CHAPTER 1

Models for Integrated-Circuit Active Devices

Resumen interesante al final del capítulo

1.1 Introduction

The analysis and design of integrated circuits depend heavily on the utilization of suitable models for integrated-circuit components. This is true in hand analysis, where fairly simple models are generally used, and in computer analysis, where more complex models are encountered. Since any analysis is only as accurate as the model used, it is essential that the circuit designer have a thorough understanding of the origin of the models commonly utilized and the degree of approximation involved in each.

This chapter deals with the derivation of large-signal and small-signal models for integrated-circuit devices. The treatment begins with a consideration of the properties of *pn* junctions, which are basic parts of most integrated-circuit elements. Since this book is primarily concerned with circuit analysis and design, no attempt has been made to produce a comprehensive treatment of semiconductor physics. The emphasis is on summarizing the basic aspects of semiconductor-device behavior and indicating how these can be modeled by equivalent circuits.

1.2 Depletion Region of a *pn* Junction

The properties of reverse-biased *pn* junctions have an important influence on the characteristics of many integrated-circuit components. For example, reverse-biased *pn* junctions exist between many integrated-circuit elements and the underlying substrate, and these junctions all contribute voltage-dependent parasitic capacitances. In addition, a number of important characteristics of active devices, such as breakdown voltage and output resistance, depend directly on the properties of the depletion region of a reverse-biased *pn* junction. Finally, the basic operation of the junction field-effect transistor is controlled by the width of the depletion region of a *pn* junction. Because of its importance and application to many different problems, an analysis of the depletion region of a reverse-biased *pn* junction is considered below. The properties of forward-biased *pn* junctions are treated in Section 1.3 when bipolar-transistor operation is described.

Consider a *pn* junction under reverse bias as shown in Fig. 1.1. Assume *constant doping densities* of N_D atoms/cm^3 in the *n*-type material and N_A atoms/cm^3 in the *p*-type material. (The characteristics of junctions with nonconstant doping densities will be described later.) Due to the difference in carrier concentrations in the *p*-type and *n*-type regions, there exists a region at the junction where the mobile holes and electrons have been removed, leaving the fixed acceptor and donor ions. Each acceptor atom carries a negative charge and each donor atom carries a positive charge, so that the region near the junction is one of significant space charge and resulting high electric field. This is called the *depletion* region or *space-charge*

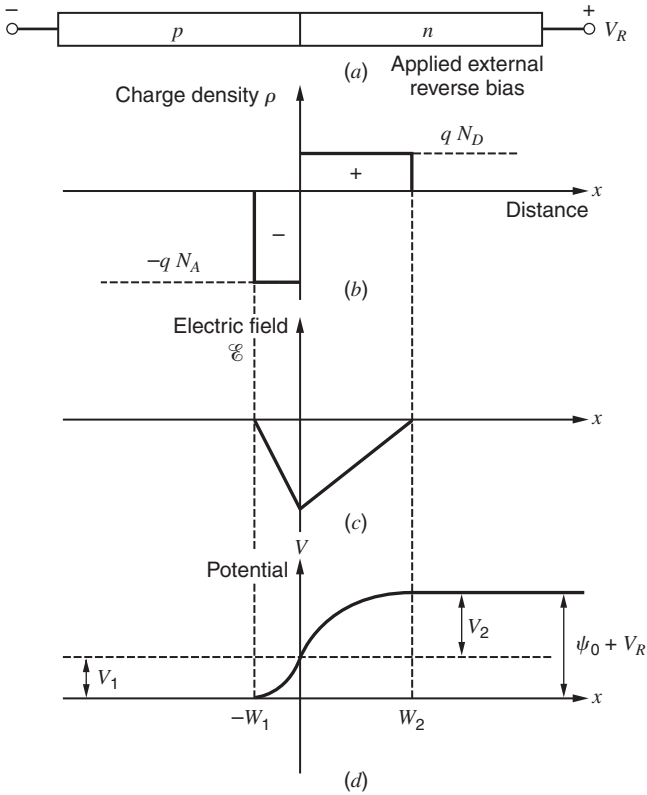


Figure 1.1 The abrupt junction under reverse bias V_R. (*a*) Schematic. (*b*) Charge density. (*c*) Electric field. (*d*) Electrostatic potential.

region. It is assumed that the edges of the depletion region are sharply defined as shown in Fig. 1.1, and this is a good approximation in most cases.

For zero applied bias, there exists a voltage ψ_0 across the junction called the *built-in potential*. This potential opposes the diffusion of mobile holes and electrons across the junction in equilibrium and has a value[1]

$$\psi_0 = V_T \ln \frac{N_A N_D}{n_i^2} \tag{1.1}$$

where

$$V_T = \frac{kT}{q} \simeq 26 \text{ mV} \quad \text{at} \quad 300^\circ\text{K}$$

the quantity n_i is the intrinsic carrier concentration in a pure sample of the semiconductor and $n_i \simeq 1.5 \times 10^{10}\text{cm}^{-3}$ at 300°K for silicon.

In Fig. 1.1 the built-in potential is augmented by the applied reverse bias, V_R, and the total voltage across the junction is $(\psi_0 + V_R)$. If the depletion region penetrates a distance W_1 into the *p*-type region and W_2 into the *n*-type region, then we require

$$W_1 N_A = W_2 N_D \tag{1.2}$$

because the total charge per unit area on either side of the junction must be equal in magnitude but opposite in sign.

Poisson's equation in one dimension requires that

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\epsilon} = \frac{qN_A}{\epsilon} \quad \text{for} \quad -W_1 < x < 0 \tag{1.3}$$

where ρ is the charge density, q is the electron charge (1.6×10^{-19} coulomb), and ϵ is the permittivity of the silicon (1.04×10^{-12} farad/cm). The permittivity is often expressed as

$$\epsilon = K_S \epsilon_0 \tag{1.4}$$

where K_S is the dielectric constant of silicon and ϵ_0 is the permittivity of free space (8.86×10^{-14} F/cm). Integration of (1.3) gives

$$\frac{dV}{dx} = \frac{qN_A}{\epsilon}x + C_1 \tag{1.5}$$

where C_1 is a constant. However, the electric field $\mathscr{E}$ is given by

$$\mathscr{E} = -\frac{dV}{dx} = -\left(\frac{qN_A}{\epsilon}x + C_1\right) \tag{1.6}$$

Since there is zero electric field outside the depletion region, a boundary condition is

$$\mathscr{E} = 0 \quad \text{for} \quad x = -W_1$$

and use of this condition in (1.6) gives

$$\mathscr{E} = -\frac{qN_A}{\epsilon}(x + W_1) = -\frac{dV}{dx} \quad \text{for} \quad -W_1 < x < 0 \tag{1.7}$$

Thus the dipole of charge existing at the junction gives rise to an electric field that varies linearly with distance.

Integration of (1.7) gives

$$V = \frac{qN_A}{\epsilon}\left(\frac{x^2}{2} + W_1 x\right) + C_2 \tag{1.8}$$

If the zero for potential is arbitrarily taken to be the potential of the neutral p-type region, then a second boundary condition is

$$V = 0 \quad \text{for} \quad x = -W_1$$

and use of this in (1.8) gives

$$V = \frac{qN_A}{\epsilon}\left(\frac{x^2}{2} + W_1 x + \frac{W_1^2}{2}\right) \quad \text{for} \quad -W_1 < x < 0 \tag{1.9}$$

At $x = 0$, we define $V = V_1$, and then (1.9) gives

$$V_1 = \frac{qN_A}{\epsilon}\frac{W_1^2}{2} \tag{1.10}$$

If the potential difference from $x = 0$ to $x = W_2$ is V_2, then it follows that

$$V_2 = \frac{qN_D}{\epsilon}\frac{W_2^2}{2} \tag{1.11}$$

and thus the total voltage across the junction is

$$\psi_0 + V_R = V_1 + V_2 = \frac{q}{2\epsilon}(N_A W_1^2 + N_D W_2^2) \tag{1.12}$$

Substitution of (1.2) in (1.12) gives

$$\psi_0 + V_R = \frac{qW_1^2 N_A}{2\epsilon}\left(1 + \frac{N_A}{N_D}\right) \tag{1.13}$$

From (1.13), the penetration of the depletion layer into the p-type region is

$$W_1 = \left[\frac{2\epsilon(\psi_0 + V_R)}{qN_A\left(1 + \frac{N_A}{N_D}\right)}\right]^{1/2} \tag{1.14}$$

Similarly,

$$W_2 = \left[\frac{2\epsilon(\psi_0 + V_R)}{qN_D\left(1 + \frac{N_D}{N_A}\right)}\right]^{1/2} \tag{1.15}$$

Equations 1.14 and 1.15 show that the depletion regions extend into the p-type and n-type regions in *inverse* relation to the impurity concentrations and in proportion to $\sqrt{\psi_0 + V_R}$. If either N_D or N_A is much larger than the other, the depletion region exists almost entirely in the *lightly doped* region.

■ **EXAMPLE**

An abrupt *pn* junction in silicon has doping densities $N_A = 10^{15}$ atoms/cm^3 and $N_D = 10^{16}$ atoms/cm^3. Calculate the junction built-in potential, the depletion-layer depths, and the maximum field with 10 V reverse bias.

From (1.1)

$$\psi_0 = 26 \ln \frac{10^{15} \times 10^{16}}{2.25 \times 10^{20}} \text{ mV} = 638 \text{ mV} \quad \text{at} \quad 300^\circ\text{K}$$

From (1.14) the depletion-layer depth in the p-type region is

$$W_1 = \left(\frac{2 \times 1.04 \times 10^{-12} \times 10.64}{1.6 \times 10^{-19} \times 10^{15} \times 1.1}\right)^{1/2} = 3.5 \times 10^{-4}\text{cm}$$
$$= 3.5\ \mu\text{m} \ \text{(where 1 } \mu\text{m} = 1 \text{ micrometer} = 10^{-6} \text{ m)}$$

The depletion-layer depth in the more heavily doped n-type region is

$$W_2 = \left(\frac{2 \times 1.04 \times 10^{-12} \times 10.64}{1.6 \times 10^{-19} \times 10^{16} \times 11}\right)^{1/2} = 0.35 \times 10^{-4} \text{ cm} = 0.35\ \mu\text{m}$$

Finally, from (1.7) the maximum field that occurs for $x = 0$ is

$$\mathscr{E}_{\max} = -\frac{qN_A}{\epsilon} W_1 = -1.6 \times 10^{-19} \times \frac{10^{15} \times 3.5 \times 10^{-4}}{1.04 \times 10^{-12}}$$
$$= -5.4 \times 10^4 \text{ V/cm}$$

■ Note the large magnitude of this electric field.

1.2.1 Depletion-Region Capacitance

Since there is a *voltage-dependent charge* Q associated with the depletion region, we can calculate a small-signal capacitance C_j as follows:

$$C_j = \frac{dQ}{dV_R} = \frac{dQ}{dW_1}\frac{dW_1}{dV_R} \tag{1.16}$$

Now

$$dQ = AqN_A dW_1 \tag{1.17}$$

where A is the cross-sectional area of the junction. Differentiation of (1.14) gives

$$\frac{dW_1}{dV_R} = \left[\frac{\epsilon}{2qN_A\left(1+\dfrac{N_A}{N_D}\right)(\psi_0 + V_R)}\right]^{1/2} \tag{1.18}$$

Use of (1.17) and (1.18) in (1.16) gives

$$C_j = A\left[\frac{q\epsilon N_A N_D}{2(N_A+N_D)}\right]^{1/2}\frac{1}{\sqrt{\psi_0 + V_R}} \tag{1.19}$$

The above equation was derived for the case of reverse bias V_R applied to the diode. However, it is valid for positive bias voltages as long as the forward current flow is small. Thus, if V_D represents the bias on the junction (positive for forward bias, negative for reverse bias), then (1.19) can be written as

$$C_j = A\left[\frac{q\epsilon N_A N_D}{2(N_A+N_D)}\right]^{1/2}\frac{1}{\sqrt{\psi_0 - V_D}} \tag{1.20}$$

$$= \frac{C_{j0}}{\sqrt{1-\dfrac{V_D}{\psi_0}}} \tag{1.21}$$

where C_{j0} is the value of C_j for $V_D = 0$.

Equations 1.20 and 1.21 were derived using the assumption of constant doping in the p-type and n-type regions. However, many practical diffused junctions more closely approach a *graded* doping profile as shown in Fig. 1.2. In this case, a similar calculation yields

$$C_j = \frac{C_{j0}}{\sqrt[3]{1-\dfrac{V_D}{\psi_0}}} \tag{1.22}$$

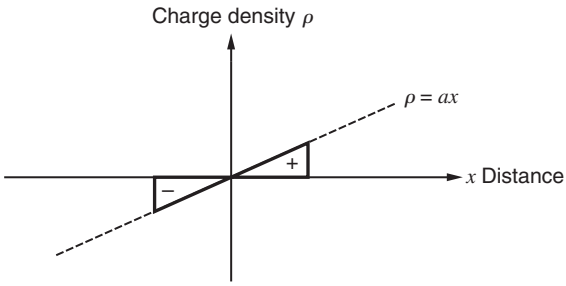


Figure 1.2 Charge density versus distance in a graded junction.

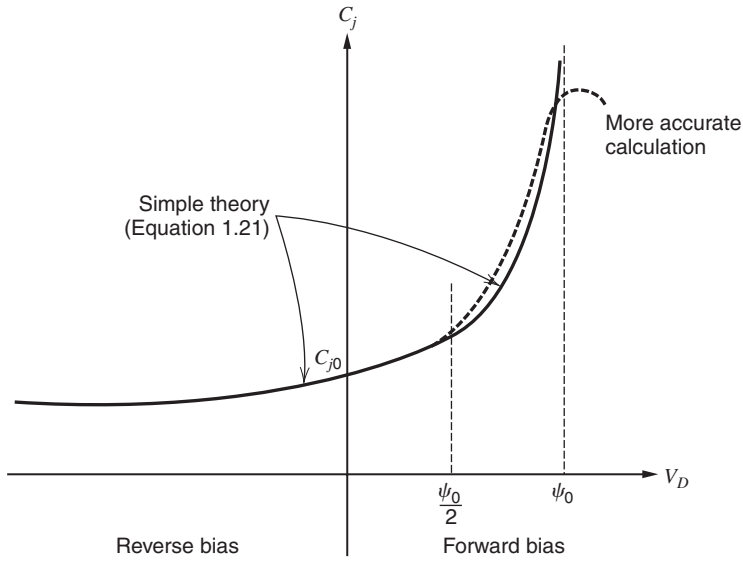


Figure 1.3 Behavior of *pn* junction depletion-layer capacitance C_j as a function of bias voltage V_D.

Note that both (1.21) and (1.22) predict values of C_j approaching infinity as V_D approaches ψ_0. However, the current flow in the diode is then appreciable and the equations are no longer valid. A more exact analysis[2,3] of the behavior of C_j as a function of V_D gives the result shown in Fig. 1.3. For forward bias voltages up to about $\psi_0/2$, the values of C_j predicted by (1.21) are very close to the more accurate value. As an approximation, some computer programs approximate C_j for $V_D > \psi_0/2$ by a linear extrapolation of (1.21) or (1.22).

■ **EXAMPLE**

If the zero-bias capacitance of a diffused junction is 3 pF and $\psi_0 = 0.5$ V, calculate the capacitance with 10 V reverse bias. Assume the doping profile can be approximated by an abrupt junction.

From (1.21)

$$C_j = \frac{3}{\sqrt{1 + \dfrac{10}{0.5}}} \text{pF} = 0.65 \text{ pF}$$

■

1.2.2 Junction Breakdown

From Fig. 1.1*c* it can be seen that the maximum electric field in the depletion region occurs at the junction, and for an abrupt junction (1.7) yields a value

$$\mathscr{E}_{\max} = -\frac{qN_A}{\epsilon} W_1 \tag{1.23}$$

Substitution of (1.14) in (1.23) gives

$$|\mathscr{E}_{\max}| = \left[\frac{2qN_A N_D V_R}{\epsilon\,(N_A + N_D)}\right]^{1/2} \tag{1.24}$$

where ψ_0 has been neglected. Equation 1.24 shows that the maximum field increases as the doping density increases and the reverse bias increases. Although useful for indicating the

functional dependence of $\mathscr{E}_{max}$ on other variables, this equation is strictly valid for an ideal plane junction only. Practical junctions tend to have edge effects that cause somewhat higher values of $\mathscr{E}_{max}$ due to a concentration of the field at the curved edges of the junction.

Any reverse-biased *pn* junction has a small reverse current flow due to the presence of minority-carrier holes and electrons in the vicinity of the depletion region. These are swept across the depletion region by the field and contribute to the leakage current of the junction. As the reverse bias on the junction is increased, the maximum field increases and the carriers acquire increasing amounts of energy between lattice collisions in the depletion region. At a critical field $\mathscr{E}_{crit}$ the carriers traversing the depletion region acquire sufficient energy to create new hole-electron pairs in collisions with silicon atoms. This is called the *avalanche process* and leads to a sudden increase in the reverse-bias leakage current since the newly created carriers are also capable of producing avalanche. The value of $\mathscr{E}_{crit}$ is about 3×10^5 V/cm for junction doping densities in the range of 10^{15} to 10^{16} atoms/cm^3, but it increases slowly as the doping density increases and reaches about 10^6 V/cm for doping densities of 10^{18} atoms/cm^3.

A typical *I-V* characteristic for a junction diode is shown in Fig. 1.4, and the effect of avalanche breakdown is seen by the large increase in reverse current, which occurs as the reverse bias approaches the breakdown voltage *BV*. This corresponds to the maximum field $\mathscr{E}_{max}$ approaching $\mathscr{E}_{crit}$. It has been found empirically[4] that if the normal reverse bias current of the diode is I_R with no avalanche effect, then the actual reverse current near the breakdown voltage is

$$I_{RA} = MI_R \tag{1.25}$$

where M is the *multiplication factor* defined by

$$M = \frac{1}{1 - \left(\dfrac{V_R}{BV}\right)^n} \tag{1.26}$$

In this equation, V_R is the reverse bias on the diode and n has a value between 3 and 6.

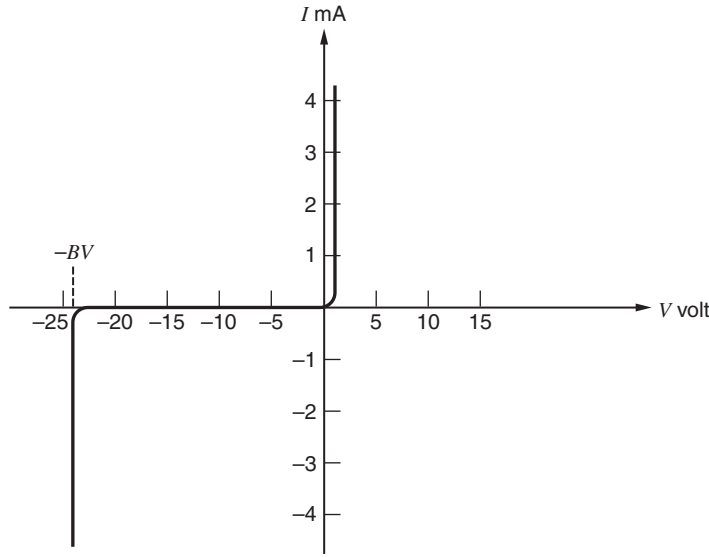


Figure 1.4 Typical *I-V* characteristic of a junction diode showing avalanche breakdown.

The operation of a *pn* junction in the breakdown region is not inherently destructive. However, the avalanche current flow must be limited by external resistors in order to prevent excessive power dissipation from occurring at the junction and causing damage to the device. Diodes operated in the avalanche region are widely used as voltage references and are called *Zener diodes*. There is another, related process called *Zener breakdown*,[5] which is different from the avalanche breakdown described above. Zener breakdown occurs only in very heavily doped junctions where the electric field becomes large enough (even with small reverse-bias voltages) to strip electrons away from the valence bonds. This process is called *tunneling*, and there is no multiplication effect as in avalanche breakdown. Although the Zener breakdown mechanism is important only for breakdown voltages below about 6 V, all breakdown diodes are commonly referred to as Zener diodes.

The calculations so far have been concerned with the breakdown characteristic of plane abrupt junctions. Practical diffused junctions differ in some respects from these results and the characteristics of these junctions have been calculated and tabulated for use by designers.[5] In particular, edge effects in practical diffused junctions can result in breakdown voltages as much as 50 percent below the value calculated for a plane junction.

■ **EXAMPLE**

An abrupt plane *pn* junction has doping densities $N_A = 5 \times 10^{15}$ atoms/cm^3 and $N_D = 10^{16}$ atoms/cm^3. Calculate the breakdown voltage if $\mathscr{E}_{\text{crit}} = 3 \times 10^5$ V/cm.

The breakdown voltage is calculated using $\mathscr{E}_{\text{max}} = \mathscr{E}_{\text{crit}}$ in (1.24) to give

$$BV = \frac{\epsilon (N_A + N_D)}{2qN_AN_D}\mathscr{E}_{\text{crit}}^2$$

$$= \frac{1.04 \times 10^{-12} \times 15 \times 10^{15}}{2 \times 1.6 \times 10^{-19} \times 5 \times 10^{15} \times 10^{16}} \times 9 \times 10^{10} \text{ V}$$

$$= 88 \text{ V}$$

■

1.3 Large-Signal Behavior of Bipolar Transistors

In this section, the large-signal or dc behavior of bipolar transistors is considered. Large-signal models are developed for the calculation of total currents and voltages in transistor circuits, and such effects as breakdown voltage limitations, which are usually not included in models, are also considered. Second-order effects, such as current-gain variation with collector current and Early voltage, can be important in many circuits and are treated in detail.

The sign conventions used for bipolar transistor currents and voltages are shown in Fig. 1.5. All bias currents for both *npn* and *pnp* transistors are assumed positive going into the device.

1.3.1 Large-Signal Models in the Forward-Active Region

A typical *npn* planar bipolar transistor structure is shown in Fig. 1.6*a*, where collector, base, and emitter are labeled C, B, and E, respectively. The method of fabricating such transistor structures is described in Chapter 2. It is shown there that the impurity doping density in the base and the emitter of such a transistor is not constant but varies with distance from the top surface. However, many of the characteristics of such a device can be predicted by analyzing the idealized transistor structure shown in Fig. 1.6*b*. In this structure, the base and emitter

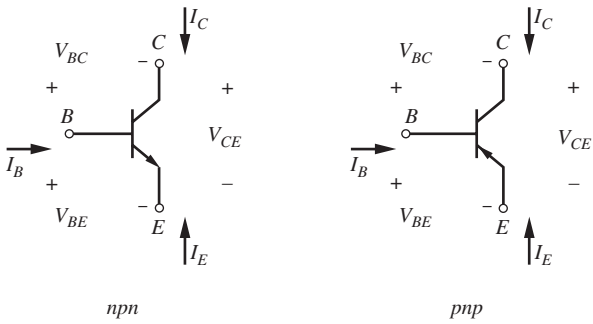
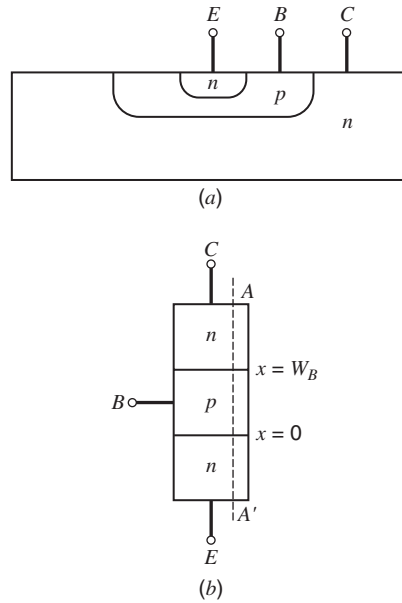


Figure 1.5 Bipolar transistor sign convention.



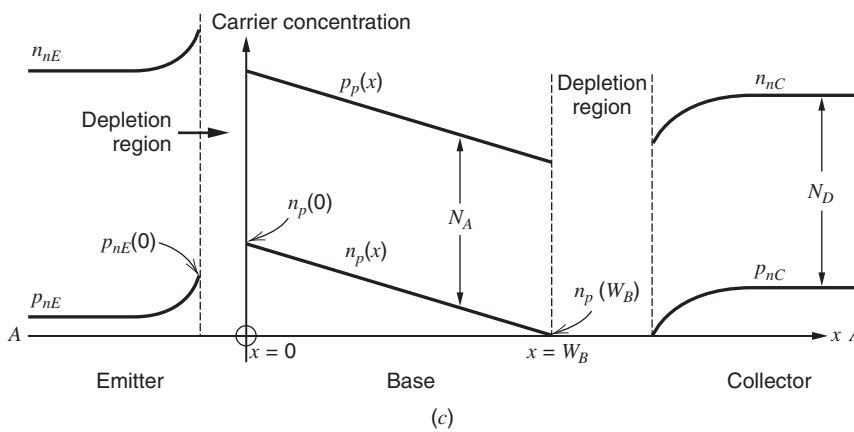


Figure 1.6 (*a*) Cross section of a typical *npn* planar bipolar transistor structure. (*b*) Idealized transistor structure. (*c*) Carrier concentrations along the cross section AA' of the transistor in (*b*). Uniform doping densities are assumed. (Not to scale.)

doping densities are assumed constant, and this is sometimes called a *uniform-base* transistor. Where possible in the following analyses, the equations for the uniform-base analysis are expressed in a form that applies also to nonuniform-base transistors.

A cross section AA' is taken through the device of Fig. 1.6*b* and carrier concentrations along this section are plotted in Fig. 1.6*c*. Hole concentrations are denoted by p and electron concentrations by n with subscripts p or n representing p-type or n-type regions. The n-type emitter and collector regions are distinguished by subscripts E and C, respectively. The carrier concentrations shown in Fig. 1.6*c* apply to a device in the *forward-active region*. That is, the base-emitter junction is forward biased and the base-collector junction is reverse biased. The minority-carrier concentrations in the base at the edges of the depletion regions can be calculated from a Boltzmann approximation to the Fermi-Dirac distribution function to give[6]

$$n_p(0) = n_{po} \exp \frac{V_{BE}}{V_T} \tag{1.27}$$

$$n_p(W_B) = n_{po} \exp \frac{V_{BC}}{V_T} \simeq 0 \tag{1.28}$$

where W_B is the width of the base from the base-emitter depletion layer edge to the base-collector depletion layer edge and n_{po} is the equilibrium concentration of electrons in the base. Note that V_{BC} is negative for an *npn* transistor in the forward-active region and thus $n_p(W_B)$ is very small. Low-level injection conditions are assumed in the derivation of (1.27) and (1.28). This means that the minority-carrier concentrations are always assumed much smaller than the majority-carrier concentration.

If *recombination* of holes and electrons in the base is small, it can be shown that[7] the minority-carrier concentration $n_p(x)$ in the base varies *linearly* with distance. Thus a straight line can be drawn joining the concentrations at $x = 0$ and $x = W_B$ in Fig. 1.6*c*.

For charge neutrality in the base, it is necessary that

$$N_A + n_p(x) = p_p(x) \tag{1.29}$$

and thus

$$p_p(x) - n_p(x) = N_A \tag{1.30}$$

where $p_p(x)$ is the hole concentration in the base and N_A is the base doping density that is assumed constant. Equation 1.30 indicates that the hole and electron concentrations are separated by a constant amount and thus $p_p(x)$ also varies linearly with distance.

Collector current is produced by minority-carrier electrons in the base diffusing in the direction of the concentration gradient and being swept across the collector-base depletion region by the field existing there. The diffusion current density due to electrons in the base is

$$J_n = qD_n \frac{dn_p(x)}{dx} \tag{1.31}$$

where D_n is the diffusion constant for electrons. From Fig. 1.6*c*

$$J_n = -qD_n \frac{n_p(0)}{W_B} \tag{1.32}$$

If I_C is the collector current and is taken as positive flowing *into* the collector, it follows from (1.32) that

$$I_C = qAD_n \frac{n_p(0)}{W_B} \tag{1.33}$$

where A is the cross-sectional area of the emitter. Substitution of (1.27) into (1.33) gives

$$I_C = \frac{qAD_n n_{po}}{W_B} \exp \frac{V_{BE}}{V_T} \tag{1.34}$$

$$= I_S \exp \frac{V_{BE}}{V_T} \tag{1.35}$$

where

$$I_S = \frac{qAD_n n_{po}}{W_B} \tag{1.36}$$

and I_S is a constant used to describe the transfer characteristic of the transistor in the forward-active region. Equation 1.36 can be expressed in terms of the base doping density by noting that[8] (see Chapter 2)

$$n_{po} = \frac{n_i^2}{N_A} \tag{1.37}$$

and substitution of (1.37) in (1.36) gives

$$I_S = \frac{qAD_n n_i^2}{W_B N_A} = \frac{qA\overline{D}_n n_i^2}{Q_B} \tag{1.38}$$

where $Q_B = W_B N_A$ is the number of doping atoms in the base per unit area of the emitter and n_i is the intrinsic carrier concentration in silicon. In this form (1.38) applies to both uniform- and nonuniform-base transistors and D_n has been replaced by $\overline{D}_n$, which is an average effective value for the electron diffusion constant in the base. This is necessary for nonuniform-base devices because the diffusion constant is a function of impurity concentration. Typical values of I_S as given by (1.38) are from 10^{-14} to 10^{-16} A.

