, those formulas that are most often needed in the formulation and solution of engineering problems. Whereas even inexpensive electronic calculators contain tabular material (e.g., tables of trigonometric and logarithmic functions) that used to be needed in this kind of handbook, most calculators do not give symbolic results. Hence, we have included formulas along with brief summaries that guide their use. In many cases we have added numerical examples, as in the discussions of matrices, their inverses, and their use in the solutions of linear systems. A table of derivatives is included, as well as key applications of the derivative in the solution of problems in maxima and minima, related rates, analysis of curvature, and finding approximate

roots by numerical methods. A list of infinite series, along with the interval of convergence of each, is also included.

Of the two branches of calculus, integral calculus is richer in its applications, as well as in its theoretical content. Though the theory is not emphasized here, important applications such as finding areas, lengths, volumes, centroids, and the work done by a nonconstant force are included. Both cylindrical and spherical polar coordinates are discussed, and a table of integrals is included. Vector analysis is summarized in a separate section and includes a summary of the algebraic formulas involving dot and cross multiplication, frequently needed in the study of fields, as well as the important theorems of Stokes and Gauss. The part on special functions includes the gamma function, hyperbolic functions, Fourier series, orthogonal functions, and both Laplace and *z*-transforms. The Laplace transform provides a basis for the solution of differential equations and is fundamental to all concepts and definitions underlying analytical tools for describing feedback control systems. The *z*-transform, not discussed in most applied mathematics books, is most useful in the analysis of discrete signals as, for example, when a computer receives data sampled at some prespecified time interval. The Bessel functions, also called cylindrical functions, arise in many physical applications, such as the heat transfer in a "long" cylinder, whereas the other orthogonal functions discussed—Legendre, Hermite, and Laguerre polynomials—are needed in quantum mechanics and many other subjects (e.g., solid-state electronics) that use concepts of modern physics.

The world of mathematics, even applied mathematics, is vast. Even the best mathematicians cannot keep up with more than a small piece of this world. The topics included in this section, however, have withstood the test of time and, thus, are truly *core* for the modern engineer.

This section also incorporates tables of physical constants and symbols widely used by engineers. While not exhaustive, the constants, conversion factors, and symbols provided will enable the reader to accommodate a majority of the needs that arise in design, test, and manufacturing functions.

Mathematics, Symbols, and Physical Constants

Greek Alphabet

Greek letter		Greek name	English equivalent	Greek letter		Greek name	English equivalent
A	α	Alpha	a	N	ν	Nu	n
B	β	Beta	b	Ξ	ξ	Xi	x
Γ	γ	Gamma	g	O	o	Omicron	ŏ
Δ	δ	Delta	d	Π	π	Pi	p
E	ε	Epsilon	ĕ	P	ρ	Rho	r
Z	ζ	Zeta	z	Σ	σ	Sigma	s
H	η	Eta	ē	T	τ	Tau	t
Θ	θ ϑ	Theta	th	Y	υ	Upsilon	u
I	ι	Iota	i	Φ	ϕ φ	Phi	ph
K	κ	Kappa	k	X	χ	Chi	ch
Λ	λ	Lambda	l	Ψ	ψ	Psi	ps
M	μ	Mu	m	Ω	ω	Omega	ō

International System of Units (SI)

The International System of units (SI) was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960. It is a coherent system of units built form seven *SI base units,* one for each of the seven dimensionally independent base quantities: they are the meter, kilogram, second, ampere, kelvin, mole, and candela, for the dimensions length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity, respectively. The definitions of the SI base units are given below. The *SI derived units* are expressed as products of powers of the base units, analogous to the corresponding relations between physical quantities but with numerical factors equal to unity.

In the International System there is only one SI unit for each physical quantity. This is either the appropriate SI base unit itself or the appropriate SI derived unit. However, any of the approved decimal prefixes, called *SI prefixes,* may be used to construct decimal multiples or submultiples of SI units.

It is recommended that only SI units be used in science and technology (with SI prefixes where appropriate). Where there are special reasons for making an exception to this rule, it is recommended always to define the units used in terms of SI units. This section is based on information supplied by IUPAC.

Definitions of SI Base Units

Meter—The meter is the length of path traveled by light in vacuum during a time interval of 1/299 792 458 of a second (17th CGPM, 1983).

Kilogram—The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).

Second—The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967).

Ampere—The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length (9th CGPM, 1948).

Kelvin—The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (13th CGPM, 1967).

Mole—The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, or other particles, or specified groups of such particles (14th CGPM, 1971).

