Soclof, S., Watson, J., Brews, J.R. “Transistors”
The Electrical Engineering Handbook
Ed. Richard C. Dorf
Boca Raton: CRC Press LLC, 2000

24

Transistors

Sidney Soclof
California State University, Los Angeles

Joseph Watson
University of Wales, Swansea

John R. Brews
The University of Arizona

24.1 Junction Field-Effect Transistors

Sidney Soclof

A junction field-effect transistor, or JFET, is a type of transistor in which the current flow through the device between the drain and source electrodes is controlled by the voltage applied to the gate electrode. A simple physical model of the JFET is shown in Fig. 24.1. In this JFET an *n*-type conducting channel exists between drain and source. The gate is a p^+ region that surrounds the *n*-type channel. The gate-to-channel *pn* junction is normally kept reverse-biased. As the reverse bias voltage between gate and channel increases, the depletion region width increases, as shown in Fig. 24.2. The depletion region extends mostly into the *n*-type channel because of the heavy doping on the p^+ side. The depletion region is depleted of mobile charge carriers and thus cannot contribute to the conduction of current between drain and source. Thus as the gate voltage increases, the cross-sectional areas of the *n*-type channel available for current flow decreases. This reduces the current flow between drain and source. As the gate voltage increases, the channel gets further constricted, and the current flow gets smaller. Finally when the depletion regions meet in the middle of the channel, as shown in Fig. 24.3, the channel is pinched off in its entirety between source and drain. At this point the current flow between drain and source is reduced to essentially zero. This voltage is called the **pinch-off voltage,** V_P. The pinch-off voltage is also represented by V_{GS} (off) as being the gate-to-source voltage that turns the drain-to-source current I_{DS} off. We have been considering here an *n*-channel JFET. The complementary device is the *p*-channel JFET that has an n^+ gate region surrounding a *p*-type channel. The operation of a *p*-channel JFET is the same as for an *n*-channel device, except the algebraic signs of all dc voltages and currents are reversed.

We have been considering the case for V_{DS} small compared to the pinch-off voltage such that the channel is essentially uniform from drain to source, as shown in Fig. 24.4(a). Now let's see what happens as V_{DS} increases. As an example let's assume an *n*-channel JFET with a pinch-off voltage of $V_P = -4$ V. We will see what happens

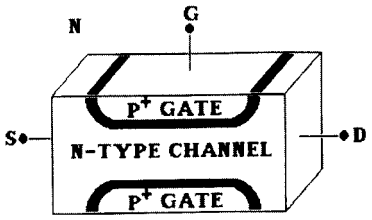


FIGURE 24.1

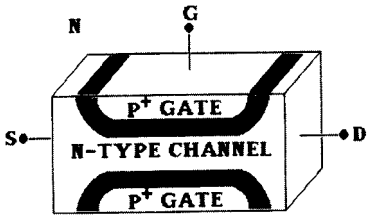


FIGURE 24.2

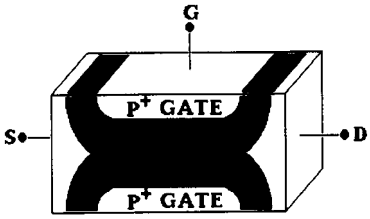


FIGURE 24.3

for the case of $V_{GS} = 0$ as V_{DS} increases. In Fig. 24.4(a) the situation is shown for the case of $V_{DS} = 0$ in which the JFET is fully "on" and there is a uniform channel from source to drain. This is at point A on the I_{DS} vs. V_{DS} curve of Fig. 24.5. The drain-to-source conductance is at its maximum value of g_{ds} (on), and the drain-to-source resistance is correspondingly at its minimum value of r_{ds} (on). Now let's consider the case of $V_{DS} = +1$ V, as shown in Fig. 24.4(b). The gate-to-channel bias voltage at the source end is still $V_{GS} = 0$. The gate-to-channel bias voltage at the drain end is $V_{GD} = V_{GS} - V_{DS} = -1$ V, so the depletion region will be wider at the drain end of the channel than at the source end. The channel will thus be narrower at the drain end than at the source end, and this will result in a decrease in the channel conductance g_{ds} and, correspondingly, an increase in the channel resistance r_{ds}. So the slope of the I_{DS} vs. V_{DS} curve that corresponds to the channel conductance will be smaller at $V_{DS} = 1$ V than it was at $V_{DS} = 0$, as shown at point B on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

In Fig. 24.4(c) the situation for $V_{DS} = +2$ V is shown. The gate-to-channel bias voltage at the source end is still $V_{GS} = 0$, but the gate-to-channel bias voltage at the drain end is now $V_{GD} = V_{GS} - V_{DS} = -2$ V, so the depletion region will now be substantially wider at the drain end of the channel than at the source end. This leads to a further constriction of the channel at the drain end, and this will again result in a decrease in the channel conductance g_{ds} and, correspondingly, an increase in the channel resistance r_{ds}. So the slope of the I_{DS} vs. V_{DS} curve will be smaller at $V_{DS} = 2$ V than it was at $V_{DS} = 1$ V, as shown at point C on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

In Fig. 24.4(d) the situation for $V_{DS} = +3$ V is shown, and this corresponds to point D on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

When $V_{DS} = +4$ V, the gate-to-channel bias voltage will be $V_{GD} = V_{GS} - V_{DS} = 0 - 4$ V $= -4$ V $= V_P$. As a result the channel is now pinched off at the drain end but is still wide open at the source end since $V_{GS} = 0$, as shown in Fig. 24.4(e). It is very important to note that the channel is pinched off just for a very short distance at the drain end so that the drain-to-source current I_{DS} can still continue to flow. This is not at all the same situation as for the case of $V_{GS} = V_P$, where the channel is pinched off in its entirety, all the way from source to drain. When this happens, it is like having a big block of insulator the entire distance between source and drain, and I_{DS} is reduced to essentially zero. The situation for $V_{DS} = +4$ V $= -V_P$ is shown at point E on the I_{DS} vs. V_{DS} curve of Fig. 24.5.

For $V_{DS} > +4$ V, the current essentially saturates and doesn't increase much with further increases in V_{DS}. As V_{DS} increases above +4 V, the pinched-off region at the drain end of the channel gets wider, which increases r_{ds}. This increase in r_{ds} essentially counterbalances the increase in V_{DS} such that I_{DS} does not increase much. This region of the I_{DS} vs. V_{DS} curve in which the channel is pinched off at the drain end is called the **active region** and is also known as the *saturated region.* It is called the active region because when the JFET is to be used as an amplifier, it should be biased and operated in this region. The saturated value of drain current up in the active region for the case of $V_{GS} = 0$ is called the **drain saturation current,** I_{DSS} (the third subscript S

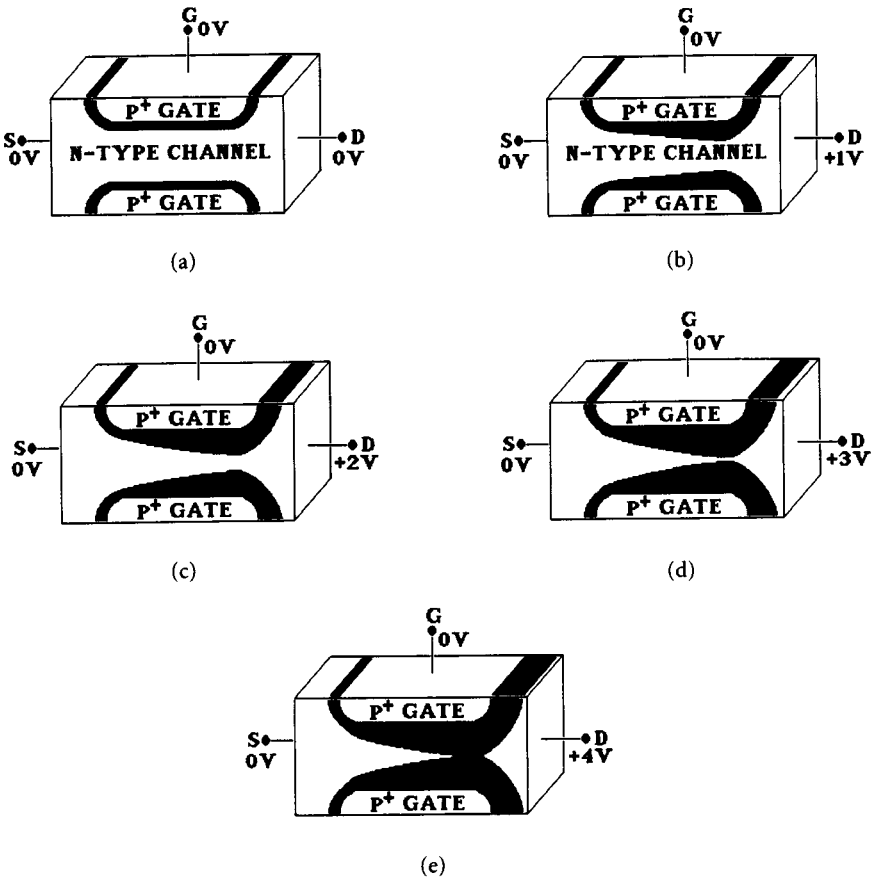


FIGURE 24.4

refers to I_{DS} under the condition of the gate *shorted* to the source). Since there is not really a true saturation of current in the active region, I_{DSS} is usually specified at some value of V_{DS}. For most JFETs, the values of I_{DSS} fall in the range of 1 to 30 mA.

The region below the active region where $V_{DS} < +4\text{ V} = -V_P$ has several names. It is called the **nonsaturated region**, the **triode region**, and the **ohmic region**. The term *triode region* apparently originates from the similarity of the shape of the curves to that of the vacuum tube triode. The term *ohmic region* is due to the variation of I_{DS} with V_{DS} as in Ohm's law, although this variation is nonlinear except for the region of V_{DS} that is small compared to the pinch-off voltage where I_{DS} will have an approximately linear variation with V_{DS}.

The upper limit of the active region is marked by the onset of the breakdown of the gate-to-channel *pn* junction. This will occur at the drain end at a voltage designated as BV_{DG}, or BV_{DS}, since $V_{GS} = 0$. This breakdown voltage is generally in the 30- to 150-V range for most JFETs.

So far we have looked at the I_{DS} vs. V_{DS} curve only for the case of $V_{GS} = 0$. In Fig. 24.6 a family of curves of I_{DS} vs. V_{DS} for various constant values of V_{GS} is presented. This is called the *drain characteristics*, also known as the *output characteristics*, since the output side of the JFET is usually the drain side. In the active region where I_{DS} is relatively independent of V_{DS}, a simple approximate equation relating I_{DS} to V_{GS} is the square-law *transfer equation* as given by $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$. When $V_{GS} = 0$, $I_{DS} = I_{DSS}$ as expected, and as $V_{GS} \rightarrow V_P$, $I_{DS} \rightarrow 0$. The lower boundary of the active region is controlled by the condition that the channel be pinched off at the drain end. To meet this condition

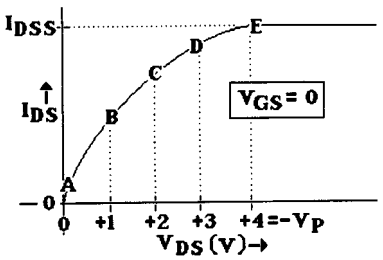


FIGURE 24.5

the basic requirement is that the gate-to-channel bias voltage at the drain end of the channel, V_{GD}, be greater than the pinch-off voltage V_P. For the example under consideration with $V_P = -4$ V, this means that $V_{GD} = V_{GS} - V_{DS}$ must be more negative than –4 V. Therefore, $V_{DS} - V_{GS} \geq +4$ V. Thus, for $V_{GS} = 0$, the active region will begin at $V_{DS} = +4$ V. When $V_{GS} = -1$ V, the active region will begin at $V_{DS} = +3$ V, for now $V_{GD} = -4$ V. When $V_{GS} = -2$ V, the active region begins at $V_{DS} = +2$ V, and when $V_{GS} = -3$ V, the active region begins at $V_{DS} = +1$ V. The dotted line in Fig. 24.6 marks the boundary between the nonsaturated and active regions.

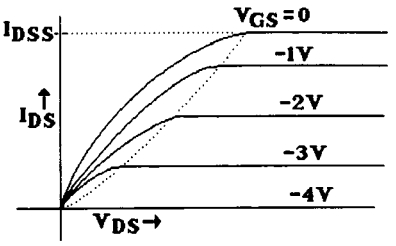


FIGURE 24.6

The upper boundary of the active region is marked by the onset of the avalanche breakdown of the gate-to-channel *pn* junction. When $V_{GS} = 0$, this occurs at $V_{DS} = BV_{DS} = BV_{DG}$. Since $V_{DG} = V_{DS} - V_{GS}$ and breakdown occurs when $V_{DG} = BV_{DG}$, as V_{GS} increases the breakdown voltage decreases, as given by $BV_{DG} = BV_{DS} - V_{GS}$. Thus $BV_{DS} = BV_{DG} + V_{GS}$. For example, if the gate-to-channel breakdown voltage is 50 V, the V_{DS} breakdown voltage will start off at 50 V when $V_{GS} = 0$ but decrease to 46 V when $V_{GS} = -4$ V.

In the nonsaturated region I_{DS} is a function of both V_{GS} and V_{DS}, and in the lower portion of the nonsaturated region where V_{DS} is small compared to V_P, I_{DS} becomes an approximately linear function of V_{DS}. This linear portion of the nonsaturated is called the *voltage-variable resistance* (VVR) region, for in this region the JFET acts like a linear resistance element between source and drain. The resistance is variable in that it is controlled by the gate voltage. This region and VVR application will be discussed in a later section. The JFET can also be operated in this region as a switch, and this will also be discussed in a later section.

JFET Biasing

Voltage Source Biasing

Now we will consider the biasing of JFETs for operation in the active region. The simplest biasing method is shown in Fig. 24.7, in which a voltage source V_{GG} is used to provide the quiescent gate-to-source bias voltage V_{GSQ}. In the active region the transfer equation for the JFET has been given as $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$, so for a quiescent drain current of I_{DSQ} the corresponding gate voltage will be given by $V_{GSQ} = V_P\,(1 - \sqrt{I_{DSQ}/I_{DSS}}$. For a Q point in the middle of the active region, we have that $I_{DSQ} = I_{DSS}/2$, so $V_{GSQ} = V_P\,(1 - \sqrt{1/2}) = 0.293\ V_P$.

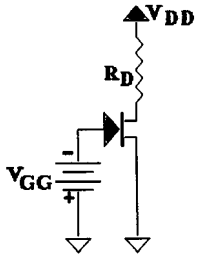


FIGURE 24.7 Voltage source biasing.

The voltage source method of biasing has several major drawbacks. Since V_P will have the opposite polarity of the drain supply voltage V_{DD}, the gate bias voltage will require a second power supply. For the case of an *n*-channel JFET, V_{DD} will come from a positive supply voltage and V_{GG} must come from a separate negative power supply voltage or battery. A second, and perhaps more serious, problem is the "open-loop" nature of this biasing method. The JFET parameters of I_{DDS} and V_P will exhibit very substantial unit-to-unit variations, often by as much as a 2:1 factor. There is also a significant temperature dependence of I_{DDS} and V_P. These variations will lead to major shifts in the position of the *Q* point and the resulting distortion of the signal. A much better biasing method is shown in Fig. 24.8.

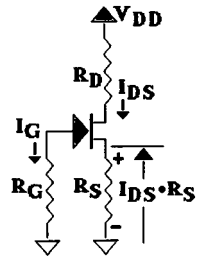


FIGURE 24.8 Self-biasing.

Self-Biasing

The biasing circuit of Fig. 24.8 is called a *self-biasing* circuit in that the gate-to-source voltage is derived from the voltage drop produced by the flow of drain current through the source biasing resistor R_S. It is a closed-loop system in that variations in the JFET parameters can be partially compensated for by the biasing circuit. The gate resistor R_G is used to provide a dc return path for the gate leakage current and is generally up in the megohm range.

The voltage drop across R_S is given by $V_S = I_{DS} \cdot R_S$. The voltage drop across the gate resistor R_G is $V_G = I_G \cdot R_G$. Since I_G is usually in the low nanoampere or even picoampere range, as long as R_G is not extremely large

the voltage drop across R_G can be neglected, so $V_G \cong 0$. Thus, we have that $V_{GS} = V_G - V_S \cong -V_S = -I_{DS} \cdot R_S$. For example, if $I_{DSS} = 10$ mA and $V_P = -4$ V, and for a Q point in the middle of the active region with $I_{DSQ} = I_{DSS}/2 = 5$ mA, we have that $V_{GSQ} = 0.293 V_P = -1.17$ V. Therefore the required value for the source biasing resistor is given by $R_S = -V_{GS}/I_{DSQ} = 1.17$ V/5 mA = 234 Ω. This produces a more stable quiescent point than voltage source biasing, and no separate negative power supply is required.

The closed-loop nature of this biasing circuit can be seen by noting that if changes in the JFET parameters were to cause I_{DS} to increase, the voltage drop across R_S would also increase. This will produce an increase in V_{GS} (in the negative direction for an *n*-channel JFET), which will act to reduce the increase in I_{DS}. Thus the net increase in I_{DS} will be less due to the feedback voltage drop produced by the flow of I_{DS} through R_S. The same basic action would, of course, occur for changes in the JFET parameters that would cause I_{DS} to decrease.

Bias Stability

Now let's examine the stability of the Q point. We will start again with the basic transfer equation as given by $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$. From this equation the change in the drain current, ΔI_{DS}, due to changes in I_{DSS}, V_{GS}, and V_P can be written as

$$\Delta I_{DS} = g_m \Delta V_{GS} - g_m \frac{V_{GS}}{V_P} \Delta V_P + \frac{I_{DS}}{I_{DSS}} \Delta I_{DSS}$$

Since $V_{GS} = -I_{DS} \cdot R_S$, $\Delta V_{GS} = -R_S \cdot \Delta I_{DS}$, we obtain that

$$\Delta I_{DS} = -g_m R_S \Delta I_{DS} - g_m \frac{V_{GS}}{V_P} \Delta V_P + \frac{I_{DS}}{I_{DSS}} \Delta I_{DSS}$$

Collecting terms in ΔI_{DS} on the left side gives

$$\Delta I_{DS}(1 + g_m R_S) = -g_m \frac{V_{GS}}{V_P} \Delta V_P + \frac{I_{DS}}{I_{DSS}} \Delta I_{DSS}$$

Now solving this for ΔI_{DS} yields

$$\Delta I_{DS} = \frac{-g_m (V_{GS}/V_P) \Delta V_P + \dfrac{I_{DS}}{I_{DSS}} \Delta I_{DSS}}{1 + g_m R_S}$$

From this we see that the shift in the quiescent drain current, ΔI_{DS}, is reduced by the presence of R_S by a factor of $1 + g_m R_S$.

If $I_{DS} = I_{DSS}/2$, then

$$g_m = \frac{2\sqrt{I_{DS} \cdot I_{DSS}}}{-V_P} = \frac{2\sqrt{I_{DS} \cdot 2I_{DS}}}{-V_P} = \frac{2\sqrt{2}\, I_{DS}}{-V_P}$$

Since $V_{GS} = 0.293 V_P$, the source biasing resistor will be $R_S = -V_{GS}/I_{DS} = -0.293\ V_P/I_{DS}$. Thus

$$g_m R_S = \frac{2\sqrt{2}\, I_{DS}}{-V_P} \times \frac{-0.293 V_P}{I_{DS}} = 2\sqrt{2} \times 0.293 = 0.83$$

so $1 + g_m R_S = 1.83$. Thus the sensitivity of I_{DS} due to changes in V_P and I_{DSS} is reduced by a factor of 1.83.

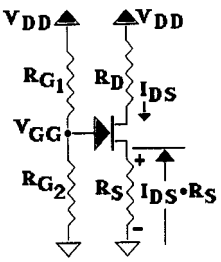


FIGURE 24.9

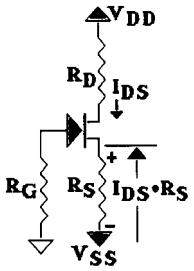


FIGURE 24.10

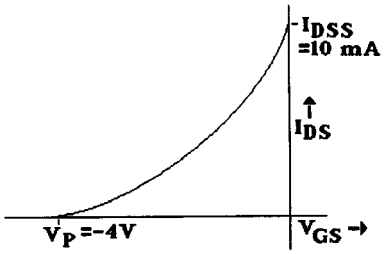


FIGURE 24.11 Transfer characteristic.

The equation for ΔI_{DS} can now be written in the following form for the fractional change in I_{DS}:

$$\frac{\Delta I_{DS}}{I_{DS}} = \frac{-0.83(\Delta V_P/V_P) + 1.41(\Delta I_{DSS}/I_{DSS})}{1.83}$$

so $\Delta I_{DS}/I_{DS} = -0.45\ (\Delta V_P/V_P) + 0.77\ (\Delta I_{DSS}/I_{DSS})$, and thus a 10% change in V_P will result in approximately a 4.5% change in I_{DS}, and a 10% change in I_{DSS} will result in an 8% change in I_{DS}. Thus, although the situation is improved with the self-biasing circuit using R_S, there will still be a substantial variation in the quiescent current with changes in the JFET parameters.

A further improvement in bias stability can be obtained by the use of the biasing methods of Figs. 24.9 and 24.10. In Fig. 24.9 a gate bias voltage V_{GG} is obtained from the V_{DD} supply voltage by means of the R_{G1}–R_{G2} voltage divider. The gate-to-source voltage is now $V_{GS} = V_G - V_S = V_{GG} - I_{DS}R_S$. So now for R_S we have $R_S = (V_{GG} - V_{GS})/I_{DS}$. Since V_{GS} is of opposite polarity to V_{GG}, this will result in a larger value for R_S than before. This in turn will result in a larger value for the $g_m R_S$ product and hence improved bias stability. If we continue with the preceding examples and now let $V_{GG} = V_{DD}/2 = +10$ V, we have that $R_S = (V_{GG} - V_{GS})/I_{DS} = [+10\text{V} - (-1.17\text{V})]/5\text{ mA} = 2.234\ \text{k}\Omega$, as compared to $R_S = 234\ \Omega$ that was obtained before. For g_m we have $g_m = 2\sqrt{I_{DS} \cdot I_{DSS}}/(-V_P) = 3.54$ mS, so $g_m R_S = 3.54\text{ mS} \cdot 2.234\ \text{k}\Omega = 7.90$. Since $1 + g_m R_S = 8.90$, we now have an improvement by a factor of 8.9 over the open-loop voltage source biasing and by a factor of 4.9 over the self-biasing method without the V_{GG} biasing of the gate.

