I

Analog Integrated Circuits

John Choma, Jr.
University of Southern California

1

Monolithic Device Models

Bogdan M. Wilamowski
Auburn University

Guofu Niu
Auburn University

John Choma, Jr.
University of Southern California

Stephen I. Long
University of California, Santa Barbara

Nhat M. Nguyen
Rambus Inc.

Martin A. Brooke
Georgia Institute of Technology

1.1 Bipolar Junction Transistor

Bogdan M. Wilamowski and Guofu Niu

The bipolar junction transistor (BJT) is historically the first solid-state analog amplifier and digital switch, and formed the basis of integrated circuits (ICs) in the 1970s. Starting in the early 1980s, the MOSFET had gradually taken over, particularly for main stream digital ICs. However, in the 1990s, the invention of silicon–germanium base heterojunction bipolar transistor (SiGe HBT) brought the bipolar transistor back into high-volume commercial production, mainly for the now widespread wireless and wire line communications applications. Today, SiGe HBTs are used to design radio-frequency (RF) ICs and systems for cell phones, wireless local area network (WLAN), automobile collision avoidance

radar, wireless distribution of cable television, millimeter wave radios, and many more applications, due to its outstanding high-frequency performance and ability to integrate with CMOS for realizing digital, analog, and RF functions on the same chip.

Below we first introduce the basic concepts of BJT using a historically important equivalent circuit model, the Ebers–Moll model. Then the Gummel–Poon model is introduced, as it is widely used for computer-aided design, and is the basis of modern BJT models like the VBIC, Mextram, and HICUM models. Current gain, high-current phenomena, fabrication technologies, and SiGe HBTs are then discussed.

1.1.1 Ebers–Moll Model

A NPN BJT consists of two closely spaced PN junctions connected back to back sharing the same p-type region, as shown in Figure 1.1a. The drawing is not drawn to scale. The emitter and base layers are thin, typically less than 1 μm, and the collector is much thicker to support a high output voltage swing. For forward mode operation, the emitter–base (EB) junction is forward biased, and the collector–base (CB) junction is reverse biased. Minority carriers are injected from emitter to base, travel across the base, and are then collected by the reverse biased CB junction. Therefore, the collector current is transported from the EB junction, and thus proportional to the EB junction current. In the forward-active mode, the current–voltage characteristic of the EB junction is described by the well-known diode equation

$$I_{EF} = I_{E0}\left[\exp\left(\frac{V_{BE}}{V_T}\right) - 1\right] \tag{1.1}$$

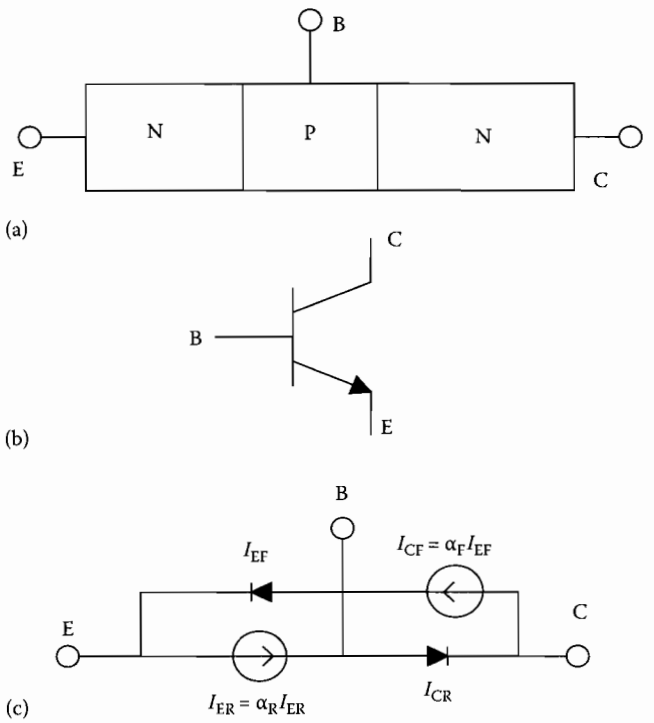

FIGURE 1.1 (a) Cross-sectional view of a NPN BJT. (b) Circuit symbol. (c) The Ebers–Moll equivalent circuit model.

where

I_{E0} is the EB junction saturation current

$V_T = kT/q$ is the thermal potential (about 25 mV at room temperature)

The collector current is typically smaller than the emitter current $I_{CF} = \alpha_F I_{EF}$, where α_F is the forward current gain.

Under reverse mode operation, the CB junction is forward biased and the EB junction is reverse biased. Like in the forward mode, the forward biased CB junction current gives the collector current

$$I_{CF} = I_{C0}\left[\exp\left(\frac{V_{BC}}{V_T}\right) - 1\right] \tag{1.2}$$

where I_{C0} is the CB junction saturation current. Similarly $I_{ER} = \alpha_R I_R$, where α_R is the reverse current gain. Under general biasing conditions, it can be proven that to first order, a superposition of the above described forward and reverse mode equivalent circuits can be used to describe transistor operation, as shown in Figure 1.1b. The forward transistor operation is described by Equation 1.1, and the reverse transistor operation is described by Equation 1.2. From the Kirchoff's current law one can write $I_C = I_{CF} - I_{CR}$, $I_E = I_{EF} - I_{ER}$, and $I_B = I_E - I_C$. Using Equations 1.1 and 1.2 the emitter and collector currents can be described as

$$\begin{aligned} I_E &= a_{11}\left(\exp\frac{V_{BE}}{V_T} - 1\right) - a_{12}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \\ I_C &= a_{21}\left(\exp\frac{V_{BE}}{V_T} - 1\right) - a_{22}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \end{aligned} \tag{1.3}$$

which are known as the Ebers–Moll equations [1]. The Ebers–Moll coefficients a_{ij} are given as

$$a_{11} = I_{E0}, \quad a_{12} = \alpha_R I_{C0}, \quad a_{21} = \alpha_F I_{E0}, \quad a_{22} = I_{C0} \tag{1.4}$$

The Ebers–Moll coefficients are a very strong function of the temperature

$$a_{ij} = K_x T^m \exp\frac{V_{go}}{V_T} \tag{1.5}$$

where

K_x is proportional to the junction area and independent of the temperature

$V_{go} = 1.21$ V is the bandgap voltage in silicon (extrapolated to 0 K)

m is a material constant with a value between 2.5 and 4

When both EB and CB junctions are forward biased, the transistor is called to be working in the saturation region. Current injection through the collector junction may activate parasitic transistors in ICs using p-type substrate, where base acts as emitter, collector as base, and substrate as collector. In typical ICs, bipolar transistors must not operate in saturation. Therefore, for the integrated bipolar transistor the Ebers–Moll equations can be simplified to the form

$$\begin{aligned} I_E &= a_{11}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \\ I_C &= a_{21}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \end{aligned} \tag{1.6}$$

where $a_{21}/a_{11} = \alpha_F$. This equation corresponds to the circuit diagram shown in Figure 1.1c.

1.1.2 Gummel–Poon Model

In real bipolar transistors the current voltage characteristics are more complex than those described by the Ebers–Moll equations. Typical current–voltage characteristics of the bipolar transistor, plotted in semilogarithmic scale, are shown in Figure 1.2. At small-base emitter voltages, due to the generation–recombination phenomena, the base current is proportional to

$$I_{BL} \propto \exp\frac{V_{BE}}{2V_T} \tag{1.7}$$

Also, due to the base conductivity modulation at high-level injections, the collector current for larger voltages can be expressed by the similar relation

$$I_{CH} \propto \exp\frac{V_{BE}}{2V_T} \tag{1.8}$$

Note, that the collector current for wide range is given by

$$I_C = I_s \exp\frac{V_{BE}}{V_T} \tag{1.9}$$

The saturation current is a function of device structure parameters

$$I_s = \frac{qAn_i^2 V_T \mu_B}{\int_0^{w_B} N_B(x)dx} \tag{1.10}$$

where

$q = 1.6 \times 10^{-19}$ C is the electron charge
A is the EB junction area
n_i is the intrinsic concentration ($n_i = 1.5 \times 10^{10}$ at 300 K)
μ_B is the mobility of the majority carriers in the transistor base
w_B is the effective base thickness
$N_B(x)$ is the distribution of impurities in the base

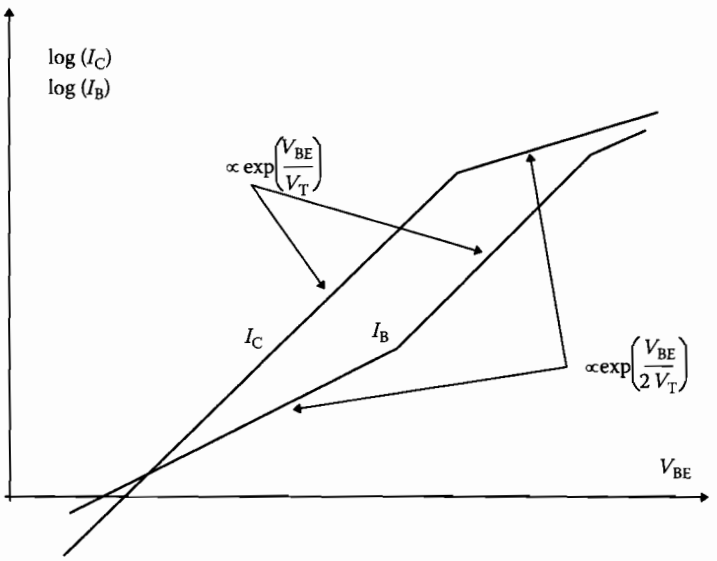

FIGURE 1.2 Collector and base currents as a function of base–emitter voltage.

Note, that the saturation current is inversely proportional to the total impurity dose in the base. In the transistor with the uniform base, the saturation current is given by

$$I_s = \frac{qAn_i^2 V_T \mu_B}{w_B N_B} \tag{1.11}$$

When a transistor operates in the reverse-active mode (emitter and collector are switched) then the current of such biased transistor is given by

$$I_E = I_s \exp \frac{V_{BC}}{V_T} \tag{1.12}$$

Note, that the I_s parameter is the same for forward and reverse mode of operation. The Gummel–Poon transistor model [2] was derived from the Ebers–Moll model using the assumption that $a_{12} = a_{21} = I_s$. For the Gummel–Poon model, Equations 1.3 are simplified to the form

$$\begin{aligned} I_E &= I_s\left(\frac{1}{\alpha_F}\exp\frac{V_{BE}}{V_T} - \exp\frac{V_{BC}}{V_T}\right) \\ I_C &= I_s\left(\exp\frac{V_{BE}}{V_T} - \frac{1}{\alpha_R}\exp\frac{V_{BC}}{V_T}\right) \end{aligned} \tag{1.13}$$

These equations require only three coefficients, while the Ebers–Moll requires four. The saturation current I_s is constant for a wide range of currents. The current gain coefficients α_F and α_R have values smaller, but close to unity. Often instead of using the current gain as $\alpha = I_C/I_E$, the current gain β as a ratio of the collector current to the base current $\beta = I_C/I_B$ is used. The mutual relationships between α and β coefficients are given by

$$\alpha_F = \frac{\beta_F}{\beta_F + 1}, \quad \beta_F = \frac{\alpha_F}{1 - \alpha_F}, \quad \alpha_R = \frac{\beta_R}{\beta_R + 1}, \quad \beta_R = \frac{\alpha_R}{1 - \alpha_R} \tag{1.14}$$

The Gummel–Poon model was implemented in Simulation Program with Integrated Circuit Emphasis (SPICE) [3] and other computer programs for circuit analysis. To make the equations more general, the material parameters η_F and η_R were introduced

$$I_C = I_s\left[\exp\frac{V_{BE}}{\eta_F V_T} - \left(1 + \frac{1}{\beta_R}\right)\exp\frac{V_{BC}}{\eta_R V_T}\right] \tag{1.15}$$

The values of η_F and η_R vary from 1 to 2.

1.1.3 Current Gains of Bipolar Transistors

The transistor current gain β is limited by two phenomena: base transport efficiency and emitter injection efficiency. The effective current gain β can be expressed as

$$\frac{1}{\beta} = \frac{1}{\beta_I} + \frac{1}{\beta_T} + \frac{1}{\beta_R} \tag{1.16}$$

where

β_I is the transistor current gain caused by emitter injection efficiency
β_T is the transistor current gain caused by base transport efficiency
β_R is the recombination component of the current gain

As one can see from Equation 1.16, smaller values of β_I, β_T, and β_R dominate. The base transport efficiency can be defined as a ratio of injected carriers into the base, to the carriers that recombine within the base. This ratio is also equal to the ratio of the minority carrier life time, to the transit time of carriers through the base. The carrier transit time can be approximated by an empirical relationship

$$\tau_{transit} = \frac{w_B^2}{V_T \mu_B (2 + 0.9\eta)}, \quad \eta = \ln\left(\frac{N_{BE}}{N_{BC}}\right) \tag{1.17}$$

where
- μ_B is the mobility of the minority carriers in base
- w_B is the base thickness
- N_{BE} is the impurity doping level at the emitter side of the base
- N_{BC} is the impurity doping level at the collector side of the base

Therefore, the current gain due to the transport efficiency is

$$\beta_T = \frac{\tau_{life}}{\tau_{transit}} = (2 + 0.9\eta)\left(\frac{L_B}{w_B}\right)^2 \tag{1.18}$$

where $L_B = \sqrt{V_T \mu_B \tau_{life}}$ is the diffusion length of minority carriers in the base.

The current gain β_I, due to the emitter injection efficiency, is given

$$\beta_I = \frac{\mu_B \int_0^{w_E} N_{Eeff}(x)dx}{\mu_E \int_0^{w_B} N_B(x)dx} \tag{1.19}$$

where
- μ_B and μ_E are minority carrier mobilities in the base and in the emitter
- $N_B(x)$ is impurity distribution in the base
- N_{Eeff} is the effective impurity distribution in the emitter

The recombination component of current gain β_R is caused by the different current–voltage relationship of base and collector currents as can be seen in Figure 1.2. The slower base current increase is due to the recombination phenomenon within the depletion layer of the base–emitter junction. Since the current gain is a ratio of the collector current to the base current, the relation for β_R can be found as

$$\beta_R = K_{R0} I_C^{1-(1/\eta_R)} \tag{1.20}$$

As it can be seen from Figure 1.2, the current gain β is a function of the current. This gain–current relationship is illustrated in Figure 1.3. The range of a constant current gain is wide for bipolar transistors with a technology characterized by a lower number of generation–recombination centers.

With an increase of CB voltage, the depletion layer penetrates deeper into the base. Therefore, the effective thickness of the base decreases. This leads to an increase of transistor current gain with applied collector voltages. Figure 1.4 illustrates this phenomenon, which is known as the Early's effect. The extensions of transistor characteristics (dotted lines in Figure 1.4) are crossing the voltage axis at

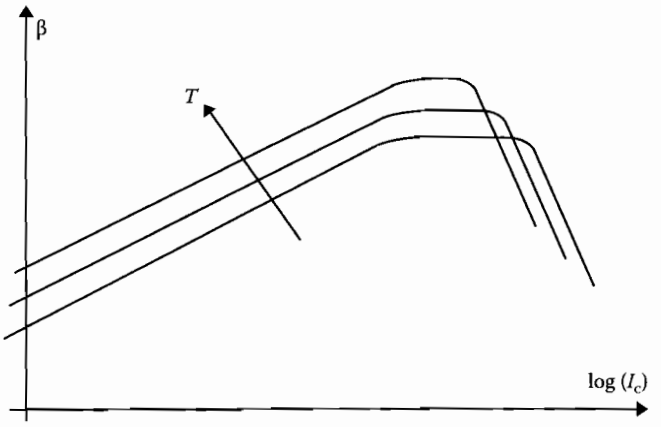

FIGURE 1.3 Current gain β as a function of collector current.

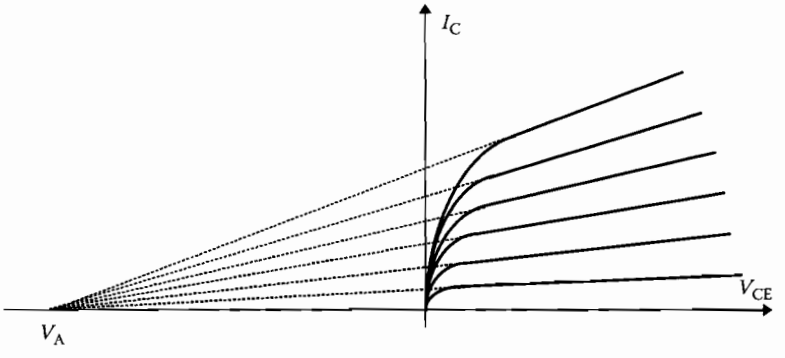

FIGURE 1.4 Current–voltage characteristics of a bipolar transistor.

the point $-V_A$, where V_A is known as the Early voltage. The current gain β, as a function of collector voltage, is usually expressed using the relation

$$\beta = \beta_o\left(1 + \frac{V_{CE}}{V_A}\right) \tag{1.21}$$

Similar equation can be defined for the reverse mode of operation.