Examples of the use of the mole:

1 mol of H_2 contains about 6.022×10^{23} H_2 molecules, or 12.044×10^{23} H atoms
1 mol of HgCl has a mass of 236.04 g
1 mol of Hg_2Cl_2 has a mass of 472.08 g
1 mol of Hg_2^{2+} has a mass of 401.18 g and a charge of 192.97 kC
1 mol of $Fe_{0.91}S$ has a mass of 82.88 g
1 mol of e^- has a mass of 548.60 μg and a charge of –96.49 kC
1 mol of photons whose frequency is 10^{14} Hz has energy of about 39.90 kJ

Candela—The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of (1/683) watt per steradian (16th CGPM, 1979).

Names and Symbols for the SI Base Units

Physical quantity	Name of SI unit	Symbol for SI unit
length	meter	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd

SI Derived Units with Special Names and Symbols

Physical quantity	Name of SI unit	Symbol for SI unit	Expression in terms of SI base units	
frequency[1]	hertz	Hz	s^{-1}	
force	newton	N	m kg s^{-2}	
pressure, stress	pascal	Pa	N m^{-2}	$=m^{-1}$ kg s^{-2}
energy, work, heat	joule	J	N m	$= m^2$ kg s^{-2}
power, radiant flux	watt	W	J s^{-1}	$= m^2$ kg s^{-3}
electric charge	coulomb	C	A s	
electric potential, electromotive force	volt	V	J C^{-1}	$= m^2$ kg s^{-3} A^{-1}
electric resistance	ohm	Ω	V A^{-1}	$= m^2$ kg s^{-3} A^{-2}
electric conductance	siemens	S	$Ω^{-1}$	$= m^{-2}$ kg^{-1} s^3 A^2
electric capacitance	farad	F	C V^{-1}	$= m^{-2}$ kg^{-1} s^4 A^2
magnetic flux density	tesla	T	V s m^{-2}	= kg s^{-2} A^{-1}
magnetic flux	weber	Wb	V s	$= m^2$ kg s^{-2} A^{-1}
inductance	henry	H	V A^{-1} s	$= m^2$ kg s^{-2} A^{-2}
Celsius temperature[2]	degree Celsius	°C	K	
luminous flux	lumen	lm	cd sr	
illuminance	lux	lx	cd sr m^{-2}	
activity (radioactive)	becquerel	Bq	s^{-1}	
absorbed dose (of radiation)	gray	Gy	J kg^{-1}	$= m^2$ s^{-2}
dose equivalent (dose equivalent index)	sievert	Sv	J kg^{-1}	$= m^2$ s^{-2}

Physical quantity	Name of SI unit	Symbol for SI unit	Expression in terms of SI base units	
plane angle	radian	rad	1	$= \mathrm{m\ m^{-1}}$
solid angle	steradian	sr	1	$= \mathrm{m^2\ m^{-2}}$

[1]For radial (circular) frequency and for angular velocity the unit rad s^{-1}, or simply s^{-1}, should be used, and this may not be simplified to Hz. The unit Hz should be used only for frequency in the sense of cycles per second.

[2]The Celsius temperature θ is defined by the equation:

$$\theta/°\mathrm{C} = T/\mathrm{K} - 273.15$$

The SI unit of Celsius temperature interval is the degree Celsius, °C, which is equal to the kelvin, K. °C should be treated as a single symbol, with no space between the ° sign and the letter C. (The symbol °K, and the symbol °, should no longer be used.)

Units in Use Together with the SI

These units are not part of the SI, but it is recognized that they will continue to be used in appropriate contexts. SI prefixes may be attached to some of these units, such as milliliter, ml; millibar, mbar; megaelectronvolt, MeV; kilotonne, ktonne.

Physical quantity	Name of unit	Symbol for unit	Value in SI units
time	minute	min	60 s
time	hour	h	3600 s
time	day	d	86 400 s
plane angle	degree	°	$(\pi/180)$ rad
plane angle	minute	′	$(\pi/10\ 800)$ rad
plane angle	second	″	$(\pi/648\ 000)$ rad
length	ångstrom[1]	Å	10^{-10} m
area	barn	b	$10^{-28}\ \mathrm{m^2}$
volume	litre	l, L	$\mathrm{dm^3} = 10^{-3}\ \mathrm{m^3}$
mass	tonne	t	$\mathrm{Mg} = 10^3$ kg
pressure	bar[1]	bar	10^5 Pa $= 10^5\ \mathrm{N\ m^{-2}}$
energy	electronvolt[2]	eV $(= e \times \mathrm{V})$	$\approx 1.60218 \times 10^{-19}$ J
mass	unified atomic mass unit[2,3]	u $(= m_a(^{12}\mathrm{C})/12)$	$\approx 1.66054 \times 10^{-27}$ kg

[1]The ångstrom and the bar are approved by CIPM for "temporary use with SI units," until CIPM makes a further recommendation. However, they should not be introduced where they are not used at present.