Another biasing method that can lead to similar results is the method shown in Fig. 24.10. In this method the bottom end of the source biasing resistor goes to a negative supply voltage V_{SS} instead of to ground. The gate-to-source bias voltage is now given by $V_{GS} = V_G - V_S = 0 - (I_{DS} \cdot R_S + V_{SS})$ so that for R_S we now have $R_S = (-V_{GS} - V_{SS})/I_{DS}$. If $V_{SS} = -10$ V, and as before $I_{DS} = 5$ mA and $V_{GS} = -1.17$ V, we have $R_S = 11.7\text{ V}/5\text{ mA} = 2.34\ \text{k}\Omega$, and thus $g_m R_S = 7.9$ as in the preceding example. So this method does indeed lead to results similar to that for the R_S and V_{GG} combination biasing. With either of these two methods the change in I_{DS} due to a 10% change in V_P will be only 0.9%, and the change in I_{DS} due to a 10% change in I_{DSS} will be only 1.6%.

The biasing circuits under consideration here can be applied directly to the common-source (CS) amplifier configuration, and can also be used for the common-drain (CD), or source-follower, and common-gate (CG) JFET configurations.

Transfer Characteristics

Transfer Equation

Now we will consider the *transfer characteristics* of the JFET, which is a graph of the output current I_{DS} vs. the input voltage V_{GS} in the active region. In Fig. 24.11 a transfer characteristic curve for a JFET with $V_P = -4$ V and $I_{DSS} = +10$ mA is given. This is approximately a square-law relationship as given by $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$. This equation is not valid for V_{GS} beyond V_P (i.e., $V_{GS} < V_P$), for in this region the channel is pinched off and $I_{DS} \cong 0$.

At $V_{GS} = 0$, $I_{DS} = I_{DSS}$. This equation and the corresponding transfer curve can actually be extended up to the point where $V_{GS} \cong +0.5$ V. In the region where $0 < V_{GS} < +0.5$ V, the gate-to-channel *pn* junction is *forward-biased* and the depletion region width is reduced below the width under zero bias conditions. This reduction in the depletion region width leads to a corresponding expansion of the conducting channel and thus an increase in I_{DS} above I_{DSS}. As long as the gate-to-channel forward bias voltage is less than about 0.5 V, the *pn* junction will be essentially "off" and very little gate current will flow. If V_{GS} is increased much above +0.5 V, however, the gate-to-channel *pn* junction will turn "on" and there will be a substantial flow of gate voltage I_G. This gate current will load down the signal source and produce a voltage drop across the signal source resistance, as shown in Fig. 24.12. This voltage drop can cause V_{GS} to be much smaller than the signal source voltage V_{in}. As V_{in} increases, V_{GS} will ultimately level off at a forward bias voltage of about +0.7 V, and the signal source will lose control over V_{GS}, and hence over I_{DS}. This can result in severe distortion of the input signal in the form of clipping, and thus this situation should be avoided. Thus, although it is possible to increase I_{DS} above I_{DSS} by allowing the gate-to-channel junction to become forward-biased by a small amount ($\leq$0.5 V), the possible benefits are generally far outweighed by the risk of signal distortion. Therefore, JFETs are almost always operated with the gate-to-channel *pn* junction reverse-biased.

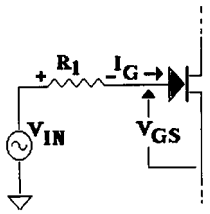


FIGURE 24.12 Effect of forward bias on V_{GS}.

Transfer Conductance

The slope of the transfer curve, dI_{DS}/dV_{GS}, is the *dynamic forward transfer conductance,* or *mutual transfer conductance,* g_m. We see that g_m starts off at zero when $V_{GS} = V_P$ and increases as I_{DS} increases, reaching a maximum when $I_{DS} = I_{DSS}$. Since $I_{DS} = I_{DSS}[1 - (V_{GS}/V_P)]^2$, g_m can be obtained as

$$g_m = \frac{dI_{DS}}{dV_{GS}} = 2I_{DSS}\frac{\left(1 - \dfrac{V_{GS}}{V_P}\right)}{-V_P}$$

Since

$$1 - \left(\frac{V_{GS}}{V_P}\right) = \sqrt{\frac{I_{DS}}{I_{DSS}}}$$

we have that

$$g_m = 2I_{DSS}\frac{\sqrt{I_{DS}/I_{DSS}}}{-V_P} = 2\frac{\sqrt{I_{DS} \cdot I_{DSS}}}{-V_P}$$

The maximum value of g_m is obtained when $V_{GS} = 0$ ($I_{DS} = I_{DSS}$) and is given by $g_m(V_{GS} = 0) = g_{m0} = 2I_{DS}/(-V_P)$.

Small-Signal AC Voltage Gain

Let's consider the CS amplifier circuit of Fig. 24.13. The input ac signal is applied between gate and source, and the output ac voltage is taken between drain and source. Thus the source electrode of this triode device is common to input and output, hence the designation of this JFET configuration as a CS amplifier.

A good choice of the dc operating point or quiescent point (Q point) for an amplifier is in the middle of the active region at $I_{DS} = I_{DSS}/2$. This allows for the maximum symmetrical drain current swing, from the quiescent level of $I_{DSQ} = I_{DSS}/2$, down to a minimum of $I_{DS} \cong 0$, and up to a maximum of $I_{DS} = I_{DSS}$. This choice for the Q point is also a good one from the standpoint of allowing for an adequate safety margin for the location

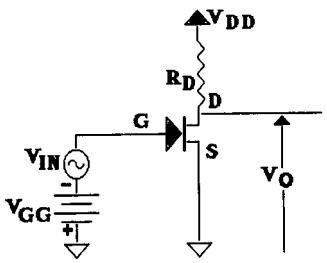


FIGURE 24.13 Common-source amplifier.

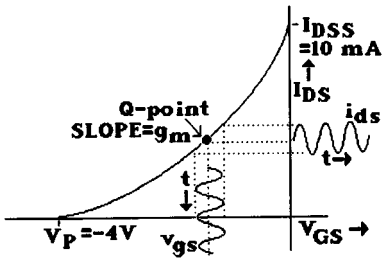


FIGURE 24.14 Transfer characteristic.

of the actual Q point due to the inevitable variations in device and component characteristics and values. This safety margin should keep the Q point well away from the extreme limits of the active region, and thus ensure operation of the JFET in the active region under most conditions. If $I_{DSS} = +10$ mA, then a good choice for the Q point would thus be around +5.0 mA. If $V_P = -4$ V, then

$$g_m = \frac{2\sqrt{I_{DS} \cdot I_{DSS}}}{-V_P} = \frac{2\sqrt{5\text{ mA} \cdot 10\text{ mA}}}{4\text{ V}} = 3.54\text{ mA/V} = 3.54\text{ mS}$$

If a small ac signal voltage v_{GS} is superimposed on the dc gate bias voltage V_{GS}, only a small segment of the transfer characteristic adjacent to the Q point will be traversed, as shown in Fig. 24.14. This small segment will be close to a straight line, and as a result the ac drain current i_{ds} will have a waveform close to that of the ac voltage applied to the gate. The ratio of i_{ds} to v_{GS} will be the slope of the transfer curve as given by $i_{ds}/v_{GS} \cong dI_{DS}/dV_{GS} = g_m$. Thus $i_{ds} \cong g_m v_{GS}$. If the net load driven by the drain of the JFET is the drain load resistor R_D, as shown in Fig. 24.13, then the ac drain current i_{ds} will produce an ac drain voltage of $v_{ds} = -i_{ds} \cdot R_D$. Since $i_{ds} = g_m v_{GS}$, this becomes $v_{ds} = -g_m v_{GS} \cdot R_D$. The ac small-signal voltage gain from gate to drain thus becomes $A_V = v_O/v_{\text{in}} = v_{ds}/v_{GS} = -g_m \cdot R_D$. The negative sign indicates signal inversion as is the case for a CS amplifier.

If the dc drain supply voltage is $V_{DD} = +20$ V, a quiescent drain-to-source voltage of $V_{DSQ} = V_{DD}/2 = +10$ V will result in the JFET being biased in the middle of the active region. Since $I_{DSQ} = +5$ mA in the example under consideration, the voltage drop across the drain load resistor R_D is 10 V. Thus $R_D = 10$ V/5 mA = 2 kΩ. The ac small-signal voltage gain A_V thus becomes $A_V = -g_m \cdot R_D = -3.54$ mS · 2 kΩ = –7.07. Note that the voltage gain is relatively modest as compared to the much larger voltage gains that can be obtained in a bipolar-junction transistor (BJT) common-emitter amplifier. This is due to the lower transfer conductance of both JFETs and MOSFETs (metal-oxide semiconductor field-effect transistors) as compared to BJTs. For a BJT the transfer conductance is given by $g_m = I_C/V_T$, where I_C is the quiescent collector current and $V_T = kT/q \cong 25$ mV is the thermal voltage. At $I_C = 5$ mA, $g_m = 5$ mA/25 mV = 200 mS, as compared to only 3.5 mS for the JFET in this example. With a net load of 2 kΩ, the BJT voltage gain will be –400 as compared to the JFET voltage gain of only 7.1. Thus FETs do have the disadvantage of a much lower transfer conductance, and therefore voltage gain, than BJTs operating under similar quiescent current levels, but they do have the major advantage of a much higher input impedance and a much lower input current. In the case of a JFET the input signal is applied to the *reverse-biased* gate-to-channel *pn* junction and thus sees a very high impedance. In the case of a common-emitter BJT amplifier, the input signal is applied to the *forward-biased* base-emitter junction, and the input impedance is given approximately by $r_{\text{in}} = r_{BE} \cong 1.5 \cdot \beta \cdot V_T/I_C$. If $I_C = 5$ mA and $\beta = 200$, for example, then $r_{\text{in}} \cong 1500\ \Omega$. This moderate input resistance value of 1.5 kΩ is certainly no problem if the signal source resistance is less than around 100 Ω. However, if the source resistance is above 1 kΩ, then there will be a substantial signal loss in the coupling of the signal from the signal source to the base of the transistor. If the source resistance is in the range of above 100 kΩ, and certainly if it is above 1 MΩ, then there will be severe signal attenuation due to the BJT input impedance, and the FET amplifier will probably offer a greater overall voltage gain. Indeed, when high-impedance signal sources are encountered, a multistage amplifier with a FET input stage followed by cascaded BJT stages is often used.

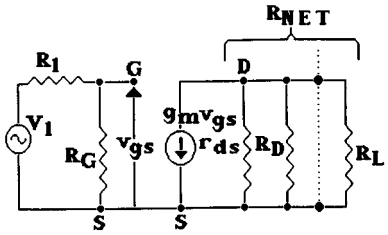


FIGURE 24.15 Effect of r_{ds} on R_{net}.

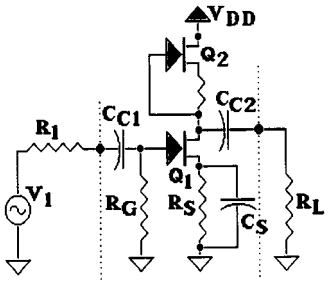


FIGURE 24.16 Active load circuit.

JFET Output Resistance

Dynamic Drain-to-Source Conductance

For the JFET in the active region the drain current I_{DS} is a strong function of the gate-to-source voltage V_{GS} but is relatively independent of the drain-to-source voltage V_{DS}. The transfer equation has previously been stated as $I_{DS} = I_{DSS}\,[1 - (V_{GS}/V_P)]^2$.

The drain current will, however, increase slowly with increasing V_{DS}. To take this dependence of I_{DS} on V_{DS} into account, the transfer equation can be modified to give

$$I_{DS} = I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2\left(1 + \frac{V_{DS}}{V_A}\right)$$

where V_A is a constant called the *Early voltage* and is a parameter of the transistor with units of volts. The early voltage V_A is generally in the range of 30 to 300 V for most JFETs. The variation of the drain current with drain voltage is the result of the *channel length modulation effect* in which the channel length decreases as the drain voltage increases. This decrease in the channel length results in an increase in the drain current. In BJTs a similar effect is the *base width modulation effect.*

The *dynamic drain-to-source conductance* is defined as $g_{ds} = dI_{DS}/dV_{DS}$ and can be obtained from the modified transfer equation $I_{DS} = I_{DSS}\,[1 - (V_{GS}/V_P)]^2\,[1 + V_{DS}/V_A]$ as simply $g_{ds} = I_{DS}/V_A$. The reciprocal of g_{ds} is *dynamic drain-to-source resistance* r_{ds}, so $r_{ds} = 1/g_{ds} = V_A/I_{DS}$. If, for example, $V_A = 100$ V, we have that $r_{ds} = 100\text{ V}/I_{DS}$. At $I_{DS} = 1$ mA, $r_{ds} = 100$ V/1 mA = 100 kΩ, and at $I_{DS} = 10$ mA, $r_{ds} = 10$ kΩ.

Equivalent Circuit Model of CS Amplifier Stage

A small-signal equivalent circuit model of a CS FET amplifier stage is shown in Fig. 24.15. The ac small-signal voltage gain is given by $A_V = -g_m \cdot R_{net}$, where $R_{net} = [r_{ds} \| R_D \| RL]$ is the net load driven by the drain for the FET and includes the dynamic drain-to-source resistance r_{ds}. Since r_{ds} is generally much larger than $[R_D \| R_L]$, it will usually be the case that $R_{net} \cong [R_D \| R_L]$, and r_{ds} can be neglected. There are, however, some cases in which r_{ds} must be taken into account. This is especially true for the case in which an active load is used, as shown in Fig. 24.16. For this case $R_{net} = [r_{ds1} \| r_{ds2} \| R_L]$, and r_{ds} can be a limiting factor in determining the voltage gain.

Consider an example for the active load circuit of Fig. 24.16 for the case of identical JFETs with the same quiescent current. Assume that $R_L >> r_{ds}$ so that $R_{net} \cong [r_{ds1} \| r_{ds2}] = \text{VA}/(2I_{DSQ})$. Let $I_{DSQ} = I_{DSS}/2$, so $g_m = -2\sqrt{I_{DSS} \cdot I_{DSQ}}/(-V_P) = 2\sqrt{2}I_{DSQ}/(-V_P)$. The voltage gain is

$$A_V = -g_m \cdot R_{net} = \frac{2\sqrt{2}I_{DSQ}}{V_P} \times \frac{V_A}{2I_{DSQ}} = \sqrt{2}\,\frac{V_A}{V_P}$$

If $V_A = 100$ V and $V_P = -2$ V, we obtain $A_V = -70$, so we see that with active loads relatively large voltage gains can be obtained with FETs.

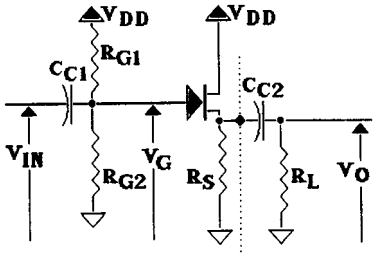


FIGURE 24.17 Source follower.

Another circuit in which the dynamic drain-to-source resistance r_{ds} is important is the constant-current source or current regulator diode. In this case the current regulation is directly proportional to the dynamic drain-to-source resistance.

Source Follower

Source-Follower Voltage Gain

We will now consider the CD JFET configuration, which is also known as the source follower. A basic CD circuit is shown in Fig. 24.17. The input signal is supplied to the gate of the JFET. The output is taken from the source of the JFET, and the drain is connected directly to the V_{DD} supply voltage, which is ac ground.

For the JFET in the active region we have that $i_{ds} = g_m v_{GS}$. For this CD circuit we also have that $v_{GS} = v_G - v_S$ and $v_S = i_{ds} R_{\text{net}}$, where $R_{\text{net}} = [R_S \| R_L]$ is the net load resistance driven by the transistor. Since $v_{GS} = i_{ds}/g_m$, we have that $i_{ds}/g_m = v_G - i_{ds} R_{\text{net}}$. Collecting terms in i_{ds} on the left side yields $i_{ds}[(1/g_m) + R_{\text{net}}] = v_G$, so

$$i_{ds} = \frac{v_G}{(1/g_m) + R_{\text{net}}} = \frac{g_m v_G}{1 + g_m R_{\text{net}}}$$

The output voltage is

$$v_O = v_S = i_{ds} R_{\text{net}} = \frac{g_m R_{\text{net}} v_G}{1 + g_m R_{\text{net}}}$$

and thus the ac small-signal voltage gain is

$$A_V = \frac{v_O}{v_G} = \frac{g_m R_{\text{net}}}{1 + g_m R_{\text{net}}}$$

Upon dividing through by g_m this can be rewritten as

$$A_V = \frac{R_{\text{net}}}{(1/g_m) + R_{\text{net}}}$$

From this we see that the voltage gain will be positive, and thus the source follower is a noninverting amplifier. We also note that A_V will always be less than unity, although for the usual case of $R_{\text{net}} >> 1/g_m$, the voltage gain will be close to unity.

The source follower can be represented as an amplifier with an open-circuit (i.e., no load) voltage transfer ratio of unity and an output resistance of $r_O = 1/g_m$. The equation for A_V can be expressed as $A_V = R_{\text{net}}/(R_{\text{net}} + r_O)$, which is the voltage division ratio of the $r_O = R_{\text{net}}$ circuit.

Source-Follower Examples

Let's consider an example of a JFET with I_{DSS} = 10 mA and V_P = –4 V. Let V_{DD} = +20 V and $I_{DSQ} = I_{DSS}/2$ = 5 mA. For $I_{DS} = I_{DSS}/2$ the value of V_{GS} is –1.17 V. To bias the JFET in the middle of the active region, we will let $V_{GQ} = V_{DD}/2$ = +10 V, so $V_{SQ} = V_{GQ} - V_{GS}$ = +10 V – (–1.17 V) = +11.17 V. Thus $R_S = V_{SQ}/I_{DSQ}$ = 11.17 V/5 mA = 2.23 kΩ.

The transfer conductance at I_{DS} = 5 mA is 3.54 mS so that $r_O = 1/g_m$ = 283 Ω. Since $g_m R_S$ = 7.9, good bias stability will be obtained. If $R_L >> R_S$, then $A_V \cong R_S/(r_O + R_S)$ = 2.23 kΩ/(283 Ω + 2.23 kΩ) = 0.887. If R_L = 1 kΩ, then R_{net} = 690 Ω, and A_V drops to 0.709, and if R_L = 300 Ω, R_{net} = 264 Ω and A_V is down to 0.483. A BJT emitter-follower circuit has the same equations for the voltage gain as the FET source follower. For the BJT case, $r_O = 1/g_m = V_T/I_O$ where V_T = thermal voltage = $kT/q \cong$ 25 mV and I_C is the quiescent collector current. For I_C = 5 mA, we get $r_O \cong$ 25 mV/5 mA = 5 Ω as compared to r_O = 283 Ω for the JFET case at the same quiescent current level. So the emitter follower does have a major advantage over the source follower since it has a much lower output resistance r_O and can thus drive very small load resistances with a voltage gain close to unity. For example, with R_L = 100 Ω, we get $A_V \cong$ 0.26 for the source follower as compared to $A_V \cong$ 0.95 for the emitter follower.

The FET source follower does, however, offer substantial advantages over the emitter follower of a much higher input resistance and a much lower input current. For the case in which a very high-impedance source, up in the megohm range, is to be coupled to a low-impedance load down in the range of 100 Ω or less, a good combination to consider is that of a cascaded FET source follower followed by a BJT emitter follower. This combination offers the very high input resistance of the source follower and the very low output resistance of the emitter follower.

For the source-follower circuit under consideration the input resistance will be $R_{in} = [R_{G1} \| R_{G2}]$ = 10 MΩ. If the JFET gate current is specified as 1 nA (max), and for good bias stability the change in gate voltage due to the gate current should not exceed $|V_P|/10$ = 0.4 V, the maximum allowable value for $[R_{G1} \| R_{G2}]$ is given by $I_G \cdot [R_{G1} \| R_{G2}] < 0.4$ V. Thus $[R_{G1} \| R_{G2}] < 0.4$ V/1 nA = 0.4 GΩ = 400 MΩ. Therefore R_{G1} and R_{G2} can each be allowed to be as large as 800 MΩ, and very large values for R_{in} can thus be obtained. At higher frequencies the input capacitance C_{in} must be considered, and C_{in} will ultimately limit the input impedance of the circuit. Since the input capacitance of the FET will be comparable to that of the BJT, the advantage of the FET source follower over the BJT emitter follower from the standpoint of input impedance will be obtained only at relatively low frequencies.

Source-Follower Frequency Response

The input capacitance of the source follower is given by $C_{in} = C_{GD} + (1 - A_V)C_{GS}$. Since A_V is close to unity, C_{in} will be approximately given by $C_{in} \cong C_{GD}$. The source-follower input capacitance can, however, be reduced below C_{GD} by a bootstrapping circuit in which the drain voltage is made to follow the gate voltage. Let's consider a representative example in which C_{GD} = 5 pF, and let the signal-source output resistance be R_1 = 100 kΩ. The input circuit is in the form of a simple *RC* low-pass network. The *RC* time constant is

$$\tau = [R \| R_{G1} \| R_{G2}] \cdot C_{in} \cong R_1 \cdot C_{in} \cong R_1 \cdot C_{GD}$$

Thus $\tau \cong$ 100 kΩ · 5 pF = 500 ns = 0.5 μs. The corresponding 3-dB or half-power frequency is $f_H = 1/(2\pi\tau)$ = 318 kHz. If R_1 = 1 MΩ, the 3-dB frequency will be down to about 30 kHz. Thus we see indeed the limitation on the frequency response that is due to the input capacitance.

Frequency and Time-Domain Response

Small-Signal CS Model for High-Frequency Response

We will now consider the frequency- and time-domain response of the JFET CS amplifier. In Fig. 24.18 an ac representation of a CS amplifier is shown, the dc biasing not being shown. In Fig. 24.19 the JFET small-signal ac equivalent circuit model is shown including the junction capacitances C_{GS} and C_{GD}. The gate-to-drain capacitance C_{GD} is a feedback capacitance in that it is connected between output (drain) and input (gate). Using

Miller's theorem for shunt feedback this feedback capacitance can be transformed into an equivalent input capacitance $C_{GD'} = (1 - A_V)C_{GD}$ and an equivalent output capacitance $C_{GD''} = (1 - 1/A_V)C_{GD}$, as shown in Fig. 24.20. The net input capacitance is now $C_{in} = C_{GS} + (1 - A_V)C_{GD}$ and the net output capacitance is $C_O = (1 - 1/A_V)C_{GD} + C_L$. Since the voltage gain A_V is given by $A_V = -g_m R_{net}$, where R_{net} represents the net load resistance, the equations for C_{in} and C_O can be written approximately as $C_{in} = C_{GS} + (1 + g_m R_{net})C_{GD}$ and $C_O = [1 + 1/(g_m R_{net})]C_{GD} + C_L$. Since usually $A_V = g_m R_{net} >> 1$, C_O can be written as $C_O \cong C_{GD} + C_L$. Note that the voltage gain given by $A_V = -g_m R_{net}$ is not valid in the higher frequency, where A_V will decrease with increasing frequency. Therefore the expressions for C_{in} and C_O will not be exact but will still be a useful approximation for the determination of the frequency- and time-domain responses. We also note that the contribution of C_{GD} to the input capacitance is increased by the Miller effect factor of $1 + g_m R_{net}$.