Equation 1.35 gives the collector current as a function of base-emitter voltage. The base current I_B is also an important parameter and, at moderate current levels, consists of two major components. One of these (I_{B1}) represents recombination of holes and electrons in the base and is proportional to the minority-carrier charge Q_e in the base. From Fig. 1.6*c*, the minority-carrier charge in the base is

$$Q_e = \frac{1}{2} n_p(0) W_B q A \tag{1.39}$$

and we have

$$I_{B1} = \frac{Q_e}{\tau_b} = \frac{1}{2} \frac{n_p(0) W_B q A}{\tau_b} \tag{1.40}$$

where τ_b is the minority-carrier lifetime in the base. I_{B1} represents a flow of majority holes from the base lead into the base region. Substitution of (1.27) in (1.40) gives

$$I_{B1} = \frac{1}{2} \frac{n_{po} W_B q A}{\tau_b} \exp \frac{V_{BE}}{V_T} \tag{1.41}$$

The second major component of base current (usually the dominant one in integrated-circuit *npn* devices) is due to injection of holes from the base into the emitter. This current component depends on the gradient of minority-carrier holes in the emitter and is[9]

$$I_{B2} = \frac{qAD_p}{L_p} p_{nE}(0) \tag{1.42}$$

where D_p is the diffusion constant for holes and L_p is the diffusion length (assumed small) for holes in the emitter. $p_{nE}(0)$ is the concentration of holes in the emitter at the edge of the

depletion region and is

$$p_{nE}(0) = p_{nEo} \exp \frac{V_{BE}}{V_T} \tag{1.43}$$

If N_D is the donor atom concentration in the emitter (assumed constant), then

$$p_{nEo} \simeq \frac{n_i^2}{N_D} \tag{1.44}$$

The emitter is deliberately doped much more heavily than the base, making N_D large and p_{nEo} small, so that the base-current component, I_{B2}, is minimized.

Substitution of (1.43) and (1.44) in (1.42) gives

$$I_{B2} = \frac{qAD_p}{L_p} \frac{n_i^2}{N_D} \exp \frac{V_{BE}}{V_T} \tag{1.45}$$

The total base current, I_B, is the sum of I_{B1} and I_{B2}:

$$I_B = I_{B1} + I_{B2} = \left(\frac{1}{2} \frac{n_{po} W_B q A}{\tau_b} + \frac{qAD_p}{L_p} \frac{n_i^2}{N_D} \right) \exp \frac{V_{BE}}{V_T} \tag{1.46}$$

Although this equation was derived assuming uniform base and emitter doping, it gives the correct functional dependence of I_B on device parameters for practical double-diffused nonuniform-base devices. Second-order components of I_B, which are important at low current levels, are considered later.

Since I_C in (1.35) and I_B in (1.46) are both proportional to $\exp(V_{BE}/V_T)$ in this analysis, the base current can be expressed in terms of collector current as

$$I_B = \frac{I_C}{\beta_F} \tag{1.47}$$

where β_F is the forward current gain. An expression for β_F can be calculated by substituting (1.34) and (1.46) in (1.47) to give

$$\beta_F = \frac{\dfrac{qAD_n n_{po}}{W_B}}{\dfrac{1}{2} \dfrac{n_{po} W_B q A}{\tau_b} + \dfrac{qAD_p n_i^2}{L_p N_D}} = \frac{1}{\dfrac{W_B^2}{2\tau_b D_n} + \dfrac{D_p}{D_n} \dfrac{W_B}{L_p} \dfrac{N_A}{N_D}} \tag{1.48}$$

where (1.37) has been substituted for n_{po}. Equation 1.48 shows that β_F is maximized by minimizing the base width W_B and maximizing the ratio of emitter to base doping densities N_D/N_A. Typical values of β_F for *npn* transistors in integrated circuits are 50 to 500, whereas lateral *pnp* transistors (to be described in Chapter 2) have values 10 to 100. Finally, the emitter current is

$$I_E = -(I_C + I_B) = -\left(I_C + \frac{I_C}{\beta_F} \right) = -\frac{I_C}{\alpha_F} \tag{1.49}$$

where

$$\alpha_F = \frac{\beta_F}{1 + \beta_F} \tag{1.50}$$

The value of α_F can be expressed in terms of device parameters by substituting (1.48) in (1.50) to obtain

$$\alpha_F = \frac{1}{1+\dfrac{1}{\beta_F}} = \frac{1}{1+\dfrac{W_B^2}{2\tau_b D_n}+\dfrac{D_p}{D_n}\dfrac{W_B}{L_p}\dfrac{N_A}{N_D}} \simeq \alpha_T\gamma \tag{1.51}$$

where

$$\alpha_T = \frac{1}{1+\dfrac{W_B^2}{2\tau_b D_n}} \tag{1.51a}$$

$$\gamma = \frac{1}{1+\dfrac{D_p}{D_n}\dfrac{W_B}{L_p}\dfrac{N_A}{N_D}} \tag{1.51b}$$

The validity of (1.51) depends on $W_B^2/2\tau_b D_n \ll 1$ and $(D_p/D_n)(W_B/L_p)(N_A/N_D) \ll 1$, and this is always true if β_F is large [see (1.48)]. The term γ in (1.51) is called the *emitter injection efficiency* and is equal to the ratio of the electron current (*npn* transistor) injected into the base from the emitter to the total hole and electron current crossing the base-emitter junction. Ideally $\gamma \to 1$, and this is achieved by making N_D/N_A large and W_B small. In that case very little reverse injection occurs from base to emitter.

The term α_T in (1.51) is called the *base transport factor* and represents the fraction of carriers injected into the base (from the emitter) that reach the collector. Ideally $\alpha_T \to 1$ and this is achieved by making W_B small. It is evident from the above development that fabrication changes that cause α_T and γ to approach unity also maximize the value of β_F of the transistor.

The results derived above allow formulation of a large-signal model of the transistor suitable for bias-circuit calculations with devices in the forward-active region. One such circuit is shown in Fig. 1.7 and consists of a base-emitter diode to model (1.46) and a controlled collector-current generator to model (1.47). Note that the collector voltage ideally has no influence on the collector current and the collector node acts as a high-impedance current source. A simpler version of this equivalent circuit, which is often useful, is shown in Fig. 1.7*b*, where the input diode has been replaced by a battery with a value $V_{BE(\text{on})}$, which is usually 0.6 to 0.7 V. This represents the fact that in the forward-active region the base-emitter voltage varies very little because of the steep slope of the exponential characteristic. In some circuits the temperature coefficient of $V_{BE(\text{on})}$ is important, and a typical value for this is -2 mV/°C. The equivalent circuits of Fig. 1.7 apply for *npn* transistors. For *pnp* devices the corresponding equivalent circuits are shown in Fig. 1.8.

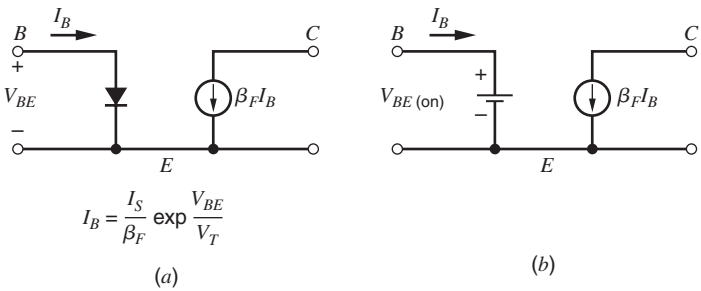


Figure 1.7 Large-signal models of *npn* transistors for use in bias calculations. (*a*) Circuit incorporating an input diode. (*b*) Simplified circuit with an input voltage source.

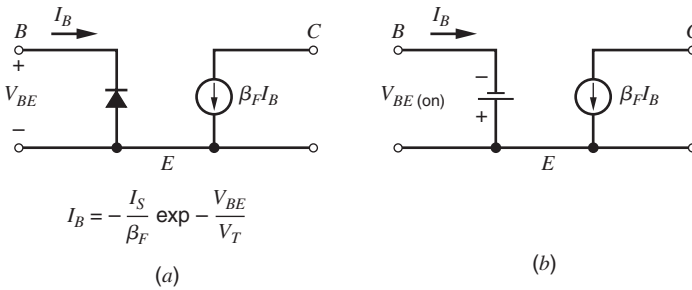


Figure 1.8 Large-signal models of *pnp* transistors corresponding to the circuits of Fig. 1.7.

1.3.2 Effects of Collector Voltage on Large-Signal Characteristics in the Forward-Active Region

In the analysis of the previous section, the collector-base junction was assumed reverse biased and ideally had no effect on the collector currents. This is a useful approximation for first-order calculations, but is not strictly true in practice. There are occasions when the influence of collector voltage on collector current is important, and this will now be investigated.

The collector voltage has a dramatic effect on the collector current in two regions of device operation. These are the saturation (V_{CE} approaches zero) and breakdown (V_{CE} very large) regions that will be considered later. For values of collector-emitter voltage V_{CE} between these extremes, the collector current increases slowly as V_{CE} increases. The reason for this can be seen from Fig. 1.9, which is a sketch of the minority-carrier concentration in the base of the transistor. Consider the effect of changes in V_{CE} on the carrier concentration for constant V_{BE}. Since V_{BE} is constant, the change in V_{CB} equals the change in V_{CE} and this causes an increase in the collector-base depletion-layer width as shown. The change in the base width of the transistor, ΔW_B, equals the change in the depletion-layer width and causes an increase ΔI_C in the collector current.

From (1.35) and (1.38) we have

$$I_C = \frac{qA\overline{D}_n n_i^2}{Q_B} \exp \frac{V_{BE}}{V_T} \tag{1.52}$$

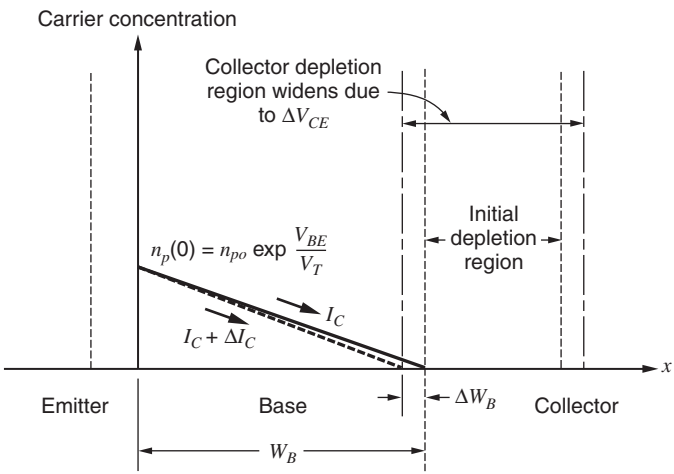


Figure 1.9 Effect of increases in V_{CE} on the collector depletion region and base width of a bipolar transistor.

Differentiation of (1.52) yields

$$\frac{\partial I_C}{\partial V_{CE}} = -\frac{qA\overline{D}_n n_i^2}{Q_B^2}\left(\exp\frac{V_{BE}}{V_T}\right)\frac{dQ_B}{dV_{CE}} \tag{1.53}$$

and substitution of (1.52) in (1.53) gives

$$\frac{\partial I_C}{\partial V_{CE}} = -\frac{I_C}{Q_B}\frac{dQ_B}{dV_{CE}} \tag{1.54}$$

For a uniform-base transistor $Q_B = W_B N_A$, and (1.54) becomes

$$\frac{\partial I_C}{\partial V_{CE}} = -\frac{I_C}{W_B}\frac{dW_B}{dV_{CE}} \tag{1.55}$$

Note that since the base width *decreases* as V_{CE} increases, dW_B/dV_{CE} in (1.55) is negative and thus $\partial I_C/\partial V_{CE}$ is positive. The magnitude of dW_B/dV_{CE} can be calculated from (1.18) for a uniform-base transistor. This equation predicts that dW_B/dV_{CE} is a function of the bias value of V_{CE}, but the variation is typically small for a reverse-biased junction and dW_B/dV_{CE} is often assumed constant. The resulting predictions agree adequately with experimental results.

Equation 1.55 shows that $\partial I_C/\partial V_{CE}$ is proportional to the collector-bias current and inversely proportional to the transistor base width. Thus narrow-base transistors show a greater dependence of I_C on V_{CE} in the forward-active region. The dependence of $\partial I_C/\partial V_{CE}$ on I_C results in typical transistor output characteristics as shown in Fig. 1.10. In accordance with the assumptions made in the foregoing analysis, these characteristics are shown for constant values of V_{BE}. However, in most integrated-circuit transistors the base current is dependent only on V_{BE} and not on V_{CE}, and thus constant-base-current characteristics can often be used in the following calculation. The reason for this is that the base current is usually dominated by the I_{B2} component of (1.45), which has no dependence on V_{CE}. Extrapolation of the characteristics of Fig. 1.10 back to the V_{CE} axis gives an intercept V_A called the Early voltage, where

$$V_A = \frac{I_C}{\dfrac{\partial I_C}{\partial V_{CE}}} \tag{1.56}$$

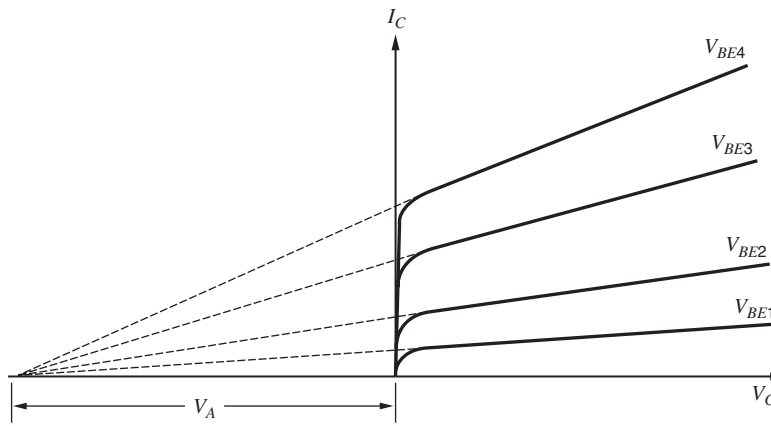


Figure 1.10 Bipolar transistor output characteristics showing the Earlyvoltage, V_A.

Substitution of (1.55) in (1.56) gives

$$V_A = -W_B \frac{dV_{CE}}{dW_B} \tag{1.57}$$

which is a constant, independent of I_C. Thus all the characteristics extrapolate to the same point on the V_{CE} axis. The variation of I_C with V_{CE} is called the Early effect, and V_A is a common model parameter for circuit-analysis computer programs. Typical values of V_A for integrated-circuit transistors are 15 to 100 V. The inclusion of Early effect in dc bias calculations is usually limited to computer analysis because of the complexity introduced into the calculation. However, the influence of the Early effect is often dominant in small-signal calculations for high-gain circuits and this point will be considered later.

Finally, the influence of Early effect on the transistor large-signal characteristics in the forward-active region can be represented approximately by modifying (1.35) to

$$I_C = I_S \left(1 + \frac{V_{CE}}{V_A}\right) \exp \frac{V_{BE}}{V_T} \tag{1.58}$$

This is a common means of representing the device output characteristics for computer simulation.

1.3.3 Saturation and Inverse-Active Regions

Saturation is a region of device operation that is usually avoided in analog circuits because the transistor gain is very low in this region. Saturation is much more commonly encountered in digital circuits, where it provides a well-specified output voltage that represents a logic state.

In saturation, both emitter-base and collector-base junctions are forward biased. Consequently, the collector-emitter voltage V_{CE} is quite small and is usually in the range 0.05 to 0.3 V. The carrier concentrations in a saturated *npn* transistor with uniform base doping are shown in Fig. 1.11. The minority-carrier concentration in the base at the edge of the depletion region is again given by (1.28) as

$$n_p(W_B) = n_{po} \exp \frac{V_{BC}}{V_T} \tag{1.59}$$

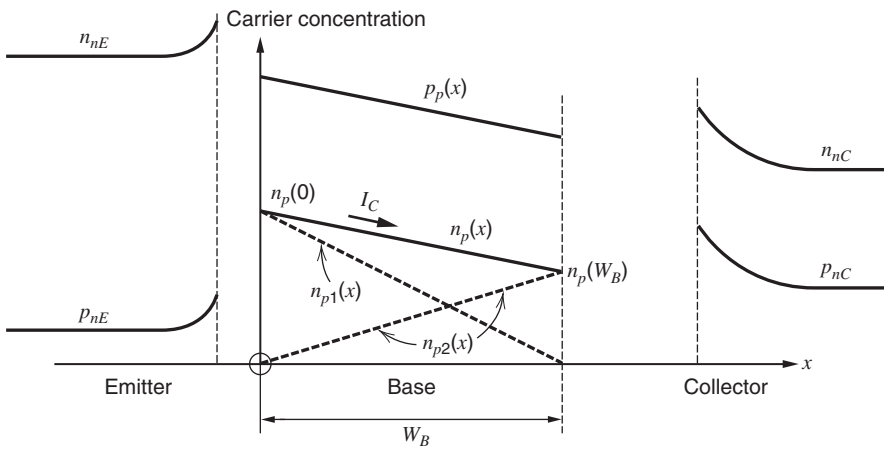


Figure 1.11 Carrier concentrations in a saturated *npn* transistor. (Not to scale.)

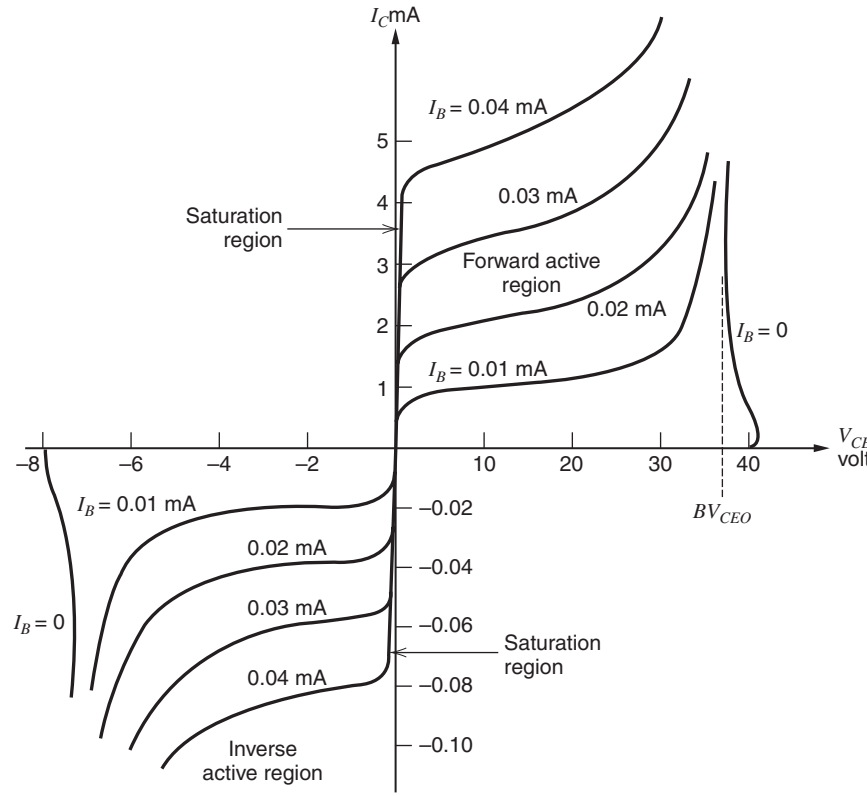


Figure 1.12 Typical I_C-V_{CE} characteristics for an *npn* bipolar transistor. Note the different scales for positive and negative currents and voltages.

but since V_{BC} is now positive, the value of $n_p(W_B)$ is no longer negligible. Consequently, changes in V_{CE} with V_{BE} held constant (which cause equal changes in V_{BC}) directly affect $n_p(W_B)$. Since the collector current is proportional to the slope of the minority-carrier concentration in the base [see (1.31)], it is also proportional to $[n_p(0) - n_p(W_B)]$ from Fig. 1.11. Thus changes in $n_p(W_B)$ directly affect the collector current, and the collector node of the transistor appears to have a *low impedance*. As V_{CE} is decreased in saturation with V_{BE} held constant, V_{BC} increases, as does $n_p(W_B)$ from (1.59). Thus, from Fig. 1.11, the collector current decreases because the slope of the carrier concentration decreases. This gives rise to the saturation region of the $I_C - V_{CE}$ characteristic shown in Fig. 1.12. The slope of the $I_C - V_{CE}$ characteristic in this region is largely determined by the resistance in series with the collector lead due to the finite resistivity of the n-type collector material. A useful model for the transistor in this region is shown in Fig. 1.13 and consists of a fixed voltage source to represent $V_{BE(\text{on})}$, and a fixed voltage source to represent the collector-emitter voltage $V_{CE(\text{sat})}$. A more accurate but more complex model includes a resistor in series with the collector. This resistor can have a value ranging from 20 to 500 Ω, depending on the device structure.

An additional aspect of transistor behavior in the saturation region is apparent from Fig. 1.11. For a given collector current, there is now a much larger amount of stored charge in the base than there is in the forward-active region. Thus the base-current contribution represented by (1.41) will be larger in saturation. In addition, since the collector-base junction is now forward biased, there is a new base-current component due to injection of carriers from the

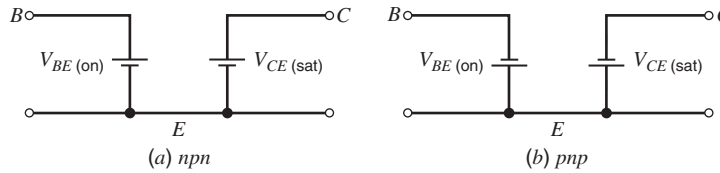


Figure 1.13 Large-signal models for bipolar transistors in the saturation region.

base to the collector. These two effects result in a base current I_B in saturation, which is larger than in the forward-active region for a given collector current I_C. Ratio I_C/I_B in saturation is often referred to as the *forced* β and is always less than β_F. As the forced β is made lower with respect to β_F, the device is said to be more *heavily saturated*.

The minority-carrier concentration in saturation shown in Fig. 1.11 is a straight line joining the two end points, assuming that recombination is small. This can be represented as a linear superposition of the two dotted distributions as shown. The justification for this is that the terminal currents depend *linearly* on the concentrations $n_p(0)$ and $n_p(W_B)$. This picture of device carrier concentrations can be used to derive some general equations describing transistor behavior. Each of the distributions in Fig. 1.11 is considered separately and the two contributions are combined. The *emitter* current that would result from $n_{p1}(x)$ above is given by the classical diode equation

$$I_{EF} = -I_{ES}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \tag{1.60}$$

where I_{ES} is a constant that is often referred to as the *saturation current* of the junction (no connection with the transistor saturation previously described). Equation 1.60 predicts that the junction current is given by $I_{EF} \simeq I_{ES}$ with a reverse-bias voltage applied. However, in practice (1.60) is applicable only in the forward-bias region, since second-order effects dominate under reverse-bias conditions and typically result in a junction current several orders of magnitude larger than I_{ES}. The junction current that flows under reverse-bias conditions is often called the *leakage current* of the junction.

Returning to Fig. 1.11, we can describe the *collector* current resulting from $n_{p2}(x)$ alone as

$$I_{CR} = -I_{CS}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \tag{1.61}$$

where I_{CS} is a constant. The total collector current I_C is given by I_{CR} plus the fraction of I_{EF} that reaches the collector (allowing for recombination and reverse emitter injection). Thus

$$I_C = \alpha_F I_{ES}\left(\exp\frac{V_{BE}}{V_T} - 1\right) - I_{CS}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \tag{1.62}$$

where α_F has been defined previously by (1.51). Similarly, the total emitter current is composed of I_{EF} plus the fraction of I_{CR} that reaches the emitter with the transistor acting in an inverted mode. Thus

$$I_E = -I_{ES}\left(\exp\frac{V_{BE}}{V_T} - 1\right) + \alpha_R I_{CS}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \tag{1.63}$$

where α_R is the ratio of emitter to collector current with the transistor operating *inverted* (i.e., with the collector-base junction forward biased and emitting carriers into the base and the emitter-base junction reverse biased and collecting carriers). Typical values of α_R are 0.5 to 0.8.

An inverse current gain β_R is also defined

$$\beta_R = \frac{\alpha_R}{1 - \alpha_R} \tag{1.64}$$

and has typical values 1 to 5. This is the current gain of the transistor when operated inverted and is much lower than β_F because the device geometry and doping densities are designed to maximize β_F. The inverse-active region of device operation occurs for V_{CE} negative in an *npn* transistor and is shown in Fig. 1.12. In order to display these characteristics adequately in the same figure as the forward-active region, the negative voltage and current scales have been expanded. The inverse-active mode of operation is rarely encountered in analog circuits.

Equations 1.62 and 1.63 describe *npn* transistor operation in the saturation region when V_{BE} and V_{BC} are both positive, and also in the forward-active and inverse-active regions. These equations are the *Ebers-Moll* equations. In the forward-active region, they degenerate into a form similar to that of (1.35), (1.47), and (1.49) derived earlier. This can be shown by putting V_{BE} positive and V_{BC} negative in (1.62) and (1.63) to obtain

$$I_C = \alpha_F I_{ES} \left(\exp \frac{V_{BE}}{V_T} - 1 \right) + I_{CS} \tag{1.65}$$

$$I_E = -I_{ES} \left(\exp \frac{V_{BE}}{V_T} - 1 \right) - \alpha_R I_{CS} \tag{1.66}$$

Equation 1.65 is similar in form to (1.35) except that leakage currents that were previously neglected have now been included. This minor difference is significant only at high temperatures or very low operating currents. Comparison of (1.65) with (1.35) allows us to identify $I_S = \alpha_F I_{ES}$, and it can be shown[10] in general that

$$\alpha_F I_{ES} = \alpha_R I_{CS} = I_S \tag{1.67}$$

where this expression represents a reciprocity condition. Use of (1.67) in (1.62) and (1.63) allows the Ebers-Moll equations to be expressed in the general form

$$I_C = I_S \left(\exp \frac{V_{BE}}{V_T} - 1 \right) - \frac{I_S}{\alpha_R} \left(\exp \frac{V_{BC}}{V_T} - 1 \right) \tag{1.62a}$$

$$I_E = -\frac{I_S}{\alpha_F} \left(\exp \frac{V_{BE}}{V_T} - 1 \right) + I_S \left(\exp \frac{V_{BC}}{V_T} - 1 \right) \tag{1.63a}$$

This form is often used for computer representation of transistor large-signal behavior.

The effect of leakage currents mentioned above can be further illustrated as follows. In the forward-active region, from (1.66)

$$I_{ES} \left(\exp \frac{V_{BE}}{V_T} - 1 \right) = -I_E - \alpha_R I_{CS} \tag{1.68}$$

Substitution of (1.68) in (1.65) gives

$$I_C = -\alpha_F I_E + I_{CO} \tag{1.69}$$

where

$$I_{CO} = I_{CS}(1 - \alpha_R \alpha_F) \tag{1.69a}$$

and I_{CO} is the collector-base leakage current with the emitter open. Although I_{CO} is given theoretically by (1.69a), in practice, surface leakage effects dominate when the collector-base junction is reverse biased and I_{CO} is typically several orders of magnitude larger than the value

given by (1.69a). However, (1.69) is still valid if the appropriate measured value for I_{CO} is used. Typical values of I_{CO} are from 10^{-10} to 10^{-12} A at 25°C, and the magnitude doubles about every 8°C. As a consequence, these leakage terms can become very significant at high temperatures. For example, consider the base current I_B. From Fig. 1.5 this is

$$I_B = -(I_C + I_E) \tag{1.70}$$

If I_E is calculated from (1.69) and substituted in (1.70), the result is

$$I_B = \frac{1 - \alpha_F}{\alpha_F} I_C - \frac{I_{CO}}{\alpha_F} \tag{1.71}$$

But from (1.50)

$$\beta_F = \frac{\alpha_F}{1 - \alpha_F} \tag{1.72}$$

and use of (1.72) in (1.71) gives

$$I_B = \frac{I_C}{\beta_F} - \frac{I_{CO}}{\alpha_F} \tag{1.73}$$

Since the two terms in (1.73) have opposite signs, the effect of I_{CO} is to *decrease* the magnitude of the external base current at a given value of collector current.

■ **EXAMPLE**

If I_{CO} is 10^{-10} A at 24°C, estimate its value at 120°C.

Assuming that I_{CO} doubles every 8°C, we have

$$\begin{aligned} I_{CO}(120^\circ\text{C}) &= 10^{-10} \times 2^{12} \\ &= 0.4\ \mu\text{A} \end{aligned}$$

■

1.3.4 Transistor Breakdown Voltages

In Section 1.2.2 the mechanism of avalanche breakdown in a *pn* junction was described. Similar effects occur at the base-emitter and base-collector junctions of a transistor and these effects limit the maximum voltages that can be applied to the device.

First consider a transistor in the common-base configuration shown in Fig. 1.14*a* and supplied with a constant emitter current. Typical $I_C - V_{CB}$ characteristics for an *npn* transistor in such a connection are shown in Fig. 1.14*b*. For $I_E = 0$ the collector-base junction breaks down at a voltage BV_{CBO}, which represents collector-base breakdown with the emitter open. For finite values of I_E, the effects of avalanche multiplication are apparent for values of V_{CB} below BV_{CBO}. In the example shown, the effective common-base current gain $\alpha_F = I_C/I_E$ becomes larger than unity for values of V_{CB} above about 60 V. Operation in this region (but below BV_{CBO}) can, however, be safely undertaken if the device power dissipation is not excessive. The considerations of Section 1.2.2 apply to this situation, and neglecting leakage currents, we can calculate the collector current in Fig. 1.14*a* as

$$I_C = -\alpha_F I_E M \tag{1.74}$$

where M is defined by (1.26) and thus

$$I_C = -\alpha_F I_E \frac{1}{1 - \left(\dfrac{V_{CB}}{BV_{CBO}}\right)^n} \tag{1.75}$$

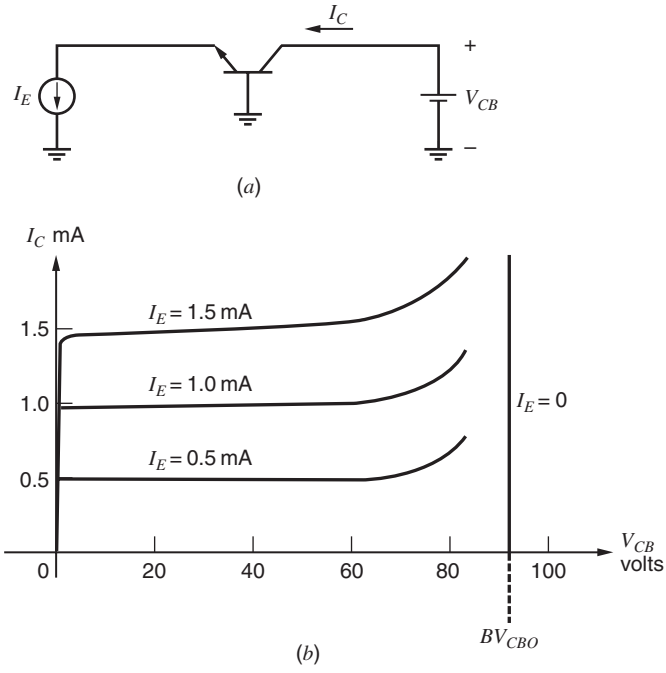


Figure 1.14 Common-base transistor connection. (*a*) Test circuit. (*b*) $I_C - V_{CB}$ characteristics.