1.1.4 High-Current Phenomena

The concentration of minority carriers increases with the rise of transistor currents. When the concentration of moving carriers exceeds a certain limit, the transistor property degenerates. Two phenomena are responsible for this limitation. The first is related to the high concentration of moving carriers (electrons in the NPN transistor) in the base–collector depletion region. This is known as the Kirk effect. The second phenomenon is caused by a high level of carriers injected into the base. When the concentration of injected minority carriers in the base exceeds the impurity concentration there, then the base conductivity modulation limits the transistor performance.

To understand the Kirk effect consider the NPN transistor in forward-active mode with the base–collector junction reversely biased. The depletion layer consists of the negative lattice charge of the base

region and the positive lattice charge of the collector region. Boundaries of the depletion layer are such that total the positive and negative charges are equal. When a collector current, carrying negatively charged electrons, flows through the junction, effective negative charge on the base side of junction increases. Also, the positive lattice charge of the collector side of the junction is compensated by negative charge of moving electrons. This way, the CB space charge region moves toward the collector, resulting in a thicker effective base. With a large current level, the thickness of the base may be doubled or tripled. This phenomenon, known as the Kirk effect, becomes very significant when the charge of moving electrons exceeds the charge of the lightly doped collector N_C. The threshold current for the Kirk effect is given by

$$I_{max} = qAv_{sat}N_C \tag{1.22}$$

where v_{sat} is the saturation velocity for electrons ($v_{sat} = 10^7$ cm/s for silicon).

The conductivity modulation in the base, or high-level injection, starts when the concentration of injected electrons into the base exceeds the lowest impurity concentration in the base N_{Bmin}. This occurs for the collector current I_{max} given by

$$I_{max} < qAN_{Bmax}, \quad v = \frac{qAV_T\mu_B N_{Bmax}(2 + 0.9\eta)}{w_B} \tag{1.23}$$

The above equation is derived using Equation 1.17 for the estimation of base transient time.

The high-current phenomena are significantly enlarged by the current crowding effect. The typical cross section of bipolar transistor is shown in Figure 1.5. The horizontal flow of the base current results in the voltage drop across the base region under the emitter. This small voltage difference on the base–emitter junction causes a significant difference in the current densities at the junction. This is due to the very nonlinear junction current–voltage characteristics. As a result, the base–emitter junction has very nonuniform current distribution across the junction. Most of the current flows through the part of the junction closest to base contact. For transistors with larger emitter areas, the current crowding effect is more significant. This nonuniform transistor current distribution makes the high-current phenomena, such as the base conductivity modulation and the Kirk effect, start for smaller currents than given by Equations 1.22 and 1.23. The current crowding effect is also responsible for the change of the effective base resistance with a current. As base current increases, the larger part of emitter current flows closer to the base contact, and the effective base resistance decreases.

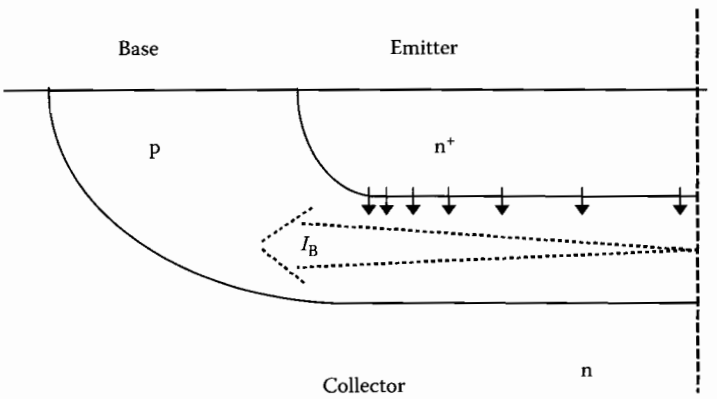

FIGURE 1.5 Current crowding effect.

1.1.5 Small-Signal Model

Small-signal transistor models are essential for AC circuit design. The small-signal equivalent circuit of the bipolar transistor is shown in Figure 1.6a. The lumped circuit shown in Figure 1.6a is only an approximation. In real transistors resistances and capacitances have a distributed character. For most design tasks, this lumped model is adequate, or even the simple equivalent transistor model shown in Figure 1.6b can be considered. The small-signal resistances, r_π and r_o, are inversely proportional to the transistor currents, and the transconductance g_m is directly proportional to the transistor currents

$$r_\pi = \frac{\eta_F V_T}{I_B} = \frac{\eta_F V_T \beta_F}{I_C}, \quad r_o = \frac{V_A}{I_C}, \quad g_m = \frac{I_C}{\eta_F V_T} \tag{1.24}$$

where

η_F is the forward emission coefficient, ranging form 1.0 to 2.0

V_T is the thermal potential ($V_T = 25$ mV at room temperature)

Similar equations to Equation 1.24 can be written for the reverse transistor operation as well.

The series base, emitter, and collector resistances R_B, R_E, and R_C are usually neglected for simple analysis (Figure 1.6b). However, for high-frequency analysis it is essential to use at least the base series resistance R_B. The series emitter resistance R_E usually has a constant, bias-independent value. The collector resistance R_C may significantly vary with the biasing current. The value of the series collector resistance may lower by one or two orders of magnitude if the collector junction becomes forward biased. A large series collector resistance may force the transistor into the saturation mode. Usually, when collector–emitter voltage is large enough, the effect of collector resistance is not significant. The SPICE model assumes constant value for the collector resistance R_C.

The series base resistance R_B may significantly limit the transistor performance at high frequencies. Due to the current crowding effect and the base conductivity modulation, the series base resistance is a function of the collector current I_C [4]

$$R_B = R_{Bmin} + \frac{R_{B0} - R_{Bmin}}{0.5 + \sqrt{0.25 + \frac{i_C}{I_{KF}}}} \tag{1.25}$$

where

I_{KF} is β_F high-current roll-off corner

R_{B0} is the base resistance at very small currents

R_{Bmin} is the minimum base resistance at high currents

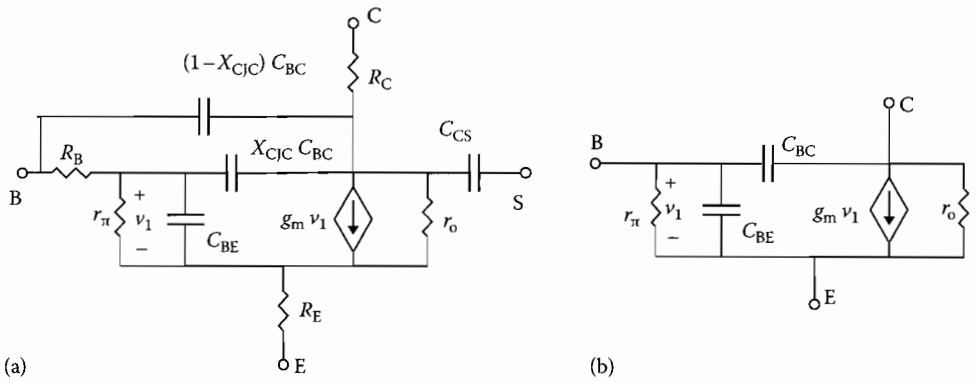

FIGURE 1.6 Bipolar transistor equivalent diagrams. (a) SPICE model and (b) simplified model.

Another possible approximation of the base series resistance R_B, as a function of the base current I_B, is [4]

$$R_B = 3(R_{B0} - R_{Bmin})\frac{\tan z - z}{z\tan^2 z} + R_{Bmin}, \quad z = \frac{\sqrt{1 + \frac{1.44 I_B}{\pi^2 I_{RB}}} - 1}{\frac{24}{\pi^2}\sqrt{\frac{I_B}{I_{RB}}}} \tag{1.26}$$

where I_{RB} is the base current for which the base resistance falls halfway to its minimum value.

The base–emitter capacitance C_{BE} is composed of two terms: the diffusion capacitance, which is proportional to the collector current, and the depletion capacitance, which is a function of the base–emitter voltage V_{BE}. The C_{BE} capacitance is given by

$$C_{BE} = \tau_F \frac{I_C}{\eta_F V_T} + C_{JE0}\left(1 - \frac{v_{BE}}{V_{JE0}}\right)^{-m_{JE}} \tag{1.27}$$

where

V_{JE0} is the base–emitter junction potential
τ_F is the base transit time for forward direction
C_{JE0} is the base–emitter zero-bias junction capacitance
m_{JE} is the base–emitter grading coefficient

The base–collector capacitance C_{BC} is given by a similar expression as Equation 1.27. In the case when the transistor operates in forward-active mode, it can be simplified to

$$C_{BC} = C_{JC0}\left(1 - \frac{v_{BC}}{V_{JC0}}\right)^{-m_{JC}} \tag{1.28}$$

where

V_{JC0} is the base–collector junction potential
C_{JC0} is the base–collector zero-bias junction capacitance
m_{JC} is the base–collector grading coefficient

In the case when the bipolar transistor is in the integrated form, the collector–substrate capacitance C_{CS} has to be considered

$$C_{CS} = C_{JS0}\left(1 - \frac{V_{CS}}{V_{JS0}}\right)^{-m_{JS}} \tag{1.29}$$

where

V_{JS0} is the collector–substrate junction potential
C_{JS0} the collector–substrate zero-bias junction capacitance
m_{JS} is the collector–substrate grading coefficient

When the transistor enters saturation, or it operates in the reverse-active mode, Equations 1.27 and 1.28 should be modified to

$$C_{BE} = \tau_F \frac{I_S \exp\left(\frac{V_{BE}}{\eta_F V_T}\right)}{\eta_F V_T} + C_{JE0}\left(1 - \frac{V_{BE}}{V_{JE0}}\right)^{-m_{JE}} \tag{1.30}$$

$$C_{BC} = \tau_R \frac{I_S \exp\left(\frac{V_{BC}}{\eta_R V_T}\right)}{\eta_R V_T} + C_{JC0}\left(1 - \frac{V_{BC}}{V_{JC0}}\right)^{-m_{JC}} \tag{1.31}$$

1.1.6 Technologies

The bipolar technology was used to fabricate the first ICs more than 40 years ago. A similar standard bipolar process is still used. In recent years, for high-performance circuits and for BiCMOS technology, the standard bipolar process was modified by using the thick selective silicon oxidation instead of the p-type isolation diffusion. Also, the diffusion process was substituted by the ion implantation process, low-temperature epitaxy, and Chemical Vapor Deposition (CVD).

1.1.6.1 Integrated NPN Bipolar Transistor

The structure of the typical integrated bipolar transistor is shown in Figure 1.7. The typical impurity profile of the bipolar transistor is shown in Figure 1.8. The emitter doping level is much higher than the base doping, so large current gains are possible (see Equation 1.19). The base is narrow and it has an impurity gradient, so the carrier transit time through the base is short (see Equation 1.17). Collector concentration near the base–collector junction is low, therefore, the transistor has a large breakdown voltage, large Early voltage V_{AF}, and CB depletion capacitance is low. High impurity concentration in the buried layer leads to a small collector series resistance. The emitter strips have to be as narrow as technology allows, reducing the base series resistance and the current crowding effect. If large emitter area is required, many narrow emitter strips interlaced with base contacts have to be used in a single

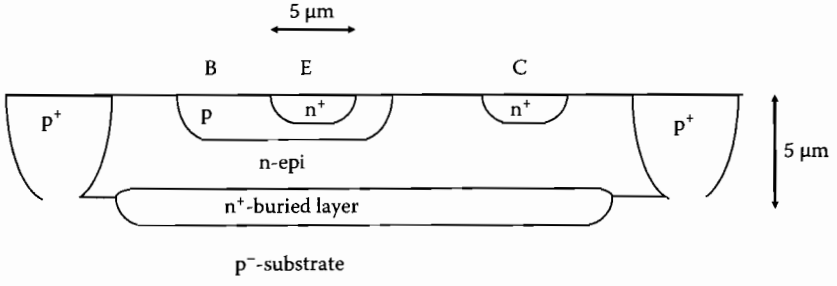

FIGURE 1.7 NPN bipolar structure.

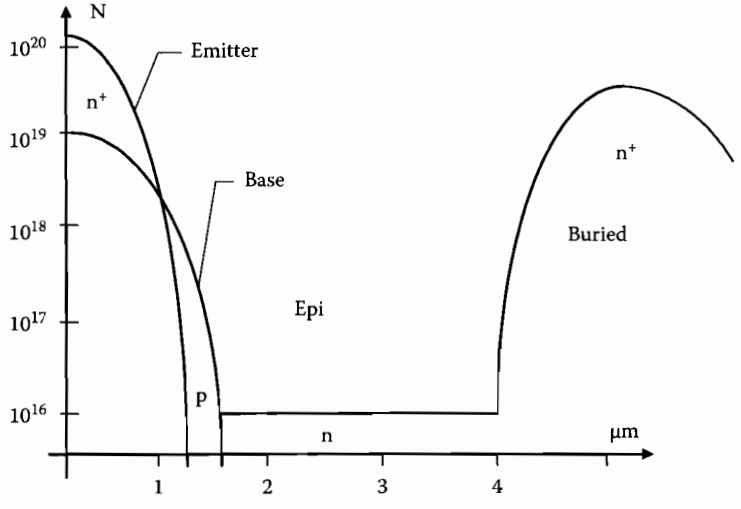

FIGURE 1.8 Cross section of a typical bipolar transistor.

transistor. Special attention has to be taken during the circuit design, so the base–collector junction is not forward biased. If the base–collector junction is forward biased, then the parasitic PNP transistors activate. This leads to undesired circuit operation. Thus, the integrated bipolar transistors must not operate in reverse or in saturation modes.

1.1.6.2 Lateral and Vertical PNP Transistors

The standard bipolar technology is oriented for fabrication of the NPN transistors with the structure shown in Figure 1.7. Using the same process, other circuit elements, such as resistors and PNP transistors, can be fabricated as well.

The lateral transistor, shown in Figure 1.9a uses the base p-type layer for both emitter and collector fabrication. The vertical transistor, shown in Figure 1.9b uses the p-type base layer for emitter, and the p-type substrate as collector. This transistor is sometimes known as the substrate transistor. In both transistors the base is made of the n-type epitaxial layer. Such transistors with a uniform and thick base are slow. Also, the current gain β of such transistors is small. Note, that the vertical transistor has the collector shorted to the substrate as Figure 1.10b illustrates. When a PNP transistor with a large current gain is required, then the concept of the composite transistor can be implemented. Such a composite transistor, known also as superbeta transistor, consists a PNP lateral transistor, and the standard NPN transistor connected as shown in Figure 1.10c. The composed transistor acts as the PNP transistor and it has a current gain β approximately equal to $\beta_{pnp}\beta_{npn}$.

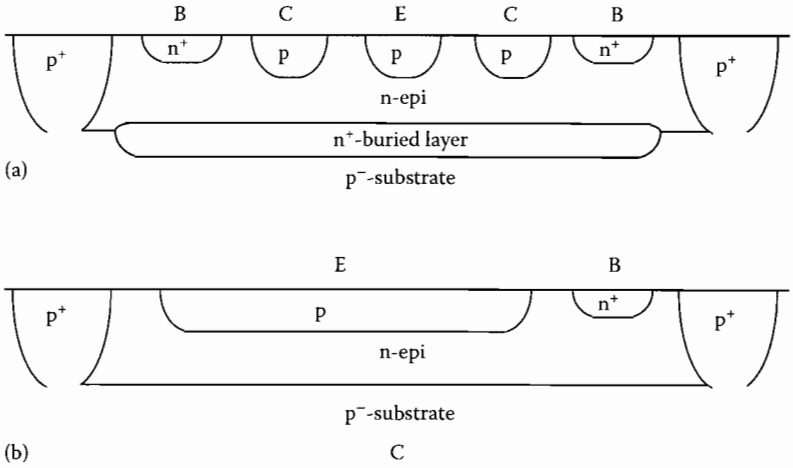

FIGURE 1.9 Integrated PNP transistors: (a) lateral PNP transistor, and (b) substrate PNP transistor.

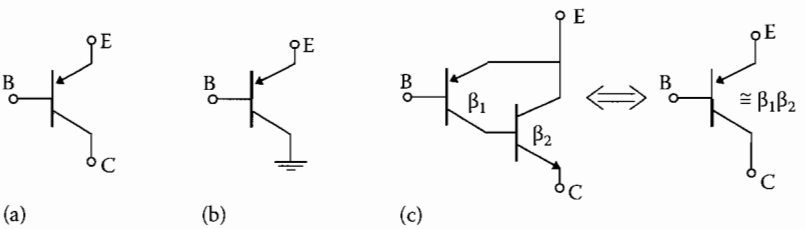

FIGURE 1.10 Integrated PNP transistors: (a) lateral transistor, (b) substrate transistor, and (c) composed transistor.

1.1.7 Model Parameters

It is essential to use proper transistor models in the computer aided design tools. The accuracy of simulation results depends on the model accuracy, and on the values of the model parameters used. In Section 1.1, the thermal and second-order effect in the transistor model are discussed. The SPICE bipolar transistor model parameters are discussed.