The circuit in Fig. 24.21 is in the form of two cascaded RC low-pass networks. The RC time constant on the input side is $\tau_1 = [R_1 \| R_G] \cdot C_{in} \cong R_1 \cdot C_{in}$, where R_1 is the signal-source resistance. The RC time constant on the output side is given by $\tau_2 = R_{net} \cdot C_O$. The corresponding breakpoint frequencies are

$$f_1 = \frac{1}{2\pi\tau_1} = \frac{1}{2\pi R_1 \cdot C_{in}}$$

and

$$f_2 = \frac{1}{2\pi\tau_2} = \frac{1}{2\pi R_{net} \cdot C_O}$$

The 3-dB or half-power frequency of this amplifier stage will be a function of f_1 and f_2. If these two breakpoint frequencies are separated by at least a decade (i.e., 10:1 ratio), the 3-dB frequency will be approximately equal to the lower of the two breakpoint frequencies. If the breakpoint frequencies are not well separated, then the 3-dB frequency can be obtained from the following approximate relationship: $(1/f_{3dB})^2 \cong (1/f_1)^2 + (1/f_2)^2$. The time-domain response as expressed in terms of the 10 to 90% rise time is related to the frequency-domain response by the approximate relationship that $t_{rise} \cong 0.35/f_{3dB}$.

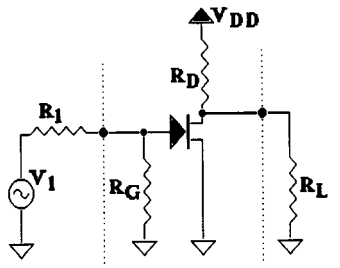


FIGURE 24.18 Common-source amplifier.

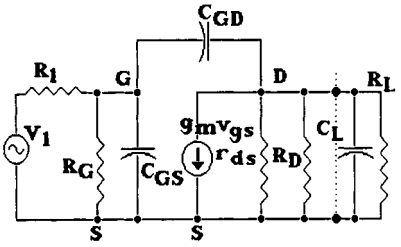


FIGURE 24.19 AC small-signal model.

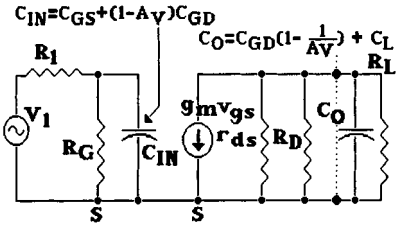


FIGURE 24.20

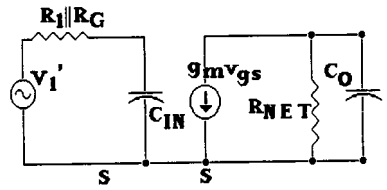


FIGURE 24.21

We will now consider a representative example. We will let $C_{GS} = 10$ pF and $C_{GD} = 5$ pF. We will assume that the net load driven by the drain of the transistors is $R_{net} = 2$ kΩ and $C_L = 10$ pF. The signal-source resistance $R_1 = 100$ Ω. The JFET will have $I_{DSS} = 10$ mA, $I_{DSQ} = I_{DSS}/2 = 5$ mA, and $V_P = -4$ V, so $g_m = 3.535$ mS. Thus the midfrequency gain is $A_V = -g_m R_{net} = -3.535$ mS · 2 kΩ = –7.07. Therefore we have that

$$C_{in} \cong C_{GS} + (1 + g_m R_{net})C_{GD} = 10 \text{ pF} + 8.07 \cdot 5 \text{ pF} = 50.4 \text{ pF}$$

and

$$C_O \cong C_{GD} + C_L = 15 \text{ pF}$$

Thus $\tau_1 = R_1 \cdot C_{in} = 100\ \Omega \cdot 50.4$ pF = 5040 ps = 5.04 ns, and $\tau_2 = R_{net} \cdot C_O = 2$ k$\Omega \cdot 15$ pF = 30 ns. The corresponding breakpoint frequencies are $f_1 = 1/(2\pi \cdot 5.04\text{ ns}) = 31.6$ MHz and $f_2 = 1/(2\pi \cdot 30\text{ ns}) = 5.3$ MHz. The 3-dB frequency of the amplifier can be obtained from $(1/f_{3dB})^2 \cong (1/f_1)^2 + (1/f_2)^2 = (1/31.6\text{ MHz})^2 + (1/5.3\text{ MHz})^2$, which gives $f_{3dB} \cong 5.2$ MHz. The 10 to 90% rise time can be obtained from $t_{rise} \cong 0.35/f_{3dB} = 0.35/5.2$ MHz = 67 ns.

In the preceding example the dominant time constant is the output circuit time constant of $\tau_2 = 30$ ns due to the combination of load resistance and output capacitance. If we now consider a signal-source resistance of 1 kΩ, the input circuit time constant will be $\tau_1 = R_1 \cdot C_{in} = 1000\ \Omega \cdot 50.4$ pF = 50.4 ns. The corresponding breakpoint frequencies are $f_1 = 1/(2\pi \cdot 50.4\text{ ns}) = 3.16$ MHz and $f_2 = 1/(2\pi \cdot 30\text{ ns}) = 5.3$ MHz. The 3-dB frequency is now $f_{3dB} \cong 2.7$ MHz, and the rise time is $t_{rise} \cong 129$ ns. If R_1 is further increased to 10 kΩ, we obtain $\tau_1 = R_1 \cdot C_{in} = 10$ k$\Omega \cdot 50.4$ pF = 504 ns, giving breakpoint frequencies of $f_1 = 1/(2\pi \cdot 504\text{ ns}) = 316$ kHz and $f_2 = 1/(2\pi \cdot 30\text{ ns}) = 5.3$ MHz. Now τ_1 is clearly the dominant time constant, the 3-dB frequency is now down to $f_{3dB} \cong f_1 = 316$ kHz, and the rise time is up to $t_{rise} \cong 1.1$ µs. Finally, for the case of $R_1 = 1$ MΩ, the 3-dB frequency will be only 3.16 kHz and the rise time will be 111 µs.

Use of Source Follower for Impedance Transformation

We see that large values of signal-source resistance can seriously limit the amplifier bandwidth and increase the rise time. In these cases, the use of an impedance transforming circuit such as an FET source follower or a BJT emitter follower can be very useful. Let's consider the use of a source follower as shown in Fig. 24.22. We will assume that both FETs are identical to the one in the preceding examples and are biased at $I_{DSQ} = 5$ mA. The source follower Q_1 will have an input capacitance of $C_{in} = C_{GD} + (1 - A_{V1})C_{GS} \cong C_{GD} = 5$ pF, since A_V will be very close to unity for a source follower that is driving a CS amplifier. The source-follower output resistance will be $r_O = 1/g_m = 1/3.535$ mS = 283 Ω. Let's again consider the case of $R_1 = 1$ MΩ. The time constant due to the combination of R_1 and the input capacitance of the source follower is $\tau_{SF} = 1$ M$\Omega \cdot 5$ pf = 5 µs. The time constant due to the combination of the source-follower output resistance r_O and the input capacitance of the CS stage is $\tau_1 = r_O \cdot C_{in} = 283\ \Omega \cdot 50.4$ pF = 14 ns, and the time constant of the output circuit is $\tau_2 = 30$ ns, as before. The breakpoint frequencies are $f_{SF} = 31.8$ kHz, $f_1 = 11$ MHz, and $f_2 = 5.3$ MHz. The 3-dB frequency of the system is now $f_{3dB} \cong f_{SF} = 31.8$ kHz, and the rise time is $t_{rise} \cong 11$ µs. The use of the source follower thus results in an improvement by a factor of 10:1 over the preceding circuit.

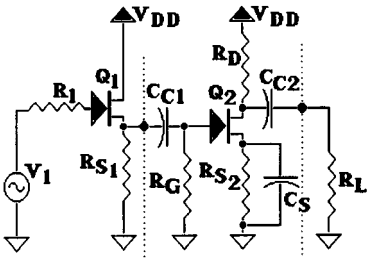


FIGURE 24.22

Voltage-Variable Resistor

Operation of a JFET as a Voltage-Variable Resistor

We will now consider the operation of a JFET as a voltage-variable resistor (VVR). A JFET can be used as a VVR in which the drain-to-source resistance r_{ds} of the JFET can be varied by variation of V_{GS}. For values of $V_{DS} << V_P$ the I_{DS} vs. V_{DS} characteristics are approximately linear, so the JFET looks like a resistor, the resistance value of which can be varied by the gate voltage as shown in Fig. 24.23.

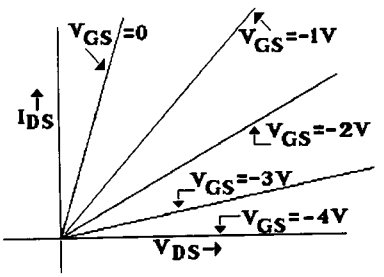
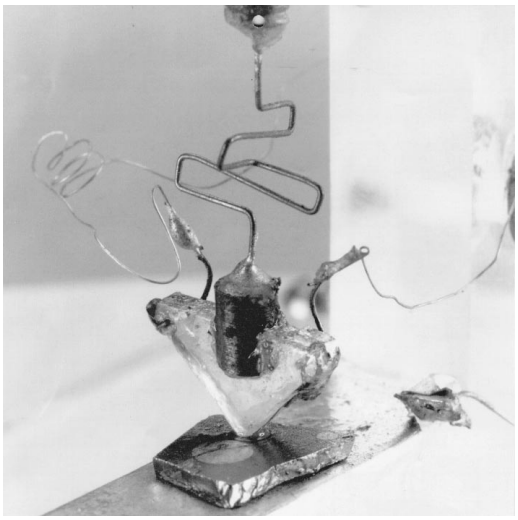


FIGURE 24.23

The channel conductance in the region where $V_{DS} << V_P$ is given by $g_{ds} = A\sigma/L = WH\sigma/L$, where the channel height H is given by $H = H_0 - 2W_D$. In this equation W_D is the depletion region width and H_0 is the value of H as $W_D \rightarrow 0$. The depletion region width is given by $W_D = K\sqrt{V_J} = K\sqrt{V_{GS} + \phi}$, where K is a constant, V_J is the junction voltage, and ϕ is the *pn*-junction contact potential (typically around 0.8 to 1.0 V). As V_{GS} increases, W_D increases and the channel height H decreases as given by $H = H_0 - 2K\sqrt{V_{GS} + \phi}$. When $V_{GS} = V_P$, the channel is completely pinched off, so $H = 0$ and thus $2K\sqrt{V_P + \phi} = H_0$. Therefore $2K = H_0/\sqrt{V_P + \phi}$, and thus

The Invention of the Transistor

In 1907, the American Telephone and Telegraph Company (AT&T) and the Western Electric Company combined their engineering departments and established the Bell Telephone Laboratory on West Street in New York City. By 1921, the laboratories constituted the largest industrial research organization in the country, occupying 400,000 square feet in a 13-story building in lower Manhattan and employing more than 1500 men and women. The organization was put on a more formal footing in 1925, when Frank B. Jewett was made President of Bell Telephone Laboratories, Inc. In the following decades, the labs distinguished themselves by contributions not only to communications technology, but to basic science as well. The awarding of the Nobel Prize in Physics to Clinton J. Davisson in 1937 was simply the most prominent recognition of the laboratories' scientific work.



The first point-contact transistor developed at Bell Labs in 1947 by John Bardeen, William Shockley, and Walter Brattain, had a thin gold foil along the sides of a polystyrene triangle. The foil was slit at the triangle's apex and was pressed against a piece of germanium by the metal piece at the top of the photo (Photo courtesy of AT&T Bell Laboratories.)

The true importance of the fusion of science and engineering in the industrial laboratory was made apparent to all in the years after World War II. In 1947, three Bell Labs physicist-engineers produced the single most significant electronic invention of the era—the transistor. John Bardeen, Walter Brattain, and William Shockley were consciously seeking to exploit new technology about the behavior of semiconducting materials when they devised a way to make a crystal of germanium do the work of a triode vacuum tube, the most basic of electronic components.

$$H = H_0 - H_0 \frac{\sqrt{V_{GS} + \phi}}{\sqrt{V_P + \phi}} = H_0 \left(1 - \frac{\sqrt{V_{GS} + \phi}}{\sqrt{V_P + \phi}} \right)$$

For g_{ds} we now have

$$g_{ds} = \frac{\sigma WH}{L} = \sigma \frac{WH_0}{L} \left(1 - \frac{\sqrt{V_{GS} + \phi}}{\sqrt{V_P + \phi}} \right)$$

When $V_{GS} = 0$, the channel is fully open or "on," and

$$g_{ds} = g_{ds}(\text{on}) = \sigma \frac{WH_0}{L} \left(1 - \frac{\sqrt{\phi}}{\sqrt{V_P + \phi}} \right)$$

AT&T Bell Laboratories

This photograph taken in 1948 is of the three Bell Labs physicist-engineers, John Bardeen, William Shockley, and Walter Brattain, who invented the first transistor. (Photo courtesy of AT&T Bell Laboratories.)

Their work built on the research of many before them, and much had to be done before the transistor and the solid-state devices that followed could become practical engineering tools, but in retrospect it is clear that the transistor gave the engineer the key to a whole new electronic world. (Courtesy of the IEEE Center for the History of Electrical Engineering.)

The drain-to-source conductance can now be expressed as

$$g_{ds} = g_{ds}(\text{on}) \frac{1 - \left(\sqrt{V_{GS} + \phi} \Big/ \sqrt{V_P + \phi}\right)}{1 - \left(\sqrt{\phi} \Big/ \sqrt{V_P + \phi}\right)}$$

The reciprocal quantity is the drain-to-source resistance r_{ds} as given by $r_{ds} = 1/g_{ds}$, and $r_{ds}(\text{on}) = 1/g_{ds}(\text{on})$, so

$$r_{ds} = r_{ds}(\text{on}) \frac{1 - \left(\sqrt{\phi} \Big/ \sqrt{V_P + \phi}\right)}{1 - \left(\sqrt{V_{GS} + \phi} \Big/ \sqrt{V_P + \phi}\right)}$$

As $V_{GS} \rightarrow 0$, $r_{ds} \rightarrow r_{ds}(\text{on})$, and as $V_{GS} \rightarrow V_P$, $r_{ds} \rightarrow \infty$. This latter condition corresponds to the channel being pinched off in its entirety all the way from source to drain. This is like having a big block of insulator (i.e., the depletion region) between source and drain. When $V_{GS} = 0$, r_{ds} is reduced to its minimum value of $r_{ds}(\text{on})$,

which for most JFETs is in the 20- to 400-Ω range. At the other extreme, when $V_{GS} > V_P$, the drain-to-source current I_{DS} is reduced to a very small value, generally down into the low nanoampere or even picoampere range. The corresponding value of r_{ds} is not really infinite but is very large, generally well up into the gigaohm (1000 MΩ) range. Thus by variation of V_{GS}, the drain-to-source resistance can be varied over a very wide range. As long as the gate-to-channel junction is reverse-biased, the gate current will be very small, generally down in the low nanoampere or even picoampere range, so the gate as a control electrode draws very little current. Since V_P is generally in the 2- to 5-V range for most JFETs, the V_{DS} values required to operate the JFET in the VVR range are generally <0.1 V. In Fig. 24.23 the VVR region of the JFET I_{DS} vs. V_{DS} characteristics is shown.

VVR Applications

Applications of VVRs include automatic gain control (AGC) circuits, electronic attenuators, electronically variable filters, and oscillator amplitude control circuits.

When using a JFET as a VVR, it is necessary to limit V_{DS} to values that are small compared to V_P to maintain good linearity. In addition V_{GS} should preferably not exceed 0.8 V_P for good linearity, control, and stability. This limitation corresponds to an r_{ds} resistance ratio of about 10:1. As V_{GS} approaches V_P, a small change in V_P can produce a large change in r_{ds}. Thus unit-to-unit variations in V_P as well as changes in V_P with temperature can result in large changes in r_{ds} as V_{GS} approaches V_P.

The drain-to-source resistance r_{ds} will have a temperature coefficient (TC) due to two causes: (1) the variation of the channel resistivity with temperature and (2) the temperature variation of V_P. The TC of the channel resistivity is positive, whereas the TC of V_P is negative due to the negative TC of the contact potential ϕ. The positive TC of the channel resistivity will contribute to a positive TC of r_{ds}. The negative TC of V_P will contribute to a negative TC of r_{ds}. At small values of V_{GS}, the dominant contribution to the TC is the positive TC of the channel resistivity, so r_{ds} will have a positive TC. As V_{GS} gets larger, the negative TC contribution of V_P becomes increasingly important, and there will be a value of V_{GS} at which the net TC of r_{ds} is zero, and above this value of V_{GS} the TC will be negative. The TC of r_{ds}(on) is typically +0.3%/°C for *n*-channel JFETs and +0.7%/°C for *p*-channel JFETs. For example, for a typical JFET with an r_{ds}(on) = 500 Ω at 25°C and V_P = 2.6 V, the zero TC point will occur at V_{GS} = 2.0 V. Any JFET can be used as a VVR, although there are JFETs that are specifically made for this application.

A simple example of a VVR application is the electronic gain control circuit of Fig. 24.24. The voltage gain is given by $A_V = 1 + (R_F/r_{ds})$. If, for example, R_F = 19 kΩ and r_{ds}(on) = 1 kΩ, then the maximum gain will be $A_{V\max} = 1 + [R_F/r_{ds}(\text{on})] = 20$. As V_{GS} approaches V_P, r_{ds} will increase and become very large such that $r_{ds} >> R_F$, so that A_V will decrease to a minimum value of close to unity. Thus the gain can be varied over a 20:1 ratio. Note that $V_{DS} \cong V_{\text{in}}$, so to minimize distortion the input signal amplitude should be small compared to V_P.

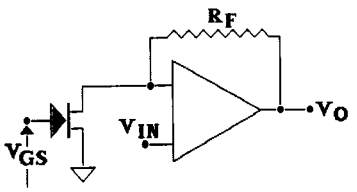


FIGURE 24.24 Electronic gain control.

Defining Terms

Active region: The region of JFET operation in which the channel is pinched off at the drain end but still open at the source end such that the drain-to-source current I_{DS} approximately saturates. The condition for this is that $|V_{GS}| < |V_P|$ and $|V_{DS}| > |V_P|$. The active region is also known as the saturated region.

Ohmic, nonsaturated, or triode region: The three terms all refer to the region of JFET operation in which a conducting channel exists all the way between source and drain. In this region the drain current varies with both V_{GS} and V_{DS}.

Drain saturation current, I_{DSS}: The drain-to-source current flow through the JFET under the conditions that $V_{GS} = 0$ and $|V_{DS}| > |V_P|$ such that the JFET is operating in the active or saturated region.

Pinch-off voltage, V_P: The voltage that when applied across the gate-to-channel *pn* junction will cause the conducting channel between drain and source to become pinched off. This is also represented as V_{GS} (off).

Related Topic

28.1 Large Signal Analysis

References

R. Mauro, *Engineering Electronics,* Englewood Cliffs, N.J.: Prentice-Hall, 1989, pp. 199–260.

J. Millman and A. Grabel, *Microelectronics,* 2nd ed., New York: McGraw-Hill, 1987, pp. 133–167, 425–429.

F. H. Mitchell, Jr. and F.H. Mitchell, Sr., *Introduction to Electronics Design,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1992, pp. 275–328.

C.J. Savant, M.S. Roden, and G.L. Carpenter, *Electronic Design,* 2nd ed., Menlo Park, Calif.: Benjamin-Cummings, 1991, pp. 171–208.

A.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., Philadelphia: Saunders, 1991, pp. 322–361.

24.2 Bipolar Transistors

Joseph Watson

Modern amplifiers abound in the form of *integrated circuits* (ICs), which contain transistors, diodes, and other structures diffused into single-crystal *dice.* As an introduction to these ICs, it is convenient to examine sing-transistor amplifiers, which in fact are also widely used in their own right as *discrete* circuits — and indeed much more complicated discrete signal-conditioning circuits are frequently found following sensors of various sorts.

There are two basic forms of transistor, the *bipolar* family and the *field-effect* family, and both appear in ICs. They differ in their modes of operation but may be incorporated into circuits in quite similar ways. To understand elementary circuits, there is no need to become too familiar with the physics of transistors, but some basic facts about their electrical properties must be known.

Consider the bipolar transistor, of which there are two types, *npn* and *pnp.* Electrically, they differ only in terms of current direction and voltage polarity. Figure 24.25(a) illustrates the idealized structure of an *npn* transistor, and diagram (b) implies that it corresponds to a pair of diodes with three leads. This representation does *not* convey sufficient information about the actual operation of the transistor, but it does make the point that the flow of conventional current (positive to negative) is easy from the *base* to the *emitter,* since it passes through a *forward-biased diode,* but difficult from the *collector* to the *base,* because flow is prevented by a *reverse-biased diode.*

Figure 24.25(c) gives the standard symbol for the *npn* transistor, and diagram (d) defines the direction of current flow and the voltage polarities observed when the device is in operation. Finally, diagram (e) shows that for the *pnp* transistor, all these directions are reversed and the polarities are inverted.

For a transistor, there is a main current flow between the collector and the emitter, and a very much smaller current flow between the base and the emitter. So, the following relations may be written:

$$I_E = I_C + I_B \tag{24.1}$$

(Note that the arrow on the transistor symbol defines the emitter and the direction of current flow—*out* for the *npn* device, and *in* for the *pnp.*) Also

$$I_C / I_B = h_{FE} \tag{24.2}$$

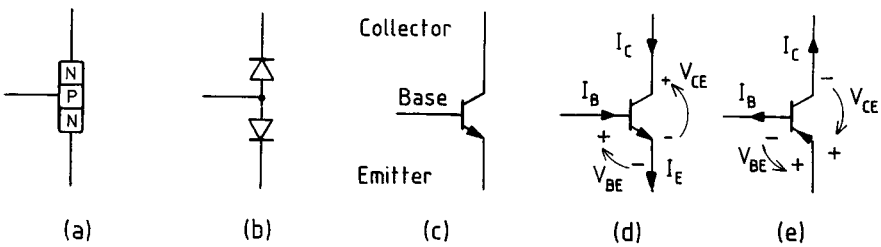


FIGURE 24.25 The bipolar transistor. (a) to (d) *npn* transistor; (e) *pnp* transistor.