[2]The values of these units in terms of the corresponding SI units are not exact, since they depend on the values of the physical constants e (for the electronvolt) and N_a (for the unified atomic mass unit), which are determined by experiment.

[3]The unified atomic mass unit is also sometimes called the dalton, with symbol Da, although the name and symbol have not been approved by CGPM.

Conversion Constants and Multipliers

Recommended Decimal Multiples and Submultiples

Multiples and submultiples	Prefixes	Symbols	Multiples and submultiples	Prefixes	Symbols
10^{18}	exa	E	10^{-1}	deci	d
10^{15}	peta	P	10^{-2}	centi	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ (Greek mu)

Multiples and submultiples	Prefixes	Symbols	Multiples and submultiples	Prefixes	Symbols
10^6	mega	M	10^{-9}	nano	n
10^3	kilo	k	10^{-12}	pico	p
10^2	hecto	h	10^{-15}	femto	f
10	deca	da	10^{-18}	atto	a

Conversion Factors—Metric to English

To obtain	Multiply	By
Inches	Centimeters	0.3937007874
Feet	Meters	3.280839895
Yards	Meters	1.093613298
Miles	Kilometers	0.6213711922
Ounces	Grams	$3.527396195 \times 10^{-2}$
Pounds	Kilogram	2.204622622
Gallons (U.S. Liquid)	Liters	0.2641720524
Fluid ounces	Milliliters (cc)	$3.381402270 \times 10^{-2}$
Square inches	Square centimeters	0.155003100
Square feet	Square meters	10.76391042
Square yards	Square meters	1.195990046
Cubic inches	Milliliters (cc)	$6.102374409 \times 10^{-2}$
Cubic feet	Cubic meters	35.31466672
Cubic yards	Cubic meters	1.307950619

Conversion Factors—English to Metric*

To obtain	Multiply	By
Microns	Mils	**25.4**
Centimeters	Inches	**2.54**
Meters	Feet	**0.3048**
Meters	Yards	**0.9144**
Kilometers	Miles	**1.609344**
Grams	Ounces	28.34952313
Kilograms	Pounds	**0.45359237**
Liters	Gallons (U.S. Liquid)	**3.785411784**
Millimeters (cc)	Fluid ounces	29.57352956
Square centimeters	Square inches	**6.4516**
Square meters	Square feet	**0.09290304**
Square meters	Square yards	**0.83612736**
Milliliters (cc)	Cubic inches	**16.387064**
Cubic meters	Cubic feet	$2.831684659 \times 10^{-2}$
Cubic meters	Cubic yards	0.764554858

Conversion Factors—General*

To obtain	Multiply	By
Atmospheres	Feet of water @ 4°C	2.950×10^{-2}
Atmospheres	Inches of mercury @ 0°C	3.342×10^{-2}
Atmospheres	Pounds per square inch	6.804×10^{-2}
BTU	Foot-pounds	1.285×10^{-3}
BTU	Joules	9.480×10^{-4}
Cubic feet	Cords	**128**
Degree (angle)	Radians	57.2958
Ergs	Foot-pounds	1.356×10^{7}

*Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

To obtain	Multiply	By
Feet	Miles	**5280**
Feet of water @ 4°C	Atmospheres	33.90
Foot-pounds	Horsepower-hours	1.98×10^{6}
Foot-pounds	Kilowatt-hours	2.655×10^{6}
Foot-pounds per min	Horsepower	3.3×10^{4}
Horsepower	Foot-pounds per sec	1.818×10^{-3}
Inches of mercury @ 0°C	Pounds per square inch	2.036
Joules	BTU	1054.8
Joules	Foot-pounds	1.35582
Kilowatts	BTU per min	1.758×10^{-2}
Kilowatts	Foot-pounds per min	2.26×10^{-5}
Kilowatts	Horsepower	0.745712
Knots	Miles per hour	0.86897624
Miles	Feet	1.894×10^{-4}
Nautical miles	Miles	0.86897624
Radians	Degrees	1.745×10^{-2}
Square feet	Acres	**43560**
Watts	BTU per min	17.5796

Temperature Factors

°F = 9/5 (°C) + 32

Fahrenheit temperature = 1.8 (temperature in kelvins) – 459.67

°C = 5/9 [(°F) – 32)]