1.1.7.1 Thermal Sensitivity

All parameters of the transistor model are temperature dependent. Some parameters are very strong functions of temperature. To simplify the model description, the temperature dependence of some parameters are often neglected. In this chapter, the temperature dependence of the transistor model is described based on the model of the SPICE program [3–5]. Deviations from the actual temperature dependence will also be discussed. The temperature dependence of junction capacitance is given by

$$C_J(T) = C_J\left\{1 + m_J\left[4.010^{-4}(T - T_{NOM}) + 1 - \frac{V_J(T)}{V_J}\right]\right\} \tag{1.32}$$

where T_{NOM} is the nominal temperature, which is specified in the SPICE program in the .OPTIONS statement. The junction potential $V_J(T)$ is a function of temperature

$$V_J(T) = V_J\frac{T}{T_{NOM}} - 3V_T\ln\left(\frac{T}{T_{NOM}}\right) - E_G(T) + E_G\frac{T}{T_{NOM}} \tag{1.33}$$

The value of 3 in the multiplication coefficient of above equation is from the temperature dependence of the effective state densities in the valence and conduction bands. The temperature dependence of the energy gap is computed in the SPICE program from

$$E_G(T) = E_G - \frac{7.0210^{-4}T^2}{T + 1108} \tag{1.34}$$

The transistor saturation current as a function of temperature is calculated as

$$I_s(T) = I_s\left(\frac{T}{T_{NOM}}\right)^{X_{TI}}\exp\left[\frac{E_G(T - T_{NOM})}{V_T T_{NOM}}\right] \tag{1.35}$$

where E_G is the energy gap at the nominal temperature. The junction leakage currents I_{SE} and I_{SC} are calculated using

$$I_{SE}(T) = I_{SE}\left(\frac{T}{T_{NOM}}\right)^{X_{TI}-X_{TB}}\exp\left[\frac{E_G(T - T_{NOM})}{\eta_E V_T T_{NOM}}\right] \tag{1.36}$$

and

$$I_{SC}(T) = I_{SC}\left(\frac{T}{T_{NOM}}\right)^{X_{TI}-X_{TB}}\exp\left[\frac{E_G(T - T_{NOM})}{\eta_C V_T T_{NOM}}\right] \tag{1.37}$$

The temperature dependence of the transistor current gains β_F and β_R are modeled in the SPICE as

$$\beta_F(T) = \beta_F\left(\frac{T}{T_{NOM}}\right)^{X_{TB}}, \quad \beta_R(T) = \beta_R\left(\frac{T}{T_{NOM}}\right)^{X_{TB}} \tag{1.38}$$

The SPICE model does not give accurate results for the temperature relationship of the current gain β at high currents. For high current levels the current gain decreases sharply with the temperature, as can be seen from Figure 1.3. Also, the knee current parameters IKF, IKR, IKB are temperature-dependent, and this is not implemented in the SPICE program.

1.1.7.2 Second-Order Effects

The current gain β is sometimes modeled indirectly by using different equations for the collector and base currents [4,5]

$$I_C = \frac{I_S(T)}{Q_b}\left(\exp\frac{V_{BE}}{\eta_F V_T} - \exp\frac{V_{BC}}{\eta_R V_T}\right) - \frac{I_S(T)}{\beta_R(T)}\left(\exp\frac{V_{BC}}{\eta_R V_T} - 1\right) - I_{SC}(T)\left(\exp\frac{V_{BC}}{\eta_C V_T} - 1\right) \tag{1.39}$$

where

$$Q_b = \frac{1 + \sqrt{1 + 4Q_X}}{2\left(1 - \frac{V_{BC}}{V_{AF}} - \frac{V_{BE}}{V_{AR}}\right)} \tag{1.40}$$

$$Q_X = \frac{I_s(T)}{I_{KF}}\left(\exp\frac{V_{BE}}{\eta_F V_T} - 1\right) + \frac{I_s(T)}{I_{KR}}\left(\exp\frac{V_{BC}}{\eta_R V_T} - 1\right) \tag{1.41}$$

and

$$I_B = \frac{I_S}{\beta_F}\left(\exp\frac{V_{BE}}{\eta_F V_T} - 1\right) + I_{SE}\left(\exp\frac{V_{BE}}{\eta_E V_T} - 1\right) + \frac{I_S}{\beta_R}\left(\exp\frac{V_{BC}}{\eta_R V_T} - 1\right) + I_{SC}\left(\exp\frac{V_{BC}}{\eta_C V_T} - 1\right) \tag{1.42}$$

where

I_{SE} is the base–emitter junction leakage current
I_{SC} is the base–collector junction leakage current
η_E is the base–emitter junction leakage emission coefficient
η_C is the base–collector junction leakage emission coefficient

The forward transit time τ_F is a function of biasing conditions. In the SPICE program the τ_F parameter is computed using

$$\tau_F = \tau_{F0}\left[1 + X_{TF}\left(\frac{I_{CC}}{I_{CC} + I_{TF}}\right)^2 \exp\frac{V_{BC}}{1.44 V_{TF}}\right], \quad I_{CC} = I_s\left(\exp\frac{V_{BE}}{\eta_F V_T} - 1\right) \tag{1.43}$$

At high frequencies the phase of the collector current shifts. This phase shift is computed in the SPICE program following way

$$I_C(\omega) = I_C \exp(j\omega P_{TF}\tau_F) \tag{1.44}$$

where P_{TF} is a coefficient for excess phase calculation.

Noise is usually modeled as the thermal noise for parasitic series resistances, and as shot and flicker noise for collector and base currents

$$\overline{i_B^2} = \frac{4kT\Delta f}{R} \tag{1.45}$$

$$\overline{i_B^2} = \left(2qI_B + \frac{K_F I_B^{A_F}}{F}\right)\Delta f \tag{1.46}$$

$$\overline{i_C^2} = 2qI_C\Delta f \tag{1.47}$$

where K_F and A_F are the flicker-noise coefficient and flicker-noise exponent. More detailed information about noise modeling is given in Section 3.2.

1.1.7.3 SPICE Model of the Bipolar Transistor

The SPICE model of bipolar transistor uses similar or identical equations as described in this chapter [3–5]. Table 1.1 shows the parameters of the bipolar transistor model and its relation to the parameters used in this chapter.

TABLE 1.1 Parameters of SPICE Bipolar Transistor Model

Name Used	Equations	SPICE Name	Parameter Description	Unit	Typical Value	SPICE Default
I_s	1.10, 1.11	IS	Saturation current	A	10^{-15}	10^{-16}
I_{SE}	1.39	ISE	B–E leakage saturation current	A	10^{-12}	0
I_{SC}	1.39	ICS	B–C leakage saturation current	A	10^{-12}	0
β_F	1.14, 1.16, 1.21	BF	Forward current gain	—	100	100
β_R	1.14, 1.16, 1.21	BF	Reverse current gain	—	0.1	1
η_F	1.15, 1.24, 1.30, 1.31, 1.39 through 1.41	NF	Forward current emission coefficient	—	1.2	1.0
η_R	1.15, 1.24, 1.30, 1.31, 1.39 through 1.42	NR	Reverse current emission coefficient	—	1.3	1.0
η_E	1.39	NE	B–E leakage emission coefficient	—	1.4	1.5
η_C	1.39	NC	B–C leakage emission coefficient	—	1.4	1.5
V_{AF}	1.21, 1.40	VAF	Forward Early voltage	V	200	∞
V_{AR}	1.21, 1.40	VAR	Reverse Early voltage	V	50	∞
I_{KF}	1.22, 1.23, 1.40	IKF	β_F high-current roll-off corner	A	0.05	∞
I_{KR}	1.22, 1.23, 1.40	IKR	β_R high-current roll-off corner	A	0.01	∞
I_{RB}	1.26	IRB	Current where base resistance falls by half	A	0.1	∞
R_B	1.25, 1.26	RB	Zero base resistance	Ω	100	0
R_{Bmin}	1.25, 1.26	RBM	Minimum base resistance	Ω	10	RB
R_E	Figure 1.6	RE	Emitter series resistance	Ω	1	0
R_C	Figure 1.6	RC	Collector series resistance	Ω	50	0
C_{JE0}	1.27	CJE	B–E zero-bias depletion capacitance	F	10^{-12}	0
C_{JC0}	1.28	CJC	B–C zero-bias depletion capacitance	F	10^{-12}	0
C_{JS0}	1.29	CJS	Zero-bias collector–substrate capacitance	F	10^{-12}	0
V_{JE0}	1.27	VJE	B–E built-in potential	V	0.8	0.75
V_{JC0}	1.28	VJC	B–C built-in potential	V	0.7	0.75
V_{JS0}	1.29	VJS	Substrate junction built-in potential	V	0.7	0.75
m_{JE}	1.27	MJE	B–E junction exponential factor	—	0.33	0.33
m_{JC}	1.28	MJC	B–C junction exponential factor	—	0.5	0.33

(continued)

TABLE 1.1 (continued) Parameters of SPICE Bipolar Transistor Model

Name Used	Equations	SPICE Name	Parameter Description	Unit	Typical Value	SPICE Default
m_{JS}	1.29	MJS	Substrate junction exponential factor	—	0.5	0
X_{CJC}	Figure 1.6	XCJC	Fraction of B–C capacitance connected to internal base node (see Figure 1.6)	—	0.5	0
τ_F	1.17, 1.28, 1.30, 1.42	TF	Ideal forward transit time	s	10^{-10}	0
τ_R	1.31	TR	Reverse transit time	s	10^{-8}	0
X_{TF}	1.43	XTF	Coefficient for bias dependence of τ_F	—		0
V_{TF}	1.43	VTF	Voltage for τ_F dependence on V_{BC}	V		∞
I_{TF}	1.43	ITF	Current where $\tau_F = f(I_C, V_{BC})$ starts	A		0
P_{TF}	1.44	PTF	Excess phase at freq $= 1/(2\pi\tau_F)$ Hz	°		0
X_{TB}	1.38	XTB	Forward and reverse beta temperature exponent			0
E_G	1.34	EG	Energy gap	eV	1.1	1.11
X_{TI}	1.35 through 1.37	XTI	Temperature exponent for effect on I_s	—	3.5	3
K_F	1.46	KF	Flicker-noise coefficient	—		0
A_F	1.46	AF	Flicker-noise exponent	—		1
F_C		FC	Coefficient for the forward biased depletion capacitance formula	—	0.5	0.5
T_{NOM}	1.32 through 1.38	TNOM	Nominal temperature specified in .OPTION statement	K	300	300

The SPICE [3] was developed mainly for analysis of ICs. During the analysis it is assumed that the temperatures of all circuit elements are the same. This is not true for power ICs where the junction temperatures may differ by 30 K or more. This is obviously not true for circuits composed of the discrete elements where the junction temperatures may differ by 100 K and more. These temperature effects, which can significantly affect the analysis results, are not implemented in the SPICE program.

Although the SPICE bipolar transistor model uses more than 40 parameters, many features of the bipolar transistor are not included in the model. For example, the reverse junction characteristics are described by Equation 1.32. This model does not give accurate results. In the real silicon junction the leakage current is proportional to the thickness of the depletion layer, which is proportional to $V^{1/m}$. Also the SPICE model of the bipolar transistor assumes that there is no junction breakdown voltage. A more accurate model of the reverse junction characteristics is described in Section 11.5 of *Fundamentals of Circuits and Filters*. The reverse transit time τ_R is very important to model the switching property of the lumped bipolar transistor, and it is a strong function of the biasing condition and temperature. Both phenomena are not implemented in the SPICE model.

1.1.8 SiGe HBTs

The performance of the Si bipolar transistor can be greatly enhanced with proper engineering of the base bandgap profile using a narrower bandgap material, SiGe, an alloy of Si and Ge. Structure wise, a SiGe

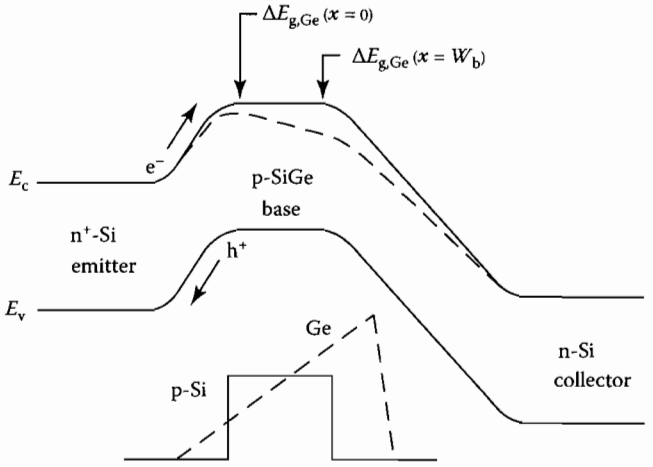

FIGURE 1.11 Energy band diagram of a graded base SiGe HBT and a comparably constructed Si BJT.

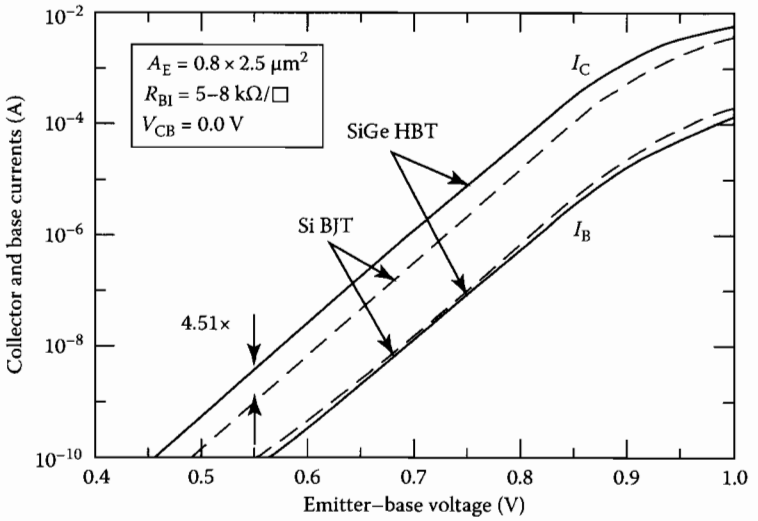

FIGURE 1.12 Experimental collector and base currents versus EB voltage for SiGe HBT and Si BJT.

HBT is essentially a Si BJT with a SiGe base. Its operation and circuit level performance advantages can be illustrated with the energy band diagram in Figure 1.11 [13]. Here the Ge content is linearly graded from emitter toward collector to create a large accelerating electric field that speeds up minority carrier transport across the base, thus making transistor speed much faster and cutoff frequency much higher. Everything else being the same, the potential barrier for electron injection into the base is reduced, thus exponentially enhancing the collector current. The base current is the same for SiGe HBT and Si BJT, as the emitter is typically made the same. Beta is thus higher in SiGe HBT. Figure 1.12 confirms these expectations experimentally with data from a typical first-generation SiGe HBT technology. The measured doping and Ge profiles are shown in Figure 1.13. The metallurgical base width is only 90 nm, and the neutral base width is around 50 nm. Figure 1.14 shows experimental cutoff frequency f_T improvement from using a graded SiGe base, which also directly translates into maximum oscillation frequency f_{max} improvement.

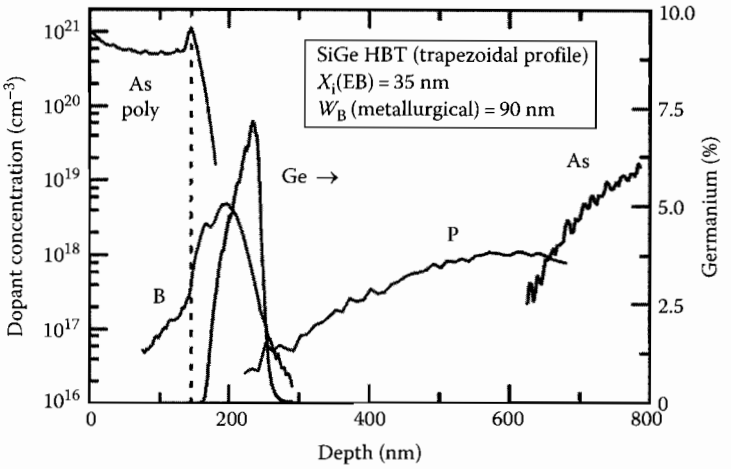

FIGURE 1.13 Measured doping and Ge profiles of a modern SiGe HBT.