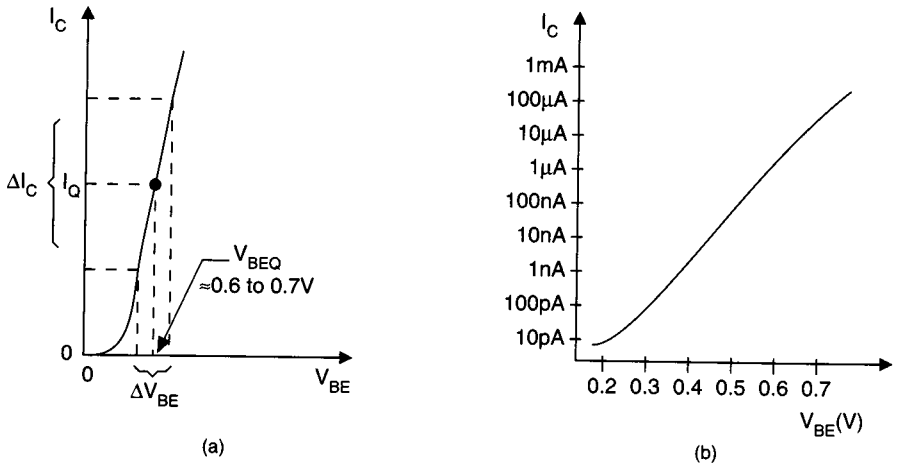


FIGURE 24.26 The transconductance curve for a transistor on (a) linear and (b) logarithmic axes.

Here, h_{FE} is called the *dc common-emitter current gain,* and because $I_C >> I_B$, then h_{FE} is large, typically 50 to 300. The implication of this may be seen immediately: if the small current I_B can be used to control the large current I_C, then the transistor may obviously be used as a current amplifier. [This is why Fig. 24.25(b) is inadequate—it completely neglects this all-important current-gain property of the transistor.] Furthermore, if a load resistance is connected into the collector circuit, it will become a voltage amplifier, too.

Unfortunately, h_{FE} is an ill-defined quantity and varies not only from transistor to transistor but also changes with temperature. The relationship between the base-emitter voltage V_{BE} and the collector current is much better defined and follows an exponential law closely over at least eight decades. This relationship is shown in both linear and logarithmic form in Fig. 24.26. Because the output current I_C is dependent upon the input voltage V_{BE}, the plot must be a transfer conductance or *transconductance* characteristic. The relevant law is

$$I_C = I_{ES}(e^{(q/kT)V_{BE}} - 1) \tag{24.3}$$

Here, I_{ES} is an extremely small leakage current internal to the transistor, q is the electronic charge, k is Boltzmann's constant, and T is the absolute temperature in kelvins. Usually, kT/q is called V_T and is about 26 mV at a room temperature of 25°C. This implies that for any value of V_{BE} over about 100 mV, then $\exp(V_{BE}/V_T) >> 1$, and for all normal operating conditions, Eq. (24.3) reduces to

$$I_C = I_{ES}e^{V_{BE}/V_T} \qquad \text{for } V_{BE} > 100 \text{ mV} \tag{24.4}$$

The term "normal operating conditions" is easily interpreted from Fig. 24.26(a), which shows that when V_{BE} has reached about 0.6 to 0.7 V, any small fluctuations in its value cause major fluctuations in I_C. This situation is illustrated by the dashed lines enclosing ΔV_{BE} and ΔI_C, and it implies that to use the transistor as an amplifier, working values of V_{BE} and I_C must be established, after which signals may be regarded as fluctuations around these values.

Under these *quiescent, operating,* or *working* conditions,

$$I_C = I_Q \qquad \text{and} \qquad V_{CE} = V_Q$$

and methods of defining these quiescent or operating conditions are called *biasing.*

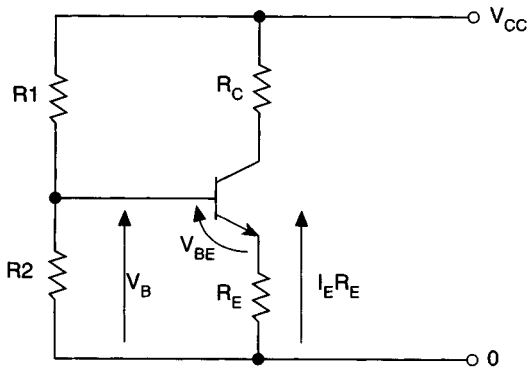


FIGURE 24.27 A transistor biasing circuit.

Biasing the Bipolar Transistor

A fairly obvious way to bias the transistor is to first establish a constant voltage V_B using a potential divider $R1$ and $R2$ as shown in the **biasing circuit** of Fig. 24.27. Here,

$$V_B \simeq \frac{V_{CC} R2}{R1 + R2}$$

if I_B is very small compared with the current through $R2$, which is usual. If it is not, this fact must be taken into account.

This voltage will be much greater than V_{BE} if a realistic power supply is used along with realistic values of $R1$ and $R2$. Hence, when the transistor is connected into the circuit, an emitter resistor must also be included so that

$$V_{BE} = V_B - I_E R_E \tag{24.5}$$

Now consider what happens when the power supply is connected. As V_B appears, a current I_B flows into the base and produces a much larger current $I_C = h_{FE} I_B$ in the collector. These currents add in the emitter to give

$$I_E = I_B + h_{FE} I_B = (1 + h_{FE}) I_B \simeq h_{FE} I_B \tag{24.6}$$

Clearly, I_E will build up until a fixed or quiescent value of base-emitter voltage V_{BEQ} appears. Should I_E try to build up further, V_{BE} will fall according to Eq. (24.5) and, hence, so will I_E. Conversely, should I_E not build up enough, V_{BE} will increase until it does so.

This is actually a case of current-derived negative feedback, and it successfully holds the collector current near the quiescent value I_Q. Furthermore, it does so in spite of different transistors with different values of h_{FE} being used and in spite of temperature variations. Actually, V_{BE} itself falls with temperature at about –2.2 mV/°C for constant I_C, and the circuit will compensate for this, too. The degree of success of the negative feedback in holding I_Q constant is called the *bias stability.*

This is one example of a **common-emitter** (CE) circuit, so-called because the emitter is the common terminal for both base and collector currents. The behavior of the transistor in such a circuit may be illustrated by superimposing a *load line* on the *output characteristics* of the transistor, as shown in Fig. 24.28.

If the collector current I_C is plotted against the collector-to-emitter voltage V_{CE}, a family of curves for various fixed values of V_{BE} or I_B results, as in Fig. 24.28. These curves show that as V_{CE} increases, I_C rises very rapidly and then turns over as it is limited by I_B. In the CE circuit, if I_B were reduced to zero, then I_C would also be

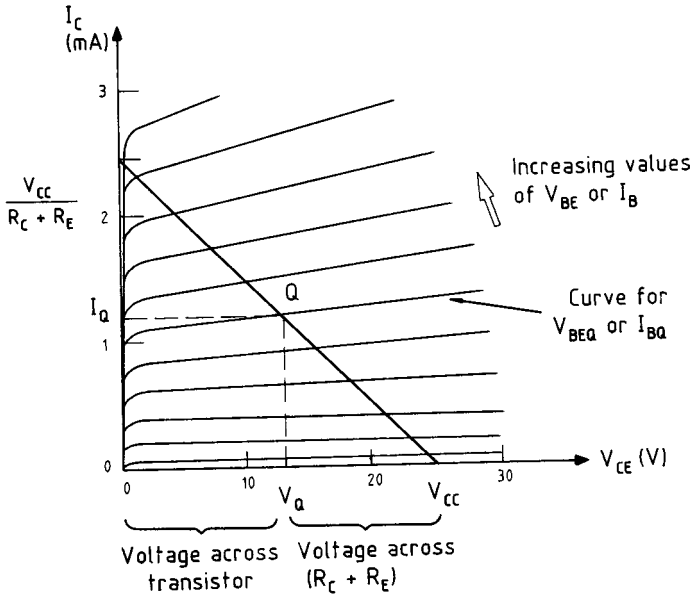


FIGURE 24.28 The load-line diagram.

zero (apart from a small leakage current I_{CE0}). Hence there would be no voltage drop in either R_C or R_E, and practically all of V_{CC} would appear across the transistor. That is, under *cut-off* conditions,

$$V_{CE} \rightarrow V_{CC} \qquad \text{for } I_B = 0 \tag{24.7}$$

Conversely, if I_B were large, I_C would be very large, almost all of V_{CC} would be dropped across $R_C + R_E$ and

$$I_C \rightarrow \frac{V_{CC}}{R_C + R_E} \qquad \text{for large } I_B \tag{24.8}$$

Actually, because the initial rise in I_C for the transistor is not quite vertical, there is always a small *saturation voltage* V_{CES} across the transistor under these conditions, where V_{CES} means the voltage across the transistor in the common-emitter mode when saturated. In this saturated condition $V_{CES} \simeq 0.3$ V for small silicon transistors. Both these conditions are shown in Fig. 24.28.

From the circuit of Fig. 24.27,

$$V_{CE} = V_{CC} - I_C(R_C + R_E) \tag{24.9a}$$

which may be rewritten as

$$I_C = -V_{CE}/(R_C + R_E) + V_{CC}/(R_C + R_E) \tag{24.9b}$$

This is the straight-line equation to the *dc load-line* (compare $y = mx + c$), showing that its slope is $-1/(R_C + R_E)$ and that it crosses the I_C axis at $V_{CC}/(R_C + R_E)$ as expected. The actual position of a point is determined by where this load line crosses the output characteristic in use, that is, by what value of V_{BE} or I_B is chosen. For example, the quiescent point for the transistor is where the load line crosses the output curve defined by $V_{BE} = V_{BEQ}$ (or $I_B = I_{BQ}$) to give $V_{CE} = V_Q$ and $I_C = I_Q$.

Note that because the transistor is *nonohmic* (that is, it does not obey Ohm's law), the voltage across it may only be determined by using the (ohmic) voltage drop across the resistors R_C and R_E according to Eq. (24.9). At the quiescent point this is

$$V_Q = V_{CC} - I_Q(R_C + R_E)$$

A design example will illustrate typical values involved with a small-transistor CE stage.

Example 1

A transistor is to be biased at a collector current of 1 mA when a 12-V power supply is applied. Using the circuit of Fig. 24.27, determine the values of $R1$, $R2$, and R_E if 3.4 V is to be dropped across R_E and if the current through $R2$ is to be 10 I_{BQ}. Assume that for the transistor used, $V_{BEQ} = 0.6$ V and $h_{FE} = 100$.

Solution. In this circuit $I_Q = 1$ mA $\simeq I_E$ (because $I_B << I_C$). Hence

$$R_E = \frac{V_{R_E}}{I_Q} = \frac{3.4}{1} = 3.4\ \text{k}\Omega$$

Also, $V_B = V_{RE} + V_{BE} = 3.4 + 0.6 = 4$ V. This gives

$$R2 = \frac{V_B}{10\ I_{BQ}}$$

where $I_{BQ} = I_Q/h_{FE} = 1/100 = 0.01$ mA, so

$$R2 = \frac{4}{10 \times 0.01} = 40\ \text{k}\Omega$$

Now $V_{R1} = V_{CC} - V_B = 12 - 4 = 8$ V, and the current through $R1$ is 10 $I_{BQ} + I_{BQ} = 11\ I_{BQ}$, so

$$R1 = \frac{V_{R1}}{I_{R1}} = \frac{8}{11 \times 0.01} = 72.7\ \text{k}\Omega$$

In the above design example, the base current I_{BQ} has been included in the current passing through $R1$. Had this not been done, $R1$ would have worked out at 80 kΩ. Usually, this difference is not very important because *discrete* (or individual) resistors are available only in a series of nominal values, and each of these is subject to a *tolerance,* including 10, 5, 2, and 1%.

In the present case, the following (5%) values could reasonably be chosen:

$$R_E = 3.3\ \text{k}\Omega \qquad R1 = 75\ \text{k}\Omega \qquad R2 = 39\ \text{k}\Omega$$

All this means that I_Q cannot be predetermined very accurately, but the circuit nevertheless settles down to a value close to the chosen one, and, most importantly, stays there almost irrespective of the transistor used and the ambient temperature encountered.

Having biased the transistor into an operating condition, it is possible to consider *small-signal operation.*

Small-Signal Operation

In the biasing circuit of Fig. 24.27, the collector resistor R_C had no discernible function, because it is simply the load resistor across which the signal output voltage is developed. However, it was included because it also drops a voltage due to the bias current flowing through it. This means that its value must not be so large that it robs the transistor of adequate operating voltage; that is, it must not be responsible for moving the operating point too far to the left in Fig. 24.28.

If the chosen bias current and voltage are I_Q and V_Q, then small signals are actually only fluctuations in these bias (or average) values that can be separated from them using coupling capacitors.

To inject an input signal to the base, causing V_{BE} and I_B to fluctuate by v_{be} and i_b, a signal source must be connected between the base and the common or zero line (also usually called ground or earth whether it is actually connected to ground or not!). However, most signal sources present a resistive path through themselves, which would shunt $R2$ and so change, or even destroy, the bias conditions. Hence, a coupling capacitor C_c must be included, as shown in Fig. 24.29, in series with a signal source represented by a Thévenin equivalent.

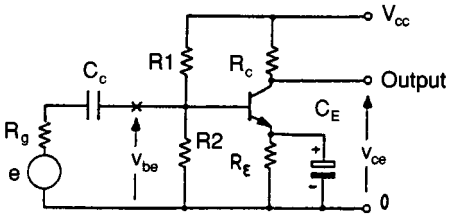


FIGURE 24.29 A complete common-emitter stage.

The emitter resistor R_E was included for biasing reasons (although there are other bias circuits that omit it), but for signal amplification purposes it must be shunted by a high-value capacitor C_E so that the signal current can flow down to ground without producing a signal voltage drop leading to negative feedback (as did the bias current). The value of C_E must be much greater than is apparent at first sight, and this point will be developed later; for the present, it will be assumed that it is large enough to constitute a short circuit at all the signal frequencies of interest. So, for ac signals R_E is short-circuited and only R_C acts as a load. This implies that a *signal* or *ac load line* comes into operation with a slope of $-1/R_C$, as shown in Fig. 24.30.

The ways in which the small-signal quantities fluctuate may now be examined. If v_{be} goes positive, this actually means that V_{BE} increases a little. This in turn implies that I_C increases by an amount i_c, so the voltage drop in R_C increases by v_{ce}. Keeping in mind that the top of R_C is held at a constant voltage, this means that the voltage at the bottom of R_C must fall by v_{ce}. This very important point shows that because v_{ce} falls as v_{be} rises, there is 180° phase shift through the stage. That is, the CE stage is an *inverting voltage amplifier.* However, because i_c increases into the collector as i_b increases into the base, it is also a *noninverting current amplifier.*

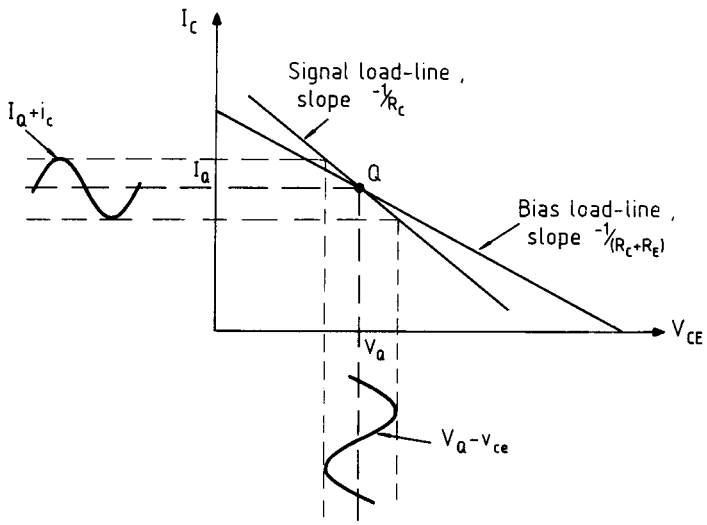


FIGURE 24.30 The signal or ac load line.

Now consider the amount by which v_{ce} changes with v_{be}, which is the *terminal voltage gain* of the stage. In Fig. 24.26, the slope of the transconductance curve at any point defines by how much I_C changes with a fluctuation in V_{BE}. That is, it gives the ratio i_c/v_{be} at any operating point Q. Equation (24.4) is

$$I_C = I_{ES} e^{V_{BE}/V_T}$$

so that

$$\frac{dI_C}{dV_{BE}} = \frac{1}{V_T} I_{ES} e^{V_{BE}/V_T}$$

or

$$\frac{i_c}{v_{be}} = \frac{I_C}{V_T} = g_m = \text{the transconductance} \tag{24.10}$$

Now the signal output voltage is

$$v_{ce} \simeq -i_c R_C$$

(Here, the approximation sign is because the collector-emitter path within the transistor does present a large resistance r_{ce} through which a very small part of i_c flows.)

The terminal voltage gain is therefore

$$A_v = \frac{v_{ce}}{v_{be}} \simeq \frac{-i_c R_C}{v_{be}} = -g_m R_C \tag{24.11}$$

where the negative sign implies signal inversion.

In practice, $V_T \simeq 26$ mV at room temperature, as has been mentioned, and this leads to a very simple numerical approximation. From Eq. (24.10) and using $I_C = I_Q$,

$$g_m = \frac{I_Q}{V_T} \simeq \frac{I_Q}{0.026} \simeq 39\, I_Q \qquad \text{mA/V}$$

if I_Q is in mA and at room temperature. This shows that irrespective of the transistor used, the transconductance may be approximated knowing only the quiescent collector current.

The magnitude and phase relationships between v_{ce} and i_c can easily be seen by including them on the signal load-line diagram as shown in Fig. 24.30, where the output characteristics of the transistor have been omitted for clarity. Sinusoidal output signals have been inserted, and either may be obtained from the other by following the signal load-line locus.

Now consider the small-signal current gain. Because the value of h_{FE} is not quite linear on the I_C/I_B graph, its slope too must be used for small-signal work. However, the departure from linearity is not great over normal working conditions, and the small-signal value h_{fe} is usually quite close to that of h_{FE}. Hence,

$$A_i = \frac{i_c}{i_b} \simeq h_{fe} \tag{24.12}$$

The small-signal or incremental input resistance to the base itself (to the right of point *X* in Fig. 24.29) may now be found:

$$R_{\text{in}} = \frac{v_{be}}{i_b} = \frac{v_{be}}{i_c}\frac{i_c}{i_b} \simeq \frac{h_{fe}}{g_m} \tag{24.13}$$

Three of the four main (midfrequency) parameters for the CE stage have now been derived, all from a rather primitive understanding of the transistor itself. The fourth, R_{out}, is the dynamic, incremental, or small-signal resistance of the transistor from collector to emitter, which is the slope of the output characteristic at the working point r_{ce}. Being associated with a reverse-biased (CB) junction, this is high—typically about 0.5 MΩ—so that the transistor acts as a current source feeding a comparatively low load resistance R_C. Summarizing, at mid frequency,

$$A_i \simeq h_{fe} \qquad A_v \simeq -g_m R_C \qquad R_{\text{in}} \simeq \frac{h_{fe}}{g_m} \qquad R_{\text{out}} \simeq r_{ce}$$

Example 2

Using the biasing values for *R*1, *R*2, and R_E already obtained in Example 1, calculate the value of R_C to give a terminal voltage gain of –150. Then determine the input resistance R_{in} if h_{fe} for the transistor is 10% higher than h_{FE}.

Solution. Because $I_Q = 1$ mA, $g_m \simeq 39 \times 1 = 39$ mA/V. Hence $A_v \simeq -g_m R_C$ or $-150 \simeq -39\, R_C$, giving

$$R_C = 150/39 \simeq 3.9\ \text{k}\Omega$$

(*Note:* This value *must* be checked to determine that it is reasonable insofar as biasing is concerned. In this case, it will drop $I_Q R_C = 1 \times 3.9 = 3.9$ V. Because $V_{RE} = 3.4$ V, this leaves 12 – 3.9 – 3.4 = 4.7 V across the transistor, which is reasonable.)

Finally,

$$R_{\text{in}} \simeq \frac{h_{fe}}{g_m} = \frac{110}{39} = 2.8\ \text{k}\Omega$$

A Small-Signal Equivalent Circuit

The conclusions reached above regarding the performance of the bipolar transistor are sufficient for the development of a basic equivalent circuit, or model, relevant *only* to small-signal operation. Taking the operating CE amplifier, this may be done by first "looking into" the base, shown as *b* in Fig. 24.31. Between this point and the actual active part of the base region *b′*, it is reasonable to suppose that the intervening (inactive) semiconductor material will present a small resistance $r_{bb'}$. This is called the *base spreading resistance*, and it is also shown in Fig. 24.31.

From *b′* to the emitter *e*, there will be a dynamic or incremental resistance given by

$$r_{b'e} = \frac{v_{b'e}}{i_b} = \frac{v_{b'e}}{i_c}\frac{i_c}{i_b} = \frac{h_{fe}}{g_m} \tag{24.14}$$

so that the full resistance from the base to the emitter must be

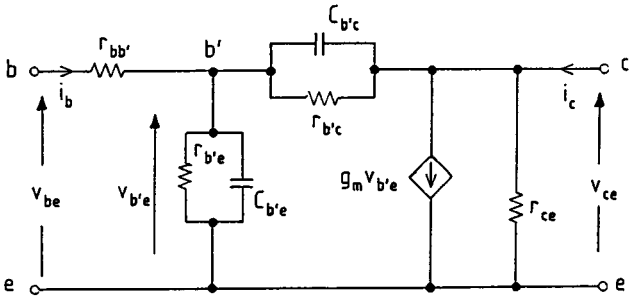


FIGURE 24.31 The hybrid-π small-signal transistor equivalent circuit or model.

$$R_{\text{in}} = r_{bb'} + r_{b'e} = r_{bb'} + \frac{h_{fe}}{g_m} \simeq \frac{h_{fe}}{g_m} \tag{24.15}$$

because $r_{bb'}$ is only about 10 to 100 Ω, which is small compared with $r_{b'e}$, this being several kilohms (as shown by the last example). It will now be understood why Eq. (24.13) gave $R_{\text{in}} \simeq h_{fe}/g_m$.

The reverse-biased junction that exists from b' to the collector ensures that the associated dynamic resistance $r_{b'c}$ will be very large indeed, which is fortunate, otherwise signal feedback from the output to the input would modify the gain characteristics of the amplifier. Typically, $r_{b'c}$ will be some tens of megohms.

However, because of transistor action, the dynamic resistance from collector to emitter, r_{ce}, will be smaller than $r_{b'c}$ and will typically be below a megohm. This "transistor action" may be represented by a current source from collector to emitter that is dependent upon either i_b or $v_{b'e}$. That is, it will be either $h_{fe}i_b$ or $g_m v_{b'e}$. The latter leads to the well-known hybrid-π model, and it is this which is shown in Fig. 24.31.

Where junctions or interfaces of any sort exist, there will always be distributed capacitances associated with them, and to make these easy to handle analytically, they may be "lumped" into single capacitances. In the present context, two lumped capacitances have been incorporated into the hybrid-π model, $C_{b'e}$ from base to emitter and $C_{b'c}$ from base to collector, respectively. These now complete the model, and it will be appreciated that they make it possible to analyze high-frequency performance. Typically, $C_{b'e}$ will be a few picofarads and will always be larger than $C_{b'c}$.