One further point to note about the common-base characteristics of Fig. 1.14*b* is that for low values of V_{CB} where avalanche effects are negligible, the curves show very little of the Early effect seen in the common-emitter characteristics. Base widening still occurs in this configuration as V_{CB} is increased, but unlike the common-emitter connection, it produces little change in I_C. This is because I_E is now fixed instead of V_{BE} or I_B, and in Fig. 1.9, this means the slope of the minority-carrier concentration at the emitter edge of the base is fixed. Thus the collector current remains almost unchanged.

Now consider the effect of avalanche breakdown on the common-emitter characteristics of the device. Typical characteristics are shown in Fig. 1.12, and breakdown occurs at a value BV_{CEO}, which is sometimes called the sustaining voltage LV_{CEO}. As in previous cases, operation near the breakdown voltage is destructive to the device only if the current (and thus the power dissipation) becomes excessive.

The effects of avalanche breakdown on the common-emitter characteristics are more complex than in the common-base configuration. This is because hole-electron *pairs* are produced by the avalanche process and the holes are swept into the base, where they effectively contribute to the base current. In a sense, the avalanche current is then *amplified* by the transistor. The base current is still given by

$$I_B = -(I_C + I_E) \tag{1.76}$$

Equation 1.74 still holds, and substitution of this in (1.76) gives

$$I_C = \frac{M\alpha_F}{1 - M\alpha_F} I_B \tag{1.77}$$

where

$$M = \frac{1}{1 - \left(\dfrac{V_{CB}}{BV_{CBO}}\right)^n} \tag{1.78}$$

Equation 1.77 shows that I_C approaches infinity as $M\alpha_F$ approaches unity. That is, the effective β approaches infinity because of the additional base-current contribution from the avalanche process itself. The value of BV_{CEO} can be determined by solving

$$M\alpha_F = 1 \tag{1.79}$$

If we assume that $V_{CB} \simeq V_{CE}$, this gives

$$\frac{\alpha_F}{1 - \left(\dfrac{BV_{CEO}}{BV_{CBO}}\right)^n} = 1 \tag{1.80}$$

and this results in

$$\frac{BV_{CEO}}{BV_{CBO}} = \sqrt[n]{1 - \alpha_F}$$

and thus

$$BV_{CEO} \simeq \frac{BV_{CBO}}{\sqrt[n]{\beta_F}} \tag{1.81}$$

Equation 1.81 shows that BV_{CEO} is less than BV_{CBO} by a substantial factor. However, the value of BV_{CBO}, which must be used in (1.81), is the *plane* junction breakdown of the collector-base junction, neglecting any edge effects. This is because it is only collector-base avalanche current actually under the emitter that is amplified as described in the previous calculation. However, as explained in Section 1.2.2, the measured value of BV_{CBO} is usually determined by avalanche in the curved region of the collector, which is remote from the active base. Consequently, for typical values of $\beta_F = 100$ and $n = 4$, the value of BV_{CEO} is about one-half of the measured BV_{CBO} and not 30 percent as (1.81) would indicate.

Equation 1.81 explains the shape of the breakdown characteristics of Fig. 1.12 if the dependence of β_F on collector current is included. As V_{CE} is increased from zero with $I_B = 0$, the initial collector current is approximately $\beta_F I_{CO}$ from (1.73); since I_{CO} is typically several picoamperes, the collector current is very small. As explained in the next section, β_F is small at low currents, and thus from (1.81) the breakdown voltage is high. However, as avalanche breakdown begins in the device, the value of I_C increases and thus β_F increases. From (1.81) this causes a *decrease* in the breakdown voltage and the characteristic bends back as shown in Fig. 1.12 and exhibits a negative slope. At higher collector currents, β_F approaches a constant value and the breakdown curve with $I_B = 0$ becomes perpendicular to the V_{CE} axis. The value of V_{CE} in this region of the curve is usually defined to be BV_{CEO}, since this is the maximum voltage the device can sustain. The value of β_F to be used to calculate BV_{CEO} in (1.81) is thus the *peak* value of β_F. Note from (1.81) that high-β transistors will thus have low values of BV_{CEO}.

The base-emitter junction of a transistor is also subject to avalanche breakdown. However, the doping density in the emitter is made very large to ensure a high value of β_F [N_D is made large in (1.45) to reduce I_{B2}]. Thus the base is the more lightly doped side of the junction and determines the breakdown characteristic. This can be contrasted with the collector-base junction, where the collector is the more lightly doped side and results in typical values of BV_{CBO} of 20 to 80 V or more. The base is typically an order of magnitude more heavily doped than the collector, and thus the base-emitter breakdown voltage is much less than BV_{CBO} and is typically about 6 to 8 V. This is designated BV_{EBO}. The breakdown voltage for inverse-active operation shown in Fig. 1.12 is approximately equal to this value because the base-emitter junction is reverse biased in this mode of operation.

The base-emitter breakdown voltage of 6 to 8 V provides a convenient reference voltage in integrated-circuit design, and this is often utilized in the form of a *Zener* diode. However,

care must be taken to ensure that all other transistors in a circuit are protected against reverse base-emitter voltages sufficient to cause breakdown. This is because, unlike collector-base breakdown, base-emitter breakdown *is* damaging to the device. It can cause a large degradation in β_F, depending on the duration of the breakdown-current flow and its magnitude.[11] If the device is used purely as a Zener diode, this is of no consequence, but if the device is an amplifying transistor, the β_F degradation may be serious.

■ **EXAMPLE**

If the collector doping density in a transistor is 2×10^{15} atoms/cm^3 and is much less than the base doping, calculate BV_{CEO} for $\beta = 100$ and $n = 4$. Assume $\mathscr{E}_{\text{crit}} = 3 \times 10^5$ V/cm.

The plane breakdown voltage in the collector can be calculated from (1.24) using $\mathscr{E}_{\max} = \mathscr{E}_{\text{crit}}$:

$$BV_{CBO} = \frac{\epsilon\,(N_A + N_D)}{2qN_AN_D}\mathscr{E}_{\text{crit}}^2$$

Since $N_D \ll N_A$, we have

$$BV_{CBO}|_{\text{plane}} = \frac{\epsilon}{2qN_D}\mathscr{E}_{\text{crit}}^2 = \frac{1.04 \times 10^{-12}}{2 \times 1.6 \times 10^{-19} \times 2 \times 10^{15}} \times 9 \times 10^{10}\ \text{V} = 146\ \text{V}$$

From (1.81)

$$BV_{CEO} = \frac{146}{\sqrt[4]{100}}\text{V} = 46\ \text{V}$$

■

1.3.5 Dependence of Transistor Current Gain β_F on Operating Conditions

Although most first-order analyses of integrated circuits make the assumption that β_F is constant, this parameter does in fact depend on the operating conditions of the transistor. It was shown in Section 1.3.2, for example, that increasing the value of V_{CE} increases I_C while producing little change in I_B, and thus the effective β_F of the transistor increases. In Section 1.3.4 it was shown that as V_{CE} approaches the breakdown voltage, BV_{CEO}, the collector current increases due to avalanche multiplication in the collector. Equation 1.77 shows that the effective current gain approaches infinity as V_{CE} approaches BV_{CEO}.

In addition to the effects just described, β_F also varies with both temperature and transistor collector current. This is illustrated in Fig. 1.15, which shows typical curves of β_F versus I_C at three different temperatures for an *npn* integrated circuit transistor. It is evident that β_F increases as temperature increases, and a typical temperature coefficient for β_F is $+7000$ ppm/°C (where ppm signifies *parts per million*). This temperature dependence of β_F is due to the effect of the extremely high doping density in the emitter,[12] which causes the emitter injection efficiency γ to increase with temperature.

The variation of β_F with collector current, which is apparent in Fig. 1.15, can be divided into three regions. Region I is the low-current region, where β_F decreases as I_C decreases. Region II is the midcurrent region, where β_F is approximately constant. Region III is the high-current region, where β_F decreases as I_C increases. The reasons for this behavior of β_F with I_C can be better appreciated by plotting base current I_B and collector current I_C on a log scale

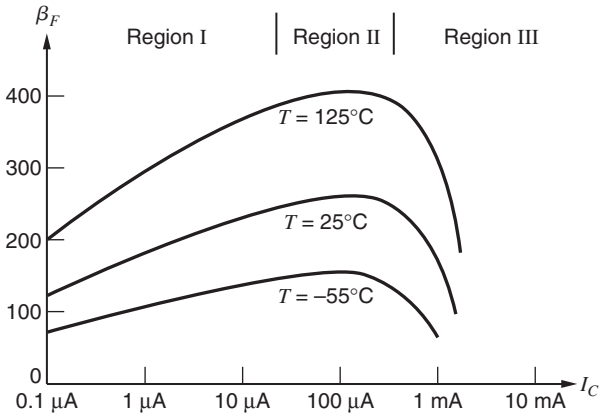


Figure 1.15 Typical curves of β_F versus I_C for an *npn* integrated-circuit transistor with 6 μm^2 emitter area.

as a function of V_{BE}. This is shown in Fig. 1.16, and because of the log scale on the vertical axis, the value of $\ln \beta_F$ can be obtained directly as the distance between the two curves.

At moderate current levels represented by region II in Figs. 1.15 and 1.16, both I_C and I_B follow the ideal behavior, and

$$I_C = I_S \exp \frac{V_{BE}}{V_T} \tag{1.82}$$

$$I_B \simeq \frac{I_S}{\beta_{FM}} \exp \frac{V_{BE}}{V_T} \tag{1.83}$$

where β_{FM} is the maximum value of β_F and is given by (1.48).

At low current levels, I_C still follows the ideal relationship of (1.82), and the decrease in β_F is due to an additional component in I_B, which is mainly due to recombination of carriers in the base-emitter depletion region and is present at any current level. However, at higher current levels the base current given by (1.83) dominates, and this additional component has

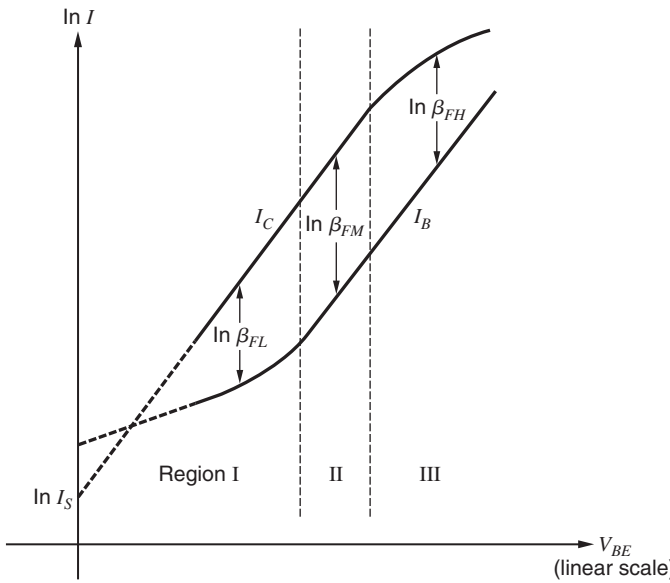


Figure 1.16 Base and collector currents of a bipolar transistor plotted on a log scale versus V_{BE} on a linear scale. The distance between the curves is a direct measure of $\ln \beta_F$.

little effect. The base current resulting from recombination in the depletion region is[5]

$$I_{BX} \simeq I_{SX} \exp \frac{V_{BE}}{m V_T} \tag{1.84}$$

where

$$m \simeq 2$$

At very low collector currents, where (1.84) dominates the base current, the current gain can be calculated from (1.82) and (1.84) as

$$\beta_{FL} \simeq \frac{I_C}{I_{BX}} = \frac{I_S}{I_{SX}} \exp \frac{V_{BE}}{V_T} \left(1 - \frac{1}{m}\right) \tag{1.85}$$

Substitution of (1.82) in (1.85) gives

$$\beta_{FL} \simeq \frac{I_S}{I_{SX}} \left(\frac{I_C}{I_S}\right)^{[1-(1/m)]} \tag{1.86}$$

If $m \simeq 2$, then (1.86) indicates that β_F is proportional to $\sqrt{I_C}$ at very low collector currents.

At high current levels, the base current I_B tends to follow the relationship of (1.83), and the decrease in β_F in region III is due mainly to a decrease in I_C below the value given by (1.82). (In practice the measured curve of I_B versus V_{BE} in Fig. 1.16 may also deviate from a straight line at high currents due to the influence of voltage drop across the base resistance.) The decrease in I_C is due partly to the effect of high-level injection, and at high current levels the collector current approaches[7]

$$I_C \simeq I_{SH} \exp \frac{V_{BE}}{2V_T} \tag{1.87}$$

The current gain in this region can be calculated from (1.87) and (1.83) as

$$\beta_{FH} \simeq \frac{I_{SH}}{I_S} \beta_{FM} \exp \left(-\frac{V_{BE}}{2V_T}\right) \tag{1.88}$$

Substitution of (1.87) in (1.88) gives

$$\beta_{FH} \simeq \frac{I_{SH}^2}{I_S} \beta_{FM} \frac{1}{I_C}$$

Thus β_F decreases rapidly at high collector currents.

In addition to the effect of high-level injection, the value of β_F at high currents is also decreased by the onset of the Kirk effect,[13] which occurs when the minority-carrier concentration in the collector becomes comparable to the donor-atom doping density. The base region of the transistor then stretches out into the collector and becomes greatly enlarged.

1.4 Small-Signal Models of Bipolar Transistors

Analog circuits often operate with signal levels that are small compared to the bias currents and voltages in the circuit. In these circumstances, *incremental* or *small-signal* models can be derived that allow calculation of circuit gain and terminal impedances without the necessity of including the bias quantities. A hierarchy of models with increasing complexity can be derived, and the more complex ones are generally reserved for computer analysis. Part of the designer's skill is knowing which elements of the model can be omitted when performing hand calculations on a particular circuit, and this point is taken up again later.

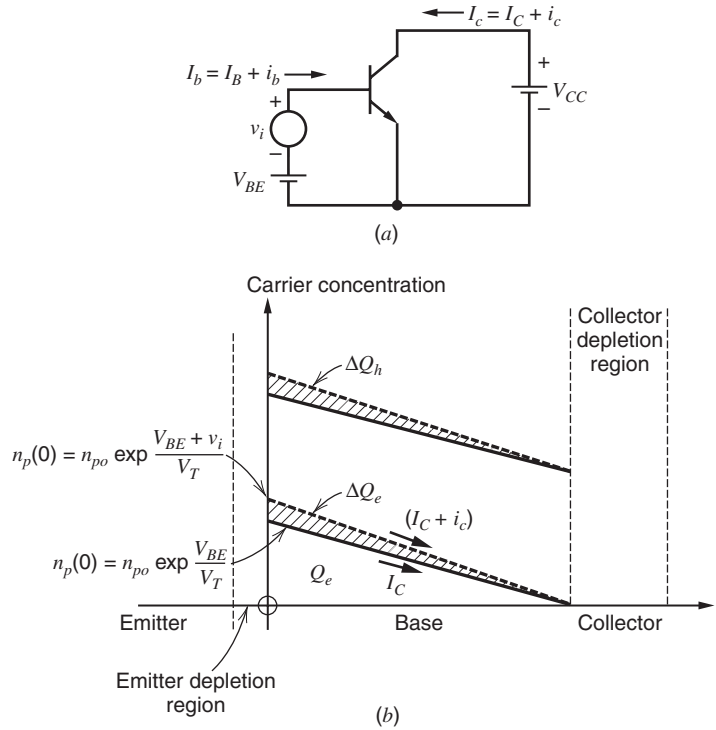


Figure 1.17 Effect of a small-signal input voltage applied to a bipolar transistor. (*a*) Circuit schematic. (*b*) Corresponding changes in carrier concentrations in the base when the device is in the forward-active region.

Consider the bipolar transistor in Fig. 1.17*a* with bias voltages V_{BE} and V_{CC} applied as shown. These produce a quiescent collector current, I_C, and a quiescent base current, I_B, and the device is in the *forward-active region*. A *small-signal* input voltage v_i is applied in series with V_{BE} and produces a small variation in base current i_b and a small variation in collector current i_c. Total values of base and collector currents are I_b and I_c, respectively, and thus $I_b = (I_B + i_b)$ and $I_c = (I_C + i_c)$. The carrier concentrations in the base of the transistor corresponding to the situation in Fig. 1.17*a* are shown in Fig. 1.17*b*. With only bias voltages applied, the carrier concentrations are given by the solid lines. Application of the small-signal voltage v_i causes $n_p(0)$ at the emitter edge of the base to increase, and produces the concentrations shown by the dotted lines. These pictures can now be used to derive the various elements in the small-signal equivalent circuit of the bipolar transistor.

1.4.1 Transconductance

The transconductance is defined as

$$g_m = \frac{dI_C}{dV_{BE}} \tag{1.89}$$

Since

$$\Delta I_C = \frac{dI_C}{dV_{BE}} \Delta V_{BE}$$

we can write

$$\Delta I_C = g_m \Delta V_{BE}$$

and thus

$$i_c = g_m v_i \tag{1.90}$$

The value of g_m can be found by substituting (1.35) in (1.89) to give

$$g_m = \frac{d}{dV_{BE}} I_S \exp \frac{V_{BE}}{V_T} = \frac{I_S}{V_T} \exp \frac{V_{BE}}{V_T} = \frac{I_C}{V_T} = \frac{qI_C}{kT} \tag{1.91}$$

The transconductance thus depends linearly on the bias current I_C and is 38 mA/V for $I_C =$ 1 mA at 25°C for any bipolar transistor of either polarity (*npn* or *pnp*), of any size, and made of any material (Si, Ge, GaAs).

To illustrate the limitations on the use of small-signal analysis, the foregoing relation will be derived in an alternative way. The total collector current in Fig. 1.17*a* can be calculated using (1.35) as

$$I_c = I_S \exp \frac{V_{BE} + v_i}{V_T} = I_S \exp \frac{V_{BE}}{V_T} \exp \frac{v_i}{V_T} \tag{1.92}$$

But the collector bias current is

$$I_C = I_S \exp \frac{V_{BE}}{V_T} \tag{1.93}$$

and use of (1.93) in (1.92) gives

$$I_c = I_C \exp \frac{v_i}{V_T} \tag{1.94}$$

If $v_i < V_T$, the exponential in (1.94) can be expanded in a power series,

$$I_c = I_C \left[1 + \frac{v_i}{V_T} + \frac{1}{2} \left(\frac{v_i}{V_T} \right)^2 + \frac{1}{6} \left(\frac{v_i}{V_T} \right)^3 + \cdots \right] \tag{1.95}$$

Now the incremental collector current is

$$i_c = I_c - I_C \tag{1.96}$$

and substitution of (1.96) in (1.95) gives

$$i_c = \frac{I_C}{V_T} v_i + \frac{1}{2} \frac{I_C}{V_T^2} v_i^2 + \frac{1}{6} \frac{I_C}{V_T^3} v_i^3 + \cdots \tag{1.97}$$

If $v_i \ll V_T$, (1.97) reduces to (1.90), and the small-signal analysis is valid. The criterion for use of small-signal analysis is thus $v_i = \Delta V_{BE} \ll 26$ mV at 25°C. In practice, if ΔV_{BE} is less than 10 mV, the small-signal analysis is accurate within about 10 percent.

1.4.2 Base-Charging Capacitance

Figure 1.17*b* shows that the change in base-emitter voltage $\Delta V_{BE} = v_i$ has caused a change $\Delta Q_e = q_e$ in the minority-carrier charge in the base. By charge-neutrality requirements, there is an equal change $\Delta Q_h = q_h$ in the majority-carrier charge in the base. Since majority carriers

are supplied by the base lead, the application of voltage v_i requires the supply of charge q_h to the base, and the device has an apparent input capacitance

$$C_b = \frac{q_h}{v_i} \tag{1.98}$$

The value of C_b can be related to fundamental device parameters as follows. If (1.39) is divided by (1.33), we obtain

$$\frac{Q_e}{I_C} = \frac{W_B^2}{2D_n} = \tau_F \tag{1.99}$$

The quantity τ_F has the dimension of time and is called the base transit time in the forward direction. Since it is the ratio of the charge in transit (Q_e) to the current flow (I_C), it can be identified as the average time per carrier spent in crossing the base. To a first order it is independent of operating conditions and has typical values 10 to 500 ps for integrated *npn* transistors and 1 to 40 ns for lateral *pnp* transistors. Practical values of τ_F tend to be somewhat lower than predicted by (1.99) for diffused transistors that have nonuniform base doping.[14] However, the functional dependence on base width W_B and diffusion constant D_n is as predicted by (1.99).

From (1.99)

$$\Delta Q_e = \tau_F \Delta I_C \tag{1.100}$$

But since $\Delta Q_e = \Delta Q_h$, we have

$$\Delta Q_h = \tau_F \Delta I_C \tag{1.101}$$

and this can be written

$$q_h = \tau_F i_c \tag{1.102}$$

Use of (1.102) in (1.98) gives

$$C_b = \tau_F \frac{i_c}{v_i} \tag{1.103}$$

and substitution of (1.90) in (1.103) gives

$$C_b = \tau_F g_m \tag{1.104}$$

$$= \tau_F \frac{qI_C}{kT} \tag{1.105}$$

Thus the small-signal, base-charging capacitance is proportional to the collector bias current.

In the inverse-active mode of operation, an equation similar to (1.99) relates stored charge and current via a time constant τ_R. This is typically orders of magnitude larger than τ_F because the device structure and doping are optimized for operation in the forward-active region. Since the saturation region is a combination of forward-active and inverse-active operation, inclusion of the parameter τ_R in a SPICE listing will model the large charge storage that occurs in saturation.

1.4.3 Input Resistance

In the forward-active region, the base current is related to the collector current by

$$I_B = \frac{I_C}{\beta_F} \tag{1.47}$$

Small changes in I_B and I_C can be related using (1.47):

$$\Delta I_B = \frac{d}{dI_C}\left(\frac{I_C}{\beta_F}\right)\Delta I_C \tag{1.106}$$

and thus

$$\beta_0 = \frac{\Delta I_C}{\Delta I_B} = \frac{i_c}{i_b} = \left[\frac{d}{dI_C}\left(\frac{I_C}{\beta_F}\right)\right]^{-1} \tag{1.107}$$

where β_0 is the *small-signal* current gain of the transistor. Note that if β_F is constant, then $\beta_F = \beta_0$. Typical values of β_0 are close to those of β_F, and in subsequent chapters little differentiation is made between these quantities. A single value of β is often assumed for a transistor and then used for both ac and dc calculations.

Equation 1.107 relates the change in base current i_b to the corresponding change in collector current i_c, and the device has a small-signal input resistance given by

$$r_\pi = \frac{v_i}{i_b} \tag{1.108}$$

Substitution of (1.107) in (1.108) gives

$$r_\pi = \frac{v_i}{i_c}\beta_0 \tag{1.109}$$

and use of (1.90) in (1.109) gives

$$r_\pi = \frac{\beta_0}{g_m} \tag{1.110}$$

Thus the small-signal input shunt resistance of a bipolar transistor depends on the current gain and is inversely proportional to I_C.

1.4.4 Output Resistance

In Section 1.3.2 the effect of changes in collector-emitter voltage V_{CE} on the large-signal characteristics of the transistor was described. It follows from that treatment that small changes ΔV_{CE} in V_{CE} produce corresponding changes ΔI_C in I_C, where

$$\Delta I_C = \frac{\partial I_C}{\partial V_{CE}}\Delta V_{CE} \tag{1.111}$$

Substitution of (1.55) and (1.57) in (1.111) gives

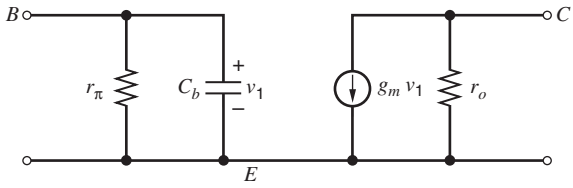
$$\frac{\Delta V_{CE}}{\Delta I_C} = \frac{V_A}{I_C} = r_o \tag{1.112}$$

where V_A is the Early voltage and r_o is the small-signal output resistance of the transistor. Since typical values of V_A are 50 to 100 V, corresponding values of r_o are 50 to 100 kΩ for $I_C = 1$mA. Note that r_o is inversely proportional to I_C, and thus r_o can be related to g_m, as are many of the other small-signal parameters.

$$r_o = \frac{1}{\eta g_m} \tag{1.113}$$

where

$$\eta = \frac{kT}{qV_A} \tag{1.114}$$



$$r_\pi = \frac{\beta}{g_m},\ r_o = \frac{1}{\eta g_m},\ g_m = \frac{qI_C}{kT},\ C_b = \tau_F g_m$$

Figure 1.18 Basic bipolar transistor small-signal equivalent circuit.

If $V_A = 100$ V, then $\eta = 2.6 \times 10^{-4}$ at 25°C. Note that $1/r_o$ is the slope of the output characteristics of Fig. 1.10.

1.4.5 Basic Small-Signal Model of the Bipolar Transistor

Combination of the above small-signal circuit elements yields the small-signal model of the bipolar transistor shown in Fig. 1.18. This is valid for both *npn* and *pnp* devices in the forward-active region and is called the *hybrid-π* model. Collector, base, and emitter nodes are labeled C, B and E, respectively. The elements in this circuit are present in the equivalent circuit of *any* bipolar transistor and are specified by relatively few parameters (β, τ_F, η, I_C). Note that in the evaluation of the small-signal parameters for *pnp* transistors, the *magnitude only* of I_C is used. In the following sections, further elements are added to this model to account for parasitics and second-order effects.

1.4.6 Collector-Base Resistance

Consider the effect of variations in V_{CE} on the minority charge in the base as illustrated in Fig. 1.9. An increase in V_{CE} causes an increase in the collector depletion-layer width and consequent reduction of base width. This causes a reduction in the total minority-carrier charge stored in the base and thus a reduction in base current I_B due to a reduction in I_{B1} given by (1.40). Since an increase ΔV_{CE} in V_{CE} causes a *decrease* ΔI_B in I_B, this effect can be modeled by inclusion of a resistor r_μ from collector to base of the model of Fig. 1.18. If V_{BE} is assumed held constant, the value of this resistor can be determined as follows.

$$r_\mu = \frac{\Delta V_{CE}}{\Delta I_{B1}} = \frac{\Delta V_{CE}}{\Delta I_C}\frac{\Delta I_C}{\Delta I_{B1}} \tag{1.115}$$

Substitution of (1.112) in (1.115) gives

$$r_\mu = r_o \frac{\Delta I_C}{\Delta I_{B1}} \tag{1.116}$$

If the base current I_B is composed entirely of component I_{B1}, then (1.107) can be used in (1.116) to give

$$r_\mu = \beta_0 r_o \tag{1.117}$$

This is a lower limit for r_μ. In practice, I_{B1} is typically less than 10 percent of I_B [component I_{B2} from (1.42) dominates] in integrated *npn* transistors, and since I_{B1} is very small, the change ΔI_{B1} in I_{B1} for a given ΔV_{CE} and ΔI_C is also very small. Thus a typical value for r_μ is greater than $10\beta_0 r_o$. For lateral *pnp* transistors, recombination in the base is more significant, and r_μ is in the range $2\beta_0 r_o$ to $5\beta_0 r_o$.

1.4.7 Parasitic Elements in the Small-Signal Model

The elements of the bipolar transistor small-signal equivalent circuit considered so far may be considered basic in the sense that they arise directly from essential processes in the device. However, technological limitations in the fabrication of transistors give rise to a number of parasitic elements that must be added to the equivalent circuit for most integrated-circuit transistors. A cross section of a typical *npn* transistor in a junction-isolated process is shown in Fig. 1.19. The means of fabricating such devices is described in Chapter 2.