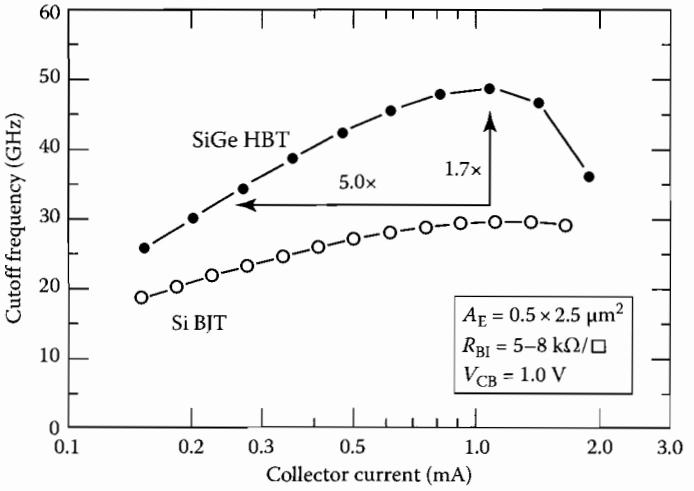

FIGURE 1.14 Experimental cutoff frequency versus collector current for SiGe HBT and Si BJT.

1.1.8.1 Operation Principle and Performance Advantages over Si BJT

In modern transistors, particularly with the use of polysilicon emitter, beta may be sufficient. If so, the higher beta potential of SiGe HBT can then be traded for reduced base resistance, through the use of higher base doping. The unique ability of simultaneously achieving high beta, low base resistance, and high cutoff frequency makes SiGe HBT attractive for many RF circuits. Broadband noise is naturally reduced, as low base resistance reduces transistor input noise voltage, and high beta as well as high f_T reduces transistor input noise current [13]. Experimentally, $1/f$ noise at the same base current was found to be approximately the same for SiGe HBT and Si BJT [14]. Consequently, $1/f$ noise is often naturally reduced in SiGe HBT circuits for the same biasing collector current, as base current is often smaller due to higher beta, as shown in Figure 1.15 using corner frequency as a figure-of-merit.

These, together with circuit-level optimization, can lead to excellent low-phase noise oscillators and frequency synthesizers suitable for both wireless and wire line communication circuits. Another less

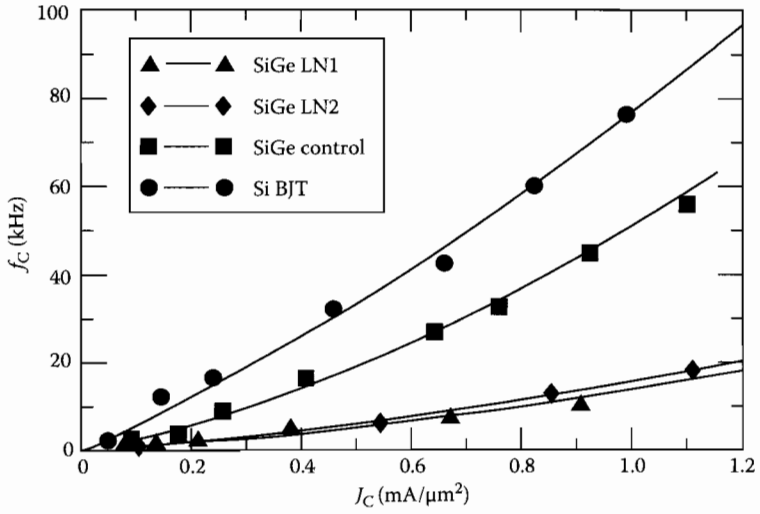

FIGURE 1.15 Experimentally measured corner frequency as a function of collector current density for three SiGe HBTs with different base SiGe designs, and a comparatively constructed Si BJT.

obvious advantage from grading Ge is the collector side of the neutral base has less impact on the collector current than the emitter side of the neutral base. Consequently, as collector voltage varies and the collector side of the neutral base is shifted toward the emitter due to increased CB junction depletion layer thickness, the collector current is increased to a much lesser extent than in a comparably constructed Si BJT, leading to a much higher output impedance or Early voltage. The $\beta \times V_A$ product is thus much higher in SiGe HBT than in Si BJT.

1.1.8.2 Industry Practice and Fabrication Technology

The standard industry practice today is to integrate SiGe HBT with CMOS, to form a SiGe BiCMOS technology. The ability to integrate with CMOS is also a significant advantage of SiGe HBT over III–V HBT. Modern SiGe BiCMOS combines the analog and RF performance advantages of the SiGe HBT, and the lower power logic, high integration level, and memory density of Si CMOS, into a single cost-effective system-on-chip (SoC) solution. Typically, SiGe HBTs with multiple breakdown voltages are offered through selective collector implantation, to provide more flexibility in circuit design.

The fabrication process of SiGe HBT and its integration with CMOS has been constantly evolving in the past two decades, and varies from company to company. Below are some common fabrication elements and modules shared by many if not all commercial first-generation (also most wide spread in manufacturing at present) SiGe technologies:

1. A starting N^+ subcollector around 5 Ω/sq on a p-type substrate at $5 \times 10^{15}/cm^3$, typically patterned to allow CMOS integration.
2. A high-temperature, lightly doped n-type collector, around 0.4–0.6 μm thick at $5 \times 10^{15}/cm^3$.
3. Polysilicon-filled deep trenches for isolation from adjacent devices, typically 1 μm wide and 7–10 μm deep.
4. Oxide filled shallow trenches or LOCOS for local device isolation, typically 0.3–0.6 μm deep.
5. An implanted collector reach through to the subcollector, typically at 10–20 $\Omega\mu m^2$.
6. A composite SiGe epi layer consisting of a 10–20 nm Si buffer, a 70–100 nm boron-doped SiGe active layer, with or without C doping to help suppress boron out diffusion, and a 10–30 nm Si cap. The integrated boron dose is typically $1–3 \times 10^{13}/cm^2$.

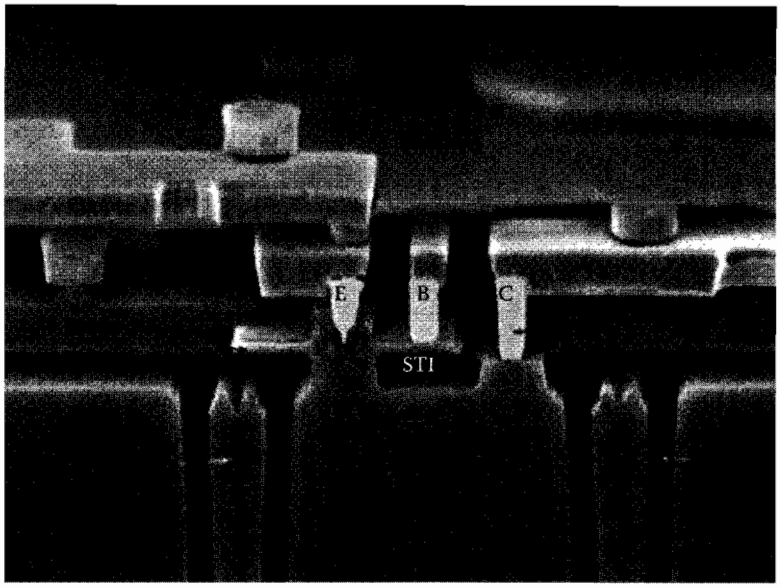

FIGURE 1.16 Structure of a modern SiGe HBT.

7. A variety of EB self-alignment scheme, depending on device structure and SiGe growth approach. All of them utilize some sort of spacer that is 100–300 nm wide.
8. Multiple self-aligned collector implantation to allow multiple breakdown voltages on the same chip.
9. Polysilicon extrinsic base, usually formed during SiGe growth over shallow trench oxide, and additional self-aligned extrinsic implantation to lower base resistance.
10. A silicided extrinsic base.
11. A 100–200 nm thick heavily doped ($>5 \times 10^{20}/cm^3$) polysilicon emitter, either implanted or in situ doped.
12. A variety of multiple level back-end-of-line metallization schemes using Al or Cu, typically borrowed from parent CMOS process.

These technological elements can also be seen in the electronic image of a second-generation SiGe HBT shown in Figure 1.16.

References

1. J. J. Ebers and J. M. Moll, Large signal behavior of bipolar transistors. *Proceedings IRE* 42, 1761–1772, December 1954.
2. H. K. Gummel and H. C. Poon, An integral charge-control model of bipolar transistors. *Bell System Technical Journal* 49, 827–852, May 1970.
3. L. W. Nagel and D. O. Pederson, SPICE (Simulation Program with Integrated Circuit Emphasis). University of California, Berkeley, ERL Memo No. ERL M382, April 1973.
4. P. Antognetti and G. Massobrio, *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York, 1988.
5. A. Vadimiresku, *The SPICE Book*, John Wiley & Sons, Hoboken, NJ, 1994.
6. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, Hoboken, NJ, 1967.
7. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 1981.
8. G. W. Neudeck, *The PN Junction Diode*, Vol II, Modular Series on Solid-State Devices, Addison-Wesley, Upper Saddle River, NJ, 1983.

9. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 1986.
10. E. S. Yang, *Microelectronic Devices*, McGraw-Hill, New York, 1988.
11. B. G. Streetman, *Solid State Electronic Devices*. 3rd ed., Prentice Hall, Upper Saddle River, NJ, 1990.
12. D. A. Neamen, *Semiconductor Physics and Devices*, Irwin, 1992.
13. J. D. Cressler and G. Niu, *Silicon-Germanium Heterojunction Bipolar Transistor*, Artech House, Norwood, MA, 2003.
14. G. Niu, Noise in SiGe HBT RF technology: Physics, modeling and circuit implications, *Proceedings of the IEEE*, pp. 1583–1597, September 2005.

1.2 Metal–Oxide–Silicon Field Effect Transistor

John Choma, Jr.

1.2.1 Introduction

Integrated electronic circuits realized in metal–oxide–silicon field effect transistor (MOSFET) technology are ubiquitous in both the commercial and military sectors of the technical community. To be sure, transistors manufactured in certain bipolar and III–V compound transistor technologies compete successfully with their MOSFET counterparts from such performance perspectives as switching speed, wideband frequency response, and insensitivity to electromagnetic interference and irradiated environments. Nevertheless, the MOSFET reigns supreme in the extant state of the electronics art for several reasons. The first of these reasons derives from the fact that the cross-section geometry of a MOSFET, when compared to that of most other solid-state transistors, is simpler. This simplicity affords a relative ease of foundry processing, which in turn promotes high device yield and therefore, cost-effective manufacturing. A second reason is that the surface area consumed on chip, or *footprint*, of a MOSFET is generally smaller than that of a comparably performing bipolar or III–V compound transistors. This feature allows increased packing density, which is particularly advantageous for digital signal processors that commonly require upwards of millions of transistors for system functionality. Third, MOSFETs can deliver acceptable circuit performance at low standby power levels, which is a laudable attribute in light of the aforementioned high device density digital architectures and the portability culture in which society is immersed presently. Finally, the native insulating oxide indigenous to the monolithic processing of silicon semiconductors renders MOSFET technologies amenable to the implementation of complex electronic systems on a single chip. No such native oxide prevails in III–V compound technologies, thereby rendering awkward the electrical isolation among the various components, subsystems, and subcircuits that comprise the overall electronic system.

The penchant toward adopting MOSFET technology for analog signal processing applications can also be rationalized. In particular, the nature of modern integrated systems is rarely exclusively digital or exclusively analog. Such systems are, in fact, "mixed signal" architectures that embody both digital and analog signal processing on the same chip. Because of the simplicity, packing density, and power dissipation attributes of MOSFETs, virtually 100% of digital architectures are realized in MOSFET technology. Prudence alone accordingly dictates a MOSFET technology realization of the analog cells implicit to a mixed signal framework if only to facilitate the electrical interface between the analog and digital units.

Aside from the operating flexibility and programmability advantages boasted by digital circuit schema, digital circuits in mixed signal architectures are often required to assure and sustain performance optimality of the analog signal flow paths in an electronic system. Unlike most digital networks, high-performance analog circuits are sensitive to specific values, or at least specific ranges of values, of several of the key physical and electrical parameters that effectively define the electrical properties of MOSFETs. Unfortunately, attaining the requisite accuracy in the numerical delineation of these parameters becomes

progressively more daunting as the performance metrics imposed on an analog network become more challenging and as device geometries scale to meet omnipresent quests for wider signal processing passbands. In these high-performance systems, digital subsystems are often deployed to sense the observable performance metrics of an analog signal flow path, compare said metrics to their respective optimal design goals, and then appropriately adjust the relevant electrical parameters or signal excitations implicit to the signal path. In effect, the combined digital controller and analog network behave as a seamless adaptive system that automatically corrects for manufacturing vagaries, increased device operating temperatures, and certain environmental effects.

The most commonly utilized MOSFETs in modern electronic systems come in two flavors: the N-channel MOSFET (NMOS), diagrammed in Figure 1.17 and the P-channel MOSFET (PMOS) shown in Figure 1.18. In the NMOS device of Figure 1.17, the bulk substrate is P-type and is doped to an average acceptor impurity concentration of N_A, for which a representative range of values is $5(10^{14})$ atoms/cm^3 $< N_A <$ 10^{16} atoms/cm^3. Its vertical depth, which is not expressly highlighted in the figure, is many times larger than the depth, Y_d, (of the order of a few tenths of microns) of either the source or drain diffusions or implants. These regions, whose widths are indicated as L_{diff} and which are connected electrically to the source (S) and drain (D) terminals of the MOSFET, are very strongly doped in that their donor impurity concentrations are $N_D = 10^{20}$ atoms/cm^3 or larger. The width, L_{diff}, is typically two- or three-times the channel length, indicated as L in the diagram. The metallization contact that forms the electrical terminal of the semiconductor bulk (B) is generally connected to the most negative potential available in the circuit into which the subject transistor is embedded. Such a connection reverse biases the PN junctions formed between the bulk and source regions and between the bulk and drain regions. This reverse biasing ensures that for at least low signal frequencies, the source and drain regions are electrically isolated from each other and from the bulk substrate. In certain types of multiwell IC processes, bulk–source and bulk–drain reverse biasing is assured simply by returning the bulk terminal directly to the source region contact.

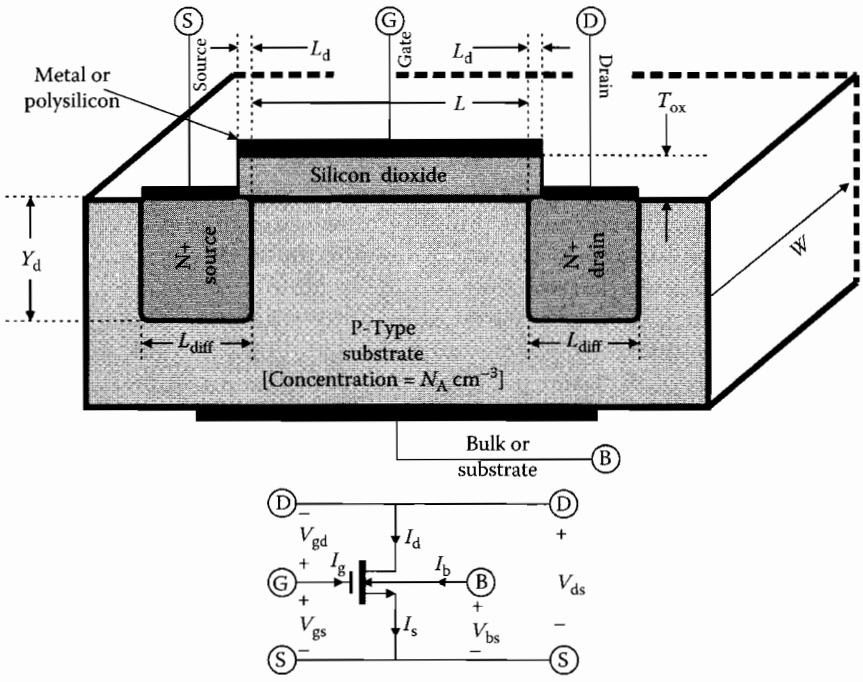

FIGURE 1.17 A simplified three-dimensional depiction of an N-channel MOSFET (NMOS) and its corresponding electrical schematic symbol. The diagram is not drawn to scale.

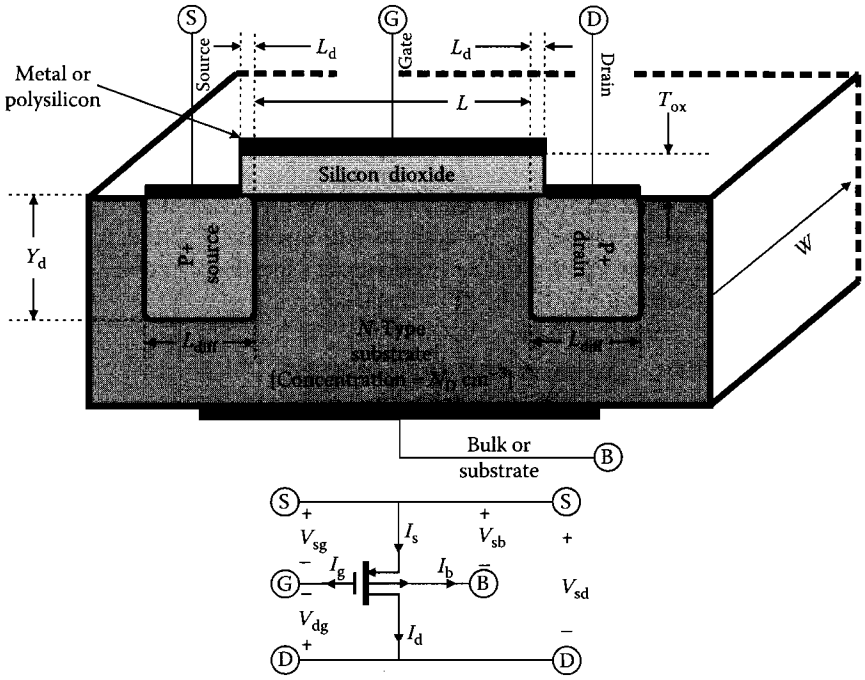

FIGURE 1.18 A simplified three-dimensional depiction of an P-channel MOSFET (PMOS) and its corresponding electrical schematic symbol. The diagram is not drawn to scale.