Figure 24.31 is the hybrid-π small-signal, dynamic, or incremental model for a bipolar transistor, and when external components are added and simplifications made, it makes possible the determination of the performance of an amplifier using that transistor not only at midfrequencies but at high and low frequencies, too.

Low-Frequency Performance

In Fig. 24.32 both a source and a load have been added to the hybrid-π equivalent circuit to model the complete CE stage of Fig. 24.29. Here, both $C_{b'e}$ and $C_{b'c}$ have been omitted because they are too small to affect the low-frequency performance, as has $r_{b'c}$ because it is large and so neither loads the source significantly compared to $r_{bb'} + r_{b'e}$ nor applies much feedback.

The signal source has been represented by a Thévenin equivalent that applies a signal via a coupling capacitor C_c. Note that this signal source has been returned to the emitter, which implies that the emitter resistor bypass capacitor C_E has been treated as a short circuit at all signal frequencies for the purposes of this analysis.

Because the top of biasing resistor $R1$ (Fig. 24.29) is taken to ground via the power supply insofar as the signal is concerned, it appears in parallel with $R2$, and the emitter is also grounded to the signal via C_E. That is, a composite biasing resistance to ground R_B appears:

$$R_B = \frac{R1 \cdot R2}{R1 + R2}$$

Finally, the collector load is taken to ground via the power supply and hence to the emitter via C_E.

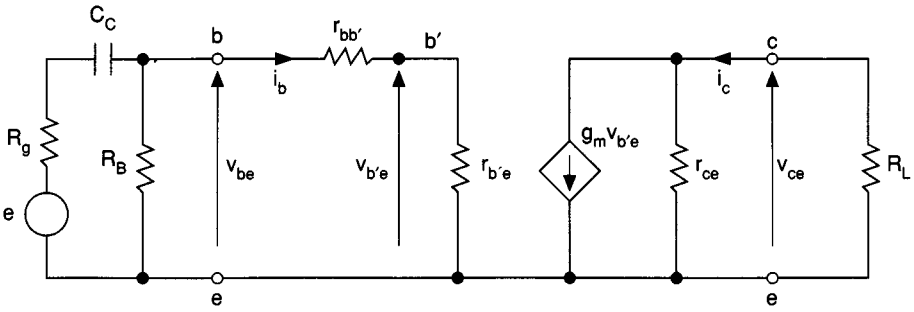


FIGURE 24.32 The loaded hybrid-π model for low frequencies.

Figure 24.32 shows that v_{be} is amplified independently of frequency, so the terminal voltage gain A_v may easily be determined:

$$A_v = \frac{v_{ce}}{v_{be}} = \frac{-i_c R_L}{v_{be}}$$

Now

$$v_{be} = \frac{v_{b'e}(r_{bb'} + r_{b'e})}{r_{b'e}} \simeq v_{b'e} \quad \text{and} \quad i_c = \frac{g_m v_{b'e} r_{ce}}{r_{ce} + R_L} \simeq g_m v_{b'e}$$

because $r_{b'e} >> r_{bb'}$ and $r_{ce} >> R_L$. So,

$$A_v \simeq -g_m R_L$$

which is as expected.

The model shows that v_{be} is amplified independently of frequency because there are no capacitances to its right, so an analysis of low-frequency response devolves down to determining v_{be} in terms of e. Here, part of e will appear across the capacitive reactance X_{Cc}, and the remainder is v_{be}. So, to make the concept of reactance valid, a sinusoidal signal E must be postulated, giving a sinusoidal value for $v_{be} = V_{be}$.

At midfrequencies, where the reactance of C_c is small, the signal input voltage is

$$V_{be}(f_m) = \frac{E \cdot R_{BP}}{R_g + R_{BP}} \tag{24.16}$$

where $R_{BP} = R_B R_{\text{in}}/(R_B + R_{\text{in}})$ and $R_{\text{in}} = r_{bb'} + r_{b'e}$ as before.

At low frequencies, where the reactance of C_c is significant,

$$V_{be}(f_{\text{low}}) = \frac{E \cdot R_{BP}}{\sqrt{(R_g + R_{BP})^2 + X_{C_c}^2}} \tag{24.17}$$

Dividing (24.16) by (24.17) gives

$$\frac{V_{be}(f_m)}{V_{be}(f_{\text{low}})} = \frac{\sqrt{(R_g + R_{BP})^2 + X_{C_c}^2}}{R_g + R_{BP}}$$

There will be a frequency f_L at which $|X_{Cc}| = R_g + R_{BP}$ given by

$$\frac{1}{2\pi f_L C_c} = R_g + R_{BP} \quad \text{or} \quad f_L = \frac{1}{2\pi C_c(R_g + R_{BP})} \tag{24.18}$$

At this frequency, $V_{be}(f_m)/V_{be}(f_L) = \sqrt{2}$ or $V_{be}(f_L)$ is 3 dB lower than $V_{be}(f_m)$.

Example 3

Using the circuit components of the previous examples along with a signal source having an internal resistance of $R_g = 5$ kΩ, find the value of a coupling capacitor that will define a low-frequency –3dB point at 42 Hz.

Solution. Using Eq. (24.18),

$$C_c = \frac{1}{2\pi(R_g + R_{BP})f_L}$$

where $R_{BP} = R1||R2||R_{\text{in}} = 75||39||2.8 = 2.5$ kΩ. That is,

$$C = \frac{10^6}{2\pi(5000 + 2500)(42)} \simeq 0.5\ \mu\text{F}$$

Since a single RC time constant is involved, the voltage gain of the CE stage will appear to fall at 6 dB/octave as the frequency is reduced because more and more of the signal is dropped across C_c. However, even if C_E is very large, it too will contribute to a fall in gain as it allows more and more of the output signal to be dropped across the $R_E||X_{CE}$ combination, this being applied also to the input loop, resulting in negative feedback. So, at very low frequencies, the gain roll-off will tend to 12 dB/octave. The question therefore arises of how large C_E should be, and this can be conveniently answered by considering a second basic form of transistor connection as follows.

The Emitter-Follower or Common-Collector (CC) Circuit

Suppose that R_C is short-circuited in the circuit of Fig. 24.29. This will not affect the biasing because the collector voltage may take any value (the output characteristic is nearly horizontal, as seen in Fig. 24.28). However, the small-signal output voltage ceases to exist because there is now no load resistor across which it can be developed, though the output current i_c will continue to flow as before.

If now C_E is removed, i_c flows entirely through R_E and develops a voltage which can be observed at the emitter $i_e R_E$ ($\simeq i_c R_E$). Consider the magnitude of this voltage. Figure 24.26(a) shows that for a normally operating transistor, the signal component of the base-emitter voltage ΔV_{BE} (or v_{be}) is very small indeed, whereas the constant component needed for biasing is normally about 0.6 to 0.7 V. That is, $v_{be} << V_{BE}$. This implies that the emitter voltage must always follow the base voltage but at a dc level about 0.6 to 0.7 V below it. So, if an output signal is taken from the emitter, it is almost the same as the input signal at the base. In other words, *the voltage gain of an* ***emitter follower*** *is almost unity.*

If this is the case, what is the use of the emitter follower? The answer is that because the signal *current gain* is unchanged at $i_e/i_b = (h_{fe} + 1) \simeq h_{fe}$, then the power gain must also be about h_{fe}. This means in turn that the output resistance must be the resistance "looking into" the transistor from the emitter, divided by h_{fe}. If the parallel combination of R_g and the bias resistors is R_G, then

$$R_{\text{out}(CC)} = \frac{R_G + r_{bb'} + r_{b'e}}{h_{fe}} \tag{24.19}$$

where $R_G = R_g||R1||R2$ (or $R_g||R_B$).

If a voltage generator with zero internal resistance ($R_g = 0$) were applied to the input, then this would become

$$R_{\text{out}(CC)} = \frac{r_{bb'} + r_{b'e}}{h_{fe}}$$

and if $r_{b'e} >> r_{bb'}$ (which is usual), then

$$R_{\text{out}(CC)} \simeq \frac{r_{b'e}}{h_{fe}} = \frac{1}{g_m} \tag{24.20}$$

Consider the numerical implications of this: if $I_C = 1$ mA, then $g_m \simeq 39$ mA/V (at room temperature), so $1/g_m \simeq 26\ \Omega$, which is a very low output resistance indeed. In fact, though it appears in parallel with R_E, it is unlikely that R_E will make any significant contribution because it is usually hundreds or thousands of ohms.

Example 4

Using the same bias resistors as for the CE examples, find the output resistance at the emitter of a CC stage.

Solution. The parallel resistances to the left of the base are

$$R_G = R_g||R1||R2 = 5||75||39 \approx 4.2\ \text{k}\Omega$$

Using Eq. (24.19),

$$R_{\text{out}} \approx \frac{R_G + r_{b'e}}{h_{fe}} = \frac{R_G}{h_{fe}} + \frac{1}{g_m} \qquad \text{(neglecting } r_{bb'}\text{)}$$

where $g_m \approx 39 I_C$, $I_Q = 1$ mA, and $h_{fe} = 110$, so

$$R_{\text{out}(CC)} \approx \frac{4200}{110} + \frac{1000}{39} \approx 63.8\ \Omega$$

From values like this, it is clear that the output of an emitter follower can be thought of as a good practical dependent voltage source of very low internal resistance.

The converse is also true: the input at the base presents a high resistance. This is simply because whereas much the same signal voltage appears at the base as at the emitter, the base signal current i_b is smaller than the emitter signal current i_e by a factor of $(h_{fe} + 1) \simeq h_{fe}$. Hence, the apparent resistance at the base must be at least $h_{fe} R_E$. To this must be added $r_{bb'} + r_{b'e}$ so that

$$R_{\text{in}(CC)} \simeq r_{bb'} + r_{b'e} + h_{fe} R_E \tag{24.21a}$$

Now h_{fe} is rarely less than about 100, so $h_{fe} R_E$ is usually predominant and

$$R_{\text{in}(CC)} \simeq h_{fe} R_E \tag{24.21b}$$

The emitter-follower circuit is therefore a *buffer stage* because it can accept a signal at a high resistance level without significant attenuation and reproduce it at a low resistance level and with *no phase shift* (except at high frequencies).

In this configuration, the unbypassed emitter resistor R_E is obviously in series with the input circuit as well as the output circuit. Hence, it is actually a feedback resistor and so may be given the alternative symbol R_F, as in Fig. 24.33. Because all the output signal voltage is fed back in series with the input, this represents 100% voltage-derived series negative feedback.

The hybrid-π model for the bipolar transistor may now be inserted into the emitter-follower circuit of Fig. 24.33, resulting in Fig. 24.34, from which the four midfrequency parameters may be obtained. As an example of the procedures involved, consider the derivation of the voltage gain expression.

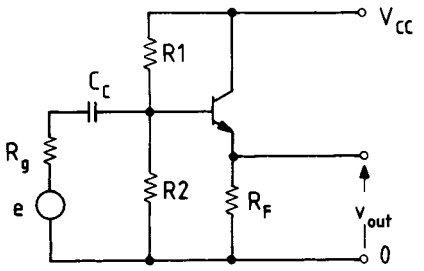


FIGURE 24.33 The emitter follower (or CC stage).

Summing signal currents at the emitter,

$$v_{\text{out}}\left(\frac{1}{R_F}+\frac{1}{r_{ce}}\right)=v_{b'e}\left(\frac{1}{r_{b'e}}+g_m\right)$$

Now $1/r_{ce} << 1/R_F$ and so may be neglected, and $v_{b'e} = v_{\text{in}} - v_{\text{out}}$, so

$$v_{\text{out}}\left(\frac{1}{R_F}\right)=(v_{\text{in}}-v_{\text{out}})\left(\frac{1}{r_{b'e}}+g_m\right)$$

or

$$v_{\text{out}}\left(\frac{1}{R_F}+\frac{1}{r_{b'e}}+g_m\right)=v_{\text{in}}\left(\frac{1}{r_{b'e}}+g_m\right)$$

giving

$$\begin{aligned}A_{v(CC)}&=\frac{v_{\text{out}}}{v_{\text{in}}}=\frac{1/r_{b'e}+g_m}{1/r_{b'e}+g_m+1/R_F}=\frac{1+g_m r_{b'e}}{1+g_m r_{b'e}+r_{b'e}/R_F}\\&\simeq\frac{g_m r_{b'e}}{g_m r_{b'e}+r_{b'e}/R_F}=\frac{g_m R_F}{g_m R_F+1}\end{aligned}\tag{24.22}$$

which is a little less than unity as expected.

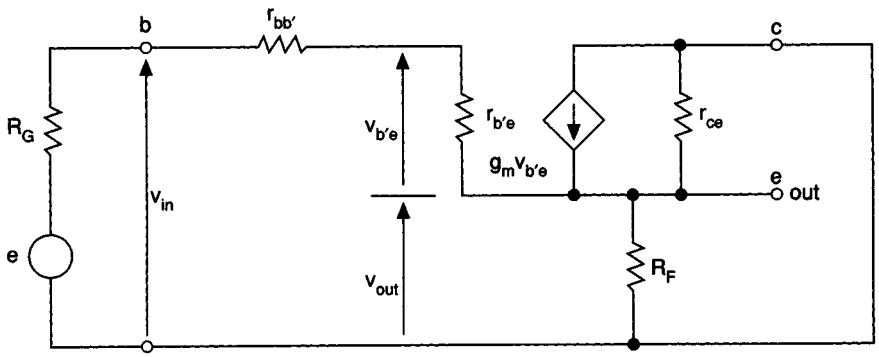


FIGURE 24.34 An emitter-follower equivalent circuit for low frequencies.

Similar derivations based on the equivalent circuit of Fig. 24.34 result in the other three basic midband operating parameters for the emitter follower, and all may be listed:

$$A_{i(CC)} \simeq h_{fe} \quad A_{v(CC)} \rightarrow +1$$
$$R_{\text{in}(CC)} \simeq r_{bb'} + r_{b'e} + h_{fe}R_F \simeq h_{fe}R_F$$

and

$$R_{\text{out}(CC)} \simeq \frac{R_G + r_{bb'} + r_{b'e}}{h_{fe}} \,||\, R_F \simeq \frac{R_G + r_{bb'} + r_{b'e}}{h_{fe}}$$
$$\simeq \frac{1}{g_m} \quad \text{if } R_g \rightarrow 0 \text{ and } r_{bb'} << r_{b'e}$$

The Common-Emitter Bypass Capacitor C_E

In a CE circuit such as that of Fig. 24.29, suppose C_c is large so that the low-frequency –3-dB point f_L is defined only by the parallel combination of the resistance at the emitter and C_E. It will now be seen why the emitter-follower work is relevant: the resistance appearing at the emitter of the CE stage is the same as the output resistance of the emitter-follower stage, and this will now appear in parallel with R_E. If this parallel resistance is renamed R_{emitter}, then, neglecting $r_{bb'}$,

$$R_{\text{emitter}} = R_{\text{out}(CC)} \,||\, R_E$$
$$\simeq \frac{R_G + r_{b'e}}{h_{fe}} \,||\, R_E$$
$$\simeq \frac{R_G + r_{b'e}}{h_{fe}}$$

and if C_E were to define f_L, then

$$f_L = \frac{1}{2\pi R_{\text{emitter}} C_E} \tag{24.23a}$$

For design purposes, C_E can be extracted for any given value of f_L:

$$C_E = \frac{1}{2\pi R_{\text{emitter}} f_L} \tag{24.23b}$$

Example 5

In Example 4, let C_c be large so that only C_E defines f_L at 42 Hz, and find the value of C_E.

Solution. In the emitter-follower example, where $R_g = 5$ kΩ, $R_{\text{out}(CC)}$ was found to be 63.8 Ω, and this is the same as R_{emitter} in the present case. Therefore,

$$C_E = \frac{10^6}{2\pi 63.8 \times 42} \simeq 60\ \mu\text{F}$$

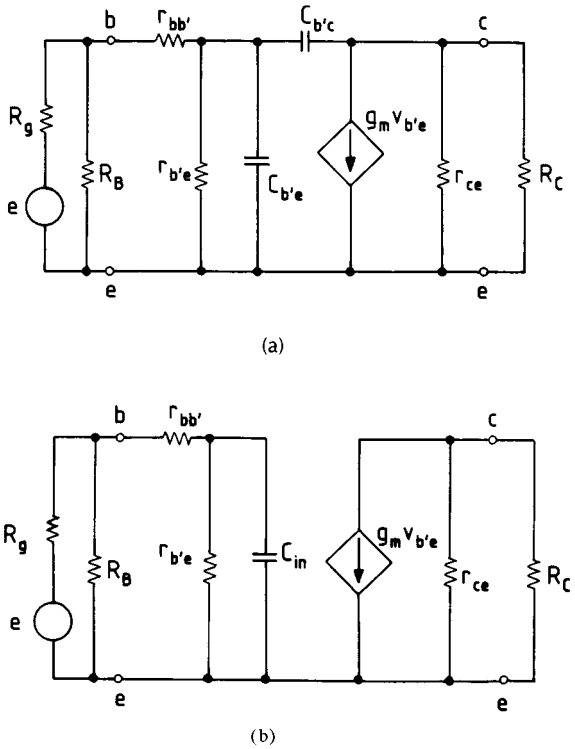


FIGURE 24.35 (a) The high-frequency hybrid-π model and (b) its simplification.

This is the value of C_E that would define f_L if C_c were large. However, if C_E is to act as a short circuit at this frequency, so allowing C_c to define f_L, then its value would have to be one or two orders of magnitude greater, that is, 600 to 6000 μF.

Summarizing, three possibilities exist:

1. If C_E is very large, C_c defines f_L and a 6-dB/octave roll-off results.
2. If C_c is large, C_E defines f_L and again a 6-dB/octave roll-off results.
3. If both C_c and C_E act together, a 12-dB/octave roll-off results.

In point of fact, at frequencies much less than f_L, both conditions (1) and (2) eventually produce 12-dB/octave roll-offs as the alternate "large" capacitors come into play at very low frequencies, but since the amplifier will not still have a useful gain at such frequencies, this is of little importance.

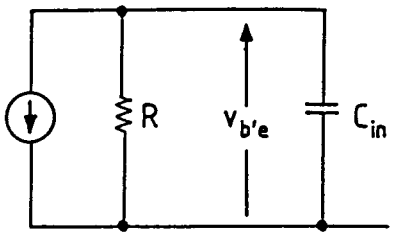
High-Frequency Response

Unlike the low-frequency response situation, the high-frequency response is governed by the small distributed capacitances inside the transistor structure, and these have been lumped together in the hybrid-π model of Fig. 24.31 as $C_{b'e}$ and $C_{b'c}$. At high frequencies, $r_{b'c}$ may be neglected in comparison with the reactance of $C_{b'c}$, so the model may be simplified as in Fig. 24.35(a). From this it will be seen that $C_{b'c}$ is a capacitance which appears from the output to the input so that it may be converted by the Miller Effect into a capacitance at the input of value:

$$C_{b'c}\,(1 - A_v) = C_{b'c}\,(1 + g_m R_C)$$

This will now add to $C_{b'e}$ to give C_{in}:

$$C_{\text{in}} = C_{b'e} + C_{b'c}\,(1 + g_m R_C) \tag{24.24}$$



$$R = \left(R_G + r_{bb'}\right) \| r_{b'e}$$

$$C_{in} = C_{b'e} + C_{b'c}\left(1 + g_m R_L\right)$$

FIGURE 24.36 Simplification of the input part of the high-frequency hybrid-π model.

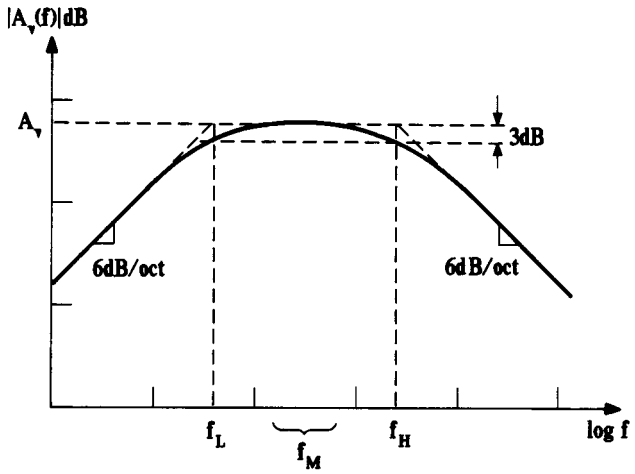


FIGURE 24.37 The complete frequency response.

This simplification is shown in Fig. 24.35(b), where C_{in} is seen to be shunted by the input parts of the model. These input parts may be reduced by sequential use of Thévenin–Norton transformations to result in Fig. 24.36, which is a simple parallel RC circuit driven by a current source. The actual value of this current source is immaterial—what matters is that the input signal to be amplified, $v_{b'e}$, will be progressively reduced as the frequency rises and the reactance of C_{in} falls.

Using a sinusoidal source, $V_{b'e}$ will be 3 dB down when $R = |X_{C_{in}}|$, which gives

$$R = \frac{1}{2\pi f_H C_{in}} \quad \text{or} \quad f_H = \frac{1}{2\pi R C_{in}} \tag{24.25}$$

where $R = (R_G + r_{bb'}) \| r_{b'e}$ from the circuit reduction.

Complete Response

Now that both the low- and high-frequency roll-offs have been related to single time constants (except when C_c and C_E act together), it is clear that the complete frequency response will look like Fig. 24.37, where the midband voltage gain is $A_v = -g_m R_C$.

Design Comments

The design of a simple single-transistor amplifier stage has now been covered in terms of both biasing and small-signal performance. These two concepts have been kept separate, but it will have been noticed that they

are bridged by the transconductance, because $g_m = (q/kT)I_Q$ ($\simeq 39I_Q$ at room temperature). That is, when I_Q has been determined, then the small-signal performance follows from expressions involving g_m.

In fact, once the quiescent voltage across the load resistor of a CE stage has been determined, the voltage gain follows from this irrespective of the values of I_Q and R_C.

If the quiescent voltage at the collector is V_{out}, then in dc biasing terms,

$$V_{RC} = I_Q R_C = (V_{CC} - V_{\text{out}})$$

and in small-signal terms,

$$A_v = -g_m R_C \cong -39 I_Q R_C \qquad \text{(at 25°C)}$$

$$= -39(V_{CC} - V_{\text{out}})$$

Thus, g_m really does act as a bridge between the bias and the small-signal conditions for the bipolar transistor amplifier stage.