Celsius temperature = temperature in kelvins – 273.15

Fahrenheit temperature = 1.8 (Celsius temperature) + 32

Conversion of Temperatures

From	To	
°Celsius	°Fahrenheit	$t_F = (t_C \times 1.8) + 32$
	Kelvin	$T_K = t_C + 273.15$
	°Rankine	$T_R = (t_C + 273.15) \times 18$
°Fahrenheit	°Celsius	$t_C = \frac{t_F - 32}{1.8}$
	Kelvin	$T_k = \frac{t_F - 32}{1.8} + 273.15$
	°Rankine	$T_R = t_F + 459.67$
Kelvin	°Celsius	$t_C = T_K - 273.15$
	°Rankine	$T_R = T_K \times 1.8$
°Rankine	Kelvin	$T_K = \frac{T_R}{1.8}$
	°Farenheit	$t_F = T_R - 459.67$

Physical Constants

General

Equatorial radius of the earth = 6378.388 km = 3963.34 miles (statute).
Polar radius of the earth, 6356.912 km = 3949.99 miles (statute).
1 degree of latitude at 40° = 69 miles.
1 international nautical mile = 1.15078 miles (statute) = 1852 m = 6076.115 ft.
Mean density of the earth = 5.522 g/cm^3 = 344.7 lb/ft^3
Constant of gravitation $(6.673 \pm 0.003) \times 10^{-8}$ cm^3 gm^{-1} s^{-2}.

Acceleration due to gravity at sea level, latitude 45° = 980.6194 cm/s^2 = 32.1726 ft/s^2.
Length of seconds pendulum at sea level, latitude 45° = 99.3575 cm = 39.1171 in.
1 knot (international) = 101.269 ft/min = 1.6878 ft/s = 1.1508 miles (statute)/h.
1 micron = 10^{-4} cm.
1 ångstrom = 10^{-8} cm.
Mass of hydrogen atom = $(1.67339 \pm 0.0031) \times 10^{-24}$ g.
Density of mercury at 0°C = 13.5955 g/ml.
Density of water at 3.98°C = 1.000000 g/ml.
Density, maximum, of water, at 3.98°C = 0.999973 g/cm^3.
Density of dry air at 0°C, 760 mm = 1.2929 g/l.
Velocity of sound in dry air at 0°C = 331.36 m/s – 1087.1 ft/s.
Velocity of light in vacuum = $(2.997925 \pm 0.000002) \times 10^{10}$ cm/s.
Heat of fusion of water 0°C = 79.71 cal/g.
Heat of vaporization of water 100°C = 539.55 cal/g.
Electrochemical equivalent of silver 0.001118 g/s international amp.
Absolute wavelength of red cadmium light in air at 15°C, 760 mm pressure = 6438.4696 Å.
Wavelength of orange-red line of krypton 86 = 6057.802 Å.

π Constants

π = 3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37511
$1/\pi$ = 0.31830 98861 83790 67153 77675 26745 02872 40689 19291 48091
π^2 = 9.8690 44010 89358 61883 44909 99876 15113 53136 99407 24079
$\log_e \pi$ = 1.14472 98858 49400 17414 34273 51353 05871 16472 94812 91531
$\log_{10} \pi$ = 0.49714 98726 94133 85435 12682 88290 89887 36516 78324 38044
$\log_{10} \sqrt{2\pi}$ = 0.39908 99341 79057 52478 25035 91507 69595 02099 34102 92128

Constants Involving *e*

e = 2.71828 18284 59045 23536 02874 71352 66249 77572 47093 69996
$1/e$ = 0.36787 94411 71442 32159 55237 70161 46086 74458 11131 03177
e^2 = 7.38905 60989 30650 22723 04274 60575 00781 31803 15570 55185
$M = \log_{10} e$ = 0.43429 44819 03251 82765 11289 18916 60508 22943 97005 80367
$1/M = \log_e 10$ = 2.30258 50929 94045 68401 79914 54684 36420 67011 01488 62877
$\log_{10} M$ = 9.63778 43113 00536 78912 29674 98645 –10