As described in Section 1.2, all *pn* junctions have a voltage-dependent capacitance associated with the depletion region. In the cross section of Fig. 1.19, three depletion-region capacitances can be identified. The base-emitter junction has a depletion-region capacitance C_{je} and the base-collector and collector-substrate junctions have capacitances C_μ and C_{cs}, respectively. The base-emitter junction closely approximates an abrupt junction due to the steep rise of the doping density caused by the heavy doping in the emitter. Thus the variation of C_{je} with bias voltage is well approximated by (1.21). The collector-base junction behaves like a graded junction for small bias voltages since the doping density is a function of distance near the junction. However, for larger reverse-bias values (more than about a volt), the junction depletion region spreads into the collector, which is uniformly doped, and thus for devices with thick collectors the junction tends to behave like an abrupt junction with uniform doping. Many modern high-speed processes, however, have very thin collector regions (of the order of one micron), and the collector depletion region can extend all the way to the buried layer for quite small reverse-bias voltages. When this occurs, both the depletion region and the associated capacitance vary quite slowly with bias voltage. The collector-base capacitance C_μ thus tends to follow (1.22) for very small bias voltages and (1.21) for large bias voltages in thick-collector devices. In practice, measurements show that the variation of C_μ with bias voltage for most devices can be approximated by

$$C_\mu = \frac{C_{\mu 0}}{\left(1 - \dfrac{V}{\psi_o}\right)^n} \tag{1.117a}$$

where V is the forward bias on the junction and n is an exponent between about 0.2 and 0.5. The third parasitic capacitance in a monolithic *npn* transistor is the collector-substrate capacitance

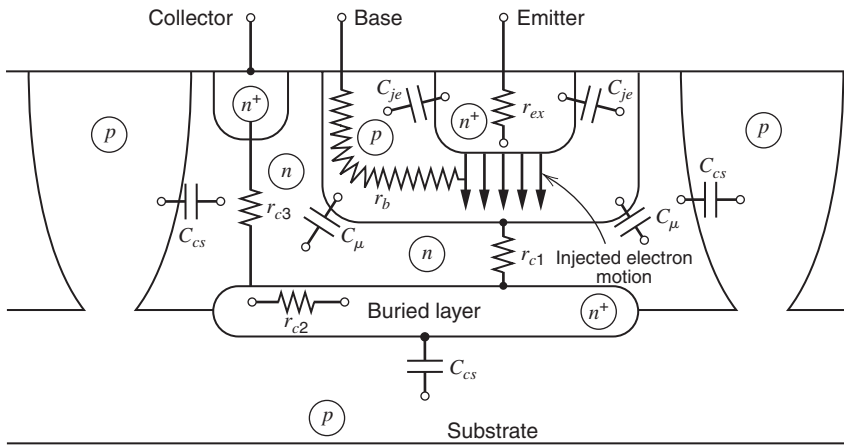


Figure 1.19 Integrated-circuit *npn* bipolar transistor structure showing parasitic elements. (Not to scale.)

C_{cs}, and for large reverse bias voltages this varies according to the abrupt junction equation (1.21) for junction-isolated devices. In the case of oxide-isolated devices, however, the deep p diffusions used to isolate the devices are replaced by oxide. The sidewall component of C_{cs} then consists of a fixed oxide capacitance. Equation 1.117a may then be used to model C_{cs}, but a value of n less than 0.5 gives the best approximation. In general, (1.117a) will be used to model all three parasitic capacitances with subscripts e, c, and s on n and ψ_0 used to differentiate emitter-base, collector-base, and collector-substrate capacitances, respectively. Typical zero-bias values of these parasitic capacitances for a minimum-size npn transistor in a modern oxide-isolated process are $C_{je0} \simeq 10$ fF, $C_{\mu 0} \simeq 10$ fF, and $C_{cs0} \simeq 20$ fF. Values for other devices are summarized in Chapter 2.

As described in Chapter 2, lateral pnp transistors have a parasitic capacitance C_{bs} from base to substrate in place of C_{cs}. Note that the substrate is always connected to the most negative voltage supply in the circuit in order to ensure that all isolation regions are separated by reverse-biased junctions. Thus the substrate is an ac ground, and all parasitic capacitance to the substrate is connected to ground in an equivalent circuit.

The final elements to be added to the small-signal model of the transistor are resistive parasitics. These are produced by the finite resistance of the silicon between the top contacts on the transistor and the active base region beneath the emitter. As shown in Fig. 1.19, there are significant resistances r_b and r_c in series with the base and collector contacts, respectively. There is also a resistance r_{ex} of several ohms in series with the emitter lead that can become important at high bias currents. (Note that the collector resistance r_c is actually composed of three parts labeled r_{c1}, r_{c2}, and r_{c3}.) Typical values of these parameters are $r_b = 50$ to 500 Ω, $r_{ex} = 1$ to 3 Ω, and $r_c = 20$ to 500 Ω. The value of r_b varies significantly with collector current because of *current crowding*.[15] This occurs at high collector currents where the dc base current produces a lateral voltage drop in the base that tends to forward bias the base-emitter junction preferentially around the edges of the emitter. Thus the transistor action tends to occur along the emitter periphery rather than under the emitter itself, and the distance from the base contact to the active base region is reduced. Consequently, the value of r_b is reduced, and in a typical npn transistor, r_b may decrease 50 percent as I_C increases from 0.1 mA to 10 mA.

The value of these parasitic resistances can be reduced by changes in the device structure. For example, a large-area transistor with multiple base and emitter stripes will have a smaller value of r_b. The value of r_c is reduced by inclusion of the low-resistance buried n^+ layer beneath the collector.

The addition of the resistive and capacitive parasitics to the basic small-signal circuit of Fig. 1.18 gives the complete small-signal equivalent circuit of Fig. 1.20. The internal base node

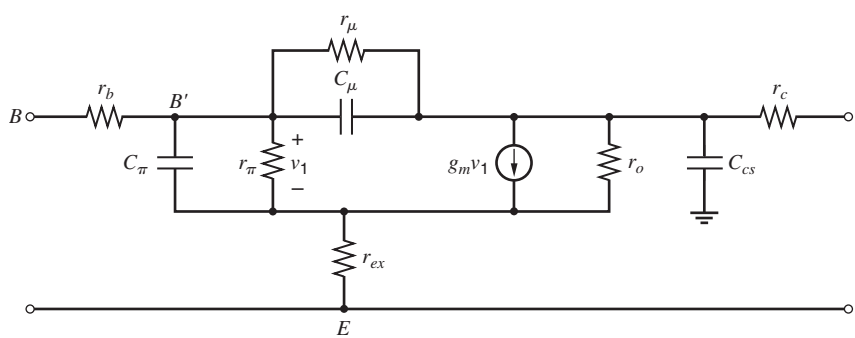


Figure 1.20 Complete bipolar transistor small-signal equivalent circuit.

is labeled B' to distinguish it from the external base contact B. The capacitance C_π contains the base-charging capacitance C_b and the emitter-base depletion layer capacitance C_{je}.

$$C_\pi = C_b + C_{je} \tag{1.118}$$

Note that the representation of parasitics in Fig. 1.20 is an approximation in that lumped elements have been used. In practice, as suggested by Fig. 1.19, C_μ is distributed across r_b and C_{cs} is distributed across r_c. This lumped representation is adequate for most purposes but can introduce errors at very high frequencies. It should also be noted that while the parasitic resistances of Fig. 1.20 can be very important at high bias currents or for high-frequency operation, they are usually omitted from the equivalent circuit for low-frequency calculations, particularly for collector bias currents less than 1 mA.

■ **EXAMPLE**

Derive the complete small-signal equivalent circuit for a bipolar transistor at $I_C = 1$ mA, $V_{CB} = 3$ V, and $V_{CS} = 5$ V. Device parameters are $C_{je0} = 10$ fF, $n_e = 0.5$, $\psi_{0e} = 0.9$ V, $C_{\mu 0} =$ 10 fF, $n_c = 0.3$, $\psi_{0c} = 0.5$ V, $C_{cs0} = 20$ fF, $n_s = 0.3$, $\psi_{0s} = 0.65$ V, $\beta_0 = 100$, $\tau_F = 10$ ps, $V_A = 20$ V, $r_b = 300\ \Omega$, $r_c = 50\ \Omega$, $r_{ex} = 5\ \Omega$, $r_\mu = 10\ \beta_0 r_o$.

Since the base-emitter junction is forward biased, the value of C_{je} is difficult to determine for reasons described in Section 1.2.1. Either a value can be determined by computer or a reasonable estimation is to double C_{je0}. Using the latter approach, we estimate

$$C_{je} = 20 \text{ fF}$$

Using (1.117a) gives, for the collector-base capacitance,

$$C_\mu = \frac{C_{\mu 0}}{\left(1 + \dfrac{V_{CB}}{\psi_{0c}}\right)^{n_c}} = \frac{10}{\left(1 + \dfrac{3}{0.5}\right)^{0.3}} = 5.6 \text{ fF}$$

The collector-substrate capacitance can also be calculated using (1.117a)

$$C_{cs} = \frac{C_{cs0}}{\left(1 + \dfrac{V_{CS}}{\psi_{0s}}\right)^{n_s}} = \frac{20}{\left(1 + \dfrac{5}{0.65}\right)^{0.3}} = 10.5 \text{ fF}$$

From (1.91) the transconductance is

$$g_m = \frac{qI_C}{kT} = \frac{10^{-3}}{26 \times 10^{-3}} \text{A/V} = 38 \text{ mA/V}$$

From (1.104) the base-charging capacitance is

$$C_b = \tau_F g_m = 10 \times 10^{-12} \times 38 \times 10^{-3} \text{ F} = 0.38 \text{ pF}$$

The value of C_π from (1.118) is

$$C_\pi = 0.38 + 0.02 \text{ pF} = 0.4 \text{ pF}$$

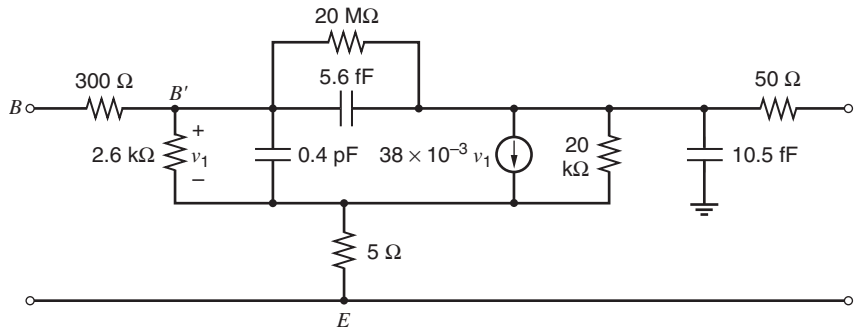


Figure 1.21 Complete small-signal equivalent circuit for a bipolar transistor at $I_C = 1$ mA, $V_{CB} = 3$ V, and $V_{CS} = 5$ V. Device parameters are $C_{je0} = 10$ fF, $n_e = 0.5$, $\psi_{0e} = 0.9$ V, $C_{\mu 0} = 10$ fF, $n_c = 0.3$, $\psi_{0c} = 0.5$ V, $C_{cs0} = 20$ fF, $n_s = 0.3$, $\psi_{0s} = 0.65$ V, $\beta_0 = 100$, $\tau_F = 10$ ps, $V_A = 20$ V, $r_b = 300\Omega$, $r_c = 50\ \Omega$, $r_{ex} = 5\ \Omega$, $r_\mu = 10\beta_0 r_0$.

The input resistance from (1.110) is

$$r_\pi = \frac{\beta_0}{g_m} = 100 \times 26\ \Omega = 2.6\ \text{k}\Omega$$

The output resistance from (1.112) is

$$r_o = \frac{20}{10^{-3}}\ \Omega = 20\ \text{k}\Omega$$

and thus the collector-base resistance is

$$r_\mu = 10\beta_0 r_o = 10 \times 100 \times 20\ \text{k}\Omega = 20\ \text{M}\Omega$$

■ The equivalent circuit with these parameter values is shown in Fig. 1.21.

1.4.8 Specification of Transistor Frequency Response

The high-frequency gain of the transistor is controlled by the capacitive elements in the equivalent circuit of Fig. 1.20. The frequency capability of the transistor is most often specified in practice by determining the frequency where the magnitude of the short-circuit, common-emitter current gain falls to unity. This is called the *transition frequency*, f_T, and is a measure of the maximum useful frequency of the transistor when it is used as an amplifier. The value of f_T can be measured as well as calculated, using the ac circuit of Fig. 1.22. A small-signal current i_i is applied to the base, and the output current i_o is measured with the collector short-circuited for ac signals. A small-signal equivalent circuit can be formed for this situation by using the equivalent circuit of Fig. 1.20 as shown in Fig. 1.23, where r_{ex} and r_μ have been neglected. If r_c is assumed small, then r_o and C_{cs} have no influence, and we have

$$v_1 \simeq \frac{r_\pi}{1 + r_\pi(C_\pi + C_\mu)s} i_i \tag{1.119}$$

If the current fed forward through C_μ is neglected,

$$i_o \simeq g_m v_1 \tag{1.120}$$

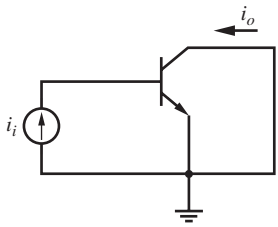


Figure 1.22 Schematic of ac circuit for measurement of f_T.

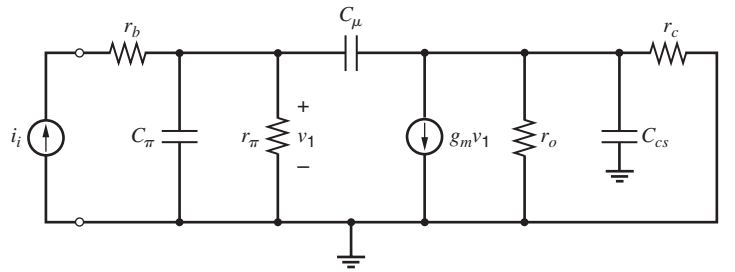


Figure 1.23 Small-signal equivalent circuit for the calculation of f_T.

Substitution of (1.119) in (1.120) gives

$$i_o \simeq i_i \frac{g_m r_\pi}{1 + r_\pi (C_\pi + C_\mu)s}$$

and thus

$$\frac{i_o}{i_i}(j\omega) = \frac{\beta_0}{1 + \beta_0 \dfrac{C_\pi + C_\mu}{g_m} j\omega} \tag{1.121}$$

using (1.110).

Now if $i_o / i_i(j\omega)$ is written as $\beta(j\omega)$ (the high-frequency, small-signal current gain), then

$$\beta(j\omega) = \frac{\beta_0}{1 + \beta_0 \dfrac{C_\pi + C_\mu}{g_m} j\omega} \tag{1.122}$$

At high frequencies the imaginary part of the denominator of (1.122) is dominant, and we can write

$$\beta(j\omega) \simeq \frac{g_m}{j\omega\,(C_\pi + C_\mu)} \tag{1.123}$$

From (1.123), $|\beta(j\omega)| = 1$ when

$$\omega = \omega_T = \frac{g_m}{C_\pi + C_\mu} \tag{1.124}$$

and thus

$$f_T = \frac{1}{2\pi} \frac{g_m}{C_\pi + C_\mu} \tag{1.125}$$

The transistor behavior can be illustrated by plotting $|\beta(j\omega)|$ using (1.122) as shown in Fig. 1.24. The frequency ω_β is defined as the frequency where $|\beta(j\omega)|$ is equal to $\beta_0/\sqrt{2}$

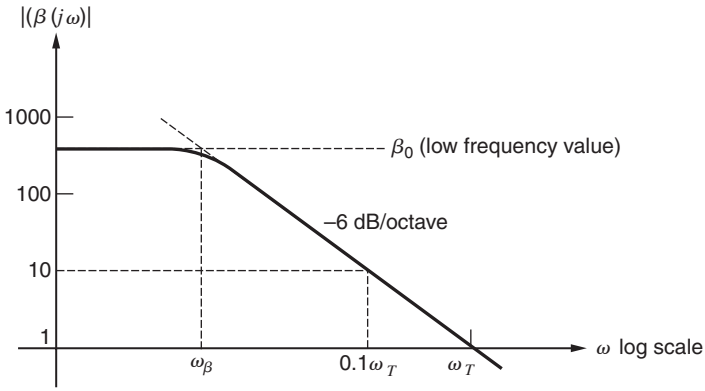


Figure 1.24 Magnitude of small-signal ac current gain $|\beta(j\omega)|$ versus frequency for a typical bipolar transistor.

(3 dB down from the low-frequency value). From (1.122) we have

$$\omega_\beta = \frac{1}{\beta_0}\frac{g_m}{C_\pi + C_\mu} = \frac{\omega_T}{\beta_0} \tag{1.126}$$

From Fig. 1.24 it can be seen that ω_T can be determined by measuring $|\beta(j\omega)|$ at some frequency ω_x where $|\beta(j\omega)|$ is falling at 6 dB/octave and using

$$\omega_T = \omega_x|\beta(j\omega_x)| \tag{1.127}$$

This is the method used in practice, since deviations from ideal behavior tend to occur as $|\beta(j\omega)|$ approaches unity. Thus $|\beta(j\omega)|$ is typically measured at some frequency where its magnitude is about 5 or 10, and (1.127) is used to determine ω_T.

It is interesting to examine the time constant, τ_T, associated with ω_T. This is defined as

$$\tau_T = \frac{1}{\omega_T} \tag{1.128}$$

and use of (1.124) in (1.128) gives

$$\tau_T = \frac{C_\pi}{g_m} + \frac{C_\mu}{g_m} \tag{1.129}$$

Substitution of (1.118) and (1.104) in (1.129) gives

$$\tau_T = \frac{C_b}{g_m} + \frac{C_{je}}{g_m} + \frac{C_\mu}{g_m} = \tau_F + \frac{C_{je}}{g_m} + \frac{C_\mu}{g_m} \tag{1.130}$$

Equation 1.130 indicates that τ_T is dependent on I_C (through g_m) and approaches a constant value of τ_F at high collector bias currents. At low values of I_C, the terms involving C_{je} and C_μ dominate, and they cause τ_T to *rise* and f_T to *fall* as I_C is decreased. This behavior is illustrated in Fig. 1.25, which is a typical plot of f_T versus I_C for an integrated-circuit *npn* transistor. The decline in f_T at high collector currents is not predicted by this simple theory and is due to an increase in τ_F caused by high-level injection and Kirk effect at high currents. These are the same mechanisms that cause a decrease in β_F at high currents as described in Section 1.3.5.

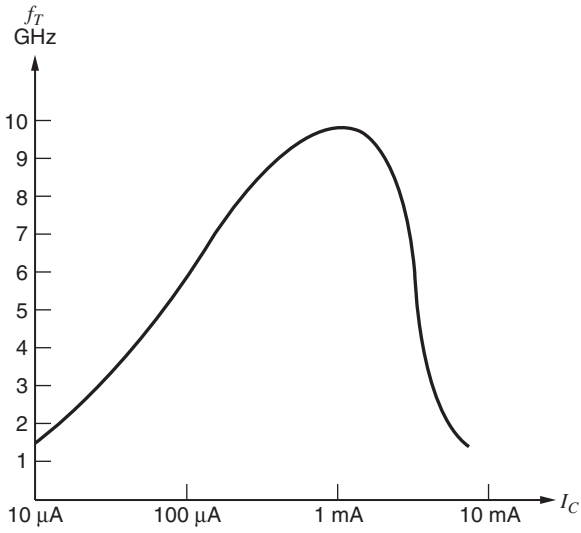


Figure 1.25 Typical curve of f_T versus I_C for an *npn* integrated-circuit transistor with 6 μm^2 emitter area in a high-speed process.

■ **EXAMPLE**

A bipolar transistor has a short-circuit, common-emitter current gain at 1 GHz of 8 with $I_C = 0.25$ mA and 9 with $I_C = 1$ mA. Assuming that high-level injection effects are negligible, calculate C_{je} and τ_F, assuming both are constant. The measured value of C_μ is 10 fF.

From the data, values of f_T are

$$f_{T1} = 8 \times 1 = 8 \text{ GHz} \quad \text{at} \quad I_C = 0.25 \text{ mA}$$
$$f_{T2} = 9 \times 1 = 9 \text{ GHz} \quad \text{at} \quad I_C = 1 \text{ mA}$$

Corresponding values of τ_T are

$$\tau_{T1} = \frac{1}{2\pi f_{T1}} = 19.9 \text{ ps}$$

$$\tau_{T2} = \frac{1}{2\pi f_{T2}} = 17.7 \text{ ps}$$

Using these data in (1.130), we have

$$19.9 \times 10^{-12} = \tau_F + 104(C_\mu + C_{je}) \tag{1.131}$$

at $I_C = 0.25$ mA. At $I_C = 1$ mA we have

$$17.7 \times 10^{-12} = \tau_F + 26(C_\mu + C_{je}) \tag{1.132}$$

Subtraction of (1.132) from (1.131) yields

$$C_\mu + C_{je} = 28.2 \text{ fF}$$

Since C_μ was measured as 10 fF, the value of C_{je} is given by

$$C_{je} \simeq 18.2 \text{ fF}$$

Substitution in (1.131) gives

$$\tau_F = 17 \text{ ps}$$

This is an example of how basic device parameters can be determined from high-frequency current-gain measurements. Note that the assumption that C_{je} is constant is a useful approximation in practice because V_{BE} changes by only 36 mV as I_C increases from 0.25 mA to 1 mA. ■

1.5 Large-Signal Behavior of Metal-Oxide-Semiconductor Field-Effect Transistors

Metal-oxide-semiconductor field-effect transistors (MOSFETs) have become dominant in the area of digital integrated circuits because they allow high density and low power dissipation. In contrast, bipolar transistors still provide many advantages in stand-alone analog integrated circuits. For example, the transconductance per unit bias current in bipolar transistors is usually much higher than in MOS transistors. So in systems where analog techniques are used on some integrated circuits and digital techniques on others, bipolar technologies are often preferred for the analog integrated circuits and MOS technologies for the digital. To reduce system cost and increase portability, both increased levels of integration and reduced power dissipation are required, forcing the associated analog circuits to use MOS-compatible technologies. One way to achieve these goals is to use a processing technology that provides both bipolar and MOS transistors, allowing great design flexibility. However, all-MOS processes are less expensive than combined bipolar and MOS processes. Therefore, economic considerations drive integrated-circuit manufacturers to use all-MOS processes in many practical cases. As a result, the study of the characteristics of MOS transistors that affect analog integrated-circuit design is important.

1.5.1 Transfer Characteristics of MOS Devices

A cross section of a typical enhancement-mode n-channel MOS (NMOS) transistor is shown in Fig. 1.26. Heavily doped n-type source and drain regions are fabricated in a p-type substrate (often called the body). A thin layer of silicon dioxide is grown over the substrate material and

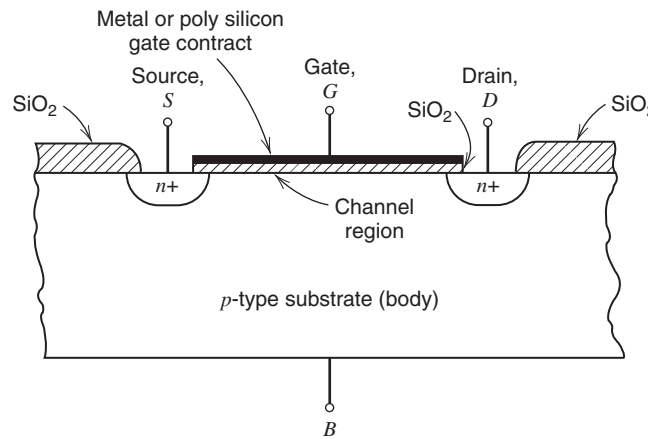


Figure 1.26 Typical enhancement-mode NMOS structure.

a conductive gate material (metal or polycrystalline silicon) covers the oxide between source and drain. Note that the gate is horizontal in Fig. 1.26, and we will use this orientation in all descriptions of the physical operation of MOS devices. In operation, the gate-source voltage modifies the conductance of the region under the gate, allowing the gate voltage to control the current flowing between source and drain. This control can be used to provide gain in analog circuits and switching characteristics in digital circuits.

The enhancement-mode NMOS device of Fig. 1.26 shows significant conduction between source and drain only when an n-type channel exists under the gate. This observation is the origin of the *n-channel* designation. The term *enhancement mode* refers to the fact that no conduction occurs for $V_{GS} = 0$. Thus, the channel must be *enhanced* to cause conduction. MOS devices can also be made by using an n-type substrate with a p-type conducting channel. Such devices are called enhancement-mode p-channel MOS (PMOS) transistors. In complementary MOS (CMOS) technology, both device types are present.

The derivation of the transfer characteristics of the enhancement-mode NMOS device of Fig. 1.26 begins by noting that with $V_{GS} = 0$, the source and drain regions are separated by back-to-back pn junctions. These junctions are formed between the n-type source and drain regions and the p-type substrate, resulting in an extremely high resistance (about $10^{12}\ \Omega$) between drain and source when the device is off.

Now consider the substrate, source, and drain grounded with a positive voltage V_{GS} applied to the gate as shown in Fig. 1.27. The gate and substrate then form the plates of a capacitor with the SiO_2 as a dielectric. Positive charge accumulates on the gate and negative charge in the substrate. Initially, the negative charge in the p-type substrate is manifested by the creation of a *depletion region* and the exclusion of holes under the gate as described in Section 1.2 for a pn-junction. The depletion region is shown in Fig. 1.27. The results of Section 1.2 can now be applied. Using (1.10), the depletion-layer width X under the oxide is

$$X = \left(\frac{2\epsilon\phi}{qN_A}\right)^{1/2} \tag{1.133}$$

where ϕ is the potential in the depletion layer at the oxide-silicon interface, N_A is the doping density (assumed constant) of the p-type substrate in atoms/cm^3, and ϵ is the permittivity of the silicon. The charge per area in this depletion region is

$$Q = qN_A X = \sqrt{2qN_A\epsilon\phi} \tag{1.134}$$

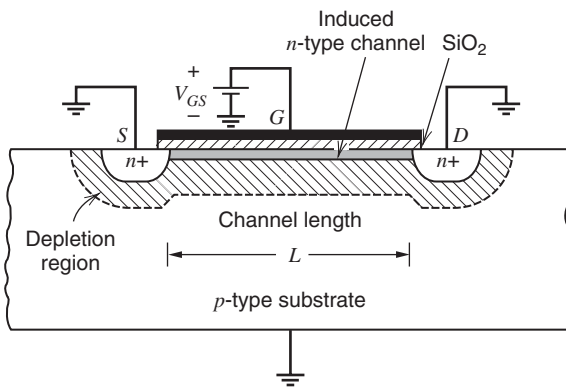


Figure 1.27 Idealized NMOS device cross section with positive V_{GS} applied, showing depletion regions and the induced channel.

When the surface potential in the silicon reaches a critical value equal to twice the Fermi level ϕ_f, a phenomenon known as *inversion* occurs.[16] The Fermi level ϕ_f is defined as

$$\phi_f = \frac{kT}{q} \ln \left[\frac{N_A}{n_i} \right] \tag{1.135}$$

where k is Boltzmann's constant. Also, n_i is the intrinsic carrier concentration, which is

$$n_i = \sqrt{N_c N_v} \exp \left(-\frac{E_g}{2kT} \right) \tag{1.136}$$

where E_g is the band gap of silicon at $T = 0°\text{K}$, N_c is the density of allowed states near the edge of the conduction band, and N_v is the density of allowed states near the edge of the valence band, respectively. The Fermi level ϕ_f is usually about 0.3 V. After the potential in the silicon reaches $2\phi_f$, further increases in gate voltage produce no further changes in the depletion-layer width but instead induce a thin layer of electrons in the depletion layer at the surface of the silicon directly under the oxide. Inversion produces a continuous n-type region with the source and drain regions and forms the conducting channel between source and drain. The conductivity of this channel can be modulated by increases or decreases in the gate-source voltage. In the presence of an inversion layer, and without substrate bias, the depletion region contains a fixed charge density

$$Q_{b0} = \sqrt{2qN_A \epsilon 2\phi_f} \tag{1.137}$$

If a substrate bias voltage V_{SB} (positive for n-channel devices) is applied between the source and substrate, the potential required to produce inversion becomes $(2\phi_f + V_{SB})$, and the charge density stored in the depletion region in general is

$$Q_b = \sqrt{2qN_A \epsilon (2\phi_f + V_{SB})} \tag{1.138}$$

The gate-source voltage V_{GS} required to produce an inversion layer is called the threshold voltage V_t and can now be calculated. This voltage consists of several components. First, a voltage $[2\phi_f + (Q_b/C_{ox})]$ is required to sustain the depletion-layer charge Q_b, where C_{ox} is the gate oxide capacitance per unit area. Second, a work-function difference ϕ_{ms} exists between the gate metal and the silicon. Third, positive charge density Q_{ss} always exists in the oxide at the silicon interface. This charge is caused by crystal discontinuities at the Si-SiO_2 interface and must be compensated by a gate-source voltage contribution of $-Q_{ss}/C_{ox}$. Thus we have a threshold voltage

$$V_t = \phi_{ms} + 2\phi_f + \frac{Q_b}{C_{ox}} - \frac{Q_{ss}}{C_{ox}} \tag{1.139}$$

$$= \phi_{ms} + 2\phi_f + \frac{Q_{b0}}{C_{ox}} - \frac{Q_{ss}}{C_{ox}} + \frac{Q_b - Q_{b0}}{C_{ox}}$$

$$= V_{t0} + \gamma \left(\sqrt{2\phi_f + V_{SB}} - \sqrt{2\phi_f} \right) \tag{1.140}$$

where (1.137) and (1.138) have been used, and V_{t0} is the threshold voltage with $V_{SB} = 0$. The parameter γ is defined as

$$\gamma = \frac{1}{C_{ox}} \sqrt{2q\epsilon N_A} \tag{1.141}$$

and

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{1.142}$$

where ϵ_{ox} and t_{ox} are the permittivity and the thickness of the oxide, respectively. A typical value of γ is 0.5 $\text{V}^{1/2}$, and $C_{ox} = 3.45$ fF/μm^2 for $t_{ox} = 100$ angstroms.

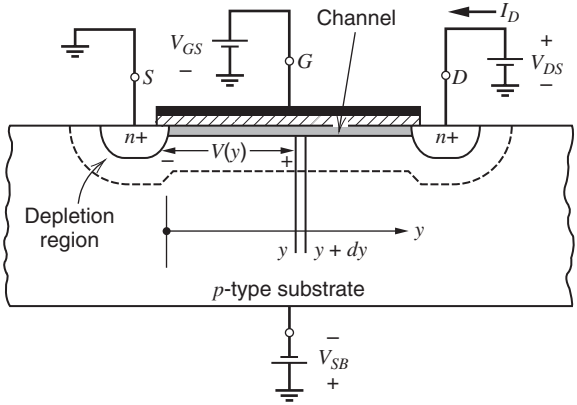


Figure 1.28 NMOS device with bias voltages applied.