Unfortunately, however, there are serious problems with such a stage from a practical viewpoint. For example, it cannot amplify down to dc because of the existence of C_c, and if a larger gain is needed, the cascading of such stages will present problems of phase shift and hence feedback stability. Furthermore, it cannot be produced in IC form because of the incorporation of large capacitances and somewhat critical and high-valued resistors. This leads to a reevaluation of the basic tenets of circuit design, and these may be summed up as follows: circuit design using *discrete* components is largely concerned with voltage drops across resistors (as has been seen), but the design of ICs depends extensively on *currents* and *current sources and sinks.*

Integrated Circuits

Monolithic ICs are fabricated on single chips of silicon or *dice* (the singular being *die*). This means that the active and passive structures on the chips are manufactured all at the same time, so it is easy to ensure that a large number of such structures are identical, or bear some fixed ratio to one another, but it is more difficult to establish precise values for such sets of structures. For example, a set of transistors may all exhibit almost the same values of h_{FE}, but the actual numerical value of h_{FE} may be subject to wider tolerances. Similarly, many pairs of resistors may bear a ratio n:1 to each other, but the actual values of these resistors are more difficult to define. So, in IC design, it is very desirable to exploit the close similarity of devices (or close ratios) rather than depend upon their having predictable absolute values. This approach has led to two ubiquitous circuit configurations, both of which depend upon device similarity: the **long-tailed pair or difference amplifier** (often called the *differential amplifier*), and the **current mirror**. This section will treat both, and the former is best introduced by considering the **degenerate common-emitter** stage.

The Degenerate Common-Emitter Stage

Consider two CE stages which are identical in every respect but which have no emitter resistor bypass capacitors, as shown in Fig. 24.38. Also, notice that in these diagrams, two power supply rails have been used, a positive one at V^+_{CC} and a negative one at V^-_{CC}. The reason for this latter, negative, rail is that the bases may be operated via signal sources referred to a common line or ground. (If, for example, V^+_{CC} and V^-_{CC} are obtained from batteries as shown, then the common line is simply the junction of the two batteries, as is also shown.) The absence of capacitors now means that amplification down to dc is possible.

It is now very easy to find the quiescent collector currents I_Q, because from a dc bias point of view the bases are connected to ground via resistances R_g, which will be taken as having low values so that they drop negligibly small voltages. Hence,

$$I_Q = \frac{|V^-_{CC}| - V_{BE}}{R_F} \tag{24.26}$$

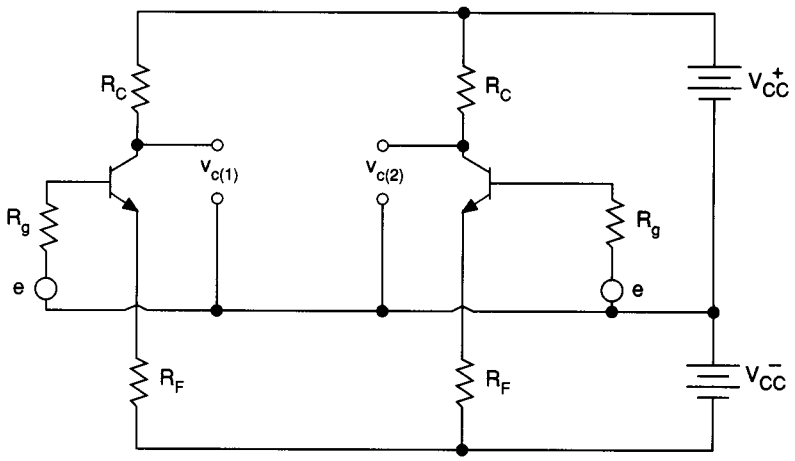


FIGURE 24.38 Two degenerate CE stages.

[For example, if industry-standard supplies of ±15 V are used, $V_{BE} = 0.6$ V, and for $R_F = 15$ kΩ, then $I_Q = (15 - 0.6)/15 = 0.96 \simeq 1$ mA.]

Now suppose that identical signals e are applied. At each collector, this will result in an output signal voltage v_c, where $v_c = -i_c R_C$. Also, at each emitter, the output signal voltage will be $v_e = i_e R_F \simeq i_c R_F$. That is,

$$\frac{v_c}{v_e} \simeq \frac{-i_c R_c}{i_c R_F} = -\frac{R_C}{R_F}$$

If the voltage gain from base to collector of a degenerate CE stage is $A_{v(dCE)}$ and the voltage gain from base to emitter is simply the emitter-follower gain $A_{v(CC)}$, then

$$v_c = A_{v(dCE)} e \qquad \text{and} \qquad v_e = A_{v(CC)} e$$

giving

$$\frac{A_{v(dCE)}}{A_{v(CC)}} \simeq -\frac{R_C}{R_F}$$

Now $A_{v(CC)}$ is known from Eq. (24.22) so that

$$A_{v(dCE)} \simeq -A_{v(CC)} \frac{R_C}{R_F} = -\frac{g_m R_C}{1 + g_m R_F} \simeq -\frac{R_C}{R_F} \tag{24.27}$$

Note that the input resistance to each base is as for the emitter-follower stage:

$$R_{\text{in}(dCE)} = r_{bb'} + r_{b'e} + h_{fe} R_F \simeq h_{fe} R_F$$

Now consider what happens if the emitters are connected together as in Fig. 24.39, where the two resistors R_F have now become R_X, where $R_X = \frac{1}{2} R_F$.

The two quiescent emitter currents now combine to give $I_X = 2I_E \simeq 2I_C$, and otherwise the circuit currents and voltages remain undisturbed. So, if the two input signals are identical, then the two output signals will also

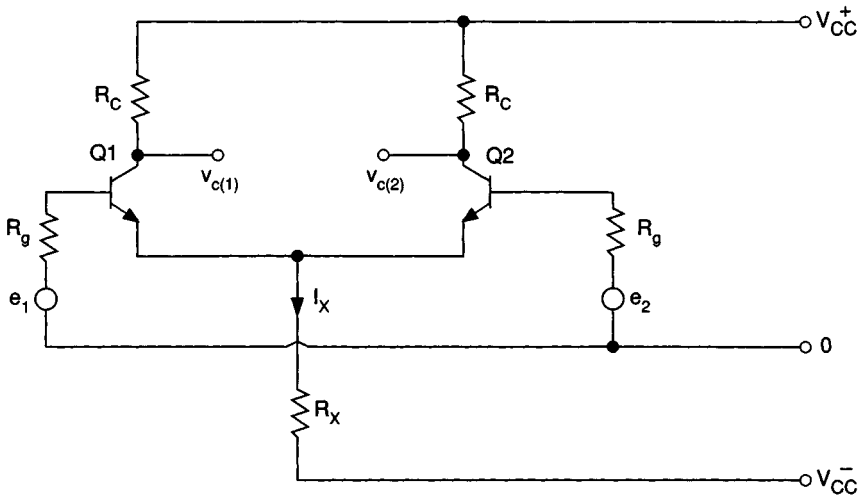


FIGURE 24.39 The difference amplifier.

be identical. This circuit is now called a *difference amplifier,* and the reason will become obvious as soon as the two input signals differ.

The Difference Amplifier

In Fig. 24.39, if $e_1 = e_2$, these are called common-mode input signals, $e_{in(CM)}$, and they will be amplified by $-R_C/R_F$ as for the degenerate CE stage. However, if $e_1 \neq e_2$, then $e_1 - e_2 = e_{in}$, the difference input signal. The following definitions now apply:

$$\frac{e_1 + e_2}{2} = e_{in(CM)} \quad \text{the common mode component}$$

and

$$\frac{\pm(e_1 - e_2)}{2} = e_{in(diff)} \quad \text{the difference component, or } \tfrac{1}{2} e_{in}$$

Hence, $e_1 = e_{in(CM)} + e_{in(diff)}$ and $e_2 = e_{in(CM)} - e_{in(diff)}$.

Consider the progress of a signal current driven by $e_1 - e_2$ and entering the base of $Q1$. It will first pass through R_g, then into the resistance R_{in} at the base of $Q1$, and will arrive at the emitter of $Q2$. Here, if R_X is large, most of this signal current will pass into the resistance presented by the $Q2$ emitter and eventually out of the $Q2$ base via another R_g to ground. The total series resistance is therefore

$$R_g + R_{in} = R_g + r_{bb'} + r_{b'e} + h_{fe} R_{emitter(2)}$$

But

$$R_{emitter(2)} = \frac{R_g + r_{bb'} + r_{b'e}}{h_{fe}}$$

so

$$R_g + R_{in} = 2(R_g + r_{bb'} + r_{b'e})$$

which is the resistance between the two signal sources. Hence,

$$i_{b(1)} = -i_{b(2)} = \frac{e_1 - e_2}{2(R_g + r_{bb'} + r_{b'e})}$$

giving

$$v_{c(1)} = -v_{c(2)} = \frac{h_{fe}R_C(e_1 - e_2)}{2(R_g + r_{bb'} + r_{b'e})}$$

so that the overall difference voltage gain to each collector is

$$A_{ov} = \frac{v_c}{e_1 - e_2} = \frac{\pm h_{fe}R_C}{2(R_g + r_{bb'} + r_{b'e})} \tag{24.28a}$$

If the voltage gain with the input signal measured between the actual bases is needed, R_g may be removed to give

$$A_v = \frac{\pm h_{fe}R_C}{2(r_{bb'} + r_{b'e})} \tag{24.28b}$$

Finally, if the output signal is measured between the collectors (which will be twice that at each collector because they are in antiphase), the difference-in–to–difference-out voltage gain will be

$$A_{v(\text{diff})} = \frac{h_{fe}R_C}{r_{bb'} + r_{b'e}} \simeq \frac{h_{fe}R_C}{r_{b'e}} = g_m R_C \tag{24.28c}$$

which is the same as for a single CE stage.

Note that this is considerably larger than the gain for a common-mode input signal; that is, the difference stage amplifies difference signals well but largely rejects common-mode signals. This common-mode rejection property is very useful, for often, small signals appear across leads, both of which may contain identical electrical noise. So, the difference stage tends to reject the noise while still amplifying the signal. Furthermore, the difference stage has the advantage that it needs no coupling or bypass capacitors and so will amplify frequencies down to zero (dc). Also, it is very stable biaswise and lends itself perfectly to realization on a monolithic IC.

To make the above derivation valid, the long-tail resistance R_X should be as large as possible so that most of the signal current enters the emitter of $Q2$. However, R_X must also carry the quiescent current, which would produce a very high quiescent voltage drop and so require a very high value of V_{CC-}. To overcome this, another transistor structure may be used within a configuration known as a current mirror.

The Current Mirror

The two transistors in Fig. 24.40 are assumed to be identical, and $Q1$ has its base and collector connected so that it acts simply as a diode (formed by the base-emitter junction). The current through it is therefore

$$I = \frac{V_{CC} - V_{BE}}{R} \tag{24.29}$$

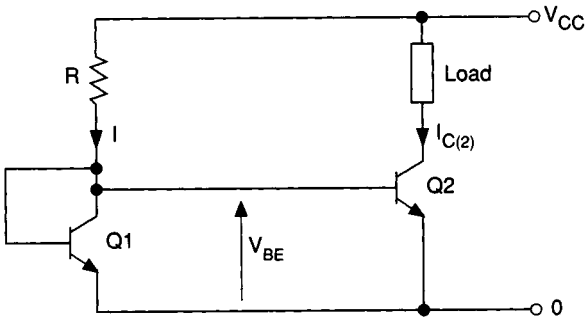


FIGURE 24.40 The current mirror.

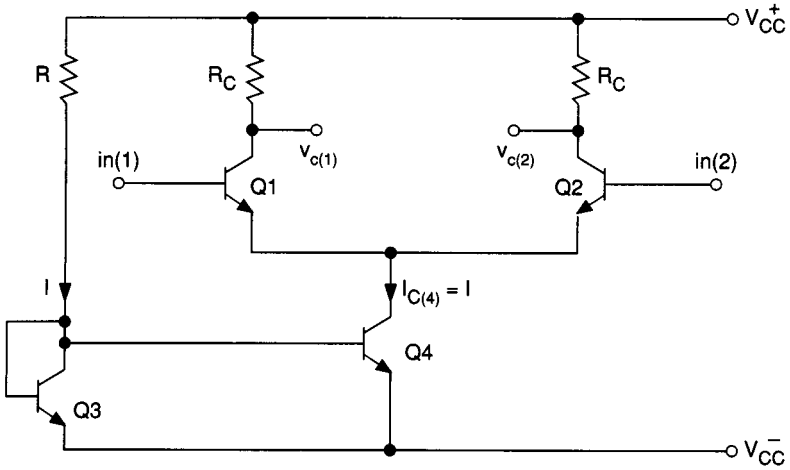


FIGURE 24.41 Current mirror biasing.

The voltage drop V_{BE} so produced is applied to $Q2$ as shown so that it is forced to carry the same collector current I; that is, it mirrors the current in $Q1$.

The transistor $Q2$ is now a device that carries a dc $I_{C(2)} = I$ but presents a large incremental resistance r_{ce} at its collector. This is exactly what is required by the difference amplifier pair, so it may be used in place of R_X.

The Difference Stage with Current Mirror Biasing

Figure 24.41 shows a complete difference stage complete with a current mirror substituting for the long-tail resistor R_X, where the emitter quiescent currents combine to give I_X:

$$I_X = \frac{V_{CC^+} + |V_{CC^-}| - V_{BE(3)}}{R} = I \tag{24.30}$$

This quiescent or bias current is very stable, because the change in $V_{BE(3)}$ due to temperature variations is exactly matched by that required by $Q4$ to produce the same current. The difference gain will be as discussed above, but the common-mode gain will be extremely low because of the high incremental resistance r_{ce} presented by the long-tail transistor.

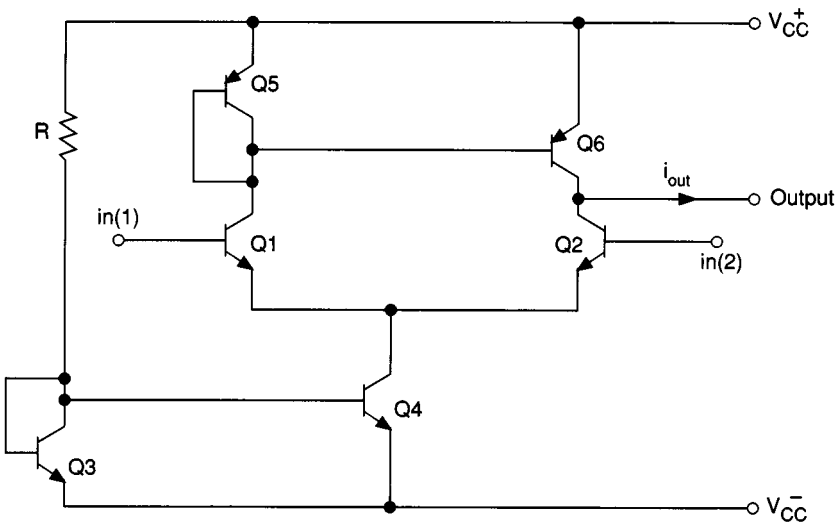


FIGURE 24.42 A complete difference amplifier stage.

The Current Mirror as a Load

A second current mirror may be used as a load for the difference amplifier, as shown in Fig. 24.42. This must utilize *pnp* transistor structures so that the *Q*6 collector loads the *Q*2 collector with a large incremental resistance $r_{ce}(6)$, making for an extremely high voltage gain. Furthermore, *Q*5 and *Q*6 combine the signal output currents of both *Q*1 and *Q*2 to perform a double-ended–to–single-ended conversion as follows. Taking signal currents,

$$i_{\text{out}} = i_c(6) - i_c(2)$$

But

$$i_c(6) = i_c(5)$$

by current mirror action, and

$$i_c(5) = i_c(1)$$

so

$$i_c(6) = i_c(1)$$

Also,

$$i_c(2) = -i_c(1)$$

by difference amplifier action, so

$$i_{\text{out}} = i_c(1) + i_c(1) = 2i_c(1) \tag{24.31}$$

Thus, both sides of the long-tailed pair are used to provide an output current that may then be applied to further stages to form a complete amplifier. Also, because no capacitors and only one resistor are needed, it is an easy circuit for monolithic integration on a single die.

Summary

It has been shown how a limited knowledge of bipolar operation can lead to properly biased amplifier stages using discrete transistors. An equivalent circuit—the hybrid-π model—was then derived, again from limited information, which made possible the analysis of such stages, and some purely practical design results were favorably compared with its predictions. Finally, the tenets of this equivalent circuit were used to evaluate the performance of the difference amplifier and current mirror circuits, which are the cornerstones of modern electronic circuit design in a very wide variety of its manifestations. These circuits are, in fact, the classic transconductance and translinear elements that are ubiquitous in modern IC signal conditioning and function networks.

It should be recognized that there are many models other than the one introduced here, from the simple but very common *h*-parameter version to complex and comprehensive versions developed for computer-aided design (CAD) methods. However, the present elementary approach has been from a design rather than an analytical direction, for it is obvious that powerful modern computer-oriented methods such as the SPICE variants become useful only when a basic circuit configuration has been established, and at the time of writing, this is still the province of the human designer.

Defining Terms

Biasing circuit: A circuit that holds a transistor in an operating condition ready to receive signals.

Common emitter: A basic transistor amplifier stage whose emitter is common to both input and output loops. It amplifies voltage, current, and hence power.

Current mirror: An arrangement of two (or more) transistors such that a defined current passing into one is mirrored in another at a high resistance level.

Degenerate common emitter: A combination of the common-emitter and emitter-follower stages with a very well-defined gain.

Difference amplifier or long-tailed pair: An arrangement of two transistors that amplifies difference signals but rejects common-mode signals. It is often called a differential pair.

Emitter follower or common collector: A basic transistor amplifier stage whose collector is common to both input and output loops. Its voltage gain is near unity, but it amplifies current and hence power. It is a high-input resistance, low-output resistance, or buffer, circuit.

Related Topic

28.2 Small Signal Analysis

Further Information

The following list of recent textbooks covers topics mainly related to analog circuitry containing both integrated and discrete semiconductor devices.

G. M. Glasford, *Analog Electronic Circuits,* Englewood Cliffs, N.J.: Prentice-Hall, 1986.
P. R. Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits,* 2nd ed., New York: Wiley, 1984.
J. Keown, *PSPICE and Circuit Analysis,* New York: Macmillan, 1991.
R.B. Northrop, *Analog Electronic Circuits,* Reading, Mass.: Addison-Wesley, 1990.
A.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., Philadelphia: Saunders, 1991.
T. Schubert and E. Kim, *Active and Non-Linear Electronics,* New York: Wiley, 1996.
J. Watson, *Analog and Switching Circuit Design,* New York: Wiley, 1989.

24.3 The Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET)

John R. Brews

The MOSFET is a transistor that uses a control electrode, the **gate**, to capacitively modulate the conductance of a surface **channel** joining two end contacts, the **source** and the **drain**. The gate is separated from the semiconductor **body** underlying the gate by a thin *gate insulator,* usually silicon dioxide. The surface channel is formed at the interface between the semiconductor body and the gate insulator (see Fig. 24.43).

The MOSFET can be understood by contrast with other field-effect devices, like the *JFET,* or junction field-effect transistor, and the *MESFET,* or metal semiconductor field-effect transistor [Hollis and Murphy, 1990]. These other transistors modulate the conductance of a *majority-carrier* path between two *ohmic* contacts by capacitive control of its cross section. (Majority carriers are those in greatest abundance in a field-free semiconductor, electrons in *n*-type material and holes in *p*-type material.) This modulation of the cross section can take place at any point along the length of the channel, so the gate electrode can be positioned anywhere and need not extend the entire length of the channel.

Analogous to these field-effect devices is the *buried-channel, depletion-mode,* or *normally on* MOSFET, which contains a surface layer of the same doping type as the source and drain (opposite type to the semiconductor body of the device). As a result, it has a built-in or normally on channel from source to drain with a conductance that is reduced when the gate depletes the majority carriers.

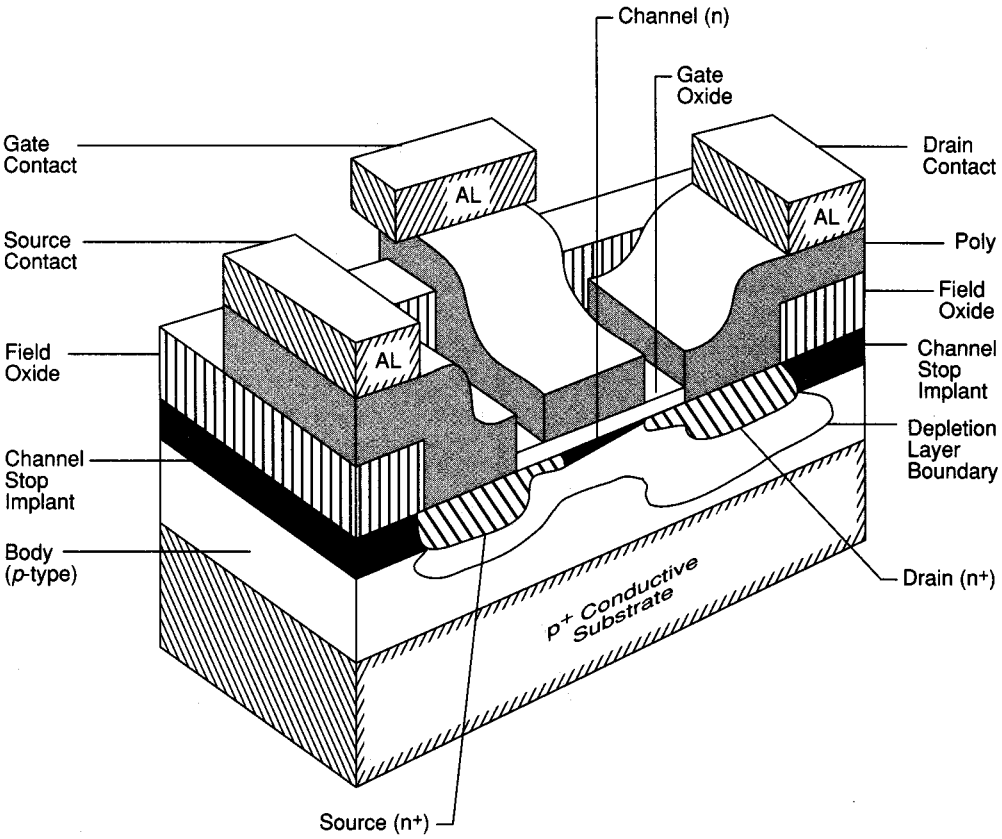


FIGURE 24.43 A high-performance *n*-channel MOSFET. The device is isolated from its neighbors by a surrounding thick *field oxide* under which is a heavily doped *channel stop* implant intended to suppress accidental channel formation that could couple the device to its neighbors. The drain contacts are placed over the field oxide to reduce the capacitance to the body, a parasitic that slows response times. These structural details are described later. (*Source:* After Brews, 1990.)