Numerical Constants

$\sqrt{2}$ = 1.41421 35623 73095 04880 16887 24209 69807 85696 71875 37695
$\sqrt[3]{2}$ = 1.25992 10498 94873 16476 72106 07278 22835 05702 51464 70151
$\log_e 2$ = 0.69314 71805 59945 30941 72321 21458 17656 80755 00134 36026
$\log_{10} 2$ = 0.30102 99956 63981 19521 37388 94724 49302 67881 89881 46211
$\sqrt{3}$ = 1.73205 08075 68877 29352 74463 41505 87236 69428 05253 81039
$\sqrt[3]{3}$ = 1.44224 95703 07408 38232 16383 10780 10958 83918 69253 49935
$\log_e 3$ = 1.09861 22886 68109 69139 52452 36922 52570 46474 90557 82275
$\log_{10} 3$ = 0.47712 12547 19662 43729 50279 03255 11530 92001 28864 19070

Symbols and Terminology for Physical and Chemical Quantities

Name	*Symbol*	*Definition*	*SI unit*
	Classical Mechanics		
mass	m		kg
reduced mass	μ	$\mu = m_1 m_2/(m_1 + m_2)$	kg
density, mass density	ρ	$\rho = M/V$	kg m^{-3}
relative density	d	$d = \rho/\rho^{\theta}$	l
surface density	ρ_A, ρ_S	$\rho_A = m/A$	kg m^{-2}

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	*Symbol*	*Definition*	*SI unit*
		Classical Mechanics (continued)	
momentum	$\boldsymbol{p}$	$\boldsymbol{p} = m\boldsymbol{v}$	kg m s^{-1}
angular momentum, action	$\boldsymbol{L}$	$\boldsymbol{l} = \boldsymbol{r} ¥ \boldsymbol{p}$	J s
moment of inertia	I, J	$I = \Sigma m_i r_i^2$	kg m^2
force	$\boldsymbol{F}$	$\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t = m\boldsymbol{a}$	N
torque, moment of a force	$\boldsymbol{T}$, ($\boldsymbol{M}$)	$\boldsymbol{T} = \boldsymbol{r} \times \boldsymbol{F}$	N m
energy	E		J
potential energy	E_p, V, Φ	$E_p = -\int \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{s}$	J
kinetic energy	E_k, T, K	$e_k = (1/2)mv^2$	J
work	W, w	$w = \int \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{s}$	J
Hamilton function	H	$H(q, p) = T(q, p) + V(q)$	J
Lagrange function	L	$L(q, \dot{q})$ T$(q, \dot{q}) - V(q)$	J
pressure	p, P	$p = F/A$	Pa, N m^{-2}
surface tension	γ, σ	$\gamma = \mathrm{d}W/\mathrm{d}A$	N m^{-1}, J m^{-2}
weight	G, (W, P)	$G = mg$	N
gravitational constant	G	$F = Gm_1m_2/r^2$	N m^2 kg^{-2}
normal stress	σ	$\sigma = F/A$	Pa
shear stress	τ	$\tau = F/A$	Pa
linear strain, relative elongation	ε, e	$\varepsilon = \Delta l/l$	l
modulus of elasticity, Young's modulus	E	$E = \sigma/\varepsilon$	Pa
shear strain	γ	$\gamma = \Delta x/d$	l
shear modulus	G	$G = \tau/\gamma$	Pa
volume strain, bulk strain	θ	$\theta = \Delta V/V_0$	l
bulk modulus, compression modulus	K	$K = -V_0(\mathrm{d}p/\mathrm{d}V)$	Pa
viscosity, dynamic viscosity	η, μ	$\tau_{x,z} = \eta(\mathrm{d}v_x/\mathrm{d}z)$	Pa s
fluidity	ϕ	$\phi = 1/\eta$	m kg^{-1} s
kinematic viscosity	ν	$\nu = \eta/\rho$	m^2 s^{-1}
friction coefficient	μ, (f)	$F_{\mathrm{frict}} = \mu F_{\mathrm{norm}}$	l
power	P	$P = \mathrm{d}W/\mathrm{d}t$	W
sound energy flux	P, P_a	$P = \mathrm{d}E/\mathrm{d}t$	W
acoustic factors			
reflection factor	ρ	$\rho = P_r/P_0$	l
acoustic absorption factor	α_a, (α)	$\alpha_a = 1 - \rho$	l
transmission factor	τ	$\tau = P_{tr}/P_0$	l
dissipation factor	δ	$\delta = \alpha_a - \tau$	l
		Electricity and Magnetism	
quantity of electricity, electric charge	Q		C
charge density	ρ	$\rho = Q/V$	C m^{-3}
surface charge density	σ	$\sigma = Q/A$	C m^{-2}
electric potential	V, ϕ	$V = \mathrm{d}W/\mathrm{d}Q$	V, J C^{-1}
electric potential difference	U, ΔV, $\Delta\phi$	$U = V_2 - V_1$	V
electromotive force	E	$E = \int(\boldsymbol{F}/Q) \cdot \mathrm{d}\boldsymbol{s}$	V
electric field strength	$\boldsymbol{E}$	$\boldsymbol{E} = \boldsymbol{F}/Q = -\mathrm{grad}\ V$	V m^{-1}
electric flux	Ψ	$\Psi = \int \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{A}$	C
electric displacement	$\boldsymbol{D}$	$\boldsymbol{D} = \varepsilon\boldsymbol{E}$	C m^{-2}
capacitance	C	$C = Q/U$	F, C V^{-1}
permittivity	ε	$D = \varepsilon E$	F m^{-1}
permittivity of vacuum	ε_0	$\varepsilon_0 = \mu_0^{-1} c_0^{-2}$	F m^{-1}
relative permittivity	ε_r	$\varepsilon_r = \varepsilon/\varepsilon_0$	l
dielectric polarization	$\boldsymbol{P}$	$\boldsymbol{P} = \boldsymbol{D} - \varepsilon_0\boldsymbol{E}$	C m^{-2}