In practice, the value of V_{t0} is usually adjusted in processing by implanting additional impurities into the channel region. Extra p-type impurities are implanted in the channel to set V_{t0} between 0.3 V and 1.5 V for n-channel enhancement devices. By implanting n-type impurities in the channel region, a conducting channel can be formed even for $V_{GS} = 0$, forming a *depletion* device with typical values of V_{t0} in the range -1 V to -4 V. If Q_i is the charge density due to the implant, then the threshold voltage given by (1.139) is shifted by approximately Q_i/C_{ox}.

The preceding equations can now be used to calculate the large-signal characteristics of an n-channel MOSFET. In this analysis, the source is assumed grounded and bias voltages V_{GS}, V_{DS}, and V_{SB} are applied as shown in Fig. 1.28. If $V_{GS} > V_t$, inversion occurs and a conducting channel exists. The channel conductivity is determined by the vertical electric field, which is controlled by the value of $(V_{GS} - V_t)$. If $V_{DS} = 0$, the current I_D that flows from drain to source is zero because the horizontal electric field is zero. Nonzero V_{DS} produces a horizontal electric field and causes current I_D to flow. The value of the current depends on both the horizontal and the vertical electric fields, explaining the term *field-effect* transistor. Positive voltage V_{DS} causes the reverse bias from the drain to the substrate to be larger than from the source to substrate, and thus the widest depletion region exists at the drain. For simplicity, however, we assume that the voltage drop along the channel itself is small so that the depletion-layer width is constant along the channel.

The drain current I_D is

$$I_D = \frac{dQ}{dt} \tag{1.143}$$

where dQ is the incremental channel charge at a distance y from the source in an incremental length dy of the channel, and dt is the time required for this charge to cross length dy. The charge dQ is

$$dQ = Q_I W dy \tag{1.144}$$

where W is the width of the device perpendicular to the plane of Fig. 1.28 and Q_I is the induced electron charge per unit area of the channel. At a distance y along the channel, the voltage with respect to the source is $V(y)$ and the gate-to-channel voltage at that point is $V_{GS} - V(y)$. We assume this voltage exceeds the threshold voltage V_t. Thus the induced electron charge per unit area in the channel is

$$Q_I(y) = C_{ox}[V_{GS} - V(y) - V_t] \tag{1.145}$$

Also,

$$dt = \frac{dy}{v_d(y)} \tag{1.146}$$

where v_d is the electron drift velocity at a distance y from the source. Combining (1.144) and (1.146) gives

$$I_D = WQ_I(y)v_d(y) \tag{1.147}$$

The drift velocity is determined by the horizontal electric field. When the horizontal electric field $\mathscr{E}(y)$ is small, the drift velocity is proportional to the field and

$$v_d(y) = \mu_n \mathscr{E}(y) \tag{1.148}$$

where the constant of proportionality μ_n is the average electron mobility in the channel. In practice, the mobility depends on both the temperature and the doping level but is almost constant for a wide range of normally used doping levels. Also, μ_n is sometimes called the surface mobility for electrons because the channel forms at the surface of the silicon. Typical values range from about 500 cm^2/(V-s) to about 700 cm^2/(V-s), which are much less than the mobility of electrons in the bulk of the silicon (about 1400 cm^2/V-s) because surface defects not present in the bulk impede the flow of electrons in MOS transistors.[17] The electric field $\mathscr{E}(y)$ is

$$\mathscr{E}(y) = \frac{dV}{dy} \tag{1.149}$$

where dV is the incremental voltage drop along the length of channel dy at a distance y from the source. Substituting (1.145), (1.148), and (1.149) into (1.147) gives

$$I_D = WC_{ox}[V_{GS} - V - V_t]\mu_n \frac{dV}{dy} \tag{1.150}$$

Separating variables and integrating gives

$$\int_0^L I_D \, dy = \int_0^{V_{DS}} W\mu_n C_{ox}(V_{GS} - V - V_t) \, dV \tag{1.151}$$

Carrying out this integration gives

$$I_D = \frac{k'}{2}\frac{W}{L}[2(V_{GS} - V_t)V_{DS} - V_{DS}^2] \tag{1.152}$$

where

$$k' = \mu_n C_{ox} = \frac{\mu_n \epsilon_{ox}}{t_{ox}} \tag{1.153}$$

When $V_{DS} \ll 2(V_{GS} - V_t)$, (1.152) predicts that I_D is approximately proportional to V_{DS}. This result is reasonable because the average horizontal electric field in this case is V_{DS}/L, and the average drift velocity of electrons is proportional to the average field when the field is small. Equation 1.152 is important and describes the *I-V* characteristics of an MOS transistor, assuming a continuous induced channel. A typical value of k' for $t_{ox} = 100$ angstroms is about 200 $\mu\text{A/V}^2$ for an *n*-channel device.

As the value of V_{DS} is increased, the induced conducting channel narrows at the drain end and (1.145) indicates that Q_I at the drain end approaches zero as V_{DS} approaches $(V_{GS} - V_t)$. That is, the channel is no longer connected to the drain when $V_{DS} > V_{GS} - V_t$. This phenomenon is called *pinch-off* and can be understood by writing a KVL equation around the transistor:

$$V_{DS} = V_{DG} + V_{GS} \tag{1.154}$$

Therefore, when $V_{DS} > V_{GS} - V_t$,

$$V_{DG} + V_{GS} > V_{GS} - V_t \tag{1.155}$$

Rearranging (1.155) gives

$$V_{GD} < V_t \tag{1.156}$$

Equation 1.156 shows that when drain-source voltage is greater than $(V_{GS} - V_t)$, the gate-drain voltage is less than a threshold, which means that the channel no longer exists at the drain. This result is reasonable because we know that the gate-to-channel voltage at the point where the channel disappears is equal to V_t by the definition of the threshold voltage. Therefore, at the point where the channel pinches off, the channel voltage is $(V_{GS} - V_t)$. As a result, the average horizontal electric field across the channel in pinch-off does not depend on the the drain-source voltage but instead on the voltage across the channel, which is $(V_{GS} - V_t)$. Therefore, (1.152) is no longer valid if $V_{DS} > V_{GS} - V_t$. The value of I_D in this region is obtained by substituting $V_{DS} = V_{GS} - V_t$ in (1.152), giving

$$I_D = \frac{k'}{2}\frac{W}{L}(V_{GS} - V_t)^2 \tag{1.157}$$

Equation 1.157 predicts that the drain current is independent of V_{DS} in the pinch-off region. In practice, however, the drain current in the pinch-off region varies slightly as the drain voltage is varied. This effect is due to the presence of a depletion region between the physical pinch-off point in the channel at the drain end and the drain region itself. If this depletion-layer width is X_d, then the *effective* channel length is given by

$$L_{\text{eff}} = L - X_d \tag{1.158}$$

If L_{eff} is used in place of L in (1.157), we obtain a more accurate formula for current in the pinch-off region

$$I_D = \frac{k'}{2}\frac{W}{L_{\text{eff}}}(V_{GS} - V_t)^2 \tag{1.159}$$

Because X_d (and thus L_{eff}) are functions of the drain-source voltage in the pinch-off region, I_D varies with V_{DS}. This effect is called *channel-length modulation*. Using (1.158) and (1.159), we obtain

$$\frac{\partial I_D}{\partial V_{DS}} = -\frac{k'}{2}\frac{W}{L_{\text{eff}}^2}(V_{GS} - V_t)^2\frac{dL_{\text{eff}}}{dV_{DS}} \tag{1.160}$$

and thus

$$\frac{\partial I_D}{\partial V_{DS}} = \frac{I_D}{L_{\text{eff}}}\frac{dX_d}{dV_{DS}} \tag{1.161}$$

This equation is analogous to (1.55) for bipolar transistors. Following a similar procedure, the Early voltage can be defined as

$$V_A = \frac{I_D}{\partial I_D/\partial V_{DS}} \tag{1.162}$$

and thus

$$V_A = L_{\text{eff}}\left(\frac{dX_d}{dV_{DS}}\right)^{-1} \tag{1.163}$$

For MOS transistors, a commonly used parameter for the characterization of channel-length modulation is the reciprocal of the Early voltage,

$$\lambda = \frac{1}{V_A} \tag{1.164}$$

As in the bipolar case, the large-signal properties of the transistor can be approximated by assuming that λ and V_A are constants, independent of the bias conditions. Thus we can include the effect of channel-length modulation in the *I-V* characteristics by modifying (1.157) to

$$I_D = \frac{k'}{2}\frac{W}{L}(V_{GS} - V_t)^2\left(1 + \frac{V_{DS}}{V_A}\right) = \frac{k'}{2}\frac{W}{L}(V_{GS} - V_t)^2(1 + \lambda V_{DS}) \tag{1.165}$$

In practical MOS transistors, variation of X_d with voltage is complicated by the fact that the field distribution in the drain depletion region is not one-dimensional. As a result, the calculation of λ from the device structure is quite difficult,[18] and developing effective values of λ from experimental data is usually necessary. The parameter λ is inversely proportional to the effective channel length and a decreasing function of the doping level in the channel. Typical values of λ are in the range 0.05 V^{-1} to 0.005 V^{-1}.

Plots of I_D versus V_{DS} with V_{GS} as a parameter are shown in Fig. 1.29 for an NMOS transistor. The device operates in the pinch-off region when $V_{DS} > (V_{GS} - V_t)$. The pinch-off region for MOS devices is often called the *saturation* region. In saturation, the output characteristics are almost flat, which shows that the current depends mostly on the gate-source voltage and only to a small extent on the drain-source voltage. On the other hand, when $V_{DS} < (V_{GS} - V_t)$, the device operates in the *Ohmic* or *triode* region, where the device can be modeled as a nonlinear voltage-controlled resistor connected between the drain and source. The resistance of this resistor is *nonlinear* because the V_{DS}^2 term in (1.152) causes the resistance to depend on V_{DS}. Since this term is small when V_{DS} is small, however, the nonlinearity is also small when V_{DS} is small, and the triode region is also sometimes called the *linear* region. The boundary between the triode and saturation regions occurs when $V_{DS} = (V_{GS} - V_t)$. On this boundary, both (1.152) and (1.157) correctly predict I_D. Since $V_{DS} = (V_{GS} - V_t)$ along the boundary between triode and saturation, (1.157) shows that the boundary is $I_D = (k'/2)(W/L)V_{DS}^2$. This parabolic function of V_{DS} is shown in Fig. 1.29. For depletion *n*-channel MOS devices, V_t is negative, and I_D is nonzero even for $V_{GS} = 0$. For PMOS devices, all polarities of voltages and currents are reversed.

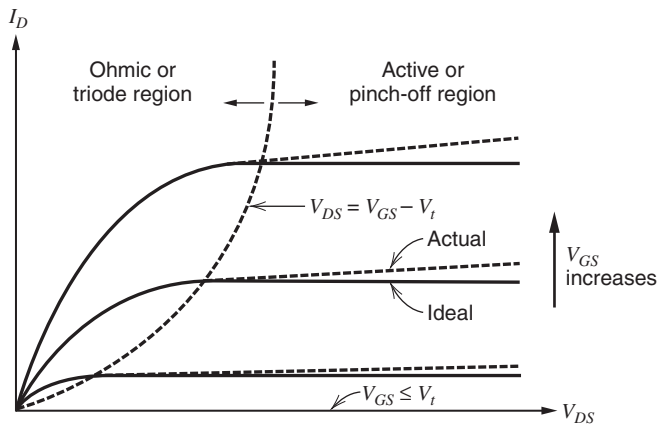


Figure 1.29 NMOS device characteristics.

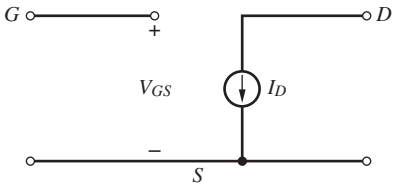


Figure 1.30 Large-signal model for the NMOS transistor.

The results derived above can be used to form a large-signal model of the NMOS transistor in saturation. The model topology is shown in Fig. 1.30, where I_D is given by (1.152) in the triode region and (1.157) in saturation, ignoring the effect of channel-length modulation. To include the effect of channel-length modulation, (1.159) or (1.165) should be used instead of (1.157) to find the drain current in saturation.

1.5.2 Comparison of Operating Regions of Bipolar and MOS Transistors

Notice that the meaning of the word *saturation* for MOS transistors is quite different than for bipolar transistors. Saturation in bipolar transistors refers to the region of operation where both junctions are forward biased and the collector-emitter voltage is approximately constant or saturated. On the other hand, saturation in MOS transistors refers to the region of operation where the channel is attached only to the source but not to the drain and the current is approximately constant or saturated. To avoid confusion, the term *active region* will be used in this book to describe the flat region of the MOS transistor characteristics, as shown in Fig. 1.29. This wording is selected to form a link between the operation of MOS and bipolar transistors. This link is summarized in the table of Fig. 1.31, which reviews the operating regions of *npn* bipolar and *n*-channel MOS transistors.

When the emitter junction is forward biased and the collector junction is reverse biased, bipolar transistors operate in the forward-active region. They operate in the reverse-active region when the collector junction is forward biased and the emitter junction is reverse biased. This distinction is important because integrated-circuit bipolar transistors are typically not symmetrical in practice; that is, the collector operates more efficiently as a collector of minority carriers than as an emitter. Similarly, the emitter operates more efficiently as an emitter of minority carriers than as a collector. One reason for this asymmetry is that the collector region surrounds the emitter region in integrated-circuit bipolar transistors, as shown in Fig. 1.19. A consequence of this asymmetry is that the current gain in the forward-active region β_F is usually much greater than the current gain in the reverse-active region β_R.

In contrast, the source and drain of MOS transistors are completely interchangeable based on the preceding description. (In practice, the symmetry is good but not perfect.) Therefore, distinguishing between the forward-active and reverse-active regions of operation of an MOS transistor is not necessary.

npn Bipolar Transistor			*n-channel MOS Transistor*		
Region	V_{BE}	V_{BC}	**Region**	V_{GS}	V_{GD}
Cutoff	$< V_{BE(\text{on})}$	$< V_{BC(\text{on})}$	Cutoff	$< V_t$	$< V_t$
Forward Active	$\geq V_{BE(\text{on})}$	$< V_{BC(\text{on})}$	Saturation(Active)	$\geq V_t$	$< V_t$
Reverse Active	$< V_{BE(\text{on})}$	$\geq V_{BC(\text{on})}$	Saturation(Active)	$< V_t$	$\geq V_t$
Saturation	$\geq V_{BE(\text{on})}$	$\geq V_{BC(\text{on})}$	Triode	$\geq V_t$	$\geq V_t$

Figure 1.31 Operating regions of *npn* bipolar and *n*-channel MOS transistors.

Figure 1.31 also shows that *npn* bipolar transistors operate in cutoff when both junctions are reversed biased. Similarly, MOS transistors operate in cutoff when the gate is biased so that inversion occurs at neither the source nor the drain. Furthermore, *npn* transistors operate in saturation when both junctions are forward biased, and MOS transistors operate in the triode region when the gate is biased so that the channel is connected to both the source and the drain. Therefore, this comparison leads us to view the voltage required to invert the surface of an MOS transistor as analogous to the voltage required to forward bias a *pn* junction in a bipolar transistor. To display this analogy, we will use the circuit symbols in Fig. 1.32*a* to represent MOS transistors. These symbols are intentionally chosen to appear similar to the symbols of the corresponding bipolar transistors. In bipolar-transistor symbols, the arrow at the emitter junction represents the direction of current flow when the emitter junction is forward biased. In MOS transistors, the *pn* junctions between the source and body and the drain and body are reverse biased for normal operation. Therefore, the arrows in Fig. 1.32*a* do not indicate *pn* junctions. Instead, they indicate the direction of current flow when the terminals are biased so that the terminal labeled as the drain operates as the drain and the terminal labeled as the source operates as the source. In NMOS transistors, the source is the source of electrons; therefore, the source operates at a lower voltage than the drain, and the current flows in a direction opposite that of the electrons in the channel. In PMOS transistors, the source is the source of holes; therefore, the source operates at a higher voltage than the drain, and the current flows in the same direction as the holes in the channel.

In CMOS technology, one device type is fabricated in the substrate, which is common to all devices, invariably connected to a dc power-supply voltage, and usually not shown on the circuit diagram. The other device type, however, is fabricated in separate isolation regions called *wells*, which may or may not be connected together and which may or may not be connected to a power-supply voltage. If these isolation regions are connected to the appropriate power supply, the symbols of Fig. 1.32*a* will be used, and the substrate connection will not

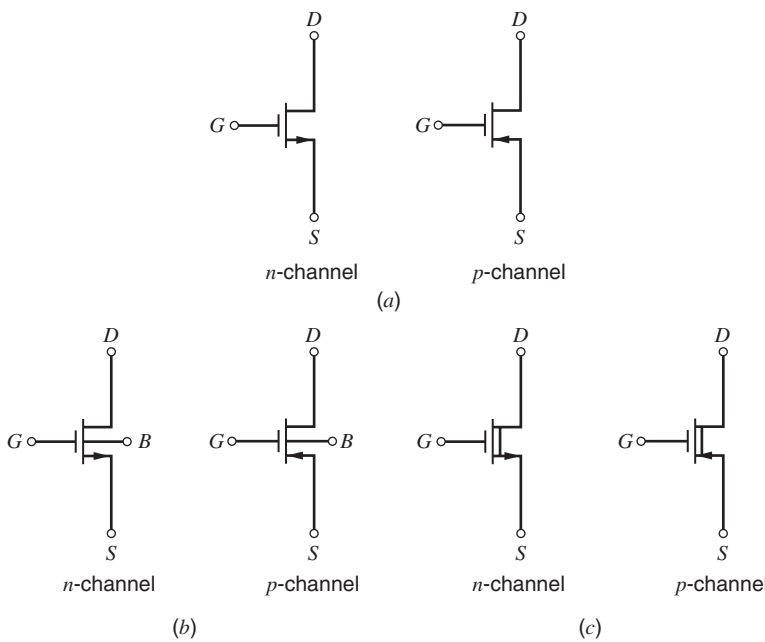


Figure 1.32 (*a*) NMOS and PMOS symbols used in CMOS circuits. (*b*) NMOS and PMOS symbols used when the substrate connection is nonstandard. (*c*) Depletion MOS device symbols.

be shown. On the other hand, if the individual isolation regions are connected elsewhere, the devices will be represented by the symbols of Fig. 1.32*b*, where the substrate is labeled *B*. Finally, symbols for depletion-mode devices, for which a channel forms for $V_{GS} = 0$, are shown in Fig. 1.32*c*.

1.5.3 Decomposition of Gate-Source Voltage

The gate-source voltage of a given MOS transistor is usually separated into two parts: the threshold, V_t, and the voltage over the threshold, $V_{GS} - V_t$. We will refer to this latter part of the gate-source voltage as the *overdrive*. This decomposition is used because these two components of the gate-source voltage have different properties. Assuming square-law behavior as in (1.157), the overdrive is

$$V_{ov} = V_{GS} - V_t = \sqrt{\frac{2I_D}{k'(W/L)}} \tag{1.166}$$

Since the transconductance parameter k' is proportional to mobility, and since mobility falls with increasing temperature, the overdrive rises with temperature. In contrast, the next section shows that the threshold falls with increasing temperature. Furthermore, (1.140) shows that the threshold depends on the source-body voltage, but not on the current; (1.166) shows that the overdrive depends directly on the current, but not on the source-body voltage.

1.5.4 Threshold Temperature Dependence

Assume that the source-body voltage is zero. Substituting (1.138) into (1.139) gives

$$V_t = \frac{\sqrt{2qN_A\epsilon(2\phi_f)}}{C_{ox}} + 2\phi_f + \phi_{ms} - \frac{Q_{ss}}{C_{ox}} \tag{1.167}$$

Assume that ϕ_{ms}, Q_{ss}, and C_{ox} are independent of temperature.[19] Then differentiating (1.167) gives

$$\frac{dV_t}{dT} = \frac{\sqrt{2qN_A\epsilon(2)}}{2C_{ox}\sqrt{\phi_f}}\frac{d\phi_f}{dT} + 2\frac{d\phi_f}{dT} = \frac{d\phi_f}{dT}\left[2 + \frac{1}{C_{ox}}\sqrt{\frac{qN_A\epsilon}{\phi_f}}\right] \tag{1.168}$$

Substituting (1.136) into (1.135) gives

$$\phi_f = \frac{kT}{q}\ln\left[\frac{N_A \exp\left(\dfrac{E_g}{2kT}\right)}{\sqrt{N_c N_v}}\right] \tag{1.169}$$

Assume both N_c and N_v are independent of temperature.[20] Then differentiating (1.169) gives

$$\frac{d\phi_f}{dT} = \frac{kT}{q}\left[-\frac{E_g}{2kT^2}\right] + \frac{k}{q}\ln\left[\frac{N_A \exp\left(\dfrac{E_g}{2kT}\right)}{\sqrt{N_c N_v}}\right] \tag{1.170}$$

Substituting (1.169) into (1.170) and simplifying gives

$$\frac{d\phi_f}{dT} = -\frac{E_g}{2qT} + \frac{\phi_f}{T} = -\frac{1}{T}\left[\frac{E_g}{2q} - \phi_f\right] \tag{1.171}$$

Substituting (1.141) and (1.171) into (1.168) gives

$$\frac{dV_t}{dT} = -\frac{1}{T}\left[\frac{E_g}{2q} - \phi_f\right]\left[2 + \frac{\gamma}{\sqrt{2\phi_f}}\right] \tag{1.172}$$

Equation 1.172 shows that the threshold voltage falls with increasing temperature if $\phi_f < E_g/(2q)$. The slope is usually in the range of -0.5 mV/°C to -4 mV/°C.[21]

■ **EXAMPLE**

Assume $T = 300°\text{K}$, $N_A = 10^{15}$ cm^{-3}, and $t_{ox} = 100$ Å. Find dV_t/dT.
From (1.135),

$$\phi_f = (25.8 \text{ mV}) \ln\left(\frac{10^{15}\text{cm}^{-3}}{1.45 \times 10^{10}\text{cm}^{-3}}\right) = 287 \text{ mV} \tag{1.173}$$

Also

$$\frac{E_g}{2q} = \frac{1.12 \text{ eV}}{2q} = 0.56 \text{ V} \tag{1.174}$$

Substituting (1.173) and (1.174) into (1.171) gives

$$\frac{d\phi_f}{dT} = -\frac{1}{300}(560 - 287)\frac{\text{mV}}{°\text{K}} = -0.91\frac{\text{mV}}{°\text{K}} \tag{1.175}$$

From (1.142),

$$C_{ox} = \frac{3.9\left(8.854 \times 10^{-14} \text{ F/cm}\right)}{100 \times 10^{-8} \text{ cm}} = 3.45 \frac{\text{fF}}{\mu\text{m}^2} \tag{1.176}$$

Also,

$$\begin{aligned}\frac{\gamma}{\sqrt{2\phi_f}} &= \frac{1}{C_{ox}}\sqrt{\frac{(2)(1.6 \times 10^{-19} \text{ C})(11.7)(8.854 \times 10^{-14} \text{ F/cm})(10^{15} \text{ cm}^{-3})}{(2)(0.287 \text{ V})}} \\ &= \frac{2.4 \times 10^{-8} \text{ F/cm}^2}{3.45 \times 10^{-15} \text{ F/}\mu\text{m}^2} = \frac{2.4 \times 10^{-16} \text{ F/}\mu\text{m}^2}{3.45 \times 10^{-15} \text{ F/}\mu\text{m}^2} = 0.07\end{aligned} \tag{1.177}$$

Substituting (1.173) – (1.177) into (1.172) gives

$$\frac{dV_t}{dT} = \left(-0.91\frac{\text{mV}}{°\text{K}}\right)(2 + 0.07) \simeq -1.9\frac{\text{mV}}{°\text{K}} = -1.9\frac{\text{mV}}{°\text{C}} \tag{1.178}$$

■

1.5.5 MOS Device Voltage Limitations

The main voltage limitations in MOS transistors are described next.[22,23] Some of these limitations have a strong dependence on the gate length L; others have little dependence on L. Also, some of the voltage limitations are inherently destructive; others cause no damage as long as overheating is avoided.

Junction Breakdown. For long channel lengths, the drain-depletion region has little effect on the channel, and the I_D-versus-V_{DS} curves closely follow the ideal curves of Fig. 1.29. For increasing V_{DS}, however, eventually the drain-substrate *pn*-junction breakdown voltage is exceeded, and the drain current increases abruptly by avalanche breakdown as described in Section 1.2.2. This phenomenon is not inherently destructive.

Punchthrough. If the depletion region around the drain in an MOS transistor touches the depletion region around the source before junction breakdown occurs, increasing the drain-source voltage increases the drain current by reducing the barrier to electron flow between the source and drain. This phenomenon is called *punchthrough*. Since it depends on the two depletion regions touching, it also depends on the gate length. Punchthrough is not inherently destructive and causes a more gradual increase in the drain current than is caused by avalanche breakdown. Punchthrough normally occurs below the surface of the silicon and is often prevented by an extra ion implantation below the surface to reduce the size of the depletion regions.

Hot Carriers. With sufficient horizontal or vertical electric fields, electrons or holes may reach sufficient velocities to be injected into the oxide, where most of them increase the gate current and some of them become trapped. Such carriers are called *hot* because the required velocity for injection into the oxide is usually greater than the random thermal velocity. Carriers trapped in the oxide shift the threshold voltage and may cause a transistor to remain on when it should turn off or vice versa. In this sense, injection of hot carriers into the oxide is a destructive process. This process is most likely to be problematic in short-channel technologies, where horizontal electric fields are likely to be high.

Oxide Breakdown. In addition to V_{DS} limitations, MOS devices must also be protected against excessive gate voltages. Typical gate oxides break down with an electric field of about 6×10^6 V/cm to 7×10^6 V/cm,[24,25] which corresponds to 6 to 7 V applied from gate to channel with an oxide thickness of 100 angstroms. Since this process depends on the vertical electrical field, it is independent of channel length. However, this process is destructive to the transistor, resulting in resistive connections between the gate and the channel. Oxide breakdown can be caused by static electricity and can be avoided by using *pn* diodes and resistors to limit the voltage range at sensitive nodes internal to the integrated circuit that connect to bonding pads.

1.6 Small-Signal Models of MOS Transistors

As mentioned in Section 1.5, MOS transistors are often used in analog circuits. To simplify the calculation of circuit gain and terminal impedances, *small-signal* models can be used. As in the case for bipolar transistors, a hierarchy of models with increasing complexity can be derived, and choosing the simplest model required to do a given analysis is important in practice.

Consider the MOS transistor in Fig. 1.33 with bias voltages V_{GS} and V_{DD} applied as shown. These bias voltages produce quiescent drain current I_D. If $V_{GS} > V_t$ and $V_{DD} > (V_{GS} - V_t)$, the device operates in the saturation or active region. A small-signal input voltage v_i is applied in series with V_{GS} and produces a small variation in drain current i_d. The total value of the drain current is $I_d = (I_D + i_d)$.

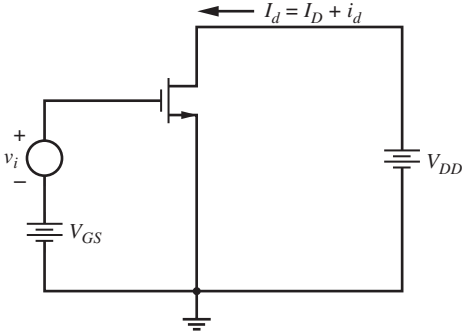


Figure 1.33 Schematic of an MOS transistor with biasing.

1.6.1 Transconductance

Assuming square-law operation, the transconductance from the gate can be determined from (1.165) by differentiating

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = k' \frac{W}{L}(V_{GS} - V_t)(1 + \lambda V_{DS}) \tag{1.179}$$

If $\lambda V_{DS} \ll 1$, (1.179) simplifies to

$$g_m = k' \frac{W}{L}(V_{GS} - V_t) = \sqrt{2k' \frac{W}{L} I_D} \tag{1.180}$$

Unlike the bipolar transistor, the transconductance of the MOS transistor is proportional to the square root of the bias current and depends on device geometry (oxide thickness via k' and W/L). Another key difference between bipolar and MOS transistors can be seen by calculating the ratio of the transconductance to the current. Using (1.157) and (1.180) for MOS transistors shows that

$$\frac{g_m}{I_D} = \frac{2}{V_{GS} - V_t} = \frac{2}{V_{ov}} \tag{1.181}$$

Also, for bipolar transistors, (1.91) shows that

$$\frac{g_m}{I_C} = \frac{q}{kT} = \frac{1}{V_T} \tag{1.182}$$

At room temperature, the thermal voltage V_T is about equal to 26 mV. In contrast, the overdrive V_{ov} for MOS transistors in many applications is chosen to be approximately several hundred mV so that MOS transistors are fast enough for the given application. (Section 1.6.8 shows that the transition frequency f_T of an MOS transistor is proportional to the overdrive.) Under these conditions, the transconductance per given current is much higher for bipolar transistors than for MOS transistors. One of the key challenges in MOS analog circuit design is designing high-quality analog circuits with a low transconductance-to-current ratio.

The transconductance calculated in (1.180) is valid for small-signal analysis. To determine the limitation on the use of small-signal analysis, the change in the drain current resulting from a change in the gate-source voltage will be derived from a large-signal standpoint. The total drain current in Fig. 1.33 can be calculated using (1.157) as

$$I_d = \frac{k'}{2}\frac{W}{L}(V_{GS} + v_i - V_t)^2 = \frac{k'}{2}\frac{W}{L}\left[(V_{GS} - V_t)^2 + 2(V_{GS} - V_t)v_i + v_i^2\right] \tag{1.183}$$

Substituting (1.157) in (1.183) gives

$$I_d = I_D + \frac{k'}{2}\frac{W}{L}\left[2(V_{GS} - V_t)v_i + v_i^2\right] \tag{1.184}$$

Rearranging (1.184) gives

$$i_d = I_d - I_D = k'\frac{W}{L}(V_{GS} - V_t)v_i\left[1 + \frac{v_i}{2(V_{GS} - V_t)}\right] \tag{1.185}$$

If the magnitude of the small-signal input $|v_i|$ is much less than twice the overdrive defined in (1.166), substituting (1.180) into (1.185) gives

$$i_d \simeq g_m v_i \tag{1.186}$$

In particular, if $|v_i| = |\Delta V_{GS}|$ is less than 20 percent of the overdrive, the small-signal analysis is accurate within about 10 percent.