Lying atop the P-type bulk substrate is an insulating silicon dioxide layer of thickness T_{ox} that extends into the page as shown by a gate width, W. The oxide thickness in the extant state of the art is of the order of several tens of angstroms, where 1 Å is 10^{-8} cm. This oxide layer entirely covers the channel length, L, that separates the source region from the drain region, and it may overlap the source and drain regions by the amount, L_d, indicated in the diagram. The overlap of the source and drain regions is undesirable in that it limits broadband frequency responses in certain types of MOSFET amplifiers. In processes boasting self-aligned gate capabilities, L_d is ideally reduced to zero. But for state of the art processes delivering channel lengths as small as 65–130 nM, gate self-alignment focused on reducing L_d to no more than 5% of L is a challenging undertaking. The gate width, W, can be no smaller than the minimum channel length that can be produced by the identified foundry process. Subject to this proviso, the gate aspect ratio, W/L, is a designable parameter selected in accordance with the operating requirements of the circuit application for which the considered MOSFET is utilized.

The gate terminal (G) is formed by a contact made of a metallic or a polycrystalline silicon layer deposited directly atop the gate oxide. The gate metal of choice is aluminum. If the MOSFET under consideration is used in high-temperature environments and/or in applications that exploit low power supply voltages, polycrystalline silicon, which is commonly referred to as polysilicon, supplants the aluminum gate.

In addition to the simplified cross-section diagram of the N-channel MOSFET, Figure 1.17 inserts the electrical schematic symbol of the NMOS transistor. Of particular interest are the positive reference conventions adopted for four device currents and four device voltages. Specifically, positive drain current, I_d, flows into the transistor, as do the gate current, I_g, and the bulk, or substrate, current, I_b, while positive source current, I_s, flows out of the transistor. It follows from Kirchhoff's current law that

$$I_s = I_d + I_g + I_b. \tag{1.48}$$

However, since the gate contact is isolated from the semiconductor bulk by an insulating oxide layer, I_g is zero at the low frequencies for which capacitive phenomena associated with the insulating gate dielectric are insignificant. Moreover, the bulk current, I_b, is likewise almost zero at low signal frequencies, provided, as is usually the case, that care is taken to ensure reverse biasing of the bulk–drain and bulk–source PN junctions. Accordingly, the source and drain currents, I_s and I_d, respectively, are essentially identical when the frequencies of signals applied to the MOSFET are low. The pages that follow demonstrate that the static and low-frequency value of the drain, and hence the source, current is controlled by the gate-to-source voltage, V_{gs}, the drain-to-source voltage, V_{ds}, and, to a somewhat lesser extent, the bulk-to-source voltage V_{bs}. Stipulating an additional dependence of drain current on gate-to-drain voltage V_{gd} is superfluous, for by Kirchhoff's voltage law,

$$V_{ds} = V_{gs} - V_{gd}. \tag{1.49}$$

The P-channel MOSFET abstracted in Figure 1.18 is architecturally identical to its N-channel counterpart. The notable differences are that the bulk substrate in PMOS is N-type and the source and drain regions are heavily doped with P-type impurities. It follows that electrical isolation between the source region and the bulk, as well as between the drain region and the bulk, requires that the bulk substrate terminal of a PMOS device be connected either to the most positive of available circuit potentials or, if the process allows, to the source terminal. All of the geometrical parameters and their representative values remain the same as stipulated in conjunction with the NMOS unit. The PMOS electrical schematic symbol, which is also shown in the figure at hand, differs from the NMOS symbol in that the directions of the source terminal and bulk terminal arrows are reversed, as are the positive reference directions of all four transistor currents. While Equation 1.48 remains applicable, the analytical expression for the drain current, I_d, which now flows out of the transistor, is more conveniently couched in terms of the source-to-gate voltage, V_{sg}, the source-to-drain voltage, V_{sd}, and the source-to-bulk voltage, V_{sb}. The drain-to-gate voltage, V_{dg}, derives from

$$V_{sd} = V_{sg} - V_{dg}, \tag{1.50}$$

which mirrors Equation 1.49 subsequent to multiplying both sides of this equation by -1.

1.2.2 Channel Charge

A fundamental understanding of the physical charge storage and charge transport mechanisms that underpin the observable volt–ampere characteristics of considered transistors facilitates the reliable and reproducible design of high-performance analog networks in MOSFET technology. Aside from establishing a foundation upon which the static characteristic curves of a MOSFET can be constructed in a physically sound framework, these charge profiles also serve to define the voltage-dependent nature of the capacitance characteristics of a MOSFET. In effect, the subject charge profiles posture the MOSFET as a plausible varactor, which is useful in the monolithic design of voltage controlled oscillators, active filters, and other electronic networks.

The profile of charge stored in the channel between the source and drain regions of a MOSFET is best examined in terms of the simple circuit given in Figure 1.19a. In this circuit, the drain terminal is short circuited to the source to pin the drain–source voltage, V_{ds}, to zero. A zero bias is applied as indicated between the bulk and source, thereby establishing a charge depletion region about the PN junction formed between the substrate and source regions. Since the source and the drain are electrically connected to one another, the zero bias applied between bulk and source establishes an identical depletion zone about the bulk–drain PN junction. These depletion layers are delineated in the companion cross-section diagram of Figure 1.19b, as are the surface potential, φ_o, and the potential, V_{ox}, dropped across the gate silicon dioxide layer. With $V_{ds} = 0$, Equation 1.49 ensures a gate–source voltage, V_{gs}, that

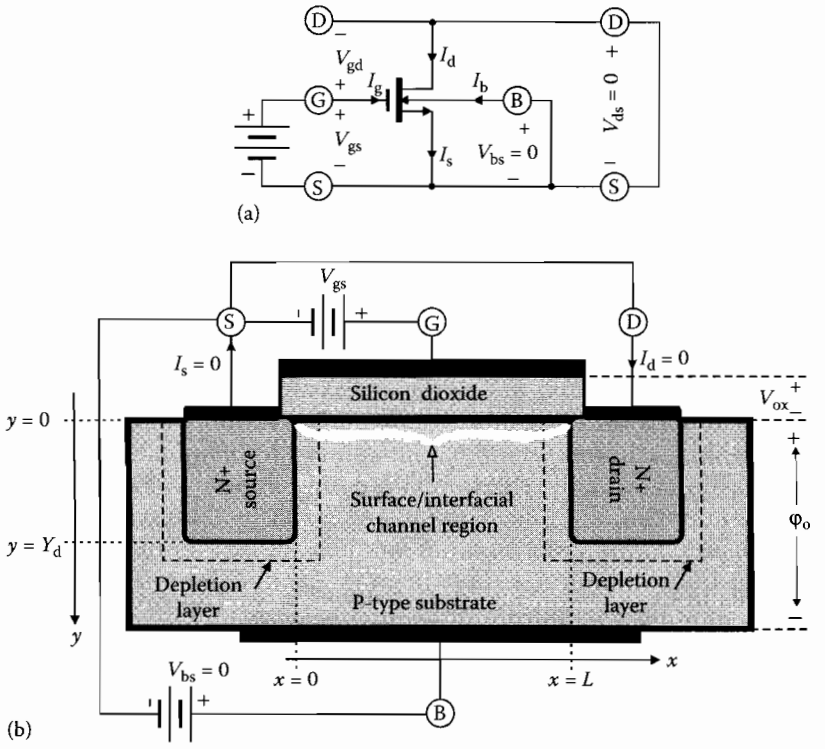

FIGURE 1.19 (a) NMOS transistor operated with $V_{ds} = 0$ and $V_{bs} = 0$. Although the battery connected between the gate and the source ensures $V_{gs} > 0$, $V_{gs} \leq 0$ is allowed in the discussion that references this circuit. (b) Cross-section diagram corresponding to the circuit in (a). Note that all applied voltages are referred to the source terminal. The diagram in (b) is not drawn to scale.

mirrors the gate–drain voltage, V_{gd}, regardless of the voltage applied between gate and source or gate and bulk terminals. In the absence of drain, source, bulk, and gate currents, $V_{ds} = 0$ also guarantees that surface potential φ_o, measured from the oxide semiconductor interface-to-the neutral zone of the bulk substrate, is the same throughout the channel region extending from $x = 0$-to-$x = L$ in the subject diagram. The aforementioned voltage, V_{ox}, includes the effects of parasitic trapped charge in the gate oxide, but it does not include the ramifications of work function differences that unavoidably prevail between the gate contact and the oxide and at the oxide–semiconductor interface. Note then that the voltage, V_{gb}, measured at the gate terminal with respect to the bulk terminal is, ignoring work function phenomena, simply

$$V_{gb} = V_{ox} + \varphi_o. \tag{1.51}$$

1.2.2.1 Surface Charge Density

A pivotally important analytical tool serving to define the charge, capacitance, and static volt–ampere characteristics of a MOSFET, is the charge density, $Q_o(\varphi_o)$, in units of coulombs per unit area, established at the semiconductor surface as a function of the surface potential, φ_o. Several authors have identified this charge profile as [1–3]

$$Q_o(\varphi_o) = -\text{sgn}(\varphi_o)\left(\frac{\sqrt{2}\varepsilon_s V_T}{D_b}\right)\sqrt{G(-\varphi_o) + G(\varphi_o)e^{-2V_F/V_T}}, \tag{1.52}$$

where $\varepsilon_s = 1.037$ pF/cm denotes the dielectric constant of silicon, and

$$\text{sgn}(\varphi_o) = \begin{cases} +1 & \text{for } \varphi_o > 0 \\ -1 & \text{for } \varphi_o < 0 \end{cases}. \tag{1.53}$$

In Equation 1.52,

$$V_T = kT/q \tag{1.54}$$

is the familiar semiconductor thermal voltage for which $k = (1.38)(10^{-23})$ J/K is Boltzmann's constant, $q = (1.60)(10^{-19})$ C is the magnitude of electron charge, and T is the absolute temperature of the semiconductor surface. The voltage, V_F, in the radical on the right-hand side of Equation 1.52 is the Fermi potential, which is given by

$$V_F \triangleq V_T \ln\left(\frac{N_A}{N_i}\right), \tag{1.55}$$

where N_A is the previously defined average acceptor impurity concentration of the bulk substrate in NMOS and $N_i = (1.45)(10^{10})$ atoms/cm^3 is the intrinsic carrier concentration of silicon at $T = 27°C$. The parameter, D_b, is known as the electron Debye length and is given by

$$D_b \triangleq \sqrt{\frac{\varepsilon_s V_T}{qN_A}}. \tag{1.56}$$

Finally, the function, $G(\varphi_o)$, in Equation 1.52 is

$$G(\varphi_o) = e^{\varphi_o/V_T} - 1 - \frac{\varphi_o}{V_T}, \tag{1.57}$$

where it is understood that the surface potential, φ_o, measured with respect to the charge neutral zone in the bulk in Figure 1.19 is established in response to an applied gate–bulk voltage, V_{gb}, or an applied gate–source voltage, V_{gs}. Observe that $G(\varphi_o) = G(-\varphi_o) = 0$ for $\varphi_o = 0$, which delivers the expected result in Equation 1.52 of $Q_o(0) = 0$. It should be understood that Equation 1.52 is premised on Poisson's equation and the Boltzmann carrier relationship,

$$p(0) = N_A e^{-\varphi_o/V_T}, \tag{1.58}$$

where $p(0)$ signifies the hole concentration at the surface if complete ionization of substrate dopant atoms is tacitly presumed. Since

$$p(0)n(0) = N_i^2, \tag{1.59}$$

the corresponding concentration of free surface electrons, $n(0)$, is

$$n(0) = \frac{N_i^2}{N_A} e^{\varphi_o/V_T} = N_A e^{(\varphi_o - 2V_F)/V_T}, \tag{1.60}$$

where Equation 1.55 has been exploited.

Because $G(\varphi_o)$ in Equation 1.57, as well as its companion relationship, $G(-\varphi_o)$, is a nonnegative number for all positive and negative values of the surface potential, the radical on the right-hand side of Equation 1.52 is a positive real number. Accordingly, Equation 1.53 forces $Q_o(\varphi_o) > 0$ for $\varphi_o < 0$ and $Q_o(\varphi_o) < 0$ for $\varphi_o > 0$. The positive nature of the surface charge density for negative surface potentials is indicative of bulk substrate holes attracted to the semiconductor surface because of the force exerted by the surface electric field established in response to negative surface potential. From Gauss' law, this field, say $E_o(\varphi_o)$ is simply

$$E_o(\varphi_o) = -\frac{Q_o(\varphi_o)}{\varepsilon_s} = \text{sgn}(\varphi_o)\left(\frac{\sqrt{2}V_T}{D_b}\right)\sqrt{G(-\varphi_o) + G(\varphi_o)e^{-2V_F/V_T}}, \tag{1.61}$$

which is indeed negative for $\varphi_o < 0$. Observe that Equation 1.58 supports the contention of an enhanced surface hole concentration when the potential established at the semiconductor surface is negative.

An equilibrium condition, which is more commonly referred to in the literature as the *flatband operating condition*, is reached when the applied gate–bulk or gate–source voltage produces a null surface potential, that is, $\varphi_o = 0$. For $\varphi_o = 0$, the net surface charge, $Q_o(\varphi_o)$, in Equation 1.52 is zero, as is the surface electric field, $E_o(\varphi_o)$, in Equation 1.61. Note further that by Equation 1.58, $p(0) = N_A$, which is the equilibrium hole concentration indicative of the NMOS bulk substrate for the transistor abstracted in Figure 1.17, assuming complete ionization of all substrate acceptor impurity atoms.

The negative surface charge prevailing for positive surface potentials, which gives rise to positive surface fields (field lines directed from the surface-to-the bulk substrate), reflects the surface charge depletion forged in response to holes repelled from the surface by $\varphi_o > 0$. Once again, Equation 1.58 is supportive of the proffered rationale in that it confirms a diminished surface hole concentration for progressively larger φ_o. Since departed holes leave in their wake a depletion zone of negative acceptor ions, the negative surface charge density resulting from positive surface potential is hardly surprising.

In addition to repelling holes from the semiconductor surface, Equation 1.60 indicates that the surface electron concentration increases as the surface potential, φ_o, rises above zero. Moreover, Equation 1.52 lends credence to this enhanced electron concentration claim since $Q_o(\varphi_o)$ is seen as becoming monotonically more negative as surface potential φ_o rises above zero. Indeed, the impact of the positive electric field associated with $\varphi_o > 0$ is to establish a force serving to attract the minority carriers (electrons) in the bulk substrate to the surface. For a surface potential in the range, $0 < \varphi_o < V_F$, the depletion charge contribution to the net surface charge continues to dominate over the charge associated with electrons cajoled to the surface, and the surface is said to operate in *depletion mode*. But as φ_o approaches and ultimately surpasses the Fermi potential, V_F, the impact on the nature of the surface charge becomes increasingly more interesting. For example, consider $\varphi_o = V_F$, for which Equations 1.58 and 1.60 yield $p(0) \equiv n(0) = N_i$, that is, the hole and electron concentrations at the surface are identically equal to the intrinsic carrier concentration. In effect, the surface region of the semiconductor changes from obviously P-type-to-intrinsic type, which is to say that the surface at $\varphi_o = V_F$ is neither P-type nor N-type.

For $V_F < \varphi_o < 2V_F$, Equations 1.58 and 1.60 project a surface electron concentration that actually exceeds the surface hole concentration, despite the originally P-type character of the semiconductor surface. In this range of surface potentials, the depletion layer at the surface continues to expand into the substrate but because of the enhanced electron concentration, the surface is said to operate in a condition of *weak inversion*. Weak inversion is significant from an engineering perspective in that it begins to establish the necessary condition for promoting observable drain and source current flow. In particular, suppose that the drain–source voltage, V_{ds}, were to be increased from its present null value to a suitably positive value. The presence of a significant mobile surface charge density in the form of free electrons allows said electrons to be transported from the source-to-the drain by the force associated with the lateral electric field established in response to $V_{ds} > 0$. In turn, this charge transport gives rise to a drain current flowing into the transistor and a source current flowing out of the device.