In contrast, the true MOSFET is an *enhancement-mode* or *normally off* device. The device is normally off because the body forms *pn* junctions with both the source and the drain, so no majority-carrier current can flow between them. Instead, *minority-carrier* current can flow, provided minority carriers are available. As discussed later, for gate biases that are sufficiently attractive, above **threshold**, minority carriers are drawn into a surface channel, forming a conducting path from source to drain. The gate and channel then form two sides of a capacitor separated by the gate insulator. As additional attractive charges are placed on the gate side, the channel side of the capacitor draws a balancing charge of minority carriers from the source and the drain. The more charges on the gate, the more populated the channel, and the larger the conductance. Because the gate *creates* the channel, to ensure electrical continuity the gate must extend over the entire length of the separation between source and drain.

The MOSFET channel is created by attraction to the gate and relies upon the insulating layer between the channel and the gate to prevent leakage of minority carriers to the gate. As a result, MOSFETs can be made only in material systems that provide very good gate insulators, and the best system known is the silicon–silicon dioxide combination. This requirement for a good gate insulator is not so important for JFETs and MESFETs, where the role of the gate is to *push away* majority carriers rather than to attract minority carriers. Thus, in GaAs systems where good insulators are incompatible with other device or fabricational requirements, MESFETs are used.

A more recent development in GaAs systems is the heterostructure field-effect transistor, or HFET [Pearton and Shaw, 1990], made up of layers of varying compositions of Al, Ga, and As or In, Ga, P, and As. These devices are made using molecular beam epitaxy or by organometallic vapor phase epitaxy, expensive methods still being refined for manufacture. HFETs include a variety of structures, the best known of which is the modulation doped FET, or MODFET. HFETs are field-effect devices, not MOSFETs, because the gate simply modulates the carrier density in a preexistent channel between ohmic contacts. The channel is formed spontaneously, regardless of the quality of the gate insulator as a condition of equilibrium between the layers, just as a depletion layer is formed in a *pn* junction. The resulting channel is created very near to the gate electrode, resulting in gate control as effective as in a MOSFET.

The silicon-based MOSFET has been successful primarily because the silicon–silicon dioxide system provides a stable interface with low trap densities, and because the oxide is impermeable to many environmental contaminants, has a high breakdown strength, and is easy to grow uniformly and reproducibly [Nicollian and Brews, 1982]. These attributes allow easy fabrication using lithographic processes, resulting in integrated circuits (ICs), with very small devices, very large device counts, and very high reliability at low cost. Because the importance of the MOSFET lies in this relationship to high-density manufacture, an emphasis of this article is to describe the issues involved in continuing miniaturization.

An additional advantage of the MOSFET is that it can be made using either electrons or holes as channel carrier. Using both types of devices in so-called complementary MOS (CMOS) technology allows circuits that draw no *dc* power if current paths include at least one series connection of both types of device because, in steady state, only one or the other type conducts, not both at once. Of course, in exercising the circuit, power is drawn during switching of the devices. This flexibility in choosing *n*- or *p*-channel devices has enabled large circuits to be made that use low power levels. Hence, complex systems can be manufactured without expensive packaging or cooling requirements.

Current-Voltage Characteristics

The derivation of the current-voltage characteristics of the MOSFET can be found in several sources [Annaratone, 1986; Brews, 1981; Pierret, 1990]. Here a qualitative discussion is provided.

Strong-Inversion Characteristics

In Fig. 24.44 the source-drain current I_D is plotted versus drain-to-source voltage V_D (the I-V curves for the MOSFET). At low V_D the current increases approximately linearly with increased V_D, behaving like a simple resistor with a resistance that is controlled by the gate voltage V_G: as the gate voltage is made more attractive for channel carriers, the channel becomes stronger, more carriers are contained in the channel, and its resistance R_{ch} drops. Hence, at larger V_G the current is larger.

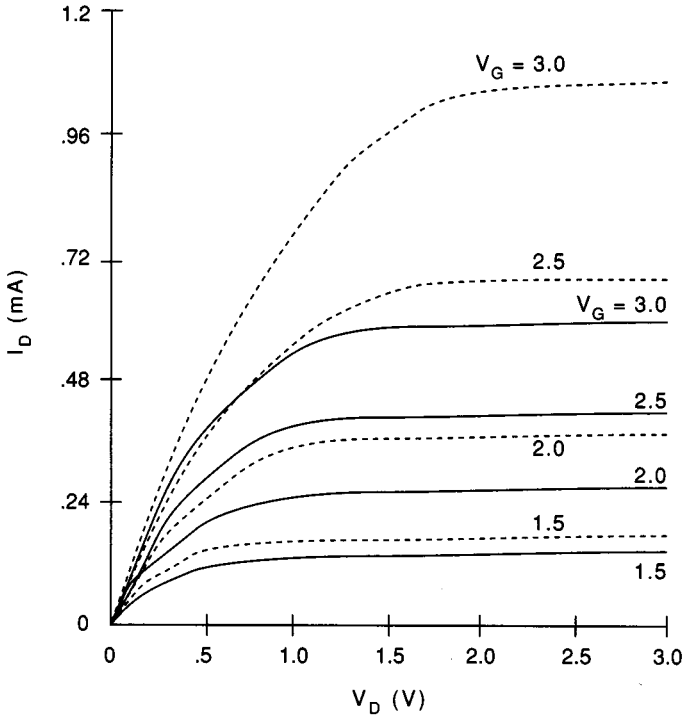


FIGURE 24.44 Drain current I_D versus drain voltage V_D for various choices of gate bias V_G. The dashed-line curves are for a long-channel device for which the current in saturation increases quadratically with gate bias. The solid-line curves are for a *short-channel* device that is approaching *velocity saturation* and thus exhibits a more linear increase in saturation current with gate bias, as discussed in the text.

At large V_D the curves flatten out, and the current is less sensitive to drain bias. The MOSFET is said to be in *saturation*. There are different reasons for this behavior, depending upon the field along the channel caused by the drain voltage. If the source-drain separation is short, near or below a micrometer, the usual drain voltage is sufficient to create fields along the channel of more than a few $\times 10^4$ V/cm. In this case the carrier energy is sufficient for carriers to lose energy by causing vibrations of the silicon atoms composing the crystal (optical phonon emission). Consequently, the carrier velocity does not increase much with increased field, saturating at a value $\upsilon_{sat} \approx 10^7$ cm/s in silicon MOSFETs. Because the carriers do not move faster with increased V_D, the current also saturates.

For longer devices the current-voltage curves saturate for a different reason. Consider the potential along the insulator-channel interface, the surface potential. Whatever the surface potential is at the source end of the channel, it varies from the source end to a value larger at the drain end by V_D because the drain potential is V_D higher than the source. The gate, on the other hand, is at the same potential everywhere. Thus, the difference in potential between the gate and the source is larger than that between the gate and the drain. Correspondingly, the oxide field at the source is larger than that at the drain, and as a result less charge can be supported at the drain. This reduction in attractive power of the gate reduces the number of carriers in the channel at the drain end, increasing channel resistance. In short, we have $I_D \approx V_D / R_{ch}$, but the channel resistance $R_{ch} = R_{ch}(V_D)$ is increasing with V_D. As a result, the current-voltage curves do not continue along the initial straight line, but bend over and saturate.

Another difference between the current-voltage curves for short devices and those for long devices is the dependence on gate voltage. For long devices, the current level in saturation, $I_{D,sat}$, increases quadratically with gate bias. The reason is that the number of carriers in the channel is proportional to $V_G - V_{TH}$ (where V_{TH} is the *threshold voltage*), as is discussed later, the channel resistance $R_{ch} \propto 1/(V_G - V_{TH})$, and the drain bias in saturation is approximately V_G. Thus, $I_{D,sat} = V_D/R_{ch} \propto (V_G - V_{TH})^2$, and we have quadratic dependence. When

the carrier velocity is saturated, however, the dependence of the current on drain bias is suppressed because the speed of the carriers is fixed at υ_{sat}, and $I_{D,sat} \propto \upsilon_{sat}/R_{ch} \propto (V_G - V_{TH})\,\upsilon_{sat}$, a linear gate-voltage dependence. As a result, the current available from a short device is not as large as would be expected if we assumed it behaved like a long device.

Subthreshold Characteristics

Quite different current-voltage behavior is seen in **subthreshold**, that is, for gate biases so low that the channel is in *weak inversion*. In this case the number of carriers in the channel is so small that their charge does not affect the potential, and channel carriers simply must adapt to the potential set up by the electrodes and the dopant ions. Likewise, in subthreshold any flow of current is so small that it causes no potential drop along the interface, which becomes an equipotential.

As there is no lateral field to move the channel carriers, they move by diffusion only, driven by a gradient in carrier density set up because the drain is effective in reducing the carrier density at the drain end of the channel. In subthreshold the current is then independent of drain bias once this bias exceeds a few tens of millivolts, enough to reduce the carrier density at the drain end of the channel to near zero.

In short devices, however, the source and drain are close enough together to begin to share control of the potential with the gate. If this effect is too strong, a drain-voltage dependence of the subthreshold characteristic then occurs, which is undesirable because it increases the MOSFET off current, and can cause a drain-bias dependent threshold voltage.

Although for a well-designed device there is no drain-voltage dependence in subthreshold, gate-bias dependence is exponential. The surface is lowered in energy relative to the semiconductor body by the action of the gate. If this *surface potential* is ϕ_S below that of the body, the carrier density is enhanced by a Boltzmann factor $\exp(q\phi_S/kT)$ relative to the body concentration, where kT/q = the thermal voltage ≈ 25 mV at 290 K. As ϕ_S is roughly proportional to V_G, this exponential dependence on ϕ_S leads to an exponential dependence upon V_G for the carrier density and, hence, for the current in subthreshold.

Important Device Parameters

A number of MOSFET parameters are important to the performance of a MOSFET. In this subsection some of these parameters are discussed, particularly from the viewpoint of digital ICs.

Threshold Voltage

The threshold voltage is vaguely defined as the gate voltage V_{TH} at which the channel begins to form. At this voltage devices begin to switch from "off" to "on," and circuits depend on a voltage swing that straddles this value. Thus, threshold voltage helps in deciding the necessary supply voltage for circuit operation, and also it helps in determining the leakage or "off" current that flows when the device is in the off state.

Threshold voltage is controlled by oxide thickness d and by body doping. To control the body doping, ion implantation is used so that the dopant-ion density is not simply a uniform extension of the bulk, background level N_B ions/unit volume but has superposed upon it an implanted-ion density. To estimate the threshold voltage, we need a picture of what happens in the semiconductor under the gate as the gate voltage is changed from its off level toward threshold.

If we imagine changing the gate bias from its off condition toward threshold, at first the result is to repel majority carriers, forming a surface *depletion layer* (refer to Fig. 24.43). In the depletion layer there are almost no carriers present, but there are dopant ions. In *n*-type material these dopant ions are positive donor impurities that cannot move under fields because they are locked in the silicon lattice, where they have been deliberately introduced to replace silicon atoms. In *p*-type material these dopant ions are negative acceptors. Thus, each charge added to the gate electrode to bring the gate voltage closer to threshold causes an increase in the depletion-layer width sufficient to balance the gate charge by an equal but opposite charge of dopant ions in the silicon depletion layer.

This expansion of the depletion layer continues to balance the addition of gate charge until threshold is reached. Then this charge response changes: above threshold any additional gate charge is balanced by an increasingly strong inversion layer or channel. The border between a depletion-layer and an inversion-layer response, threshold, should occur when

$$\frac{dqN_{\text{inv}}}{d\phi_S} = \frac{dQ_D}{d\phi_S} \tag{24.32}$$

where $d\phi_S$ is the small change in surface potential that corresponds to our incremental change in gate charge, qN_{inv} is the inversion-layer charge/unit area, and Q_D is the depletion-layer charge/unit area. According to Eq. (24.32), the two types of response are equal at threshold, so one is larger than the other on either side of this condition. To be more quantitative, the rate of increase in qN_{inv} is exponential; that is, its rate of change is proportional to qN_{inv}, so as qN_{inv} increases, so does the left side of Eq. (24.32). On the other hand, Q_D has a square-root dependence on ϕ_S, which means its rate of change becomes smaller as Q_D increases. Thus, as surface potential is increased, the left side of Eq. (24.32) increases proportional to qN_{inv} until, at threshold, Eq. (24.32) is satisfied. Then, beyond threshold, the exponential increase in qN_{inv} with ϕ_S swamps Q_D, making change in qN_{inv} the dominant response. Likewise, below threshold, the exponential decrease in qN_{inv} with decreasing ϕ_S makes qN_{inv} negligible and change in Q_D becomes the dominant response. The abruptness of this change in behavior is the reason for the term *threshold* to describe MOSFET switching.

To use Eq. (24.32) to find a formula for threshold voltage, we need expressions for N_{inv} and Q_D. Assuming the interface is held at a lower energy than the bulk due to the charge on the gate, the minority-carrier density at the interface is larger than in the bulk semiconductor, even below threshold. Below threshold and even up to the threshold of Eq. (24.32), the number of charges in the channel/unit area N_{inv} is given for *n*-channel devices approximately by [Brews, 1981]:

$$N_{\text{inv}} \approx d_{\text{INV}} \frac{n_i^2}{N_B} e^{q(\phi_S - V_S)/kT} \tag{24.33}$$

where the various symbols are defined as follows: n_i = intrinsic carrier density/unit volume $\approx 10^{10}/\text{cm}^3$ in silicon at 290 K, and V_S = body reverse bias, if any. The first factor, d_{INV}, is an effective depth of minority carriers from the interface given by

$$d_{\text{INV}} = \frac{\varepsilon_s kT/q}{Q_D} \tag{24.34}$$

where Q_D = depletion-layer charge/unit area due to charged dopant ions in the region where there are no carriers and ε_s is the dielectric permittivity of the semiconductor.

Equation (24.33) expresses the net minority-carrier density/unit area as the product of the bulk minority-carrier density/unit volume, n_i^2/N_B, with the depth of the minority-carrier distribution d_{INV} multiplied in turn by the customary Boltzmann factor $\exp[q(\phi_S - V_S)/kT]$ expressing the enhancement of the interface density over the bulk due to lower energy at the interface. The depth d_{INV} is related to the carrier distribution near the interface using the approximation (valid in *weak inversion*) that the minority-carrier density decays exponentially with distance from the oxide-silicon surface. In this approximation, d_{INV} is the *centroid* of the minority-carrier density. For example, for a uniform bulk doping of 10^{16} dopant ions/cm^3 at 290 K, using Eq. (24.33) and the surface potential at threshold from Eq. (24.38) below ($\phi_{TH} = 0.69$ V), there are $Q_D/q = 3 \times 10^{11}$ charges/cm^2 in the depletion layer at threshold. This Q_D corresponds to a $d_{\text{INV}} = 5.4$ nm and a carrier density at threshold of $N_{\text{inv}} = 5.4 \times 10^9$ charges/cm^2.

The next step in using the definition of threshold, Eq. (24.32), is to introduce the depletion-layer charge/unit area Q_D. For the ion-implanted case, Q_D is made up of two terms [Brews, 1981]:

$$Q_D = qN_B L_B \left\{ 2\left[q\phi_{TH}/(kT) - m_1 - 1 \right] \right\}^{1/2} + qD_I \tag{24.35}$$

where the first term is Q_B, the depletion-layer charge from bulk dopant atoms in the depletion layer with a width that has been reduced by the first moment of the implant, namely, m_1 given in terms of the centroid of the implant x_C by

$$m_1 = \frac{D_I x_C}{N_B L_B^2} \tag{24.36}$$

The second term is the additional charge due to the implanted-ion density within the depletion layer, D_I ions per unit area. The Debye length L_B is defined as

$$L_B^2 \equiv \frac{kT}{q} \frac{\varepsilon_s}{qN_B} \tag{24.37}$$

where ε_s is the dielectric permittivity of the semiconductor. The Debye length is a measure of how deeply a variation of surface potential penetrates into the body when $D_I = 0$ and the depletion layer is of zero width.

Approximating qN_{inv} by Eq. (24.33) and Q_D by Eq. (24.35), Eq. (24.32) determines the surface potential at threshold, ϕ_{TH}, to be

$$\phi_{TH} = 2\frac{kT}{q}\ln\frac{N_B}{n_i} + \frac{kT}{q}\ln\left(1 + \frac{qD_I}{Q_B}\right) \tag{24.38}$$

where the new symbols are defined as follows: Q_B = depletion-layer charge/unit area due to bulk body dopant N_B in the depletion layer, and qD_I = depletion-layer charge/unit area due to implanted ions in the depletion layer between the inversion-layer edge and the depletion-layer edge. Because even a small increase in ϕ_S above ϕ_{TH} causes a large increase in qN_{inv}, which can balance a rather large change in gate charge or gate voltage, ϕ_S does not increase much as $V_G - V_{TH}$ increases. Nonetheless, in strong inversion $N_{\text{inv}} \approx 10^{12}$ charges/cm^2, so in strong inversion ϕ_S will be about 10 kT/q larger than ϕ_{TH}.

Equation (24.38) indicates for uniform doping (no implant, $D_I = 0$) that threshold occurs approximately for $\phi_S = \phi_{TH} = 2(kT/q)\ln(N_B/n_i) \equiv 2\phi_B$, but for the nonuniformly doped case a larger surface potential is needed, assuming the case of a normal implant where D_I is positive, increasing the dopant density. The implant increases the required surface potential because the field at the surface is larger, narrowing the inversion layer, and reducing the channel strength for $\phi_S = 2\phi_B$. Hence, a somewhat larger surface potential is needed to increase qN_{inv} to the point that Eq. (24.32) is satisfied. Equation (24.38) would not apply if a significant fraction of the implant were confined to lie within the inversion layer itself. However, no realistic implant can be confined within a distance comparable to an inversion-layer thickness (a few tens of nanometers), so Eq. (24.38) covers practical cases.

With the surface potential ϕ_{TH} known, the potential on the gate at threshold Φ_{TH} can be found if we know the oxide field F_{ox} by simply adding the potential drop across the semiconductor to that across the oxide. That is, $\Phi_{TH} = \phi_{TH} + F_{\text{ox}}\, d$, with d = oxide thickness and F_{ox} given by Gauss's law as

$$\varepsilon_{\text{ox}} F_{\text{ox}} = Q_D \tag{24.39}$$

There are two more complications in finding the threshold voltage. First, the *gate voltage* V_{TH} usually differs from the gate potential Φ_{TH} at threshold because of a work-function difference between the body and the gate material. This difference causes a spontaneous charge exchange between the two materials as soon as the MOSFET is placed in a circuit allowing charge transfer to occur. Thus, even before any *voltage* is applied to the device, a *potential* difference exists between the gate and the body due to spontaneous charge transfer. The

second complication affecting threshold voltage is the existence of charges in the insulator and at the insulator-semiconductor interface. These nonideal contributions to the overall charge balance are due to traps and fixed charges incorporated during the device processing.

Ordinarily interface-trap charge is negligible ($<10^{10}$/cm^2 in silicon MOSFETs), and the other nonideal effects upon threshold voltage are accounted for by introducing the *flatband voltage* V_{FB}, which corrects the gate bias for these contributions. Then, using Eq. (24.39) with $F_{ox} = (V_{TH} - V_{FB} - \phi_{TH})/d$ we find

$$V_{TH} = V_{FB} + \phi_{TH} + Q_D \frac{d}{\varepsilon_{ox}} \tag{24.40}$$

which determines V_{TH} even for the nonuniformly doped case, using Eq. (24.38) for ϕ_{TH} and Q_D at threshold from Eq. (24.35). If interface-trap charge/unit area is not negligible, then terms in the interface-trap charge/unit area Q_{IT} must be added to Q_D in Eq. (24.40).

From Eqs. (24.35) and (24.38), the threshold voltage depends upon the implanted dopant-ion profile only through two parameters, the net charge introduced by the implant in the region between the inversion layer and the depletion-layer edge qD_p and the centroid of this portion of the implanted charge x_C. As a result, a variety of implants can result in the same threshold, ranging from the extreme of a δ-function spike implant of dose D_I/unit area located at the centroid x_O to a box-type rectangular distribution with the same dose and centroid, namely, a rectangular distribution of width $x_W = 2x_C$ and volume density D_I/x_W. (Of course, x_W must be no larger than the depletion-layer width at threshold for this equivalence to hold true, and x_C must not lie within the inversion layer.) This weak dependence on the details of the profile leaves flexibility to satisfy other requirements, such as control of off current.

As already said, for gate biases $V_G > V_{TH}$ any gate charge above the threshold value is balanced mainly by inversion-layer charge. Thus, the additional oxide field, given by $(V_G - V_{TH})/d$, is related by Gauss's law to the inversion-layer carrier density approximately by

$$\varepsilon_{ox} \frac{V_G - V_{TH}}{d} \approx qN_{inv} \tag{24.41}$$

which shows that channel strength above threshold is proportional to $V_G - V_{TH}$, an approximation often used in this article. Thus, the switch in balancing gate charge from the depletion layer to the inversion layer causes N_{inv} to switch from an exponential gate-voltage dependence in subthreshold to a linear dependence above threshold.

For circuit analysis Eq. (24.41) is a convenient *definition* of V_{TH} because it fits current-voltage curves. If this definition is chosen instead of the charge-balance definition of Eq. (24.32), then Eqs. (24.32) and (24.38) result in an *approximation* to ϕ_{TH}.

Driving Ability and $I_{D,sat}$

The driving ability of the MOSFET is proportional to the current it can provide at a given gate bias. One might anticipate that the larger this current, the faster the circuit. Here this current is used to find some response times governing MOSFET circuits.

MOSFET current is dependent upon the carrier density in the channel, or upon $V_G - V_{TH}$, see Eq. (24.41). For a long-channel device, driving ability depends also on channel length. The shorter the channel length, L, the greater the driving ability, because the channel resistance is directly proportional to the channel length. Although it is an oversimplification, let us suppose that the MOSFET is primarily in saturation during the driving of its load. This simplification will allow a clear discussion of the issues involved in making faster MOSFET's without complicated mathematics. Assuming the MOSFET to be saturated over most of the switching period, driving ability is proportional to current in saturation, or to

$$I_{D,sat} = \frac{\varepsilon_{ox} Z\mu}{2dL} \left(V_G - V_{TH}\right)^2 \tag{24.42}$$

where the factor of two results from the saturating behavior of the *I-V* curves at large drain biases and Z is the width of the channel normal to the direction of current flow. Evidently, for long devices driving ability is quadratic in $V_G - V_{TH}$, and inversely proportional to d.