1.6.2 Intrinsic Gate-Source and Gate-Drain Capacitance

If C_{ox} is the oxide capacitance per unit area from gate to channel, then the total capacitance under the gate is $C_{ox}WL$. This capacitance is intrinsic to the device operation and models the gate control of the channel conductance. In the triode region of device operation, the channel exists continuously from source to drain, and the gate-channel capacitance is usually lumped into two equal parts at the drain and source with

$$C_{gs} = C_{gd} = \frac{C_{ox}WL}{2} \tag{1.187}$$

In the saturation or active region, however, the channel pinches off before reaching the drain, and the drain voltage exerts little influence on either the channel or the gate charge. As a consequence, the intrinsic portion of C_{gd} is essentially zero in the saturation region. To calculate the value of the intrinsic part of C_{gs} in the saturation or active region, we must calculate the total charge Q_T stored in the channel. This calculation can be carried out by substituting (1.145) into (1.144) and integrating to obtain

$$Q_T = WC_{ox}\int_0^L [V_{GS} - V(y) - V_t]dy \tag{1.188}$$

Solving (1.150) for dy and substituting into (1.188) gives

$$Q_T = \frac{W^2C_{ox}^2\mu_n}{I_D}\int_0^{V_{GS}-V_t} (V_{GS} - V - V_t)^2 dV \tag{1.189}$$

where the limit $y = L$ corresponds to $V = (V_{GS} - V_t)$ in the saturation or active region. Solution of (1.189) and use of (1.153) and (1.157) gives

$$Q_T = \frac{2}{3}WLC_{ox}(V_{GS} - V_t) \tag{1.190}$$

Therefore, in the saturation or active region,

$$C_{gs} = \frac{\partial Q_T}{\partial V_{GS}} = \frac{2}{3}WLC_{ox} \tag{1.191}$$

and

$$C_{gd} = 0 \tag{1.192}$$

1.6.3 Input Resistance

The gate of an MOS transistor is insulated from the channel by the SiO_2 dielectric. As a result, the low-frequency gate current is essentially zero and the input resistance is essentially infinite. This characteristic is important in some circuits such as sample-and-hold amplifiers, where the gate of an MOS transistor can be connected to a capacitor to sense the voltage on the capacitor without leaking away the charge that causes that voltage. In contrast, bipolar transistors have small but nonzero base current and finite input resistance looking into the base, complicating the design of bipolar sample-and-hold amplifiers.

1.6.4 Output Resistance

In Section 1.5.1, the effect of changes in drain-source voltage on the large-signal characteristics of the MOS transistor was described. Increasing drain-source voltage in an n-channel MOS transistor increases the width of the depletion region around the drain and reduces the effective channel length of the device in the saturation or active region. This effect is called channel-length modulation and causes the drain current to increase when the drain-source voltage is increased. From that treatment, we can calculate the change in the drain current ΔI_D arising from changes in the drain-source voltage ΔV_{DS} as

$$\Delta I_D = \frac{\partial I_D}{\partial V_{DS}} \Delta V_{DS} \tag{1.193}$$

Substitution of (1.161), (1.163), and (1.164) in (1.193) gives

$$\frac{\Delta V_{DS}}{\Delta I_D} = \frac{V_A}{I_D} = \frac{1}{\lambda I_D} = r_o \tag{1.194}$$

where V_A is the Early voltage, λ is the channel-length modulation parameter, I_D is the drain current without channel-length modulation given by (1.157), and r_o is the small-signal output resistance of the transistor.

1.6.5 Basic Small-Signal Model of the MOS Transistor

Combination of the preceding small-signal circuit elements yields the small-signal model of the MOS transistor shown in Fig. 1.34. This model was derived for n-channel transistors in the saturation or active region and is called the *hybrid-π* model. Drain, gate, and source nodes are labeled D, G, and S, respectively. When the gate-source voltage is increased, the model predicts that the incremental current i_d flowing from drain to source increases. Since the dc drain current I_D also flows from drain to source in an n-channel transistor, increasing the gate-source voltage also increases the total drain current I_d. This result is reasonable physically because increasing the gate-source voltage in an n-channel transistor increases the channel conductivity and drain current.

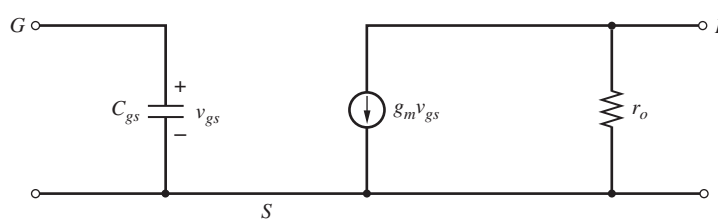


Figure 1.34 Basic small-signal model of an MOS transistor in the saturation or active region.

The model shown in Fig. 1.34 is also valid for p-channel devices. Therefore, the model again shows that increasing the gate-source voltage increases the incremental current i_d flowing from drain to source. Unlike in the n-channel case, however, the dc current I_D in a p-channel transistor flows from source to drain because the source acts as the source of holes. Therefore, the incremental drain current flows in a direction opposite to the dc drain current when the gate-source voltage increases, reducing the total drain current I_d. This result is reasonable physically because increasing the gate-source voltage in a p-channel transistor reduces the channel conductivity and drain current.

1.6.6 Body Transconductance

The drain current is a function of both the gate-source and body-source voltages. On the one hand, the gate-source voltage controls the vertical electric field, which controls the channel conductivity and therefore the drain current. On the other hand, the body-source voltage changes the threshold, which changes the drain current when the gate-source voltage is fixed. This effect stems from the influence of the substrate acting as a second gate and is called the *body effect*. Note that the body of an MOS transistor is usually connected to a constant power-supply voltage, which is a small-signal or ac ground. However, the source connection can have a significant ac voltage impressed on it, which changes the body-source voltage when the body voltage is fixed. Therefore, when the body-source voltage is not constant, two transconductance terms are required to model MOS transistors: one associated with the main gate and the other associated with the body or second gate.

Using (1.165), the transconductance from the body or second gate is

$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = -k'\frac{W}{L}(V_{GS} - V_t)(1 + \lambda V_{DS})\frac{\partial V_t}{\partial V_{BS}} \tag{1.195}$$

From (1.140)

$$\frac{\partial V_t}{\partial V_{BS}} = -\frac{\gamma}{2\sqrt{2\phi_f + V_{SB}}} = -\chi \tag{1.196}$$

This equation defines a factor χ, which is the rate of change of threshold voltage with body bias voltage. Substitution of (1.141) in (1.196) and use of (1.20) gives

$$\chi = \frac{C_{js}}{C_{ox}} \tag{1.197}$$

where C_{js} is the capacitance per unit area of the depletion region under the channel, assuming a one-sided step junction with a built-in potential $\psi_0 = 2\phi_f$. Substitution of (1.196) in (1.195) gives

$$g_{mb} = \frac{\gamma k'(W/L)(V_{GS} - V_t)(1 + \lambda V_{DS})}{2\sqrt{2\phi_f + V_{SB}}} \tag{1.198}$$

If $\lambda V_{DS} \ll 1$, we have

$$g_{mb} = \frac{\gamma k'(W/L)(V_{GS} - V_t)}{2\sqrt{2\phi_f + V_{SB}}} = \gamma\sqrt{\frac{k'(W/L)I_D}{2(2\phi_f + V_{SB})}} \tag{1.199}$$

The ratio g_{mb}/g_m is an important quantity in practice. From (1.179) and (1.198), we find

$$\frac{g_{mb}}{g_m} = \frac{\gamma}{2\sqrt{2\phi_f + V_{SB}}} = \chi \tag{1.200}$$

The factor χ is typically in the range 0.1 to 0.3; therefore, the transconductance from the main gate is typically a factor of about 3 to 10 times larger than the transconductance from the body or second gate.

1.6.7 Parasitic Elements in the Small-Signal Model

The elements of the small-signal model for MOS transistors described above may be considered basic in the sense that they arise directly from essential processes in the device. As in the case of bipolar transistors, however, technological limitations in the fabrication of the devices give rise to a number of parasitic elements that must be added to the equivalent circuit for most integrated-circuit transistors. A cross section and top view of a typical n-channel MOS transistor are shown in Fig. 1.35. The means of fabricating such devices is described in Chapter 2.

All pn junctions in the MOS transistor should be reverse biased during normal operation, and each junction exhibits a voltage-dependent parasitic capacitance associated with its depletion region. The source-body and drain-body junction capacitances shown in Fig. 1.35a are C_{sb} and C_{db}, respectively. If the doping levels in the source, drain, and body regions are assumed to be constant, (1.21) can be used to express these capacitances as follows:

$$C_{sb} = \frac{C_{sb0}}{\left(1 + \dfrac{V_{SB}}{\psi_0}\right)^{1/2}} \tag{1.201}$$

$$C_{db} = \frac{C_{db0}}{\left(1 + \dfrac{V_{DB}}{\psi_0}\right)^{1/2}} \tag{1.202}$$

These capacitances are proportional to the source and drain region areas (including sidewalls). Since the channel is attached to the source in the saturation or active region, C_{sb} also includes

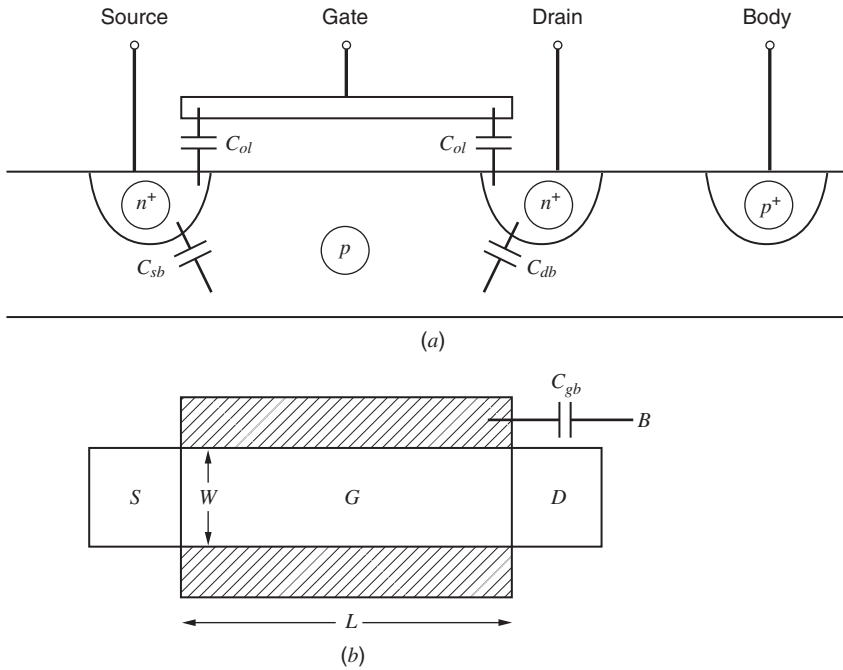


Figure 1.35 (a) Cross section and (b) top view of an n-channel MOS transistor.

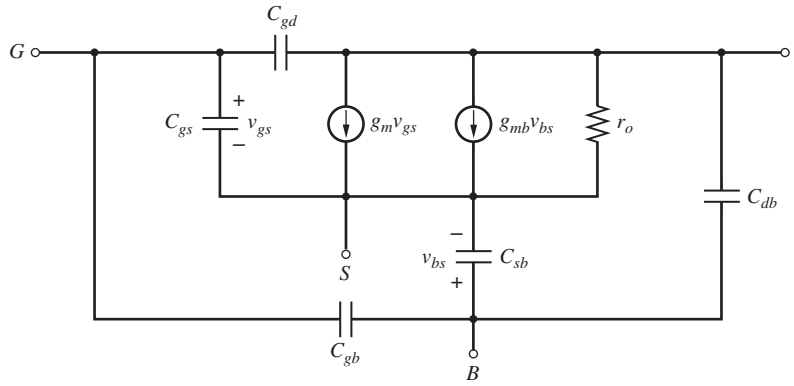


Figure 1.36 Small-signal MOS transistor equivalent circuit.

depletion-region capacitance from the induced channel to the body. A detailed analysis of the channel-body capacitance is given by Tsividis.[26]

In practice, C_{gs} and C_{gd}, given in (1.187) for the triode region of operation and in (1.191) and (1.192) for the saturation or active region, are increased due to parasitic oxide capacitances arising from gate overlap of the source and drain regions. These overlap capacitances C_{ol} are shown in Fig. 1.35*a*, and their values are calculated in Chapter 2.

Capacitance C_{gb} between gate and body or substrate models parasitic oxide capacitance between the gate-contact material and the substrate outside the active-device area. This capacitance is independent of the gate-body voltage and models coupling from polysilicon and metal interconnects to the underlying substrate, as shown by the shaded regions in the top view of Fig. 1.35*b*. Parasitic capacitance of this type underlies all polysilicon and metal traces on integrated circuits. Such parasitic capacitance should be taken into account when simulating and calculating high-frequency circuit and device performance. Typical values depend on oxide thicknesses. With a silicon dioxide thickness of 100 Å, the capacitance is about 3.45 fF per square micron. Fringing capacitance becomes important for lines narrower in width than several microns.

Parasitic resistance in series with the source and drain can be used to model the nonzero resistivity of the contacts and diffusion regions. In practice, these resistances are often ignored in hand calculations for simplicity but included in computer simulations. These parasitic resistances have an inverse dependence on channel width W. Typical values of these resistances are 50 Ω to 100 Ω for devices with W of about 1 μm. Similar parasitic resistances in series with the gate and body terminals are sometimes included but often ignored because very little current flows in these terminals, especially at low frequencies. The small-signal model including capacitive parasitics but ignoring resistive parasitics is shown in Fig. 1.36.

1.6.8 MOS Transistor Frequency Response

As for a bipolar transistor, the frequency capability of an MOS transistor is usually specified by finding the transition frequency f_T. For an MOS transistor, f_T is defined as the frequency where the magnitude of the short-circuit, common-source current gain falls to unity. Although the dc gate current of an MOS transistor is essentially zero, the high-frequency behavior of the transistor is controlled by the capacitive elements in the small-signal model, which cause the gate current to increase as frequency increases. To calculate f_T, consider the ac circuit of Fig. 1.37*a* and the small-signal equivalent of Fig. 1.37*b*. Since $v_{sb} = v_{ds} = 0$, g_{mb},

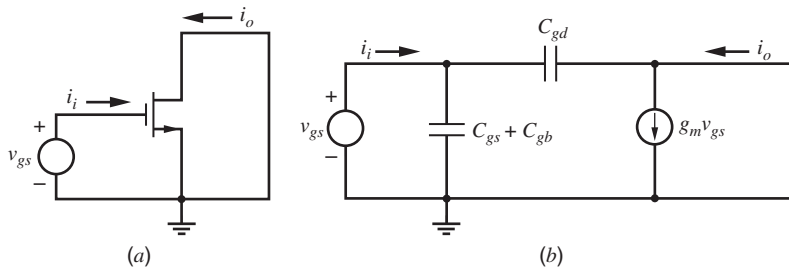


Figure 1.37 Circuits for calculating the f_T of an MOS transistor: (*a*) ac schematic and (*b*) small-signal equivalent.

r_o, C_{sb}, and C_{db} have no effect on the calculation and are ignored. The small-signal input current i_i is

$$i_i = s(C_{gs} + C_{gb} + C_{gd})v_{gs} \tag{1.203}$$

If the current fed forward through C_{gd} is neglected,

$$i_o \simeq g_m v_{gs} \tag{1.204}$$

Solving (1.203) for v_{gs} and substituting into (1.204) gives

$$\frac{i_o}{i_i} \simeq \frac{g_m}{s(C_{gs} + C_{gb} + C_{gd})} \tag{1.205}$$

To find the frequency response, we set $s = j\omega$. Then

$$\frac{i_o}{i_i} \simeq \frac{g_m}{j\omega(C_{gs} + C_{gb} + C_{gd})} \tag{1.206}$$

The magnitude of the small-signal current gain is unity when

$$\omega = \omega_T = \frac{g_m}{C_{gs} + C_{gb} + C_{gd}} \tag{1.207}$$

Therefore,

$$f_T = \frac{1}{2\pi}\omega_T = \frac{1}{2\pi}\frac{g_m}{C_{gs} + C_{gb} + C_{gd}} \tag{1.208}$$

Assume the intrinsic device capacitance C_{gs} is much greater than $(C_{gb} + C_{gd})$. Then substituting (1.180) and (1.191) into (1.208) gives

$$f_T = 1.5\frac{\mu_n}{2\pi L^2}(V_{GS} - V_t) \tag{1.209}$$

Comparison of this equation with the intrinsic f_T of a bipolar transistor when parasitic depletion-layer capacitance is neglected leads to an interesting result. From (1.128) and (1.130) with $\tau_F \gg (C_{je} + C_\mu)/g_m$,

$$f_T = \frac{1}{2\pi\tau_F} \tag{1.210}$$

Substituting from (1.99) for τ_F and using the Einstein relationship $D_n/\mu_n = kT/q = V_T$, we find for a bipolar transistor

$$f_T = 2\frac{\mu_n}{2\pi W_B^2} V_T \tag{1.211}$$

The similarity in form between (1.211) and (1.209) is striking. In both cases, the intrinsic device f_T increases as the inverse square of the critical device dimension across which carriers are in transit. The voltage $V_T = 26$ mV is fixed for a bipolar transistor, but the f_T of an MOS transistor can be increased by operating at high values of $(V_{GS} - V_t)$. Note that the base width W_B in a bipolar transistor is a vertical dimension determined by diffusions or implants and can typically be made much smaller than the channel length L of an MOS transistor, which depends on surface geometry and photolithographic processes. Thus bipolar transistors generally have higher f_T than MOS transistors made with comparable processing. Finally, (1.209) was derived assuming that the MOS transistor exhibits square-law behavior as in (1.157). However, as described in Section 1.7, submicron MOS transistors depart significantly from square-law characteristics, and we find that for such devices f_T is proportional to L^{-1} rather than L^{-2}.

■ **EXAMPLE**

Derive the complete small-signal model for an NMOS transistor with $I_D = 100$ μA, $V_{SB} = 1$ V, $V_{DS} = 2$ V. Device parameters are $\phi_f = 0.3$ V, $W = 10$ μm, $L = 1$ μm, $\gamma = 0.5$ V$^{1/2}$, $k' = 200$ μA/V^2, $\lambda = 0.02$ V^{-1}, $t_{ox} = 100$ angstroms, $\psi_0 = 0.6$ V, $C_{sb0} = C_{db0} = 10$ fF. Overlap capacitance from gate to source and gate to drain is 1 fF. Assume $C_{gb} = 5$ fF.

From (1.166),

$$V_{ov} = V_{GS} - V_t = \sqrt{\frac{2I_D}{k'(W/L)}} = \sqrt{\frac{2 \times 100}{200 \times 10}} \simeq 0.316 \text{ V}$$

Since $V_{DS} > V_{ov}$, the transistor operates in the saturation or active region. From (1.180),

$$g_m = \sqrt{2k'\frac{W}{L} I_D} = \sqrt{2 \times 200 \times 10 \times 100}\ \mu\text{A/V} \simeq 632\ \mu\text{A/V}$$

From (1.199),

$$g_{mb} = \gamma\sqrt{\frac{k'(W/L)I_D}{2(2\phi_f + V_{SB})}} = 0.5\sqrt{\frac{200 \times 10 \times 100}{2 \times 1.6}} \simeq 125\ \mu\text{A/V}$$

From (1.194),

$$r_o = \frac{1}{\lambda I_D} = \frac{1000}{0.02 \times 100}\ \text{k}\Omega = 500\ \text{k}\Omega$$

Using (1.201) with $V_{SB} = 1$ V, we find

$$C_{sb} = \frac{10}{\left(1 + \frac{1}{0.6}\right)^{1/2}}\ \text{fF} \simeq 6\ \text{fF}$$

The voltage from drain to body is

$$V_{DB} = V_{DS} + V_{SB} = 3 \text{ V}$$

and substitution in (1.202) gives

$$C_{db} = \frac{10}{\left(1 + \dfrac{3}{0.6}\right)^{1/2}} \text{ fF} \simeq 4 \text{ fF}$$

From (1.142), the oxide capacitance per unit area is

$$C_{ox} = \frac{3.9 \times 8.854 \times 10^{-14} \dfrac{\text{F}}{\text{cm}} \times \dfrac{100 \text{ cm}}{10^6 \ \mu\text{m}}}{100 \text{ Å} \times \dfrac{10^6 \ \mu\text{m}}{10^{10} \text{ Å}}} \simeq 3.45 \ \frac{\text{fF}}{\mu\text{m}^2}$$

The intrinsic portion of the gate-source capacitance can be calculated from (1.191), giving

$$C_{gs} \simeq \frac{2}{3} \times 10 \times 1 \times 3.45 \text{ fF} \simeq 23 \text{ fF}$$

The addition of overlap capacitance gives

$$C_{gs} \simeq 24 \text{ fF}$$

Finally, since the transistor operates in the saturation or active region, the gate-drain capacitance consists of only overlap capacitance and is

$$C_{gd} = 1 \text{ fF}$$

The complete small-signal equivalent circuit is shown in Fig. 1.38. The f_T of the device can be calculated from (1.208) as

$$f_T = \frac{1}{2\pi} \frac{g_m}{C_{gs} + C_{gb} + C_{gd}} = \frac{1}{2\pi} \times 632 \times 10^{-6} \times \frac{10^{15}}{24 + 5 + 1} \text{ Hz} = 3.4 \text{ GHz}$$

■

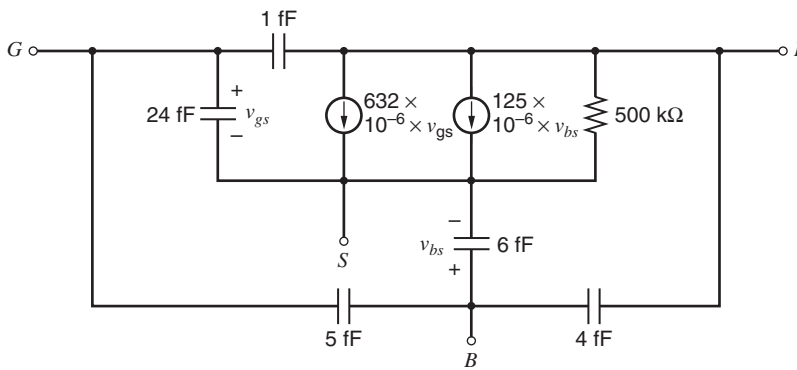


Figure 1.38 Complete small-signal equivalent circuit for an NMOS transistor with $I_D = 100$ μA, $V_{SB} = 1$ V, $V_{DS} = 2$ V. Device parameters are $W = 10$ μm, $L = 1$ μm, $\gamma = 0.5 \text{ V}^{1/2}$, $k' = 200 \ \mu\text{A/V}^2$, $\lambda = 0.02 \text{ V}^{-1}$, $t_{ox} = 100$ Å, $\psi_0 = 0.6$ V, $C_{sb0} = C_{db0} = 10$ fF, $C_{gd} = 1$ fF, and $C_{gb} = 5$ fF.

1.7 Short-Channel Effects in MOS Transistors

The evolution of integrated-circuit processing techniques has led to continuing reductions in both the horizontal and vertical dimensions of the active devices. (The minimum allowed dimensions of passive devices have also decreased.) This trend is driven primarily by economics in that reducing dimensions increases the number of devices and circuits that can be processed at one time on a given wafer. A second benefit has been that the frequency capability of the active devices continues to increase, as intrinsic f_T values increase with smaller dimensions while parasitic capacitances decrease.

Vertical dimensions such as the base width of a bipolar transistor in production processes may now be on the order of 0.05 μm or less, whereas horizontal dimensions such as bipolar emitter width or MOS transistor gate length may be significantly less than 1 μm. Even with these small dimensions, the large-signal and small-signal models of bipolar transistors given in previous sections remain valid. However, significant short-channel effects become important in MOS transistors at channel lengths of about 1 μm or less and require modifications to the MOS models given previously. The primary effect is to modify the classical MOS square-law transfer characteristic in the saturation or active region to make the device voltage-to-current transfer characteristic more linear. However, even in processes with submicron capability, many of the MOS transistors in a given analog circuit may be deliberately designed to have channel lengths larger than the minimum and may be well approximated by the square-law model.

1.7.1 Velocity Saturation from the Horizontal Field

The most important short-channel effect in MOS transistors stems from velocity saturation of carriers in the channel.[27] When an MOS transistor operates in the triode region, the average horizontal electric field along the channel is V_{DS}/L. When V_{DS} is small and/or L is large, the horizontal field is low, and the linear relation between carrier velocity and field assumed in (1.148) is valid. At high fields, however, the carrier velocities approach the thermal velocities, and subsequently the slope of the carrier velocity decreases with increasing field. This effect is illustrated in Fig. 1.39, which shows typical measured electron drift velocity v_d versus horizontal electric field strength magnitude $\mathscr{E}$ in an NMOS surface channel. While the velocity at low field values is proportional to the field, the velocity at high field values approaches a constant called the *scattering-limited* velocity v_{scl}. A first-order analytical approximation to this curve is

$$v_d = \frac{\mu_n \mathscr{E}}{1 + \mathscr{E}/\mathscr{E}_c} \tag{1.212}$$

where $\mathscr{E}_c \simeq 1.5 \times 10^6$ V/m and $\mu_n \simeq 0.07$ m^2/V-s is the low-field mobility close to the gate. Equation 1.212 is also plotted in Fig. 1.39. From (1.212), as $\mathscr{E} \to \infty$, $v_d \to v_{scl} = \mu_n \mathscr{E}_c$. At the critical field value $\mathscr{E}_c$, the carrier velocity is a factor of 2 less than the low-field formula would predict. In a device with a channel length $L = 0.5$ μm, we need a voltage drop of only 0.75 V along the channel to produce an average field equal to $\mathscr{E}_c$, and this condition is readily achieved in short-channel MOS transistors. Similar results are found for PMOS devices.

Substituting (1.212) and (1.149) into (1.147) and rearranging gives

$$I_D \left(1 + \frac{1}{\mathscr{E}_c}\frac{dV}{dy}\right) = WQ_I(y)\mu_n \frac{dV}{dy} \tag{1.213}$$

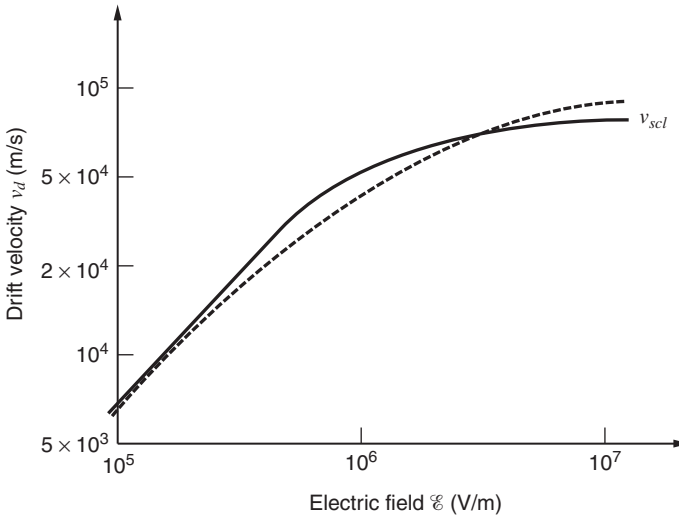


Figure 1.39 Typical measured electron drift velocity v_d versus horizontal electric field $\mathscr{E}$ in an MOS surface channel (solid plot). Also shown (dashed plot) is the analytical approximation of Eq. 1.212 with $\mathscr{E}_c = 1.5 \times 10^6$ V/m and $\mu_n = 0.07$ m^2/V-s.

Note that as $\mathscr{E}_c \to \infty$ and velocity saturation becomes negligible, (1.213) approaches the original equation (1.147). Integrating (1.213) along the channel, we obtain

$$\int_0^L I_D \left(1 + \frac{1}{\mathscr{E}_c}\frac{dV}{dy}\right) dy = \int_0^{V_{DS}} WQ_I(y)\mu_n \, dV \tag{1.214}$$

and thus

$$I_D = \frac{\mu_n C_{ox}}{2\left(1 + \dfrac{V_{DS}}{\mathscr{E}_c L}\right)} \frac{W}{L}[2(V_{GS} - V_t)V_{DS} - V_{DS}^2] \tag{1.215}$$

In the limit as $\mathscr{E}_c \to \infty$, (1.215) is the same as (1.152), which gives the drain current in the triode region without velocity saturation. The quantity V_{DS}/L in (1.215) can be interpreted as the average horizontal electric field in the channel. If this field is comparable to $\mathscr{E}_c$, the drain current for a given V_{DS} is less than the simple expression (1.152) would predict.