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	*Symbol*	*Definition*	*SI unit*
Electricity and Magnetism (continued)			
(dipole moment per volume)			
electric susceptibility	χ_e	$\chi_e = \varepsilon_r - 1$	l
electric dipole moment	$\boldsymbol{p}$, μ	$\boldsymbol{p} = Q\boldsymbol{r}$	C m
electric current	I	$I = \mathrm{d}Q/\mathrm{d}t$	A
electric current density	$\boldsymbol{j}$, $\boldsymbol{J}$	$I = \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A}$	A m^{-2}
magnetic flux density, magnetic induction	$\boldsymbol{B}$	$\boldsymbol{F} = Q\boldsymbol{v} \times \boldsymbol{B}$	T
magnetic flux	Φ	$\Phi = \int \boldsymbol{B} \cdot \mathrm{d}A$	Wb
magnetic field strength	$\boldsymbol{H}$	$\boldsymbol{B} = \mu\boldsymbol{H}$	A M^{-1}
permeability	μ	$\boldsymbol{B} = \mu\boldsymbol{H}$	N A^{-2}, H m^{-1}
permeability of vacuum	μ_0		H m^{-1}
relative permeability	μ_r	$\mu_r = \mu/\mu_0$	l
magnetization (magnetic dipole moment per volume)	$\boldsymbol{M}$	$\boldsymbol{M} = \boldsymbol{B}/\mu_0 - \boldsymbol{H}$	A m^{-1}
magnetic susceptibility	χ, κ, (χ_m)	$\chi = \mu_r - 1$	l
molar magnetic susceptibility	χ_m	$\chi_m = V_m\chi$	m^3 mol^{-1}
magnetic dipole moment	$\boldsymbol{m}$, μ	$E_p = -\boldsymbol{m} \cdot \boldsymbol{B}$	A m^2, J T^{-1}
electrical resistance	R	$\boldsymbol{P} = \boldsymbol{Y}/\boldsymbol{I}$	Ω
conductance	G	$G = 1/R$	S
loss angle	δ	$\delta = (\pi/2) + \phi_I - \phi_U$	1, rad
reactance	X	$X = (U/I)\sin\delta$	Ω
impedance (complex impedance)	Z	$Z = R + \mathrm{i}X$	Ω
admittance (complex admittance)	Y	$Y = 1/Z$	S
susceptance	B	$Y = G + \mathrm{i}B$	S
resistivity	ρ	$\rho = E/j$	Ω m
conductivity	κ, γ, σ	$\kappa = 1/\rho$	S m^{-1}
self-inductance	L	$E = -L(\mathrm{d}I/\mathrm{d}t)$	H
mutual inductance	M, L_{12}	$E_1 = L_{12}(\mathrm{D}i_2/\mathrm{d}t)$	H
magnetic vector potential	$\boldsymbol{A}$	$\boldsymbol{B} = \nabla \times \boldsymbol{A}$	Wb m^{-1}
Poynting vector	$\boldsymbol{S}$	$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$	W m^{-2}
Electromagnetic Radiation			
wavelength	λ		m
speed of light			m s^{-1}
in vacuum	c_0		
in a medium	c	$c = c_0/n$	
wavenumber in vacuum	$\bar{\nu}$	$\bar{\nu} = \nu/c_0 = 1/n\lambda$	m^{-1}
wavenumber (in a medium)	σ	$\sigma = 1/\lambda$	m^{-1}
frequency	ν	$\nu = c/\lambda$	Hz
circular frequency, pulsatance	ω	$\omega = 2\pi\nu$	s^{-1}, rad s^{-1}
refractive index	n	$n = c_0/c$	l
Planck constant	h		J s
Planck constant/2π	$\hbar$	$\hbar = h/2\pi$	J s
radiant energy	Q, W		J
radiant energy density	ρ, w	$\rho = Q/V$	J m^{-3}
spectral radiant energy density			
in terms of frequency	ρ_ν, w_ν	$\rho_\nu = \delta\rho/\mathrm{d}\nu$	J m^{-3} Hz^{-1}
in terms of wavenumber	$\rho_{\bar{\nu}}$, $w_{\bar{\nu}}$	$\rho_{\bar{\nu}} = \mathrm{d}\rho/\mathrm{d}\bar{\nu}$	J m^{-2}
in terms of wavelength	ρ_λ, w_λ	$\rho_\lambda = \delta\rho/\mathrm{d}\lambda$	J m^{-4}
Einstein transition probabilities			
spontaneous emission	A_{nm}	$\mathrm{d}N_n/\mathrm{d}t = -A_{nm}N_n$	s^{-1}
stimulated emission	B_{nm}	$\mathrm{d}n_n/\mathrm{d}t = -\rho_{\bar{\nu}}(\bar{\nu}_{nm}) \times B_{nm}N_n$	s kg^{-1}
radiant power, radiant energy per time	Φ, P	$\Phi = \mathrm{d}Q/\mathrm{d}t$	W
radiant intensity	I	$I = \mathrm{d}\Phi/\mathrm{d}\Omega$	W sr^{-1}
radiant exitance (emitted radiant flux)	M	$M = \mathrm{d}\Phi/\mathrm{d}A_{\mathrm{source}}$	W m^{-2}