When φ_o rises to the value, $2V_F$, Equation 1.60 confirms a surface electron concentration that is numerically equal to the substrate doping concentration, N_A. In other words, the surface electron concentration precipitated by the strong positive electric fields implicit to $\varphi_o = 2V_F$ is identical to the equilibrium hole concentration evidenced in a silicon mass whose impurity concentration of completely ionized acceptor atoms is N_A. The surface has effectively changed its sex from its former P-type state to a field-induced (hence the terminology, "field-effect," in the FET nomenclature) N-type state. Since the resultant surface electron concentration, $n(0)$, is rendered substantive, appreciable drain and source currents can flow for even modest values of applied drain–source voltages. In effect, the transistor can be said to be "turned on" when φ_o rises to twice the Fermi potential in the sense that a capability for substantial drain current flow is forged. When $\varphi_o \geq 2V_F$, the semiconductor surface is *strongly inverted*, or simply *inverted*.

Figure 1.20 displays a representative surface charge density profile as a function of the surface potential. Since a logarithmic charge scale is required to display all salient features of the charge density, the negative nature of the surface charge for positive surface potentials compels plotting the magnitude of the surface charge density on the vertical (charge) scale in the subject figure. The horizontal (voltage) scale is normalized to the thermal voltage, V_T. The plot invokes the presumptions of a 27°C semiconductor surface temperature and a substrate impurity concentration of $N_A = 10^{15}$ atoms/cm^3. For these stipulations, the thermal voltage is $V_T = 25.89$ mV, and the Fermi potential is $V_F = 288.4$ mV, whence $V_F/V_T = 11.14$. The plot displayed in Figure 1.20 clearly identifies the regions of hole accumulation ($\varphi_o < 0$), surface depletion ($0 < \varphi_o < 2V_F$), weak inversion, as typified by the increased concentration of free electrons at the surface ($V_F < \varphi_o < 2V_F$), and strong inversion, for which $\varphi_o \geq 2V_F$.

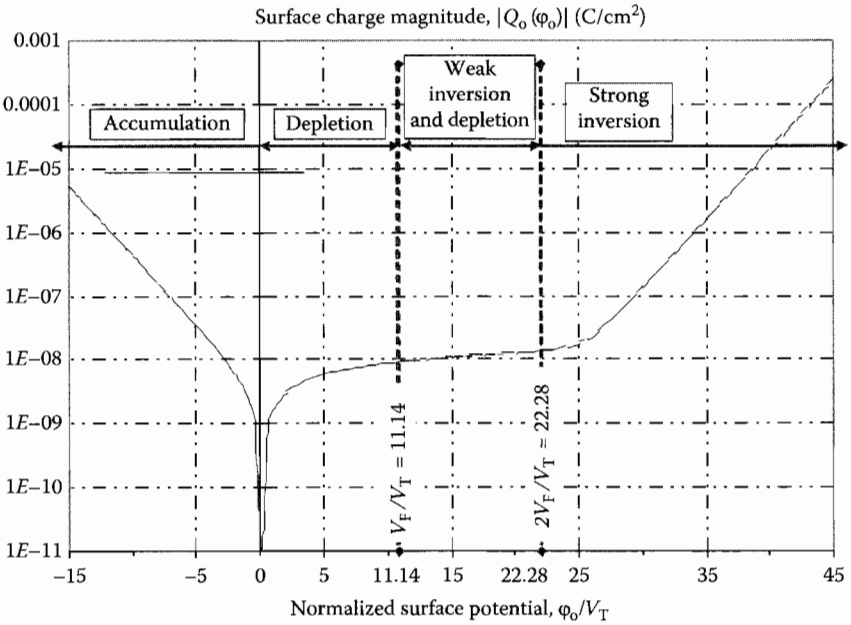

FIGURE 1.20 The magnitude of the surface charge density in the channel interfacial region for the MOSFET configured as shown in Figure 1.19b. A surface temperature of 27°C is assumed, as is a substrate impurity concentration of $N_A = 10^{15}$ atoms/cm^2.

1.2.2.2 Gate–Bulk Capacitance

The density of the net gate-to-bulk capacitance, $C_{gb}(\varphi_o)$, of the MOSFET whose cross-section diagram appears in Figure 1.19b is a series combination of the oxide capacitance density, C_{ox}, and the density of capacitance $C_d(\varphi_o)$, which is established between the oxide–substrate interface and the charge neutral region of the bulk. The pertinent equivalent circuit for $V_{ds} = 0$ is the structure depicted in Figure 1.21b, for which

$$C_{gb}(\varphi_o) = \frac{C_{ox}C_d(\varphi_o)}{C_{ox} + C_d(\varphi_o)}. \tag{1.62}$$

In Equation 1.62,

$$C_{ox} = \frac{\varepsilon_{ox}}{T_{ox}}, \tag{1.63}$$

where $\varepsilon_{ox} = 345$ fF/cm is the dielectric constant of silicon dioxide. Moreover,

$$C_d(\varphi_o) = \frac{d|Q_o(\varphi_o)|}{d\varphi_o}, \tag{1.64}$$

where the surface charge density, $Q_o(\varphi_o)$, is defined by Equation 1.52. After a trifle of differential calculus pain, it can be shown that

$$C_d(\varphi_o) = \text{sgn}(\varphi_o)\frac{\varepsilon_s}{\sqrt{2}D_b}\left[\frac{\left(e^{\varphi_o/V_T} - 1\right)e^{-2V_F/V_T} - \left(e^{-\varphi_o/V_T} - 1\right)}{\sqrt{G(-\varphi_o) + G(\varphi_o)e^{-2V_F/V_T}}}\right]. \tag{1.65}$$

The result at hand defines the surface capacitance density for all values of the surface potential, φ_o. A problem arises for $\varphi_o = 0$ in that the right-hand side becomes an indeterminate 0/0 form. This problem is circumvented by supplanting the exponential terms on the right-hand side of Equation 1.65, inclusive of those embedded in the functions, $G(\varphi_o)$ and $G(-\varphi_o)$, by their second order MacLaurin series expansions. Upon replacement of these exponential terms by said expansions, the surface capacitance density at the flatband condition, $\varphi_o = 0$, is found to be

$$C_d(0) = \frac{\varepsilon_s}{D_b}\sqrt{1 + e^{-2V_F/V_T}} \approx \frac{\varepsilon_s}{D_b} \triangleq C_{FB}, \tag{1.66}$$

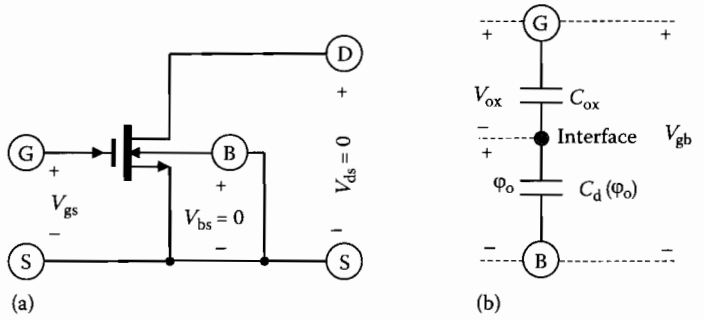

FIGURE 1.21 (a) NMOS transistor of Figure 1.19a operated with $V_{ds} = 0$ and $V_{bs} = 0$. (b) Circuit model between the gate and bulk terminals of the transistor in (a).

where C_{FB} is termed the surface flatband capacitance. The indicated approximation exploits the presumption that the impurity concentration, N_A, in the bulk substrate is significantly larger than the intrinsic carrier concentration, N_i, of silicon. It follows from Equation 1.62 that

$$\frac{C_{gb}(\varphi_o)}{C_{ox}} = \frac{1}{1 + \frac{C_{ox}}{C_d(\varphi_o)}}, \tag{1.67}$$

for which

$$\frac{C_{gb}(0)}{C_{ox}} \approx \frac{1}{1 + \left(\frac{\varepsilon_{ox}}{\varepsilon_s}\right)\left(\frac{D_b}{T_{ox}}\right)}. \tag{1.68}$$

Figure 1.22 displays a plot of the normalized gate–bulk capacitance, $C_{gb}(\varphi_o)/C_{ox}$, as a function of the normalized surface potential, φ_o/V_T, at room temperature (27°C) conditions. The relevant MOSFET is presumed to have an acceptor impurity concentration, N_A, in the bulk of 10^{15} atoms/cm^2, and a gate silicon dioxide thickness, T_{ox}, of 30 Å. The curve shows that in strong accumulation where $\varphi_o \ll 0$, the gate–bulk capacitance per unit area approaches the density of the oxide capacitance, C_{ox}. This observation reflects engineering expectations in that $\varphi_o \ll 0$ attracts a very large concentration of holes to the surface, for which the associated charge density serves to increase dramatically the surface density of capacitance, $C_d(\varphi_o)$. Indeed, for $\varphi_o \ll 0$, it is a simple matter to show that Equation 1.65 collapses to

$$C_d(\varphi_o)|_{\varphi_o \ll 0} \approx \frac{\varepsilon_s}{\sqrt{2}D_b}\sqrt{e^{-\varphi_o/V_T} - 1} \approx \frac{\varepsilon_s}{\sqrt{2}D_b} e^{|\varphi_o|/2V_T}, \tag{1.69}$$

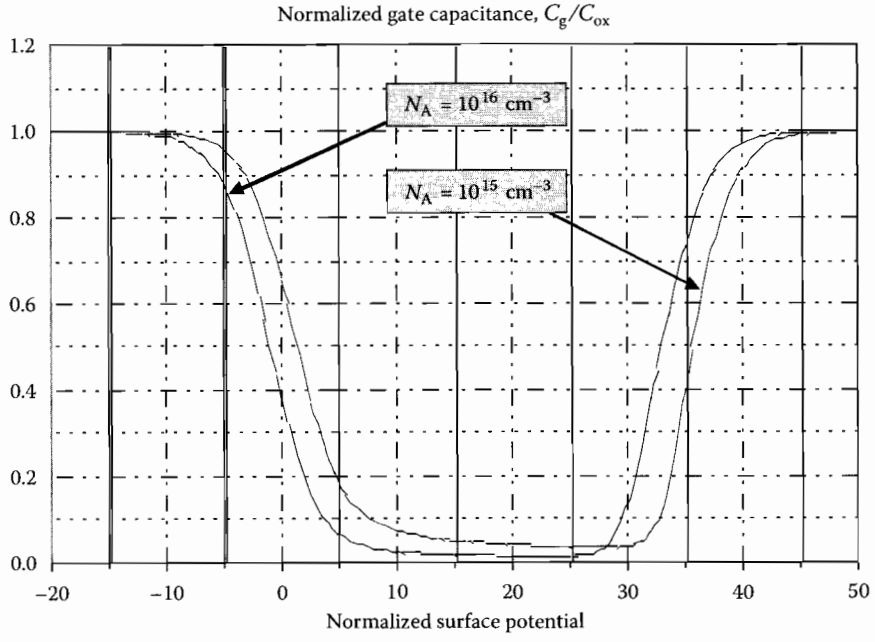

FIGURE 1.22 The normalized gate-to-bulk capacitance of the N-channel MOSFET shown in Figure 1.21a as a function of the indicated normalized surface potential. The temperature of the oxide–semiconductor interface is taken to be $T = 27$°C, the oxide thickness is $T_{ox} = 30$ Å, and the acceptor impurity concentration in the bulk is $N_A = 10^{15}$ atoms/cm^2. The curve is applicable to only low signal frequencies.

which clearly suggests a sharp rise in capacitance density with the absolute value of the negative surface potential. Since the surface capacitance density can be viewed as a ratio of the silicon dielectric constant, ε_s, to an effective and voltage-dependent dielectric thickness, say $y(\varphi_o)$, observe a dielectric thickness associated with Equation 1.69 of

$$y(\varphi_o)|_{\varphi_o \ll 0} \approx \sqrt{2} D_b e^{-|\varphi_o|/2V_T}, \tag{1.70}$$

which diminishes rapidly with progressively more negative surface potentials.

As the surface potential increases toward and beyond zero, the normalized capacitance plotted in Figure 1.22 decreases because a depletion layer begins to form at the interface. This depletion layer acts as a dielectric whose thickness increases as a nominal square root function of the surface potential. Under the depletion condition, the surface capacitance given by Equation 1.65 can be approximated by

$$C_d(\varphi_o)|_{\text{Depletion}} \approx \frac{\varepsilon_s}{\sqrt{2} D_b \sqrt{\frac{\varphi_o}{V_T} - 1}}, \tag{1.71}$$

which implies a depletion layer thickness, say y_d, (not to be confused with Y_d, the depth of the source and drain regions) of

$$y_d \triangleq y(\varphi_o)|_{\text{Depletion}} \approx D_b \sqrt{2\left(\frac{\varphi_o}{V_T} - 1\right)}. \tag{1.72}$$

Equation 1.71 approximates the actual surface capacitance density to within an error magnitude of nominally less than 10% for $3V_T \leq \varphi_o \leq 20V_T$. This allowable range of surface potential can actually be extended to embrace $\varphi_o \leq 2V_F$ since for $\varphi_o > V_F$, an appreciable portion of the charge observed at the interface can be attributed to free electrons, and not simply to the ionic charge in the depletion layer forged by holes repelled from the interface.

As φ_o continues to increase, $C_d(\varphi_o)$, and hence $C_{gb}(\varphi_o)$, continues decreasing toward a minimum value that is achieved at a value close to a surface potential of $2V_F$, which is the threshold of strong surface inversion. At this potential, the thickness of the depletion layer implicit to the interfacial capacitance density is maximized since further increases in the surface charge density derive dominantly from electrons attracted to the surface. Rather than indulge in the academic propriety of using Equation 1.65 to compute the exact surface potential commensurate with minimal surface capacitance density, engineering prudence encourages the simplified approach of presuming $\varphi_o = 2V_F$ to be a sufficiently accurate requirement for minimal depletion capacitance. Upon adoption of this stance, Equation 1.72 is suitable for computing the maximum thickness, say W_d, of the depletion layer. Accordingly,

$$W_d \triangleq y(2V_F) \approx D_b \sqrt{2\left(\frac{2V_F}{V_T} - 1\right)}, \tag{1.73}$$

and since $2V_F$ is invariably much larger than the thermal voltage, V_T, Equation 1.56 allows this result to be written as

$$W_d \approx 2\sqrt{\frac{\varepsilon_s V_F}{qN_A}}. \tag{1.74}$$

The resultant minimum density of surface depletion capacitance is

$$C_d(2V_F) = \frac{\varepsilon_s}{W_d} \approx \frac{1}{2}\sqrt{\frac{qN_A\varepsilon_s}{V_F}}. \tag{1.75}$$

Using Equations 1.67 and 1.63, the corresponding density, C_{min}, of minimum gate–bulk capacitance is

$$C_{min} \approx C_{gb}(2V_F) \approx \frac{\varepsilon_{ox}/T_{ox}}{1 + 2\left(\frac{\varepsilon_{ox}}{T_{ox}}\right)\sqrt{\frac{V_F}{qN_A\varepsilon_s}}}. \tag{1.76}$$

Observe that the maximum factor by which the effective gate–bulk capacitance can be reduced is

$$\frac{C_{ox}}{C_{min}} \approx 1 + 2\left(\frac{\varepsilon_{ox}}{T_{ox}}\right)\sqrt{\frac{V_F}{qN_A\varepsilon_s}}, \tag{1.77}$$

which involves parameters that are largely out of the control of the circuit designer. By inspection of Figure 1.22, this capacitance perturbation requires a surface potential swing extending from roughly $-15V_T$ (about -400 mV at 27°C) to $2V_F$ (generally smaller than 600 mV). Although the requisite surface potential excursion is somewhat large for maximal capacitance modulation, it should be noted that the maximum capacitance change factor predicted by Equation 1.77 can be as large as almost 100.

As φ_o increases beyond twice the Fermi potential, the interface charge density increases robustly as the semiconductor surface begins to invert strongly. Figure 1.22 resultantly displays an increasing bulk–gate capacitance density, not unlike the increased capacitance prevailing in strong accumulation because of holes attracted to the interface. Under actual measurement conditions, however, the indicated increased capacitance for $\varphi_o > 2V_F$ is observed only when the frequencies of signals established between the gate and bulk are below a few tens of hertz [4–6]. The problem is that for most practical signal frequencies, the recombination–generation rates of electrons in NMOS devices are unable to track with the signal-induced exchanges in charge between the neutral bulk and the inversion layer. Figure 1.23 displays the true gate–bulk capacitance characteristics for practical signal frequencies, wherein the dashed segment drawn for $\varphi_o > 2V_F$ is the applicable high-voltage capacitance trace for frequencies above a few tens of hertz.