Equation 1.215 is valid in the triode region. Let $V_{DS(\text{act})}$ represent the maximum value of V_{DS} for which the transistor operates in the triode region, which is equivalent to the minimum value of V_{DS} for which the transistor operates in the active region. In the active region, the current should be independent of V_{DS} because channel-length modulation is not included here. Therefore, $V_{DS(\text{act})}$ is the value of V_{DS} that sets $\partial I_D/\partial V_{DS} = 0$. From (1.215),

$$\frac{\partial I_D}{\partial V_{DS}} = \frac{k'}{2}\frac{W}{L}\left[\frac{\left(1 + \dfrac{V_{DS}}{\mathscr{E}_c L}\right)[2(V_{GS} - V_t) - 2V_{DS}] - \dfrac{[2(V_{GS} - V_t)V_{DS} - V_{DS}^2]}{\mathscr{E}_c L}}{\left(1 + \dfrac{V_{DS}}{\mathscr{E}_c L}\right)^2}\right] \tag{1.216}$$

where $k' = \mu_n C_{ox}$ as given by (1.153). To set $\partial I_D/\partial V_{DS} = 0$,

$$\left(1 + \frac{V_{DS}}{\mathscr{E}_c L}\right)[2(V_{GS} - V_t) - 2V_{DS}] - \frac{[2(V_{GS} - V_t)V_{DS} - V_{DS}^2]}{\mathscr{E}_c L} = 0 \tag{1.217}$$

Rearranging (1.217) gives

$$\frac{V_{DS}^2}{\mathscr{E}_c L} + 2V_{DS} - 2(V_{GS} - V_t) = 0 \tag{1.218}$$

Solving the quadratic equation gives

$$V_{DS(\text{act})} = V_{DS} = -\mathscr{E}_c L \pm \mathscr{E}_c L\sqrt{1 + \frac{2(V_{GS} - V_t)}{\mathscr{E}_c L}} \tag{1.219}$$

Since the drain-source voltage must be greater than zero,

$$V_{DS(\text{act})} = V_{DS} = \mathscr{E}_c L\left(\sqrt{1 + \frac{2(V_{GS} - V_t)}{\mathscr{E}_c L}} - 1\right) \tag{1.220}$$

To determine $V_{DS(\text{act})}$ without velocity-saturation effects, let $\mathscr{E}_c \to \infty$ so that the drift velocity is proportional to the electric field, and let $x = (V_{GS} - V_t)/(\mathscr{E}_c L)$. Then $x \to 0$, and a Taylor series can be used to show that

$$\sqrt{1 + 2x} = 1 + x - \frac{x^2}{2} + \cdots \tag{1.221}$$

Using (1.221) in (1.220) gives

$$V_{DS(\text{act})} = (V_{GS} - V_t)\left(1 - \frac{V_{GS} - V_t}{2\mathscr{E}_c L} + \cdots\right) \tag{1.222}$$

When $\mathscr{E}_c \to \infty$, (1.222) shows that $V_{DS(\text{act})} \to (V_{GS} - V_t)$, as expected.[28] This observation is confirmed by plotting the ratio of $V_{DS(\text{act})}$ to the overdrive V_{ov} versus $\mathscr{E}_c L$ in Fig. 1.40. When $\mathscr{E}_c \to \infty$, $V_{DS(\text{act})} \to V_{ov} = V_{GS} - V_t$, as predicted by (1.222). On the other hand, when $\mathscr{E}_c$ is small enough that velocity saturation is significant, Fig. 1.40 shows that $V_{DS(\text{act})} < V_{ov}$.

To find the drain current in the active region with velocity saturation, substitute $V_{DS(\text{act})}$ in (1.220) for V_{DS} in (1.215). After rearranging, the result is

$$I_D = \frac{\mu_n C_{ox}}{2}\frac{W}{L}[V_{DS(\text{act})}]^2 \tag{1.223}$$

Equation 1.223 is in the same form as (1.157), where velocity saturation is neglected, except that $V_{DS(\text{act})}$ is less than $(V_{GS} - V_t)$ when velocity saturation is significant, as shown in Fig. 1.40. Therefore, the current predicted by (1.157) overestimates the current that really flows when the carrier velocity saturates. To examine the limiting case when the velocity is completely

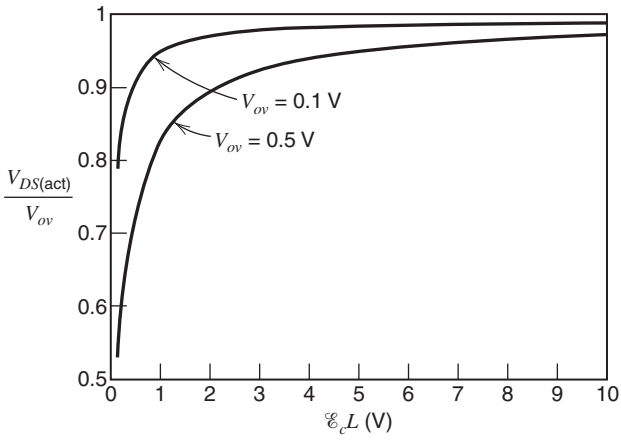


Figure 1.40 Ratio of the minimum drain-source voltage required for operation in the active region to the overdrive versus the product of the critical field and the channel length. When $\mathscr{E}_c \to \infty$, velocity saturation is not a factor, and $V_{DS(\text{act})} \to V_{ov} = V_{GS} - V_t$, as expected. When velocity saturation is significant, $V_{DS(\text{act})} < V_{ov}$.

saturated, let $\mathscr{E}_c \to 0$. Then (1.212) shows that the drift velocity approaches the scattering-limited velocity $v_d \to v_{scl} = \mu_n \mathscr{E}_c$. Substituting (1.220) into (1.223) gives

$$\lim_{\mathscr{E}_c \to 0} I_D = \mu_n C_{ox} W (V_{GS} - V_t) \mathscr{E}_c = W C_{ox} (V_{GS} - V_t) v_{scl} \tag{1.224}$$

In contrast to the square-law behavior predicted by (1.157), (1.224) shows that the drain current is a *linear* function of the overdrive $(V_{GS} - V_t)$ when the carrier velocity saturates. Also, (1.224) shows that the drain current is independent of the channel length when the carrier velocity saturates. In this case, both the charge in the channel and the time required for the charge to cross the channel are proportional to L. Since the current is the ratio of the charge in the channel to the time required to cross the channel, the current does not depend on L as long as the channel length is short enough to produce an electric field that is high enough for velocity saturation to occur.[29] In contrast, when the carrier velocity is proportional to the electric field instead of being saturated, the time required for channel charge to cross the channel is proportional to L^2 because increasing L both reduces the carrier velocity and increases the distance between the source and the drain. Therefore, when velocity saturation is not significant, the drain current is inversely proportional to L, as we have come to expect through (1.157). Finally, (1.224) shows that the drain current in the active region is proportional to the scattering-limited velocity $v_{scl} = \mu_n \mathscr{E}_c$ when the velocity is saturated.

Substituting (1.222) into (1.223) gives

$$\begin{aligned} I_D &= \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{GS} - V_t)^2 \left(1 - \frac{V_{GS} - V_t}{2\mathscr{E}_c L} + \cdots\right)^2 \\ &= \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{GS} - V_t)^2 \left(1 - \frac{x}{2} + \cdots\right)^2 \\ &= \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{GS} - V_t)^2 (1 - x + \cdots) \\ &= \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{GS} - V_t)^2 \left(1 - \frac{V_{GS} - V_t}{\mathscr{E}_c L} + \cdots\right) \end{aligned} \tag{1.225}$$

where $x = (V_{GS} - V_t)/(\mathscr{E}_c L)$ as defined for (1.221). If $x \ll 1$, $(1 - x) \simeq 1/(1 + x)$, and

$$I_D \simeq \frac{\mu_n C_{ox}}{2\left(1 + \dfrac{V_{GS} - V_t}{\mathscr{E}_c L}\right)} \frac{W}{L} (V_{GS} - V_t)^2 \tag{1.226}$$

Equation 1.226 is valid without velocity saturation and at its onset, where $(V_{GS} - V_t) \ll \mathscr{E}_c L$. The effect of velocity saturation on the current in the active region predicted by (1.226) can be modeled with the addition of a resistance in series with the source of an ideal square-law device, as shown in Fig. 1.41. Let V'_{GS} be the gate-source voltage of the ideal square-law

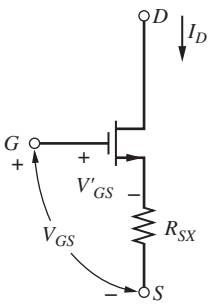


Figure 1.41 Model of velocity saturation in an MOSFET by addition of series source resistance to an ideal square-law device.

transistor. From (1.157),

$$I_D = \frac{\mu_n C_{ox}}{2}\frac{W}{L}(V'_{GS} - V_t)^2 \tag{1.227}$$

Let V_{GS} be the sum of V'_{GS} and the voltage drop on R_{SX}. Then

$$V_{GS} = V'_{GS} + I_D R_{SX} \tag{1.228}$$

This sum models the gate-source voltage of a real MOS transistor with velocity saturation. Substituting (1.228) into (1.227) gives

$$I_D = \frac{\mu_n C_{ox}}{2}\frac{W}{L}(V_{GS} - I_D R_{SX} - V_t)^2$$

$$I_D = \frac{\mu_n C_{ox}}{2}\frac{W}{L}\left((V_{GS} - V_t)^2 - 2(V_{GS} - V_t)I_D R_{SX} + (I_D R_{SX})^2\right) \tag{1.229}$$

Rearranging (1.229) while ignoring the $(I_D R_{SX})^2$ term gives

$$I_D \simeq \frac{\mu_n C_{ox}}{2\left(1 + \mu_n C_{ox}\dfrac{W}{L}R_{SX}(V_{GS} - V_t)\right)}\frac{W}{L}(V_{GS} - V_t)^2 \tag{1.230}$$

Equation 1.230 has the same form as (1.226) if we identify

$$\mu_n C_{ox}\frac{W}{L}R_{SX} = \frac{1}{\mathscr{E}_c L} \tag{1.231}$$

Rearranging (1.231) gives

$$R_{SX} = \frac{1}{\mathscr{E}_c \mu_n C_{ox} W} \tag{1.232}$$

Thus the influence of velocity saturation on the large-signal characteristics of an MOS transistor can be modeled to first order by a resistor R_{SX} in series with the source of an ideal square-law device. Note that R_{SX} varies inversely with W, as does the intrinsic physical series resistance due to the source and drain contact regions. Typically, R_{SX} is larger than the physical series resistance. For $W = 2$ μm, $k' = \mu_n C_{ox} = 200$ μA/V^2, and $\mathscr{E}_c = 1.5 \times 10^6$ V/m, we find $R_{SX} \simeq 1700\ \Omega$.

1.7.2 Transconductance and Transition Frequency

The values of all small-signal parameters can change significantly in the presence of short-channel effects.[30] One of the most important changes is to the transconductance. Substituting (1.220) into (1.223) and calculating $\partial I_D/\partial V_{GS}$ gives

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = WC_{ox}v_{scl}\frac{\sqrt{1 + \dfrac{2(V_{GS} - V_t)}{\mathscr{E}_c L}} - 1}{\sqrt{1 + \dfrac{2(V_{GS} - V_t)}{\mathscr{E}_c L}}} \tag{1.233}$$

where $v_{scl} = \mu_n \mathscr{E}_c$ as in Fig. 1.39. To determine g_m without velocity saturation, let $E_c \to \infty$ and $x = (V_{GS} - V_t)/(\mathscr{E}_c L)$. Then substituting (1.221) into (1.233) and rearranging gives

$$\lim_{\mathscr{E}_c \to \infty} g_m = k'\frac{W}{L}(V_{GS} - V_t) \tag{1.234}$$

as predicted by (1.180). In this case, the transconductance increases when the overdrive increases or the channel length decreases. On the other hand, letting $\mathscr{E}_c \to 0$ to determine g_m when the velocity is saturated gives

$$\lim_{\mathscr{E}_c \to 0} g_m = WC_{ox}v_{scl} \tag{1.235}$$

Equation 1.235 shows that further decreases in L or increases in $(V_{GS} - V_t)$ do not change the transconductance when the velocity is saturated.

From (1.223) and (1.233), the ratio of the transconductance to the current can be calculated as

$$\frac{g_m}{I} = \frac{2}{(\mathscr{E}_c L)\sqrt{1 + \dfrac{2(V_{GS} - V_t)}{\mathscr{E}_c L}}\left(\sqrt{1 + \dfrac{2(V_{GS} - V_t)}{\mathscr{E}_c L}} - 1\right)} \tag{1.236}$$

As $\mathscr{E}_c \to 0$, the velocity saturates and

$$\lim_{\mathscr{E}_c \to 0} \frac{g_m}{I} = \frac{1}{V_{GS} - V_t} \tag{1.237}$$

Comparing (1.237) to (1.181) shows that velocity saturation reduces the transconductance-to-current ratio for a given overdrive.

On the other hand, when $x = (V_{GS} - V_t)/(\mathscr{E}_c L) \ll 1$, substituting (1.221) into (1.236) gives

$$\frac{g_m}{I} \simeq \frac{2}{(V_{GS} - V_t)(1 + x)} \tag{1.238}$$

Therefore, as $\mathscr{E}_c \to \infty$, $x \to 0$, and (1.238) collapses to

$$\lim_{\mathscr{E}_c \to \infty} \frac{g_m}{I} = \frac{2}{V_{GS} - V_t} \tag{1.239}$$

as predicted by (1.181). Equation 1.238 shows that if $x < 0.1$, the error in using (1.181) to calculate the transconductance-to-current ratio is less than about 10 percent. Therefore, we will conclude that velocity-saturation effects are insignificant in hand calculations if

$$(V_{GS} - V_t) < 0.1(\mathscr{E}_c L) \tag{1.240}$$

Figure 1.42 plots the transconductance-to-current ratio versus the overdrive for three cases. The highest and lowest ratios come from (1.239) and (1.237), which correspond to asymptotes where velocity saturation is insignificant and dominant, respectively. In practice, the transition between these extreme cases is gradual and described by (1.236), which is plotted in Fig. 1.42 for an example where $\mathscr{E}_c = 1.5 \times 10^6$ V/m and $L = 0.5$ μm.

One reason the change in transconductance caused by velocity saturation is important is because it affects the transition frequency f_T. Assuming that $C_{gs} \gg C_{gb} + C_{gd}$, substituting (1.235) into (1.208) shows that

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_{gs}} \propto \frac{WC_{ox}v_{scl}}{WLC_{ox}} \propto \frac{v_{scl}}{L} \tag{1.241}$$

One key point here is that the transition frequency is independent of the overdrive once velocity saturation is reached. In contrast, (1.209) shows that increasing $(V_{GS} - V_t)$ increases f_T before the velocity saturates. Also, (1.241) shows that the transition frequency is inversely proportional to the channel length when the velocity is saturated. In contrast, (1.209) predicts that f_T is

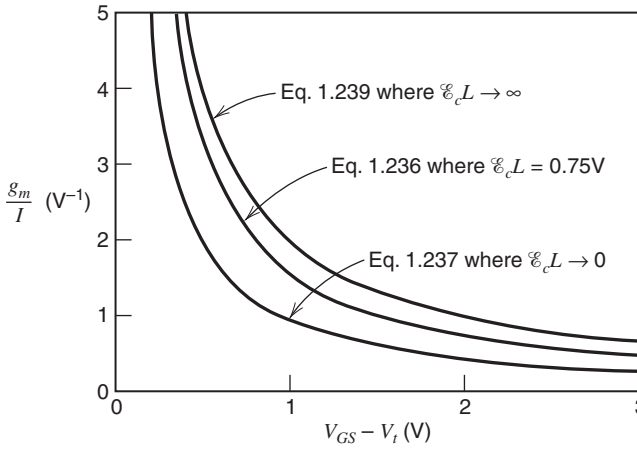


Figure 1.42
Transconductance-to-current ratio versus overdrive ($V_{GS} - V_t$) where velocity saturation is insignificant ($\mathscr{E}_c L \to \infty$), dominant ($\mathscr{E}_c L = 0$), and of gradually increasing importance ($\mathscr{E}_c L = 0.75$ V).

inversely proportional to the square of the channel length before the velocity saturates. As a result, velocity saturation reduces the speed improvement that can be achieved through reductions in the minimum channel length.

1.7.3 Mobility Degradation from the Vertical Field

Thus far, we have considered only the effects of the horizontal field due to the V_{DS} along the channel when considering velocity saturation. However, a vertical field originating from the gate voltage also exists and influences carrier velocity. A physical reason for this effect is that increasing the vertical electric field forces the carriers in the channel closer to the surface of the silicon, where surface imperfections impede their movement from the source to the drain, reducing mobility.[31] The vertical field at any point in the channel depends on the gate-channel voltage. Since the gate-channel voltage is not constant from the source to the drain, the effect of the vertical field on mobility should be included within the integration in (1.214) in principle.[32] For simplicity, however, this effect is often modeled after integration by changing the mobility in the previous equations to an effective mobility given by

$$\mu_{\text{eff}} = \frac{\mu_n}{1 + \theta(V_{GS} - V_t)} \tag{1.242}$$

where μ_n is the mobility with zero vertical field, and θ is inversely proportional to the oxide thickness. For $t_{ox} = 100$ Å, θ is typically in the range from 0.1 V^{-1} to 0.4 V^{-1}.[33] In practice, θ is determined by a best fit to measured device characteristics.

1.8 Weak Inversion in MOS Transistors

The MOSFET analysis of Section 1.5 considered the normal region of operation for which a well-defined conducting channel exists under the gate. In this region of *strong inversion*, changes in the gate-source voltage are assumed to cause only changes in the channel charge and not in the depletion-region charge. In contrast, for gate-source voltages less than the extrapolated threshold voltage V_t but high enough to create a depletion region at the surface of the silicon, the device operates in *weak inversion*. In the weak-inversion region, the channel charge is much less than the charge in the depletion region, and the drain current arising from the drift of majority carriers is negligible. However, the total drain current in weak inversion

is larger than that caused by drift because a gradient in minority-carrier concentration causes a diffusion current to flow. In weak inversion, an n-channel MOS transistor operates as an npn bipolar transistor, where the source acts as the emitter, the substrate as the base, and the drain as the collector.[34]

1.8.1 Drain Current in Weak Inversion

To analyze this situation, assume that the source and the body are both grounded. Also assume that $V_{DS} > 0$. (If $V_{DS} < 0$, the drain acts as the emitter and the source as the collector.)[35] Then increasing the gate-source voltage increases the surface potential ψ_s, which tends to reduce the reverse bias across the source-substrate (emitter-base) junction and to exponentially increase the concentration of electrons in the p-type substrate at the source $n_p(0)$. From (1.27),

$$n_p(0) = n_{po} \exp \frac{\psi_s}{V_T} \tag{1.243}$$

where n_{po} is the equilibrium concentration of electrons in the substrate (base). Similarly, the concentration of electrons in the substrate at the drain $n_p(L)$ is

$$n_p(L) = n_{po} \exp \frac{\psi_s - V_{DS}}{V_T} \tag{1.244}$$

From (1.31), the drain current due to the diffusion of electrons in the substrate is

$$I_D = qAD_n \frac{n_p(L) - n_p(0)}{L} \tag{1.245}$$

where D_n is the diffusion constant for electrons, and A is the cross-sectional area in which the diffusion current flows. The area A is the product of the transistor width W and the thickness X of the region in which I_D flows. Substituting (1.243) and (1.244) into (1.245) and rearranging gives

$$I_D = \frac{W}{L} qXD_n n_{po} \exp\left(\frac{\psi_s}{V_T}\right)\left[1 - \exp\left(-\frac{V_{DS}}{V_T}\right)\right] \tag{1.246}$$

In weak inversion, the surface potential is approximately a linear function of the gate-source voltage.[36] Assume that the charge stored at the oxide-silicon interface is independent of the surface potential. Then, in weak inversion, changes in the surface potential $\Delta\psi_s$ are controlled by changes in the gate-source voltage ΔV_{GS} through a voltage divider between the oxide capacitance C_{ox} and the depletion-region capacitance C_{js}. Therefore,

$$\frac{d\psi_s}{dV_{GS}} = \frac{C_{ox}}{C_{js} + C_{ox}} = \frac{1}{n} = \frac{1}{1 + \chi} \tag{1.247}$$

in which $n = (1 + C_{js}/C_{ox})$ and $\chi = C_{js}/C_{ox}$, as defined in (1.197). Separating variables in (1.247) and integrating gives

$$\psi_s = \frac{V_{GS}}{n} + k_1 \tag{1.248}$$

where k_1 is a constant. Equation 1.248 is valid only when the transistor operates in weak inversion. When $V_{GS} = V_t$ with $V_{SB} = 0$, $\psi_s = 2\phi_f$ by definition of the threshold voltage. For $V_{GS} > V_t$, the inversion layer holds the surface potential nearly constant and (1.248) is not valid. Since (1.248) is valid only when $V_{GS} \leq V_t$, (1.248) is rewritten as follows:

$$\psi_s = \frac{V_{GS} - V_t}{n} + k_2 \tag{1.249}$$

where $k_2 = k_1 + V_t/n$. Substituting (1.249) into (1.246) gives

$$I_D = \frac{W}{L} qXD_n n_{po} \exp\left(\frac{k_2}{V_T}\right) \exp\left(\frac{V_{GS} - V_t}{nV_T}\right) \left[1 - \exp\left(-\frac{V_{DS}}{V_T}\right)\right] \tag{1.250}$$

Let

$$I_t = qXD_n n_{po} \exp\left(\frac{k_2}{V_T}\right) \tag{1.251}$$

represent the drain current with $V_{GS} = V_t$, $W/L = 1$, and $V_{DS} \gg V_T$. Then

$$I_D = \frac{W}{L} I_t \exp\left(\frac{V_{GS} - V_t}{nV_T}\right) \left[1 - \exp\left(-\frac{V_{DS}}{V_T}\right)\right] \tag{1.252}$$

Figure 1.43 plots the drain current versus the drain-source voltage for three values of the overdrive, with $W = 20$ μm, $L = 20$ μm, $n = 1.5$, and $I_t = 0.1$ μA. Notice that the drain current is almost constant when $V_{DS} > 3V_T$ because the last term in (1.252) approaches unity in this case. Therefore, unlike in strong inversion, the minimum drain-source voltage required to force the transistor to operate as a current source in weak inversion is independent of the overdrive.[37] Figure 1.43 and Equation 1.252 also show that the drain current is not zero when $V_{GS} \leq V_t$. To further illustrate this point, we show measured NMOS characteristics plotted on two different scales in Fig. 1.44. In Fig. 1.44*a*, we show $\sqrt{I_D}$ versus V_{GS} in the active region plotted on linear scales. For this device, $W = 20$ μm, $L = 20$ μm, and short-channel effects are negligible. (See Problem 1.21 for an example of a case in which short-channel effects are important.) The resulting straight line shows that the device characteristic is close to an ideal square law. Plots like the one in Fig. 1.44*a* are commonly used to obtain V_t by extrapolation (0.7 V in this case) and also k' from the slope of the curve (54 μA/V^2 in this case). Near the threshold voltage, the curve deviates from the straight line representing the square law. This region is weak inversion. The data are plotted a second time in Fig. 1.44*b* on log-linear scales. The straight line obtained for $V_{GS} < V_t$ fits (1.252) with $n = 1.5$. For $I_D < 10^{-12}$ A, the slope decreases because leakage currents are significant and do not follow (1.252).

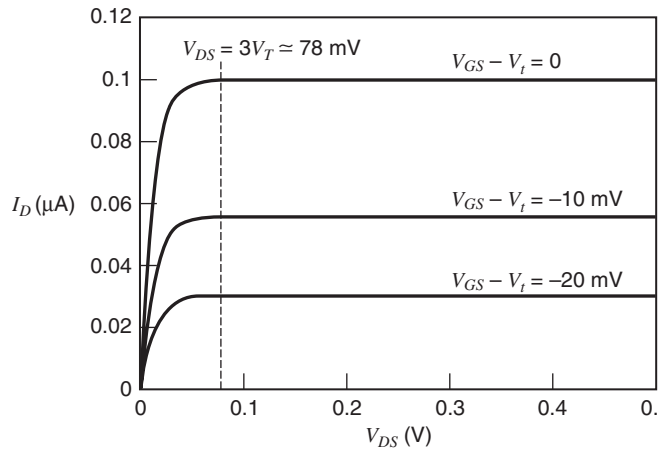
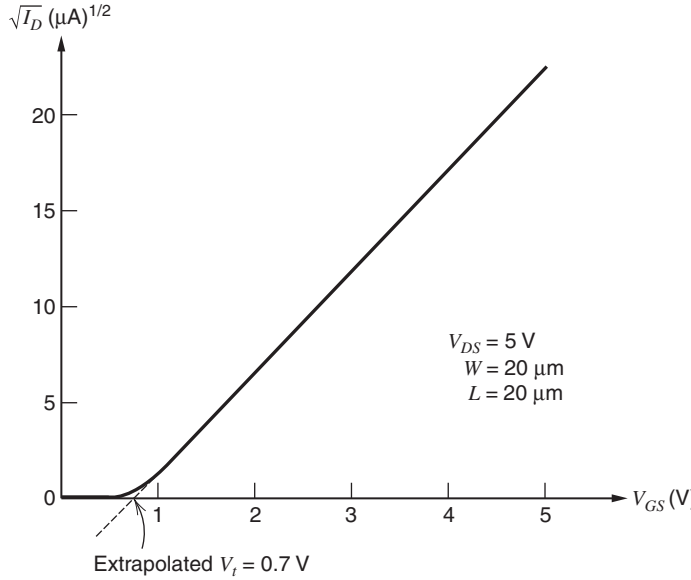
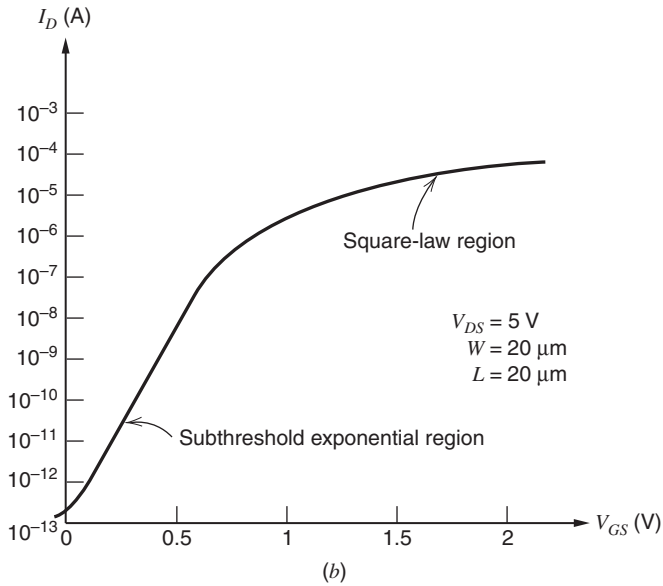


Figure 1.43 Drain current versus drain-source voltage in weak inversion.



(*a*)

Figure 1.44 (*a*) Measured NMOS transfer characteristic in the active region plotted on linear scales as $\sqrt{I_D}$ versus V_{GS}, showing the square-law characteristic.



(*b*)

Figure 1.44 (*b*) Data from Fig. 1.44*a* plotted on log-linear scales showing the exponential characteristic in the subthreshold region.

The major use of transistors operating in weak inversion is in very low power applications at relatively low signal frequencies. The limitation to low signal frequencies occurs because the MOSFET f_T becomes very small. This result stems from the fact that the small-signal g_m calculated from (1.252) becomes proportional to I_D and therefore very small in weak inversion, as shown next.

1.8.2 Transconductance and Transition Frequency in Weak Inversion

Calculating $\partial I_D/\partial V_{GS}$ from (1.252) and using (1.247) gives

$$g_m = \frac{W}{L}\frac{I_t}{nV_T}\exp\left(\frac{V_{GS}-V_t}{nV_T}\right)\left[1-\exp\left(-\frac{V_{DS}}{V_T}\right)\right] = \frac{I_D}{nV_T} = \frac{I_D}{V_T}\frac{C_{ox}}{C_{js}+C_{ox}} \tag{1.253}$$

The transconductance of an MOS transistor operating in weak inversion is identical to that of a corresponding bipolar transistor, as shown in (1.182), except for the factor of $1/n = C_{ox}/(C_{js}+C_{ox})$. This factor stems from a voltage divider between the oxide and depletion capacitors in the MOS transistor, which models the indirect control of the gate on the surface potential.

From (1.253), the ratio of the transconductance to the current of an MOS transistor in weak inversion is

$$\frac{g_m}{I} = \frac{1}{nV_T} = \frac{1}{V_T}\frac{C_{ox}}{C_{js}+C_{ox}} \tag{1.254}$$

Equation 1.254 predicts that this ratio is independent of the overdrive. In contrast, (1.181) predicts that the ratio of transconductance to current is inversely proportional to the overdrive. Therefore, as the overdrive approaches zero, (1.181) predicts that this ratio becomes infinite. However, (1.181) is valid only when the transistor operates in strong inversion. To estimate the overdrive required to operate the transistor in strong inversion, we will equate the g_m/I ratios calculated in (1.254) and (1.181). The result is

$$V_{ov} = V_{GS} - V_t = 2nV_T \tag{1.255}$$

which is about 78 mV at room temperature with $n = 1.5$. Although this analysis implies that the transition from weak to strong inversion occurs abruptly, a nonzero transition width occurs in practice. Between weak and strong inversion, the transistor operates in a region of *moderate* inversion, where both diffusion and drift currents are significant.[38]

Figure 1.45 plots the transconductance-to-current ratio versus overdrive for an example case with $n = 1.5$. When the overdrive is negative but high enough to cause depletion at the surface, the transistor operates in weak inversion and the transconductance-to-current ratio is constant, as predicted by (1.254). When $V_{GS} - V_t = 0$, the surface potential is $2\psi_f$, which means that the surface concentration of electrons is equal to the bulk concentration of holes. This point is usually defined as the upper bound on the region of weak inversion. When

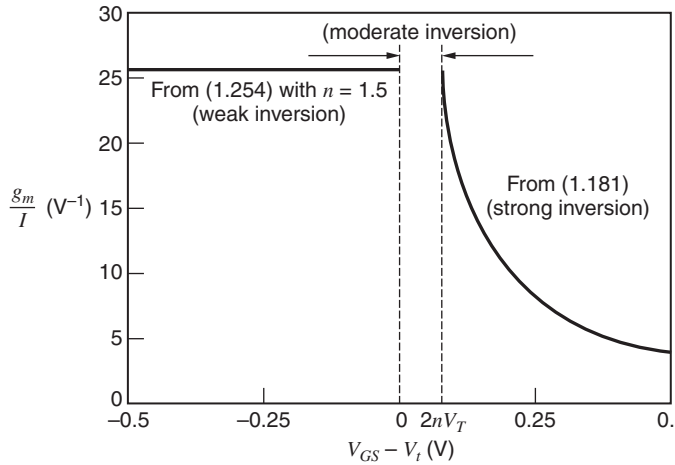


Figure 1.45 Transconductance-to-current ratio versus overdrive.