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	*Symbol*	*Definition*	*SI unit*
		Electromagnetic Radiation (continued)	
irradiance (radiant flux received)	E, (I)	$E = \mathrm{d}\Phi/\delta A$	W m^{-2}
emittance	ε	$\varepsilon = M/M_{\mathrm{bb}}$	l
Stefan-Boltzmann constant	σ	$M_{\mathrm{bb}} = \sigma T^4$	W m^{-2} K^{-4}
first radiation constant	c_1	$c_1 = 2\pi h c_0^2$	W m^2
second radiation constant	c_2	$c_2 = hc_0/k$	K m
transmittance, transmission factor	τ, T	$\tau = \Phi_{\mathrm{tr}}/\Phi_0$	l
absorptance, absorption factor	α	$\alpha = \phi_{\mathrm{abs}}/\phi_0$	l
reflectance, reflection factor	ρ	$\rho = \phi_{\mathrm{refl}}/\Phi_0$	l
(decadic) absorbance	A	$A = \lg(1 - \alpha_{\mathrm{i}})$	l
napierian absorbance	B	$B = \ln(1 - \alpha_{\mathrm{i}})$	l
absorption coefficient			
(linear) decadic	a, K	$a = A/l$	m^{-1}
(linear) napierian	α	$\alpha = B/l$	m^{-1}
molar (decadic)	ε	$\varepsilon = a/c = A/cl$	m^2 mol^{-1}
molar napierian	κ	$\kappa = \alpha/c = B/cl$	m^2 mol^{-1}
absorption index	k	$k = \alpha/4\pi\tilde{\nu}$	l
complex refractive index	$\hat{n}$	$\hat{n} = n + \mathrm{i}k$	l
molar refraction	R, R_{m}	$R = \dfrac{(n^2-1)}{(n^2+2)}\ V_{\mathrm{m}}$	m^3 mol^{-1}
angle of optical rotation	α		l, rad
		Solid State	
lattice vector	$\boldsymbol{R}$, $\boldsymbol{R}_0$		m
fundamental translation vectors for the crystal lattice	$\boldsymbol{a}_1$; $\boldsymbol{a}_2$; $\boldsymbol{a}_3$, $\boldsymbol{a}$; $\boldsymbol{b}$; $\boldsymbol{c}$	$\mathrm{R} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3$	m
(circular) reciprocal lattice vector	$\boldsymbol{G}$	$\boldsymbol{G} \cdot \boldsymbol{R} = 2\pi m$	m^{-1}
(circular) fundamental translation vectors for the reciprocal lattice	$\boldsymbol{b}_1$; $\boldsymbol{b}_2$; $\boldsymbol{b}_3$, $\boldsymbol{a}^*$; $\boldsymbol{b}^*$; $\boldsymbol{c}^*$	$\boldsymbol{a}_i \cdot \boldsymbol{b}_k = 2\pi\delta_{ik}$	m^{-1}
lattice plane spacing	d		m
Bragg angle	θ	$n\lambda = 2d \sin\theta$	l, rad
order of reflection	n		l
order parameters			
short range	σ		l
long range	s		1
Burgers vector	b		m
particle position vector	r, R_j		m
equilibrium position vector of an ion	R_{o}		m
displacement vector of an ion	$\boldsymbol{u}$	$\boldsymbol{u} = \boldsymbol{R} - \boldsymbol{R}_0$	m
Debye-Waller factor	B, D		l
Debye circular wavenumber	q_{D}		m^{-1}
Debye circular frequency	ω_{D}		s^{-1}
Grüneisen parameter	γ, Γ	$\gamma = \alpha V/\kappa C_V$	l
Madelung constant	α, $\mathbf{M}$	$E_{\mathrm{coul}} = \dfrac{\alpha N_A z_+ z_- e^2}{4\pi\varepsilon_0 \mathrm{R}_0}$	l
density of states	N_E	$N_E = \mathrm{d}N(E)/\mathrm{d}E$	J^{-1} m^{-3}
(spectral) density of vibrational modes	N_ω, g	$N_\omega = \mathrm{d}N(\omega)/\mathrm{d}\omega$	s m^{-3}
resistivity tensor	ρ_{ik}	$E = \rho \cdot j$	Ω m
conductivity tensor	σ_{ik}	$\sigma = \rho^{-1}$	S m^{-1}
thermal conductivity tensor	λ_{ik}	$J_q = -\lambda \cdot \mathrm{grad}\ T$	W m^{-1} K^{-1}
residual resistivity	ρ_{R}		Ω m
relaxation time	τ	$\tau = l/v_{\mathrm{F}}$	s
Lorenz coefficient	L	$L = \lambda/\sigma T$	V^2 K^{-2}
Hall coefficient	A_{H}, R_{H}	$\boldsymbol{E} = \rho \cdot \boldsymbol{j} + R_{\mathrm{H}}(\boldsymbol{B} \times \boldsymbol{j})$	m^3 C^{-1}
thermoelectric force	E		V
Peltier coefficient	Π		V
Thomson coefficient	μ, (τ)		V K^{-1}