1.2.2.3 Approximate Depletion Zone Analysis

As delineated in the discussion pertaining to the surface charge density defined by Equation 1.52, the MOSFET in Figure 1.21a exhibits depletion at the oxide–semiconductor interface for surface potentials satisfying the constraint, $0 < \varphi_o < 2V_F$. The formation of the surface depletion zone is critically important to the establishment of the volt–ampere characteristics of a MOSFET because it serves as a precursor to the surface inversion that comprises the necessary condition for drain and source current conduction. Recall, for example, that weak inversion is said to initiate at $\varphi_o = V_F$, in the sense that the original P-type character of the interfacial semiconductor is transformed to intrinsic material. When φ_o is elevated to $2V_F$, the surface is strongly inverted in that the concentration of free electrons at the surface increases to a value that is identical to the average impurity concentration in the substrate. Although the concentration of surface electrons begins to increase for φ_o barely above zero, as is highlighted by Equation 1.60, an electron concentration commensurate with the possibility of substantial drain and source current flow does not materialize until the surface potential, φ_o, reaches the immediate neighborhood of twice the Fermi potential. An attribute of Equation 1.52 is that this relationship does not explicitly distinguish between immobile depletion charge and mobile electron charge, both of which contribute to the observed surface charge density. A shortfall of Equation 1.52 is that its analytically cumbersome nature all but precludes the development of mathematically tractable expressions for the volt–ampere characteristics of

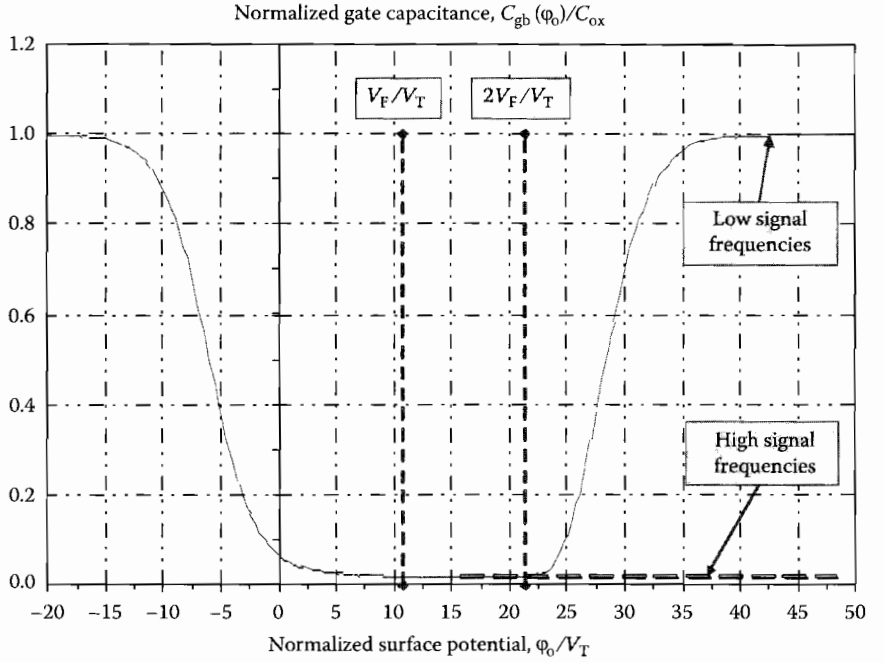

FIGURE 1.23 The capacitance characteristics of Figure 1.22 for the conditions of both low and high signal frequencies.

a MOSFET. Fortunately, the awkwardness of Equation 1.52 is mitigated if the reasonable approximation is made that for $0 < \varphi_o < 2V_F$, the charge in the surface channel region derives exclusively from depletion phenomena, that is, a substantive density of free electron charge does not materialize at the surface until $\varphi_o = 2V_F$.

To the extent that the entire substrate region is uniformly doped at the indicated impurity concentration of N_A and assuming complete ionization of all substrate impurity atoms, the resultant concentration, say $\rho(y)$, of immobile ionic charge in the depletion zone throughout the channel region from source-to-drain is nominally constant at the value, $-qN_A$. Of course, the electron concentration at the surface increases in proportion to the decreased hole population therein but as long as φ_o remains smaller than $2V_F$, Equation 1.60 confirms that the free electron concentration is significantly smaller than N_A.

Figure 1.24a depicts the depletion charge density, $\rho(y)$, beneath the oxide–semiconductor interface, where W_d represents the depth of the depletion layer established at the interface. Using Gauss' law, the electric field, $E(y)$, promoted by this charge concentration profile derives from

$$\frac{dE(y)}{dy} = \frac{\rho(y)}{\varepsilon_s}. \tag{1.78}$$

Since $\rho(y) = -qN_A$ for $0 \leq y \leq W_d$, Equation 1.78 implies

$$\int_{E(y)}^{E(W_d)} dE(y) = -\frac{qN_A}{\varepsilon_s} \int_{y}^{W_d} dy. \tag{1.79}$$

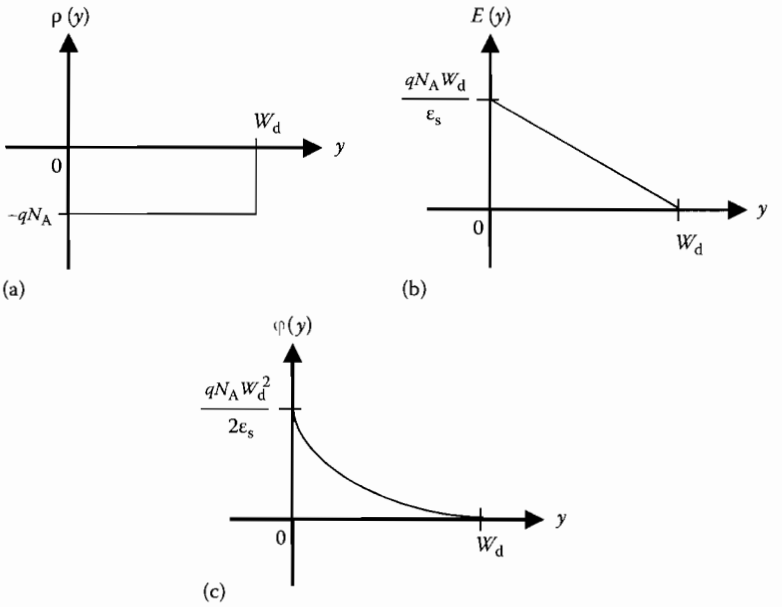

FIGURE 1.24 (a) The approximate profile of the depletion charge concentration at the surface of the MOSFET depicted in Figure 1.21a. (b) The electric field intensity as a function of bulk substrate depth measured with respect to the interfacial surface, corresponding to the charge profile in (a). (c) The potential implied by the electric field plot in (b).

In view of the fact that $E(W_d)$ is zero in the undepleted, charge neutral substrate region corresponding to $W \geq W_d$, Equation 1.79 produces the linear electric field relationship,

$$E(y) = \frac{qN_A W_d}{\varepsilon_s}\left(1 - \frac{y}{W_d}\right) = \frac{V_T W_d}{D_b^2}\left(1 - \frac{y}{W_d}\right), \tag{1.80}$$

where Equation 1.56 for the Debye length is invoked. Equation 1.80 is sketched as a function of the substrate depth variable, y, in Figure 1.24b. Observe that maximum field intensity prevails at surface where

$$E(0) \triangleq E_o = \frac{V_T}{D_b}\left(\frac{W_d}{D_b}\right). \tag{1.81}$$

The potential, $\varphi(y)$, corresponding to the field intensity, $E(y)$, stipulated by Equation 1.80 satisfies

$$-\frac{d\varphi(y)}{dy} = E(y) = \frac{qN_A}{\varepsilon_s}(W_d - y). \tag{1.82}$$

If zero reference potential is ascribed to the substrate depth, W_d, beyond which the substrate is charge neutral, Equation 1.82 sets forth

$$\int_{\varphi(y)}^{0} d\varphi(y) = \frac{qN_A}{\varepsilon_s}\int_{y}^{W_d}(W_d - y)dy, \tag{1.83}$$

whence the bulk substrate potential, $\varphi(y)$, referenced to the potential evidenced at the substrate depletion depth, W_d, is

$$\varphi(y) = \frac{qN_A W_d^2}{2\varepsilon_s}\left(1 - \frac{y}{W_d}\right)^2 = \frac{1}{2}V_T\left(\frac{W_d}{D_b}\right)^2\left(1 - \frac{y}{W_d}\right)^2, \tag{1.84}$$

whose functional dependence on variable y is sketched in Figure 1.24c. Equation 1.84 suggests that the surface potential, $\varphi(0)$, which is effectively the net voltage dropped across the depleted region of the bulk substrate, is

$$\varphi(0) \triangleq \varphi_o = \frac{qN_A W_d^2}{2\varepsilon_s} = \frac{1}{2}V_T\left(\frac{W_d}{D_b}\right)^2 \equiv \frac{E_o W_d}{2}, \tag{1.85}$$

Since the interface potential, φ_o, and hence the depletion depth, W_d, is controlled externally by the applied gate-to-bulk voltage, V_{gb}, it is of interest to determine φ_o as an explicit function of V_{gb}. To this end, the electric field intensity, E_{ox}, in the silicon dioxide layer of the structure of Figure 1.19b is uniform throughout the oxide thickness by virtue of the insulating nature of the oxide. Ignoring work function phenomena prevailing between the gate contact and gate oxide, as well as between the oxide and semiconductor surface, this field is simply

$$E_{ox} = \frac{V_{gb} - \varphi_o}{T_{ox}}. \tag{1.86}$$

Equation 1.86 also invokes the approximation that the voltage dropped from the bottom of the depletion region-to-the bulk terminal is essentially zero. This assumption is reasonable in that the substrate is ultimately reverse biased to preclude substantive bulk current flow. Moreover, the holes displaced from the interface region-to-the neutral bulk render the neutral substrate zone a low resistivity volume. Because the field immediately below the interface is E_o, as defined by Equation 1.81, continuity constraints mandate

$$\varepsilon_{ox}E_{ox} = \varepsilon_s E_o, \tag{1.87}$$

or

$$\varepsilon_{ox}\left(\frac{V_{gb} - \varphi_o}{T_{ox}}\right) = \varepsilon_s\left(\frac{V_T W_d}{D_b^2}\right) \approx C_{FB}\left(\frac{W_d}{D_b}\right)V_T. \tag{1.88}$$

Armed with Equations 1.86, 1.63, and 1.85, parameter W_d in Equation 1.88 can be eliminated to arrive at the utilitarian expression,

$$V_{gb} = \varphi_o + \sqrt{2V_\theta\varphi_o}, \tag{1.89}$$

where the voltage metric, V_θ, termed the *body effect voltage*, is given by

$$V_\theta = V_T\left(\frac{C_{FB}}{C_{ox}}\right)^2 = V_T\left(\frac{\varepsilon_s}{\varepsilon_{ox}}\right)^2\left(\frac{T_{ox}}{D_b}\right)^2 = \frac{qN_A\varepsilon_s}{C_{ox}^2}. \tag{1.90}$$

Parameter V_θ is generally of the order of the mid-tens of microvolts.* Observe that V_θ is proportional to the square of the oxide thickness, T_{ox} and is therefore reduced sharply with diminishing gate oxide thickness.

The final step to the problem of determining the dependence of interface potential φ_o on applied gate-to-bulk voltage V_{gb} involves a straightforward solution of Equation 1.89 for φ_o. The result is

$$\varphi_o = V_{gb} + V_\theta - \sqrt{V_\theta(2V_{gb} + V_\theta)}. \tag{1.91}$$

As expected, $\varphi_o = 0$ for $V_{gb} = 0$. Ordinarily, V_θ is much smaller than practical values of the gate-to-bulk voltage, V_{gb}, so that Equation 1.91 can be approximated as

$$\varphi_o \approx V_{gb} - \sqrt{2V_\theta V_{gb}}, \tag{1.92}$$

which is similar in form to Equation 1.89.

1.2.2.4 Threshold

The approximate depletion regime analysis executed in Section 1.2.2.3 conveniently precipitates an analytical definition of the threshold condition, that is, the condition whereby strong inversion materializes at the oxide–semiconductor interface. It has been demonstrated that the onset of the threshold condition corresponds to a surface potential, φ_o, of twice the Fermi potential. Accordingly, threshold requires that the gate–bulk voltage V_{gb}, in Equation 1.89 rise to a value, say V_{gbh}, such that

$$V_{gbh} = 2V_F + 2\sqrt{V_\theta V_F}. \tag{1.93}$$

It is to be understood that a gate-to-bulk voltage, V_{gb}, satisfying the constraint, $V_{gb} \geq V_{gbh}$, is commensurate with instilling strong inversion at the surface of the MOSFET depicted in Figure 1.21a. On the presumption that the depth, W_d, of the depletion layer, corresponding to $\varphi_o = 2V_F$ is unaltered by gate-bulk voltage increases beyond the threshold value, V_{gbh}, the resultant charge profile offered in Figure 1.24a changes into the form diagrammed in Figure 1.25. In this diagram, N_s is the concentration of free

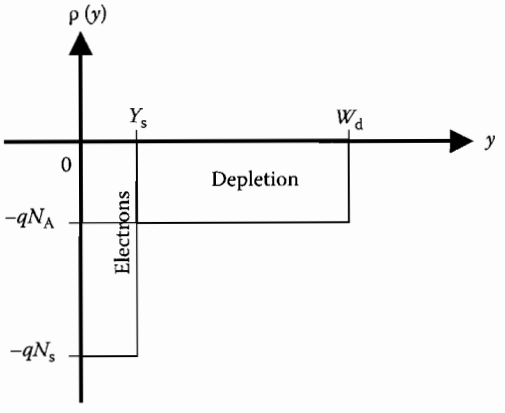

FIGURE 1.25 The approximate profile of the charge concentration for a strongly inverted surface in the MOSFET depicted in Figure 1.21a.

* Most HSPICE and other SPICE simulators use a *body effect parameter* to compute the extent to which the bulk–source voltage perturbs the gate–source threshold voltage. This body effect parameter, γ, derives from $\gamma^2 = 2V_\theta$.

electrons in the surface inversion layer, and Y_s is the thickness of the inversion layer. Depending on the value of the gate–bulk voltage excess, $(V_{gb} - V_{gbh})$, Y_s is typically 20%–50% larger than the electron Debye length.

Two circumstances limit the utility of Equation 1.93. The first of these is that MOSFETs are often operated with nonzero bulk–source bias, as opposed to the zero bias presumed to this juncture. If a bulk–source voltage, V_{bs}, is applied to the MOSFET in Figure 1.21a, the potentials at both the oxide–semiconductor interface and the bottom of the depletion layer are elevated by an amount, V_{bs}, which, in concert with earlier admonitions, is invariably a negative voltage to ensure reverse biasing of the bulk–source and bulk–drain PN junctions. Thus, the results documented in Section 1.2.2.3 remain valid because the voltage developed across the depletion layer is still the surface potential, φ_o, exploited therein. However, since the surface potential rises by V_{bs}, one of the necessary modifications to results disclosed earlier is that Equation 1.60 for the free electron concentration at the surface must be modified as

$$n(0) = \frac{N_i^2}{N_A} e^{(\varphi_o + V_{bs})/V_T} = N_A e^{(\varphi_o + V_{bs} - 2V_F)/V_T}. \tag{1.94}$$

Recall that the measure for the onset of strong inversion in a MOSFET is a surface electron concentration, $n(0)$, that equals to the hole concentration, N_A, in the equilibrium substrate. In order to effect this strong inversion condition, Equation 1.94 suggests the necessity of a surface potential that is at least as large as $(2V_F - V_{bs})$, as opposed to merely $2V_F$. Accordingly, the effect of bulk–source biasing on the gate–bulk threshold voltage can be embraced by replacing the voltage, $2V_F$, in Equation 1.93 by the voltage $(2V_F - V_{bs})$ so that

$$V_{gbh} = (2V_F - V_{bs}) + \sqrt{2V_\theta(2V_F - V_{bs})}. \tag{1.95}$$

The second shortfall of Equation 1.93 stems from the fact in actual circuit design environments, it is far more convenient to stipulate the minimum gate–source voltage, V_{gs}, and not the minimum gate–bulk voltage, V_{bs}, that establishes the onset of strong inversion. Since V_{gs} is the voltage sum, $(V_{gb} + V_{bs})$, adding V_{bs} to both sides of Equation 1.95 delivers a gate-to-source threshold voltage, say V_h, of the form,

$$V_h = V_{ho} + 2\sqrt{V_\theta V_F}\left(\sqrt{1 - \frac{V_{bs}}{2V_F}} - 1\right), \tag{1.96}$$

where V_{ho}, which represents the zero bias ($V_{bs} = 0$) value of the gate–source threshold potential, is

$$V_{ho} = 2\left(V_F + \sqrt{V_\theta V_F}\right). \tag{1.97}$$

In practice, the zero bias value, V_{ho}, of gate–source threshold voltage is best evaluated through measurement since it is strongly influenced by gate region work function phenomena and parasitic charges trapped in the gate oxide layer, whose engineering effects are difficult to quantify accurately and reliably. Observe in Equation 1.96 that the effect of an increasing bulk-to-source reverse bias ($V_{bs} < 0$) is to increase the threshold voltage above its zero bias value, V_{ho}, as a square root function of V_{bs}. This bulk-induced modulation of the threshold potential is rendered small by small values of the square root of parameter V_θ, which Equation 1.90 projects as directly dependent on the gate oxide thickness, T_{ox}. One reason for the current penchant toward progressively decreased oxide thickness is the minimization of threshold voltage modulation, which is generally an undesirable effect in MOSFETs deployed in analog network applications.