$V_{GS} - V_t > 2nV_T$, the transconductance-to-current ratio is given by (1.181), assuming that velocity saturation is negligible. If velocity saturation is significant, (1.236) should be used instead of (1.181) both to predict the transconductance-to-current ratio and to predict the overdrive required to operate in strong inversion. For $0 \leq V_{GS} - V_t \leq 2nV_T$, the transistor operates in moderate inversion. Because simple models for moderate inversion are not known in practice, we will ignore this region in the remainder of this book and assume that MOS transistors operate in weak inversion for overdrives less than the bound given in (1.255).

Equation 1.208 can be used to find the transition frequency. In weak inversion, $C_{gs} \simeq C_{gd} \simeq 0$ because the inversion layer contains little charge.[39] However, C_{gb} can be thought of as the series combination of the oxide and depletion capacitors. Therefore,

$$C_{gs} + C_{gb} + C_{gd} \simeq C_{gb} = WL\left(\frac{C_{ox}C_{js}}{C_{ox} + C_{js}}\right) \tag{1.256}$$

Substituting (1.253) and (1.256) into (1.208) gives

$$f_T = \frac{1}{2\pi}\omega_T = \frac{1}{2\pi}\frac{\dfrac{I_D}{V_T}\dfrac{C_{ox}}{C_{js} + C_{ox}}}{WL\dfrac{C_{ox}C_{js}}{C_{ox} + C_{js}}} = \frac{1}{2\pi}\frac{I_D}{V_T}\frac{1}{WLC_{js}} \tag{1.257}$$

Let I_M represent the maximum drain current that flows in the transistor in weak inversion. Then

$$I_M = \frac{W}{L}I_t \tag{1.258}$$

where I_t is given in (1.251). Multiplying numerator and denominator in (1.257) by I_M and using (1.258) gives

$$f_T = \frac{1}{2\pi}\frac{\dfrac{W}{L}I_t}{V_T}\frac{1}{WLC_{js}}\frac{I_D}{I_M} = \frac{1}{2\pi}\frac{I_t}{V_T}\frac{1}{C_{js}}\frac{1}{L^2}\frac{I_D}{I_M} \tag{1.259}$$

From (1.251), $I_t \propto D_n$. Using the Einstein relationship $D_n = \mu_n V_T$ gives

$$f_T \propto \frac{D_n}{L^2}\frac{I_D}{I_M} \propto \frac{\mu_n V_T}{L^2}\frac{I_D}{I_M} \tag{1.260}$$

Equation 1.260 shows that the transition frequency for an MOS transistor operating in weak inversion is inversely proportional to the square of the channel length. This result is consistent with (1.209) for strong inversion without velocity saturation. In contrast, when velocity saturation is significant, the transition frequency is inversely proportional to the channel length, as predicted by (1.241). Equation 1.260 also shows that the transition frequency in weak inversion is independent of the overdrive, unlike the case in strong inversion without velocity saturation, but like the case with velocity saturation. Finally, a more detailed analysis shows that the constant of proportionality in (1.260) is approximately unity.[39]

■ **EXAMPLE**

Calculate the overdrive and the transition frequency for an NMOS transistor with $I_D = 1$ μA, $I_t = 0.1$ μA, and $V_{DS} \gg V_T$. Device parameters are $W = 10$ μm, $L = 1$ μm, $n = 1.5$, $k' = 200$ μA/V^2, and $t_{ox} = 100$ Å. Assume that the temperature is 27°C.

From (1.166), if the transistor operates in strong inversion,

$$V_{ov} = V_{GS} - V_t = \sqrt{\frac{2I_D}{k'(W/L)}} = \sqrt{\frac{2 \times 1}{200 \times 10}} \simeq 32 \text{ mV}$$

Since the value of the overdrive calculated by (1.166) is less than $2nV_T \simeq 78$ mV, the overdrive calculated previously is not valid except to indicate that the transistor does not operate in strong inversion. From (1.252), the overdrive in weak inversion with $V_{DS} \gg V_T$ is

$$V_{ov} = nV_T \ln\left(\frac{I_D}{I_t}\frac{L}{W}\right) = (1.5)(26\text{ mV})\ln\left(\frac{1}{0.1}\frac{1}{10}\right) = 0$$

From (1.253),

$$g_m = \frac{1\ \mu\text{A}}{1.5(26\text{ mV})} \simeq 26\ \frac{\mu\text{A}}{\text{V}}$$

From (1.247),

$$C_{js} = (n-1)C_{ox} = (0.5)C_{ox}$$

From (1.256),

$$\begin{aligned} C_{gs} + C_{gb} + C_{gd} \simeq C_{gb} &= WL\frac{C_{ox}(0.5C_{ox})}{C_{ox} + 0.5C_{ox}} = WL\frac{C_{ox}}{3} \\ &= \frac{10\ \mu\text{m}^2}{3}\frac{3.9 \times 8.854 \times 10^{-14}\dfrac{\text{F}}{\text{cm}} \times \dfrac{100\text{ cm}}{10^6\mu\text{m}}}{100\ \text{Å} \times \dfrac{10^6\ \mu\text{m}}{10^{10}\text{Å}}} \\ &\simeq 11.5\text{ fF} \end{aligned}$$

From (1.208),

$$f_T = \frac{1}{2\pi}\omega_T = \frac{1}{2\pi}\frac{26\ \mu\text{A/V}}{11.5\text{ fF}} \simeq 360\text{ MHz}$$

Although 360 MHz may seem to be a high transition frequency at first glance, this result should be compared with the result of the example at the end of Section 1.6, where the same transistor operating in strong inversion with an overdrive of 316 mV had a transition frequency of 3.4 GHz. ■

1.9 Substrate Current Flow in MOS Transistors

In Section 1.3.4, the effects of avalanche breakdown on bipolar transistor characteristics were described. As the reverse-bias voltages on the device are increased, carriers traversing the depletion regions gain sufficient energy to create new electron-hole pairs in lattice collisions by a process known as *impact ionization*. Eventually, at sufficient bias voltages, the process results in large avalanche currents. For collector-base bias voltages well below the breakdown value, a small enhanced current flow may occur across the collector-base junction due to this process, with little apparent effect on the device characteristics.

Impact ionization also occurs in MOS transistors but has a significantly different effect on the device characteristics. This difference is because the channel electrons (for the NMOS case) create electron-hole pairs in lattice collisions in the drain depletion region, and some of the resulting holes then flow to the substrate, creating a substrate current. (The electrons created in the process flow out the drain terminal.) The carriers created by impact ionization are therefore not confined within the device as in a bipolar transistor. The effect of this phenomenon can be modeled by inclusion of a controlled current generator I_{DB} from drain to substrate, as shown

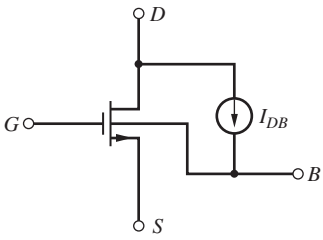


Figure 1.46 Representation of impact ionization in an MOSFET by a drain-substrate current generator.

in Fig. 1.46 for an NMOS device. The magnitude of this substrate current depends on the voltage across the drain depletion region (which determines the energy of the ionizing channel electrons) and also on the drain current (which is the rate at which the channel electrons enter the depletion region). Empirical investigation has shown that the current I_{DB} can be expressed as

$$I_{DB} = K_1(V_{DS} - V_{DS(\text{act})})I_D \exp\left(-\frac{K_2}{V_{DS} - V_{DS(\text{act})}}\right) \tag{1.261}$$

where K_1 and K_2 are process-dependent parameters and $V_{DS(\text{act})}$ is the minimum value of V_{DS} for which the transistor operates in the active region.[40] Typical values for NMOS devices are $K_1 = 5\ \text{V}^{-1}$ and $K_2 = 30$ V. The effect is generally much less significant in PMOS devices because the holes carrying the charge in the channel are much less efficient in creating electron-hole pairs than energetic electrons.

The major impact of this phenomenon on circuit performance is that it creates a parasitic resistance from drain to substrate. Because the common substrate terminal must always be connected to the most negative supply voltage in the circuit, the substrate of an NMOS device in a *p*-substrate process is an ac ground. Therefore, the parasitic resistance shunts the drain to ac ground and can be a limiting factor in many circuit designs. Differentiating (1.261), we find that the drain-substrate small-signal conductance is

$$g_{db} = \frac{\partial I_{DB}}{\partial V_D} = \frac{I_{DB}}{V_{DS} - V_{DS(\text{act})}}\left(\frac{K_2}{V_{DS} - V_{DS(\text{act})}} + 1\right) \simeq \frac{K_2 I_{DB}}{(V_{DS} - V_{DS(\text{act})})^2} \tag{1.262}$$

where the gate and the source are assumed to be held at fixed potentials.

■ **EXAMPLE**

Calculate $r_{db} = 1/g_{db}$ for $V_{DS} = 2$ V and 4 V, and compare with the device r_o. Assume $I_D = 100\ \mu\text{A}$, $\lambda = 0.05\ \text{V}^{-1}$, $V_{DS(\text{act})} = 0.3$ V, $K_1 = 5\ \text{V}^{-1}$, and $K_2 = 30$ V.

For $V_{DS} = 2$ V, we have from (1.261)

$$I_{DB} = 5 \times 1.7 \times 100 \times 10^{-6} \times \exp\left(-\frac{30}{1.7}\right) \simeq 1.8 \times 10^{-11}\ \text{A}$$

From (1.262),

$$g_{db} \simeq \frac{30 \times 1.8 \times 10^{-11}}{1.7^2} \simeq 1.9 \times 10^{-10}\ \frac{\text{A}}{\text{V}}$$

and thus

$$r_{db} = \frac{1}{g_{db}} \simeq 5.3 \times 10^9\ \Omega = 5.3\ \text{G}\Omega$$

This result is negligibly large compared with

$$r_o = \frac{1}{\lambda I_D} = \frac{1}{0.05 \times 100 \times 10^{-6}} = 200\ \text{k}\Omega$$

However, for $V_{DS} = 4$ V,

$$I_{DB} = 5 \times 3.7 \times 100 \times 10^{-6} \times \exp\left(-\frac{30}{3.7}\right) \simeq 5.6 \times 10^{-7}\ \text{A}$$

The substrate leakage current is now about 0.5 percent of the drain current. More important, we find from (1.262)

$$g_{db} \simeq \frac{30 \times 5.6 \times 10^{-7}}{3.7^2} \simeq 1.2 \times 10^{-6}\ \frac{\text{A}}{\text{V}}$$

and thus

$$r_{db} = \frac{1}{g_{db}} \simeq 8.15 \times 10^5\ \Omega = 815\ \text{k}\Omega$$

This parasitic resistor is now comparable to r_o and can have a dominant effect on high-output-impedance MOS current mirrors, as described in Chapter 4. ■

APPENDIX

A.1.1 SUMMARY OF ACTIVE-DEVICE PARAMETERS

(a) *npn* Bipolar Transistor Parameters

Quantity	Formula
Large-Signal Forward-Active Operation	
Collector current	$I_c = I_S \exp \frac{V_{be}}{V_T}$
Small-Signal Forward-Active Operation	
Transconductance	$g_m = \frac{qI_C}{kT} = \frac{I_C}{V_T}$
Transconductance-to-current ratio	$\frac{g_m}{I_C} = \frac{1}{V_T}$
Input resistance	$r_\pi = \frac{\beta_0}{g_m}$
Output resistance	$r_o = \frac{V_A}{I_C} = \frac{1}{\eta g_m}$
Collector-base resistance	$r_\mu = \beta_0 r_o$ to $5\beta_0 r_o$ → 101
Base-charging capacitance	$C_b = \tau_F\ g_m$
Base-emitter capacitance	$C_\pi = C_b + C_{je}$
Emitter-base junction depletion capacitance	$C_{je} \simeq 2C_{je0}$

(*continued*)

Quantity	Formula
Small-Signal Forward-Active Operation	
Collector-base junction capacitance	$C_\mu = \dfrac{C_{\mu 0}}{\left(1 - \dfrac{V_{BC}}{\psi_{0c}}\right)^{n_c}}$
Collector-substrate junction capacitance	$C_{cs} = \dfrac{C_{cs0}}{\left(1 - \dfrac{V_{SC}}{\psi_{0s}}\right)^{n_s}}$
Transition frequency	$f_T = \dfrac{1}{2\pi}\dfrac{g_m}{C_\pi + C_\mu}$
Effective transit time	$\tau_T = \dfrac{1}{2\pi f_T} = \tau_F + \dfrac{C_{je}}{g_m} + \dfrac{C_\mu}{g_m}$
Maximum gain	$g_m r_o = \dfrac{V_A}{V_T} = \dfrac{1}{\eta}$

(b) NMOS Transistor Parameters

Quantity	Formula
Large-Signal Operation	
Drain current (active region)	$I_d = \dfrac{\mu C_{ox}}{2}\dfrac{W}{L}(V_{gs} - V_t)^2$
Drain current (triode region)	$I_d = \dfrac{\mu C_{ox}}{2}\dfrac{W}{L}[2(V_{gs} - V_t)V_{ds} - {V_{ds}}^2]$
Threshold voltage	$V_t = V_{t0} + \gamma\left[\sqrt{2\phi_f + V_{sb}} - \sqrt{2\phi_f}\right]$
Threshold voltage parameter	$\gamma = \dfrac{1}{C_{ox}}\sqrt{2q\epsilon N_A}$
Oxide capacitance	$C_{ox} = \dfrac{\epsilon_{ox}}{t_{ox}} = 3.45\ \text{fF}/\mu\text{m}^2$ for $t_{ox} = 100\ \text{Å}$
Small-Signal Operation (Active Region)	
Top-gate transconductance	$g_m = \mu C_{ox}\dfrac{W}{L}(V_{GS} - V_t) = \sqrt{2I_D\mu C_{ox}\dfrac{W}{L}}$
Transconductance-to-current ratio	$\dfrac{g_m}{I_D} = \dfrac{2}{V_{GS} - V_t}$
Body-effect transconductance	$g_{mb} = \dfrac{\gamma}{2\sqrt{2\phi_f + V_{SB}}}g_m = \chi g_m$
Channel-length modulation parameter	$\lambda = \dfrac{1}{V_A} = \dfrac{1}{L_{\text{eff}}}\dfrac{dX_d}{dV_{DS}}$

(*continued*)

Quantity	Formula
Small-Signal Operation (Active Region)	
Output resistance	$r_o = \dfrac{1}{\lambda I_D} = \dfrac{L_{\text{eff}}}{I_D}\left(\dfrac{dX_d}{dV_{DS}}\right)^{-1}$
Effective channel length	$L_{\text{eff}} = L_{\text{drwn}} - 2L_d - X_d$
Maximum gain	$g_m r_o = \dfrac{1}{\lambda}\dfrac{2}{V_{GS} - V_t} = \dfrac{2V_A}{V_{GS} - V_t}$
Source-body depletion capacitance	$C_{sb} = \dfrac{C_{sb0}}{\left(1 + \dfrac{V_{SB}}{\psi_0}\right)^{0.5}}$
Drain-body depletion capacitance	$C_{db} = \dfrac{C_{db0}}{\left(1 + \dfrac{V_{DB}}{\psi_0}\right)^{0.5}}$
Gate-source capacitance	$C_{gs} = \dfrac{2}{3}WLC_{ox}$
Transition frequency	$f_T = \dfrac{g_m}{2\pi(C_{gs} + C_{gd} + C_{gb})}$

PROBLEMS

1.1(a) Calculate the built-in potential, depletion-layer depths, and maximum field in a plane-abrupt *pn* junction in silicon with doping densities $N_A = 8 \times 10^{15}$ atoms/cm^3 and $N_D = 10^{17}$ atoms/cm^3. Assume a reverse bias of 5 V.

(b) Repeat (a) for zero external bias and 0.3 V forward bias.

1.2 Calculate the zero-bias junction capacitance for the example in Problem 1.1, and also calculate the value at 5 V reverse bias and 0.3 V forward bias. Assume a junction area of 2×10^{-5} cm^2.

1.3 Calculate the breakdown voltage for the junction of Problem 1.1 if the critical field is $\mathscr{E}_{\text{crit}} = 4 \times 10^5$ V/cm.

1.4 If junction curvature causes the maximum field at a practical junction to be 1.5 times the theoretical value, calculate the doping density required to give a breakdown voltage of 150 V with an abrupt *pn* junction in silicon. Assume that one side of the junction is much more heavily doped than the other and $\mathscr{E}_{\text{crit}} = 3 \times 10^5$ V/cm.

1.5 If the collector doping density in a transistor is 6×10^{15} atoms/cm^3, and is much less than the base doping, find BV_{CEO} for $\beta_F = 200$ and $n = 4$. Use $\mathscr{E}_{\text{crit}} = 3 \times 10^5$ V/cm.

1.6 Repeat Problem 1.5 for a doping density of 10^{15} atoms/cm^3 and $\beta_F = 400$.

1.7(a) Sketch the I_C-V_{CE} characteristics in the forward-active region for an *npn* transistor with $\beta_F = 100$ (measured at low V_{CE}), $V_A = 50$ V, $BV_{CBO} = 120$ V, and $n = 4$. Use

$$I_C = \left(1 + \frac{V_{CE}}{V_A}\right)\frac{M\alpha_F}{1 - M\alpha_F} I_B$$

where M is given by (1.78). Plot I_C from 0 to 10 mA and V_{CE} from 0 to 50 V. Use $I_B = 1\ \mu$A, 10 μA, 30 μA, and 60 μA.

(b) Repeat (a), but sketch V_{CE} from 0 to 10 V.

1.8 Derive and sketch the complete small-signal equivalent circuit for a bipolar transistor at $I_C = 0.2$ mA, $V_{CB} = 3$ V, $V_{CS} = 4$ V. Device parameters are $C_{je0} = 20$ fF, $C_{\mu 0} = 10$ fF, $C_{cs0} = 20$ fF, $\beta_0 = 100$, $\tau_F = 15$ ps, $\eta = 10^{-3}$, $r_b = 200\ \Omega$, $r_c = 100\ \Omega$, $r_{ex} = 4\ \Omega$, and $r_\mu = 5\beta_0 r_o$. Assume $\psi_0 = 0.55$ V for all junctions.

1.9 Repeat Problem 1.8 for $I_C = 1$ mA, $V_{CB} = 1$ V, and $V_{CS} = 2$ V.

1.10 Sketch the graph of small-signal, common-emitter current gain versus frequency on log scales from 0.1 MHz to 1000 MHz for the examples of

Problems 1.8 and 1.9. Calculate the f_T of the device in each case.

1.11 An integrated-circuit *npn* transistor has the following measured characteristics: $r_b = 100\ \Omega$, $r_c = 100\ \Omega$, $\beta_0 = 100$, $r_o = 50\ \text{k}\Omega$ at $I_C = 1$ mA, $f_T = 600$ MHz with $I_C = 1$ mA and $V_{CB} = 10$ V, $f_T = 1$ GHz with $I_C = 10$ mA and $V_{CB} = 10$ V, $C_\mu = 0.15$ pF with $V_{CB} = 10$ V, and $C_{cs} = 1$ pF with $V_{CS} = 10$ V. Assume $\psi_0 = 0.55$ V for all junctions, and assume C_{je} is constant in the forward-bias region. Use $r_\mu = 5\beta_0 r_o$.

(a) Form the *complete* small-signal equivalent circuit for this transistor at $I_C = 0.1$ mA, 1 mA, and 5 mA with $V_{CB} = 2$ V and $V_{CS} = 15$ V.

(b) Sketch the graph of f_T versus I_C for this transistor on log scales from 1 μA to 10 mA with $V_{CB} = 2$ V.

1.12 A lateral *pnp* transistor has an effective base width of 10 μm (1 μm $= 10^{-4}$ cm).

(a) If the emitter-base depletion capacitance is 2 pF in the forward-bias region and is constant, calculate the device f_T at $I_C = -0.5$ mA. (Neglect C_μ.) Also, calculate the minority-carrier charge stored in the base of the transistor at this current level. *Data*: $D_P = 13\ \text{cm}^2/\text{s}$ in silicon.

(b) If the collector-base depletion layer width changes 0.11 μm per volt of V_{CE}, calculate r_o for this transistor at $I_C = -0.5$ mA.

1.13 If the area of the transistor in Problem 1.11 is effectively doubled by connecting two transistors in parallel, which model parameters in the small-signal equivalent circuit of the composite transistor would differ from those of the original device if the *total* collector current is unchanged? What is the relationship between the parameters of the composite and original devices?

1.14 An integrated *npn* transistor has the following characteristics: $\tau_F = 0.25$ ns, small-signal, short-circuit current gain is 9 with $I_C = 1$ mA at $f = 50$ MHz, $V_A = 40$ V, $\beta_0 = 100$, $r_b = 150\ \Omega$, $r_c = 150\ \Omega$, $C_\mu = 0.6$ pF, $C_{cs} = 2$ pF at the bias voltage used. Determine all elements in the small-signal equivalent circuit at $I_C = 2$ mA and sketch the circuit.

1.15 An NMOS transistor has parameters $W = 10$ μm, $L = 1$ μm, $k' = 194\ \mu\text{A/V}^2$, $\lambda = 0.024\ \text{V}^{-1}$, $t_{ox} = 80$ Å, $\phi_f = 0.3$ V, $V_{t0} = 0.6$ V, and $N_A = 5 \times 10^{15}\ \text{atoms/cm}^3$. Ignore velocity saturation effects.

(a) Sketch the I_D-V_{DS} characteristics for V_{DS} from 0 to 3 V and $V_{GS} = 0.5$ V, 1.5 V, and 3 V. Assume $V_{SB} = 0$.

(b) Sketch the I_D-V_{GS} characteristics for $V_{DS} = 2$ V as V_{GS} varies from 0 to 2 V with $V_{SB} = 0$, 0.5 V, and 1 V.

1.16 Derive and sketch the complete small-signal equivalent circuit for the device of Problem 1.15 with $V_{GS} = 1$ V, $V_{DS} = 2$ V, and $V_{SB} = 1$ V. Use $\psi_0 = 0.7$ V, $C_{sb0} = C_{db0} = 20$ fF, and $C_{gb} = 5$ fF. Overlap capacitance from gate to source and gate to drain is 2 fF.

1.17 Use the device data of Problems 1.15 and 1.16 to calculate the frequency of unity current gain of this transistor with $V_{DS} = 3$ V, $V_{SB} = 0$ V, $V_{GS} = 1$ V, 1.5 V, and 2 V.

1.18 Examine the effect of velocity saturation on MOSFET characteristics by plotting $I_D - V_{DS}$ curves for $V_{GS} = 1$ V, 2 V, and 3 V, and $V_{DS} = 0$ to 3 V in the following cases, and by comparing the results with and without inclusion of velocity saturation effects. Assume $V_{SB} = 0$, $V_{t0} = 0.6$ V, $k' = 194\ \mu\text{A/V}^2$, $\lambda = 0$, and $\mathscr{E}_c = 1.5 \times 10^6$ V/m.

(a) $W = 100$ μm and $L = 10$ μm.

(b) $W = 10$ μm and $L = 1$ μm.

(c) $W = 5$ μm and $L = 0.5$ μm.

1.19 Consider an NMOS transistor with $W = 2$ μm, $L = 0.5$ μm, $k' = 194\ \mu\text{A/V}^2$, $\lambda = 0$, $V_{t0} = 0.6$ V, and $\mathscr{E}_c = 1.5 \times 10^6$ V/m. Compare the drain current predicted by the model of Fig. 1.41 to the drain current predicted by direct calculation using the equations including velocity saturation for V_{GS} from 0 to 3 V. Assume $V_{DS} = 3$ V and $V_{SB} = 0$. For what range of V_{GS} is the model of Fig. 1.41 accurate within 10 percent?

1.20 Calculate the transconductance of an *n*-channel MOSFET with $W = 10$ μm, $\mu_n = 450\ \text{cm}^2/(\text{V-s})$, and $\mathscr{E}_c = 1.5 \times 10^6$ V/m using channel lengths from 10 μm to 0.4 μm. Assume that $t_{ox} = L/50$ and that the device operates in the active region with $V_{GS} - V_t = 0.1$ V. Compare the result to a calculation that ignores velocity saturation. For what range of channel lengths is the model without velocity saturation accurate within 10 percent?

1.21 Plot $\sqrt{I_D}$ versus V_{GS} for an *n*-channel MOSFET with $W = 1$ μm, $L = 1$ μm, $k' = 54\ \mu\text{A/V}^2$, $\lambda = 0$, $V_{DS} = 5$ V, $V_{SB} = 0$, $V_{t0} = 0.7$ V, and $\mathscr{E}_c = 1.5 \times 10^6$ V/m. Ignore subthreshold conduction. Compare the plot with Fig. 1.44*a* and explain the main difference for large V_{GS}.

1.22 Calculate the transconductance of an *n*-channel MOSFET at $I_D = 10$ nA and $V_{DS} = 1$ V, assuming subthreshold operation and $n = 1.5$. Assuming $(C_{gs} + C_{gd} + C_{gb}) = 10$ fF, calculate the corresponding device f_T.

REFERENCES

1. P. E. Gray, D. DeWitt, A. R. Boothroyd, J. F. Gibbons. *Physical Electronics and Circuit Models of Transistors.* Wiley, New York, 1964, p. 20.
2. H. C. Poon and H. K. Gummel. "Modeling of Emitter Capacitance," *Proc. IEEE,* Vol. 57, pp. 2181–2182, December 1969.
3. B. R. Chawla and H. K. Gummel. "Transition Region Capacitance of Diffused pn Junctions," *IEEE Trans. Electron Devices,* Vol. ED-18, pp. 178–195, March 1971.
4. S. L. Miller. "Avalanche Breakdown in Germanium," *Phys. Rev.,* Vol. 99, p. 1234, 1955.
5. A. S. Grove. *Physics and Technology of Semiconductor Devices.* Wiley, New York, 1967, Ch. 6.
6. A. S. Grove. Op. cit., Ch. 4.
7. A. S. Grove. Op. cit., Ch. 7.
8. P. E. Gray et al. Op. cit., p. 10.
9. P. E. Gray et al. Op. cit., p. 129.
10. P. E. Gray et al. Op. cit., p. 180.
11. B. A. McDonald. "Avalanche Degradation of h_{FE}," *IEEE Trans. Electron Devices,* Vol. ED-17, pp. 871–878, October 1970.
12. H. DeMan. "The Influence of Heavy Doping on the Emitter Efficiency of a Bipolar Transistor," *IEEE Trans. Electron Devices,* Vol. ED-18, pp. 833–835, October 1971.
13. R. J. Whittier and D. A. Tremere. "Current Gain and Cutoff Frequency Falloff at High Currents," *IEEE Trans. Electron Devices,* Vol. ED-16, pp. 39–57, January 1969.
14. J. L. Moll and I. M. Ross. "The Dependence of Transistor Parameters on the Distribution of Base Layer Resistivity," *Proc. IRE,* Vol. 44, p. 72, 1956.
15. P. E. Gray et al. Op. cit., Ch. 8.
16. R. S. Muller and T. I. Kamins. *Device Electronics for Integrated Circuits.* Second Edition, Wiley, New York, 1986, p. 386.
17. Y. P. Tsividis. *Operation and Modeling of the MOS Transistor.* McGraw-Hill, New York, 1987, p. 141.
18. D. Frohman-Bentchkowsky and A. S. Grove. "Conductance of MOS Transistors in Saturation," *IEEE Trans. Electron Devices,* Vol. ED-16, pp. 108–113, January 1969.
19. S. M. Sze. *Physics of Semiconductor Devices.* Second Edition, Wiley, New York, 1981, pp. 451– 452.
20. R. S. Muller and T. I. Kamins. Op. cit., p. 17.
21. Y. P. Tsividis. Op. cit., p. 148.
22. R. S. Muller and T. I. Kamins. Op. cit., pp. 490–496.
23. Y. P. Tsividis. Op. cit., pp. 150–151 and 198–200.
24. R. S. Muller and T. I. Kamins. Op. cit., p. 496.
25. Y. P. Tsividis. Op. cit., p. 151.
26. Y. P. Tsividis. Op. cit., pp. 310–328.
27. R. S. Muller and T. I. Kamins. Op. cit., p. 480.
28. R. S. Muller and T. I. Kamins. Op. cit., p. 482.
29. Y. P. Tsividis. Op. cit., p. 181.
30. Y. P. Tsividis. Op. cit., p. 294.
31. Y. P. Tsividis. Op. cit., p. 142.
32. R. S. Muller and T. I. Kamins. Op. cit., p. 484.
33. Y. P. Tsividis. Op. cit., p. 146.
34. S. M. Sze. Op. cit., p. 446.
35. Y. P. Tsividis. Op. cit., p. 136.
36. Y. P. Tsividis. Op. cit., p. 83.
37. Y. P. Tsividis. Op. cit., p. 139.
38. Y. P. Tsividis. Op. cit., p. 137.
39. Y. P. Tsividis. Op. cit., p. 324.
40. K. Y. Toh, P. K. Ko, and R. G. Meyer. "An Engineering Model for Short-Channel MOS Devices," *IEEE Journal of Solid-State Circuits,* Vol. 23, pp. 950–958, August 1988.

GENERAL REFERENCES

I. Getreu. *Modeling the Bipolar Transistor.* Tektronix Inc., 1976.

P. E. Gray and C. L. Searle, *Electronic Principles.* Wiley, New York, 1969.

R. S. Muller and T. I Kamins. *Device Electronics for Integrated Circuits.* Wiley, New York, 1986.

Y. P. Tsividis. *Operation and Modeling of the MOS Transistor.* McGraw-Hill, New York, 1987.