Symbols and Terminology for Physical and Chemical Quantities (continued)

Name	*Symbol*	*Definition*	*SI unit*
		Solid State (continued)	
work function	Φ	$\Phi = E_\infty - E_F$	J
number density, number concentration	n, (p)		m^{-3}
gap energy	E_γ		J
donor ionization energy	E_δ		J
acceptor ionization energy	E_α		J
Fermi energy	E_Φ, ε_F		J
circular wave vector, propagation vector	$\boldsymbol{k}$, $\boldsymbol{q}$	$k = 2\pi/\lambda$	m^{-1}
Bloch function	$u_k(\boldsymbol{r})$	$\psi(\boldsymbol{r}) = u_k(\boldsymbol{r}) \exp(i\boldsymbol{k} \cdot \boldsymbol{r})$	$m^{-3/2}$
charge density of electrons	ρ	$\rho(\boldsymbol{r}) = -e\psi^*(\boldsymbol{r})\dot{\psi}(\boldsymbol{r})$	C m^{-3}
effective mass	m^*		kg
mobility	μ	$\mu = v_{drift}/E$	$m^2\ V^{-1}\ s^{-1}$
mobility ratio	b	$b = \mu_n/\mu_p$	l
diffusion coefficient	D	$dN/dt = -DA(dn/dx)$	$m^2\ s^{-1}$
diffusion length	L	$L = \sqrt{D\tau}$	m
characteristic (Weiss) temperature	ϕ, ϕ_W		K
Curie temperature	T_C		K
Néel temperature	T_N		K

Credits

Material in Section XII was reprinted from the following sources:

D. R. Lide, Ed., *CRC Handbook of Chemistry and Physics,* 76th ed., Boca Raton, Fla.: CRC Press, 1992: International System of Units (SI), conversion constants and multipliers (conversion of temperatures), symbols and terminology for physical and chemical quantities, fundamental physical constants, classification of electromagnetic radiation.

D. Zwillinger, Ed., *CRC Standard Mathematical Tables and Formulae,* 30th ed., Boca Raton, Fla.: CRC Press, 1996: Greek alphabet, conversion constants and multipliers (recommended decimal multiples and submultiples, metric to English, English to metric, general, temperature factors), physical constants, series expansion.