The result of Eq. (24.42) holds for long devices. For short-channel devices, as explained for Fig. 24.44, the larger fields exerted by the drain electrode cause *velocity saturation* and, as a result, $I_{D,\text{sat}}$ is given roughly by [Einspruch and Gildenblat, 1989]

$$I_{D,\text{sat}} \approx \frac{\varepsilon_{\text{ox}} Z \upsilon_{\text{sat}}}{d} \frac{(V_G - V_{TH})^2}{V_G - V_{TH} + F_{\text{sat}} L} \tag{24.43}$$

where υ_{sat} is the carrier saturation velocity, about 10^7 cm/s for silicon at 290 K, F_{sat} is the field at which velocity saturation sets in, about 5×10^4 V/cm for electrons and not well established as $\gtrsim 10^5$ V/cm for holes in silicon MOSFETs. For Eq. (24.43) to agree with Eq. (24.42) at long L, we need $\mu \approx 2\upsilon_{\text{sat}}/F_{\text{sat}} \approx 400$ cm^2(V·s) for electrons in silicon MOSFETs, which is only roughly correct. Nonetheless, we can see that for devices in the submicron channel length regime, $I_{D,\text{sat}}$ tends to become independent of channel length L and becomes more linear with $V_G - V_{TH}$ and less quadratic (see Fig. 24.44). Equation (24.43) shows that velocity saturation is significant when $V_G/L \gtrsim F_{\text{sat}}$ for example, when $L \lesssim 0.5$ μm if $V_G - V_{TH} = 2.5$ V.

To relate $I_{D,\text{sat}}$ to a gate response time, τ_G, consider one MOSFET driving an identical MOSFET as load capacitance. Then the current from Eq. (24.43) charges this capacitance to a voltage V_G in a gate response time τ_G given by [Shoji, 1988]

$$\begin{aligned} \tau_G &= \frac{C_G V_G}{I_{D,\text{sat}}} \\ &= \frac{L}{\upsilon_{\text{sat}}} \left(1 + \frac{C_{\text{par}}}{C_{\text{ox}}}\right) \frac{V_G (V_G - V_{TH} + F_{\text{sat}} L)}{(V_G - V_{TH})^2} \end{aligned} \tag{24.44}$$

where C_G is the MOSFET gate capacitance $C_G = C_{\text{ox}} + C_{\text{par}}$, with $C_{\text{ox}} = \varepsilon_{\text{ox}} ZL/d$ the MOSFET oxide capacitance, and C_{par} the parasitic component of the gate capacitance [Chen, 1990]. The parasitic capacitance C_{par} is due mainly to overlap of the gate electrode over the source and drain and partly to fringing-field and channel-edge capacitances. For short-channel lengths, C_{par} is a significant part of C_G, and keeping C_{par} under control as L is reduced is an objective of gate-drain alignment technology. Typically, $V_{TH} \approx V_G/4$, so

$$\tau_G \approx \left(\frac{L}{\upsilon_{\text{sat}}}\right)\left(1 + \frac{C_{\text{par}}}{C_{\text{ox}}}\right)\left(1.3 + 1.8 \frac{F_{\text{sat}} L}{V_G}\right) \tag{24.45}$$

Thus, on an intrinsic level, the gate response time is closely related to the transit time of an electron from source to drain, which is L/υ_{sat} in velocity saturation. At shorter L, a linear reduction in delay with L is predicted, while for longer devices the improvement can be quadratic in L, depending upon how V_G is scaled as L is reduced.

The gate response time is not the only delay in device switching, because the drain-body *pn* junction also must charge or discharge for the MOSFET to change state [Shoji, 1988]. Hence, we must also consider a drain response time τ_D. Following Eq. (24.44), we suppose that the drain capacitance C_D is charged by the supply voltage through a MOSFET in saturation so that

$$\tau_D = \frac{C_D V_G}{I_{D,\text{sat}}} = \frac{C_D}{C_G} \tau_G \tag{24.46}$$

Equation (24.46) suggests that τ_D will show a similar improvement to τ_G as L is reduced, provided that C_D/C_G does not increase as L is reduced. However, $C_{ox} \propto L/d$, and the major component of C_{par}, namely, the overlap capacitance contribution, leads to $C_{par} \propto L_{ovlp}/d$ where L_{ovlp} is roughly three times the length of overlap of the gate over the source or drain [Chen, 1990]. Then $C_G \propto (L + L_{ovlp})/d$ and, to keep the C_D/C_G ratio from increasing as L is reduced, either C_D or oxide-thickness d must be reduced along with L.

Clever design can reduce C_D. For example, various *raised-drain* designs reduce the drain-to-body capacitance by separating much of the drain area from the body using a thick oxide layer. The contribution to drain capacitance stemming from the sidewall depletion-layer width next to the channel region is more difficult to handle, because the sidewall depletion layer is deliberately reduced during miniaturization to avoid *short-channel* effects, that is, drain influence upon the channel in competition with gate control. As a result this sidewall contribution to the drain capacitance tends to increase with miniaturization unless junction depth can be shrunk.

Equations (24.45) and (24.46) predict reduction of response times by reduction in channel length L. Decreasing oxide thickness leads to no improvement in τ_G, but Eq. (24.46) shows a possibility of improvement in τ_D, because C_D is independent of d while C_G increases as d decreases. The *ring oscillator*, a closed loop of an odd number of inverters, is a test circuit whose performance depends primarily on τ_G and τ_D. Gate delay/stage for ring oscillators is found to be near 12 ps/stage at 0.1 μm channel length, and 60 ps/stage at 0.5 μm.

For circuits, interconnection capacitances and fan-out (multiple MOSFET loads) will increase response times beyond the device response time, even when parasitics are taken into account. Thus, we are led to consider interconnection delay, τ_{INT}. Although a lumped model suggests, as with Eq. (24.46), that $\tau_{INT} \approx (C_{INT}/C_G)\,\tau_G$, the length of interconnections requires a *distributed* model. Interconnection delay is then

$$\tau_{INT} = \frac{R_{INT}C_{INT}}{2} + R_{INT}C_G + \left(1 + \frac{C_{INT}}{C_G}\right)\tau_G \tag{24.47}$$

where the new symbols are R_{INT} = interconnection resistance, C_{INT} = interconnection capacitance, and we have assumed that the interconnection joins a MOSFET driver in saturation to a MOSFET load C_G. For small R_{INT}, τ_{INT} is dominated by the last term, which resembles Eqs. (24.44) and (24.46). However, unlike the ratio C_D/C_G in Eq. (24.46), it is difficult to reduce or even maintain the ratio C_{INT}/C_G in Eq. (24.47) as L is reduced. Remember, $C_G \propto Z\,(L + L_{ovlp})/d$. Reduction of L therefore tends to increase C_{INT}/C_G, especially because interconnect cross sections cannot be reduced without impractical increases in R_{INT}. What is worse, along with reduction in L, chip sizes usually increase, making line lengths longer, increasing R_{INT} even at constant cross section. As a result, interconnection delay becomes a major problem as L is reduced. The obvious way to keep C_{INT}/C_G under control is to increase the device width Z so that $C_G \propto Z\,(L + L_{ovlp})/d$ remains constant as L is reduced. A better way is to cascade drivers of increasing Z [Chen, 1990; Shoji, 1988]. Either solution requires extra area, however, reducing the packing density that is a major objective in decreasing L in the first place. An alternative is to reduce the oxide thickness d, a major technology objective today.

Transconductance

Another important device parameter is the small-signal transconductance g_m [Sedra and Smith, 1991; Haznedar, 1991], which determines the amount of output current swing at the drain that results from a given input voltage variation at the gate, that is, the small-signal gain:

$$g_m = \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D=\text{const}} \tag{24.48}$$

Using the chain rule of differentiation, the transconductance in saturation can be related to the small-signal *transition* or *unity-gain frequency*, which determines at how high a frequency ω the small-signal current gain $|\imath_{out}/\imath_{in}| = g_m/(\omega C_G)$ drops to unity. Using the chain rule,

$$g_m = \frac{\partial I_{D,\text{sat}}}{\partial Q_G}\frac{\partial Q_G}{\partial V_G} = \omega_T C_G \tag{24.49}$$

where C_G is the oxide capacitance of the device, $C_G = \partial Q_G/\partial V_G|V_D$ with Q_G = the charge on the gate electrode. The frequency ω_T is a measure of the small-signal, high-frequency speed of the device, neglecting parasitic resistances. Using Eq. (24.43) in Eq. (24.49) we find that the transition frequency also is related to the transit time L/υ_{sat} of Eq. (24.45), so that both the digital and small-signal circuit speeds are related to this parameter.

Output Resistance and Drain Conductance

For small-signal circuits the output resistance r_o of the MOSFET [Sedra and Smith, 1991] is important in limiting the gain of amplifiers. This resistance is related to the small-signal drain conductance g_D in saturation by

$$r_o = \frac{1}{g_D} = \frac{\partial V_D}{\partial I_{D,\text{sat}}}\bigg|_{V_G = \text{const}} \tag{24.50}$$

If the MOSFET is used alone as a simple amplifier with a load line set by a resistor R_L, the gain becomes

$$\left|\frac{\upsilon_o}{\upsilon_{\text{in}}}\right| = g_m \frac{R_L r_o}{R_L + r_o} \le g_m R_L \tag{24.51}$$

showing how gain is reduced if r_o is reduced to a value approaching R_L.

As devices are miniaturized, r_o is decreased, g_D increased, due to several factors. At moderate drain biases, the main factor is channel-length modulation, the reduction of the channel length with increasing drain voltage that results when the depletion region around the drain expands toward the source, causing L to become drain-bias dependent. At larger drain biases, a second factor is drain control of the inversion-layer charge density, which can compete with gate control in short devices. This is the same mechanism discussed later in the context of subthreshold behavior. At rather high drain bias, carrier multiplication further lowers r_o.

In a digital inverter, a lower r_o widens the voltage swing needed to cause a transition in output voltage. This widening increases power loss due to current spiking during the transition, and reduces noise margins [Annaratone, 1986]. It is not, however, a first-order concern in device miniaturization for digital applications. Because small-signal circuits are more sensitive to r_o than digital circuits, MOSFETs designed for small-signal applications cannot be made as small as those for digital applications.

Limitations upon Miniaturization

A major factor in the success of the MOSFET has been its compatibility with processing useful down to very small dimensions. Today channel lengths (source-to-drain spacings) of 0.5 μm are manufacturable, and further reduction to 0.1 μm has been achieved for limited numbers of devices in test circuits such as ring oscillators. In this section some of the limits that must be considered in miniaturization are outlined [Brews, 1990].

Subthreshold Control

When a MOSFET is in the "off" condition, that is, when the MOSFET is in *subthreshold,* the off current drawn with the drain at supply voltage must not be too large in order to avoid power consumption and discharge of ostensibly isolated nodes [Shoji, 1988]. In small devices, however, the source and drain are closely spaced, so there exists a danger of direct interaction of the drain with the source, rather than an interaction mediated by the gate and channel. In an extreme case, the drain may draw current directly from the source, even though the gate is "off" (*punchthrough*). A less extreme but also undesirable case occurs when the drain and gate jointly control the carrier density in the channel (*drain-induced barrier lowering*, or drain control of threshold voltage).

In such a case, the on–off behavior of the MOSFET is not controlled by the gate alone, and switching can occur over a range of gate voltages dependent on the drain voltage. Reliable circuit design under these circumstances is very complicated, and testing for design errors is prohibitive. Hence, in designing MOSFETs, a drain-bias independent subthreshold behavior is necessary.

A measure of the range of influence of the source and drain is the depletion-layer width of the associated *pn* junctions. The depletion layer of such a junction is the region in which all carriers have been depleted, or pushed away, due to the potential drop across the junction. This potential drop includes the applied bias across the junction and a spontaneous *built-in* potential drop induced by spontaneous charge exchange when *p* and *n* regions are brought into contact. The depletion-layer width W of an abrupt junction is related to potential drop V and dopant-ion concentration/unit volume N by

$$W = \left(\frac{2\varepsilon_s V}{qN} \right)^{1/2} \tag{24.52}$$

To avoid subthreshold problems, a commonly used rule of thumb is to make sure that the channel length is longer than a minimum length $L_{\min}$ related to the junction depth r_j, the oxide thickness d, and the depletion-layer widths W_S and W_D of the source and drain by [Brews, 1990]

$$L_{\min} = A[r_j d(W_S + W_D)^2]^{1/3} \tag{24.53}$$

where the empirical constant $A = 0.88\ \text{nm}^{-1/3}$ if r_j, W_S, and W_D are in micrometers and d is in nanometers.

Equation (24.53) shows that smaller devices require shallower junctions (smaller r_j), thinner oxides (smaller d), or smaller depletion-layer widths (smaller voltage levels or heavier doping). These requirements introduce side effects that are difficult to control. For example, if the oxide is made thinner while voltages are not reduced proportionately, then oxide fields increase, requiring better oxides. If junction depths are reduced, better control of processing is required, and the junction resistance is increased due to smaller cross sections. To control this resistance, various *self-aligned contact* schemes have been developed to bring the source and drain contacts closer to the gate [Brews, 1990; Einspruch and Gildenblat, 1989], reducing the resistance of these connections. If depletion-layer widths are reduced by increasing the dopant-ion density the *driving ability* of the MOSFET suffers because the threshold voltage increases. That is, Q_D increases in Eq. (24.40), reducing $V_G - V_{TH}$. Thus, for devices that are not velocity-saturated, that is, devices where $V_G/L \lesssim F_{\text{sat}}$, increasing V_{TH} results in slower circuits.

As secondary consequences of increasing dopant-ion density, channel conductance is further reduced due to the combined effects of increased scattering of electrons from the dopant atoms and increased oxide fields that pin carriers in the inversion layer closer to the insulator–semiconductor interface, increasing scattering at the interface. These effects also reduce driving ability, although for shorter devices they are important only in the linear region (that is, below saturation), assuming that mobility μ is more strongly affected than saturation velocity υ_{sat}.

Hot-Electron Effects

Another limit upon how small a MOSFET can be made is a direct result of the larger fields in small devices. Let us digress to consider why proportionately larger voltages, and thus larger fields, are used in smaller devices. First, according to Eq. (24.45), τ_G is shortened if voltages are increased, at least so long as $V_G/L \lesssim F_{\text{sat}}$ 5×10^4 V/cm. If τ_G is shortened this way, then so are τ_D and τ_{INT}, Eqs. (24.46) and (24.47). Thus, faster response is gained by increasing voltages into the velocity saturation region. Second, the fabricational control of smaller devices has not improved proportionately as L has shrunk, so there is a larger percentage variation in device parameters with smaller devices. Thus, disproportionately larger voltages are needed to ensure that all devices operate in the circuit, to overcome this increased fabricational "noise." Thus, to increase speed and to cope with fabricational variations, fields go up in smaller devices.

As a result of these larger fields along the channel direction, a small fraction of the channel carriers have enough energy to enter the insulating layer near the drain. In silicon-based *p*-channel MOSFETs, energetic holes can become trapped in the oxide, leading to a positive oxide charge near the drain that reduces the strength of the channel, degrading device behavior. In *n*-channel MOSFETs, energetic electrons entering the oxide create interface traps and oxide wear-out, eventually leading to gate-to-drain shorts [Pimbley et al., 1989].

To cope with these problems "drain-engineering" has been tried, the most common solution being the *lightly doped drain* [Chen, 1990; Einspruch and Gildenblat, 1989; Pimbley et al., 1989]. In this design, a lightly doped extension of the drain is inserted between the channel and the drain proper. To keep the field moderate and reduce any peaks in the field, the lightly doped drain extension is designed to spread the drain-to-channel voltage drop as evenly as possible. The aim is to smooth out the field at a value close to F_{sat} so that energetic carriers are kept to a minimum. The expense of this solution is an increase in drain resistance and a decreased gain. To increase packing density, this lightly doped drain extension can be stacked vertically alongside the gate, rather than laterally under the gate, to control the overall device area.

Thin Oxides

According to Eq. (24.53), thinner oxides allow shorter devices and therefore higher packing densities for devices. In addition, driving ability is increased, shortening response times for capacitive loads, and output resistance and transconductance are increased. There are some basic limitations upon how thin the oxide can be made. For instance, there is a maximum oxide field that the insulator can withstand. It is thought that the intrinsic breakdown voltage of SiO_2 is of the order of 10^7 V/cm, a field that can support $\approx 2 \times 10^{13}$ charges/cm^2, a large enough value to make this field limitation secondary. Unfortunately, as they are presently manufactured, the intrinsic breakdown of MOSFET oxides is much less likely to limit fields than defect-related leakage or breakdown, and control of these defects has limited reduction of oxide thicknesses in manufacture to about 5 nm to date.

If defect-related problems could be avoided, the thinnest useful oxide would probably be about 3 nm, limited by direct tunneling of channel carriers to the gate. This tunneling limit is not well established, and also is subject to oxide-defect enhancement due to tunneling through intermediate defect levels. Thus, the manufacture of thin oxides is a very active area of exploration.

Dopant-Ion Control

As devices are made smaller, the precise positioning of dopant inside the device is critical. At high temperatures during processing, dopant ions can move. For example, source and drain dopants can enter the channel region, causing position dependence of threshold voltage. Similar problems occur in isolation structures that separate one device from another [Pimbley et al., 1989; Einspruch and Gildenblat, 1989; Wolf, 1995].

To control these thermal effects, process sequences are carefully designed to limit high-temperature steps. This design effort is shortened and improved by the use of computer modeling of the processes. Dopant-ion movement is complex, however, and its theory is made more difficult by the growing trend to use *rapid thermal processing* that involves short-time heat treatments. As a result, dopant response is not steady state, but transient. Computer models of transient response are primitive, forcing further advance in small-device design to be more empirical.

Other Limitations

Besides limitations directly related to the MOSFET, there are some broader difficulties in using MOSFETs of smaller dimension in chips involving even greater numbers of devices. Already mentioned is the increased delay due to interconnections that are lengthening due to increasing chip area and increasing complexity of connection. The capacitive loading of MOSFETs that must drive signals down these lines can slow circuit response, requiring extra circuitry to compensate. Another limitation is the need to isolate devices from each other [Brews, 1990; Chen 1990; Einspruch and Gildenblat, 1989; Pimbley et al., 1989; Wolf, 1995], so their actions remain uncoupled by parasitics. As isolation structures are reduced in size to increase device densities, new parasitics are discovered. A developing solution to this problem is the manufacture of circuits on insulating substrates, silicon-on-insulator technology [Colinge, 1991]. To succeed, this approach must deal with new problems, such as the electrical quality of the underlying silicon–insulator interface and the defect densities in the silicon layer on top of this insulator.

Defining Terms

Channel: The conducting region in a MOSFET between source and drain. In an *enhancement*-mode (or normally off) MOSFET, the channel is an inversion layer formed by attraction of minority carriers toward the gate. These carriers form a thin conducting layer that is prevented from reaching the gate by a thin *gate-oxide* isulating layer when the gate bias exceeds *threshold*. In a *buried-channel*, or *depletion-mode* (or normally on) MOSFET, the channel is present even at zero gate bias, and the gate serves to increase the channel resistance when its bias is nonzero. Thus, this device is based on majority-carrier modulation, like a MESFET.

Gate: The control electrode of a MOSFET. The voltage on the gate capacitively modulates the resistance of the connecting channel between the source and drain.

Source, drain: The two output contacts of a MOSFET, usually formed as *pn* junctions with the *substrate* or *body* of the device.

Strong inversion: The range of gate biases corresponding to the "on" condition of the MOSFET. At a fixed gate bias in this region, for low drain-to-source biases the MOSFET behaves as a simple gate-controlled resistor. At larger drain biases, the channel resistance can increase with drain bias, even to the point that the current *saturates*, or becomes independent of drain bias.

Substrate or body: The portion of the MOSFET that lies between the *source* and *drain* and under the *gate*. The gate is separated from the body by a thin *gate insulator*, usually silicon dioxide. The gate modulates the conductivity of the body, providing a gate-controlled resistance between the source and drain. The body is sometimes *dc*-biased to adjust overall circuit operation. In some circuits the body voltage can swing up and down as a result of input signals, leading to "body-effect" or "back-gate bias" effects that must be controlled for reliable circuit response.

Subthreshold: The range of gate biases corresponding to the "off" condition of the MOSFET. In this regime the MOSFET is not perfectly "off" but conducts a leakage current that must be controlled to avoid circuit errors and power consumption.

Threshold: The gate bias of a MOSFET that marks the boundary between "on" and "off" conditions.

Related Topic

13.2 Parameter Extraction for Analog Circuit Simulation

References

The following references are not to the original sources of the ideas discussed in this article, but have been chosen to be generally useful to the reader.

M. Annaratone, *Digital CMOS Circuit Design,* Boston: Kluwer Academic, 1986.

J. R. Brews, "Physics of the MOS transistor" in *Applied Solid State Science*, *Supplement 2A*, D. Kahng, Ed., New York: Academic, 1981.

J. R. Brews, "The submicron MOSFET" in *High-Speed Semiconductor Devices,* S. M. Sze, Ed., New York: Wiley, 1990, pp. 139–210.

J. Y. Chen, *CMOS Devices and Technology for VLSI,* Englewood Cliffs, N.J.: Prentice-Hall, 1990.

J.-P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI,* Boston: Kluwer Academic, 1991.

H. Haznedar, *Digital Microelectronics,* Redwood City, Calif.: Benjamin-Cummings, 1991.

M.A. Hollis and R.A. Murphy, "Homogeneous field-effect transistors," in *High-Speed Semiconductor Devices,* S. M. Sze, Ed., New York: Wiley, 1990, pp. 211–282.

N.G. Einspruch and G. Sh. Gildenblat, Eds., *VLSI Microstructure Science*, vol. 18, *Advanced MOS Device Physics*, New York: Academic, 1989.

N.R. Malik, *Electronic Circuits: Analysis, Simulation, and Design,* Englewood Cliffs, N.J.: Prentice-Hall, 1995.

E.H. Nicollian and J.R. Brews, *MOS Physics and Technology*, New York: Wiley, 1982, chap. 1.

S.J. Pearton and N.J. Shaw, "Heterostructure field-effect transistors" in *High-Speed Semiconductor Devices,* S.M. Sze, Ed., New York: Wiley, 1990, pp. 283–334.

R.F. Pierret, *Modular Series on Solid State Devices, Field Effect Devices,* 2nd ed., vol. 4, Reading, Mass.: Addison-Wesley, 1990.
J.M. Pimbley, M. Ghezzo, H.G. Parks, and D.M. Brown, *VLSI Electronics Microstructure Science, Advanced CMOS Process Technology,* vol. 19, N. G. Einspruch, Ed., New York: Academic, 1989.
S.S. Sedra and K.C. Smith, *Microelectronic Circuits,* 3rd ed., Philadelphia: Saunders, 1991.
M. Shoji, *CMOS Digital Circuit Technology,* Englewood Cliffs, N.J.: Prentice-Hall, 1988.
S. Wolf, Silicon Processing for the VLSI era: volume 3 — the submicron MOSFET, Sunset Beach, CA: Lattice Press, 1995.

Further Information

The references given in this section have been chosen to provide more detail than is possible to provide in the limited space of this article. In particular, Annaratone [1986] and Shoji [1988] provide much more detail about device and circuit behavior. Chen [1990], Pimbley et al. [1989], and Wolf [1995] provide many technological details of processing and its device impact. Haznedar [1991], Sedra and Smith [1991], and Malik [1995] provide much information about circuits. Brews [1981] and Pierret [1990] provide good discussions of the derivation of the device current-voltage curves and device behavior in all bias regions.