1.2.3 Volt–Ampere Characteristics

The static volt–ampere characteristics of the N-channel MOSFET in Figure 1.17 stipulate the dependence of the static drain current, I_d, on the static values of the gate–source voltage, V_{gs}, the drain–source voltage, V_{ds}, and the bulk–source voltage, V_{bs}, which is invariably a nonpositive voltage. It is convenient to partition these characteristics into three segments; namely, the *cutoff regime*, the *ohmic regime*, and the *saturation regime*.

1.2.3.1 Cutoff Regime

The cutoff regime is the most boring of the three MOSFET operating domains in that no drain current flows in cutoff, despite all reasonable positive value of the drain–source voltage, V_{ds}. Since drain current conduction requires surface inversion at the oxide–substrate interface, zero current is assured when no such charge inversion prevails. In turn, no inversion layer is formed when the gate–source voltage, V_{gs}, lies below its threshold value, V_h. Thus, in cutoff,

$$I_d = 0, \quad \text{if } V_{gs} < V_h. \tag{1.98}$$

As V_{gs} rises above zero but remains below V_h, a subthreshold current is induced from the drain-to-the source regions for $V_{ds} > 0$ [7]. This current is manifested by the fact that, as is conveyed by Equation 1.94, the interfacial free electron concentration increases slightly for surface potentials above 0 V. However, the current evidenced in the subthreshold regime is small because of the limited availability of free surface electrons. As a result, the gain and frequency response, in addition to the actual drain current, that the transistor is capable of mustering are limited. Although subthreshold operation enjoys utility in certain types of low-power system applications, such as hearing aids, where neither gain nor bandwidth are daunting requirements, it is rarely exploited in broadband and other high-performance applications.

1.2.3.2 Ohmic Regime

In the ohmic regime, which is sometimes called the *triode regime*, $V_{gs} \geq V_h$ and $V_{ds} \leq (V_{gs} - V_h)$. Thus, the interfacial surface of a MOSFET is strongly inverted in the ohmic domain, and simultaneously, a relatively small voltage is applied from the drain-to-the source. The voltage difference, $(V_{gs} - V_h)$, is commonly referenced as the *drain saturation voltage*, V_{dsat}, that is,

$$V_{dsat} \triangleq V_{gs} - V_h. \tag{1.99}$$

Because of Equation 1.49, observe that the provision, $V_{ds} \leq (V_{gs} - V_h)$, is equivalent to the requirement, $V_{gd} \geq V_h$. This is to say that a MOSFET operates in the ohmic regime if and only if both the gate–source and the gate–drain voltages are larger than the threshold potential. Viewed in yet another fashion, $V_{gs} \geq V_h$ and $V_{gd} \geq V_h$ ensure that both the source and the drain ends of the interfacial surface between the source and the drain regions are strongly inverted. A conduit, or channel, of free electrons that electrically couples the source to the drain is thereby established.

The aforementioned channel of free electrons is highlighted in the device cross section abstracted in Figure 1.26. Because $V_{gs} \geq V_h$ and $V_{gd} \geq V_h$, the electron inversion layer extends throughout the entire surface region from the source-to-the drain. But since $V_{gd} = (V_{gs} - V_{ds})$ and $V_{ds} > 0$, the gate-to-drain bias, V_{gd}, is necessarily smaller than its gate-to-source counterpart, V_{gs}. It follows that the surface potential in the neighborhood of the drain region is smaller than that prevailing near the source region, whence the electron concentration near the drain is smaller than it is at the source. Accordingly, the channel of electrons depicted in the figure at hand does not have a uniform depth (y-direction) and is, in fact, deeper at the source site, where $x = 0$, than it is at the drain site, which is typified analytically by $x = L$. For analogous reasons, the depletion region established about the source, at the interface, and at the drain is widest near the drain.

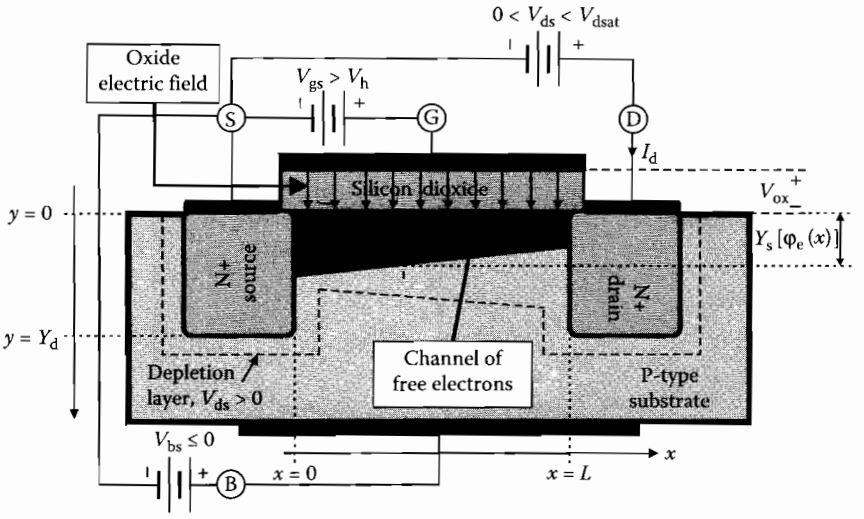

FIGURE 1.26 Cross section of the N-channel MOSFET operated in its ohmic regime. Note that all applied voltages are referred to the source terminal. The diagram is not drawn to scale.

An additionally important point is that the channel potential, symbolized in Figure 1.26 as $\varphi_c(x)$, is measured with respect to the source site. This notation is not to be confused with the previously invoked variable, $\varphi_o(y)$, which measures the potential at the interfacial surface with respect to the neutral region of the bulk substrate. The change in symbolism is reasonable and is encouraged by two issues addressed in the Section 1.2.3.3. The first of these issues is that the bulk–source biasing voltage, V_{bs}, has been absorbed into the threshold voltage metric stipulated by Equation 1.96. Second, this threshold voltage has been defined in terms of the gate voltage, V_{gs}, measured with respect to the source, as opposed to the gate voltage, V_{gb}, referenced to the bulk terminal. Because V_{ds} is nonzero, the channel potential, $\varphi_c(x)$, is not a constant but instead, it varies continuously from $\varphi_c(0) = 0$ at the source site where $x = 0$ to $\varphi_c(L) = V_{ds}$ at the drain site where $x = L$.

An applied drain–source voltage, V_{ds}, launches a lateral electric field, say $E_x(\varphi_c(x))$, that is directed from the drain site-to-the source site and is functionally dependent on the channel potential, $\varphi_c(x)$. This electric field is given by the familiar relationship,

$$E_x(\varphi_c(x)) = -\frac{d\varphi_c(x)}{dx}. \tag{1.100}$$

If μ_n denotes the mobility of electrons, whose concentration within the surface inversion layer postulated in Figures 1.26 and 1.25 is $N_s(\varphi_c(x))$, the static drain current, I_d, promoted by this lateral field is

$$\begin{aligned} I_d &= -q\mu_n W[Y_s(\varphi_c(x))][N_s(\varphi_c(x))]E_x(\varphi_c(x)) \\ &= q\mu_n W[Y_s(\varphi_c(x))][N_s(\varphi_c(x))]\frac{d\varphi_c(x)}{dx}, \end{aligned} \tag{1.101}$$

where $Y_s(\varphi_c(x))$ is the inversion layer thickness introduced in Figure 1.25 and depicted as dependent on the channel potential, $\varphi_c(x)$, in Figure 1.26. It is worthwhile noting that the product, $\mu_n E_x(\varphi_c(x))$, is the velocity of electrons propagated through the inversion layer. This velocity is zero, thereby implying zero drain current, if the gradient, $d\varphi_c(x)/dx$, of channel potential is null. In turn, the channel potential gradient is zero if the applied drain–source voltage, V_{ds}, is zero.

If it is assumed that increases in the channel potential over and above the threshold level incur no change in the geometry of the interfacial depletion region and instead, only cause electrons to be attracted to the surface, Gauss's law predicts

$$q[Y_s(\varphi_c(x))][N_s(\varphi_c(x))] = \varepsilon_{ox}E_{ox}(\varphi_c(x)), \tag{1.102}$$

with $E_{ox}(\varphi_c(x))$ symbolizing the oxide electric field, which is given by

$$E_{ox}(\varphi_c(x)) = \frac{V_{ox}}{T_{ox}} = \frac{V_{gs} - V_h - \varphi_c(x)}{T_{ox}}. \tag{1.103}$$

Recalling Equations 1.63, 1.103, and 1.102 combine with Equation 1.101 to deliver

$$I_d dx = \mu_n C_{ox} W\left[V_{gs} - V_h - \varphi_c(x)\right] d\varphi_c(x). \tag{1.104}$$

An integration of the left-hand side of this result from $x = 0$ to $x = L$ is tantamount to integrating the right-hand side of said result from $\varphi_c(0) = 0$ to $\varphi_c(L) = V_{ds}$. Assuming constant electron mobility through the channel, the requisite integration is straightforward and leads to the desired volt–ampere relationship,

$$I_d = K_n\left(\frac{W}{L}\right)V_{ds}\left(V_{gs} - V_h - \frac{V_{ds}}{2}\right), \tag{1.105}$$

where

$$K_n \triangleq \mu_n C_{ox} \tag{1.106}$$

is the so-called *transconductance coefficient* of the MOSFET. Although K_n is termed a transconductance coefficient, it is not actually a transconductance in that its physical dimension is that of siemens/volt, or mhos/volt.

Several interesting and enlightening features are advanced by Equation 1.105. The first of these is that zero drain current prevails if $V_{ds} = 0$, which is reassuring in that a current flow for null drain–source voltage violates engineering reason, if not the minor issue of conservation of energy. A second, and more significant, point is that the drain current is directly proportional to the gate aspect ratio, W/L. Thus, for fixed gate–source and drain–source voltages, the drain current can be increased or decreased in proportion to this geometric ratio. This controllability over the drain current renders the gate aspect ratio a designable circuit parameter, subject to the proviso that the circuit designer not attempt to make the gate width, W, smaller than the minimum channel length, L, that the process foundry is capable of producing. Thus, if the foundry boasts a 130 nm channel length process, the smallest practical value of W is, in fact, also 130 nm.

A third important feature of Equation 1.105 is the existence of a value of V_{ds} for which drain current I_d is maximized. By setting to zero the partial derivative of I_d in Equation 1.105 with respect to V_{ds}, this extremum is determined to lie at $V_{ds} = (V_{gs} - V_h) = V_{dsat}$, for which the corresponding maximum current, say I_{dsat}, is

$$I_{dsat} = \frac{K_n}{2}\left(\frac{W}{L}\right)(V_{gs} - V_h)^2 = \frac{K_n}{2}\left(\frac{W}{L}\right)V_{dsat}^2. \tag{1.107}$$

Recalling Equation 1.99, $V_{ds} = V_{dsat}$ corresponds to a gate–drain voltage, V_{gd}, of $V_{gd} = V_h$, which implies that the surface potential at the drain end of the channel barely sustains the onset of strong inversion. In effect, the depth of the electron channel is reduced to zero at the drain site for $V_{ds} = V_{dsat}$,

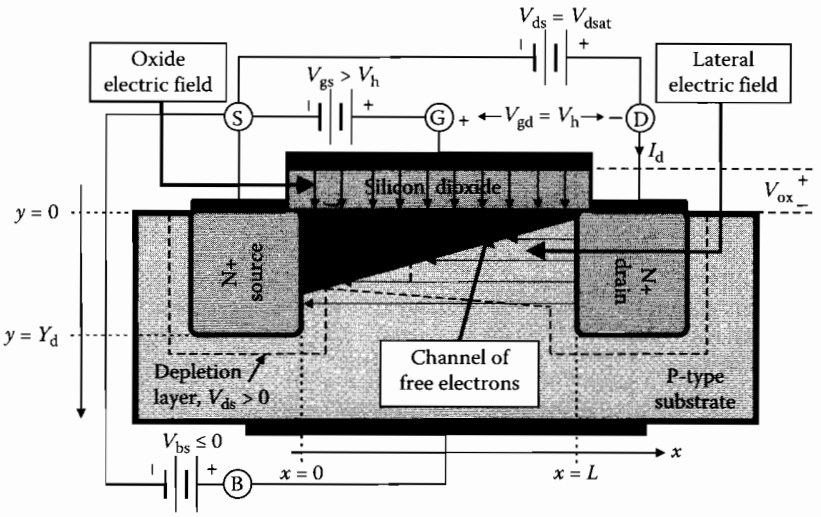

FIGURE 1.27 Cross section of the N-channel MOSFET operated in channel pinch off. The diagram is not drawn to scale.

which justifies the common vernacular of a channel that is *pinched off* at the drain. The situation at hand is diagrammed in Figure 1.27.

At first blush, it may appear incongruous that a pinched off channel, which might be viewed as a means to cut off the supply of electrons to the drain site, can sustain a drain current, yet alone the maximum drain current postulated by Equation 1.107. The current is indeed sustained because of two prevailing phenomena. First, the electric field, $-d\varphi_c(x)/dx$, within the inversion layer encourages the transit of electrons toward the tapered edge of the channel at the drain site. Second, electrons reaching the channel edge are influenced immediately by the lateral electric field established by the applied positive drain-to-source voltage. This field, which is abstracted in Figure 1.27 by the indicated horizontal vectors directed from the drain region-to-the source region, sweeps those electrons at the inversion layer boundary into the drain region. The resultant current arising from the transport of electrons across the depletion zone between the tapered channel edge and the drain is, like the current within the inversion layer, proportional to the mobility of electrons. In the depletion zone, this carrier mobility is minority carrier mobility, which is inversely proportional to the background impurity concentration of the bulk substrate. A fundamental reason for maintaining relatively low impurity concentration in the bulk is the assurance of relatively high minority carrier (electron) mobility therein so that carriers are swept across the depletion zone at high velocity, thereby facilitating fast transistor switching and broadband circuit responses.

The fourth interesting feature surrounding Equation 1.105 lends credence to the term, "ohmic," as a descriptive for the operating regime at hand. In particular, Equation 1.105 can be expressed in the form,

$$I_d = K_n\left(\frac{W}{L}\right)V_{ds}\left(V_{gs} - V_h - \frac{V_{ds}}{2}\right) = \frac{V_{ds}}{R_{ds}(V_{gs})}, \tag{1.108}$$

with

$$R_{ds}(V_{gs}) = \frac{1}{K_n(W/L)\left(V_{gs} - V_h - \frac{V_{ds}}{2}\right)}. \tag{1.109}$$

In other words, and as is proffered in Figure 1.28, a MOSFET operated in its ohmic regime, where $V_{gs} \geq V_h$ and $V_{ds} \leq V_{dsat}$, behaves as a drain-to-source resistance, $R_{ds}(V_{gs})$, whose resistance value is

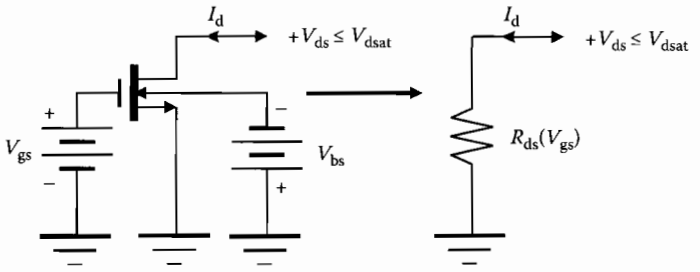

FIGURE 1.28 Static circuit model of an N-channel MOSFET operated in its ohmic regime. The transistor can be operated in such a way that its drain–source terminals emulate a voltage-controlled, nominally linear resistance.

controlled by the applied gate–source voltage, V_{gs}. Moreover, the synthesized resistance is approximately independent of the voltage, V_{ds}, developed across its terminals, and therefore emulates a linear resistance, if $V_{ds} \ll 2(V_{gs} - V_h) \equiv 2V_{dsat}$. In effect, the ohmic regime MOSFET is an electronic approximation of a linear potentiometer whose resistance setting is inversely proportional to the applied gate–source voltage.

1.2.3.3 Saturation Regime

In saturation, which is the volt-ampere domain in which MOSFETs embedded in high-performance analog circuits function, $V_{gs} \geq V_h$ and $V_{ds} \geq (V_{gs} - V_h)$. To first order, the drain current in saturation is taken to be the drain saturation current given by Equation 1.107, which is independent of drain–source voltage, V_{ds}, that is,

$$I_d = \frac{K_n}{2}\left(\frac{W}{L}\right)(V_{gs} - V_h)^2 \quad \text{for } V_{gs} > V_h, \quad \text{and} \quad V_{ds} \geq (V_{gs} - V_h). \tag{1.110}$$

The logic underlying this approximation is that the drain current in saturation is determined by the surface electron concentration established for the drain–source voltage, $V_{ds} = (V_{gs} - V_h) = V_{dsat}$, which barely allows for an electron channel spanning the entire source-to-drain spacing. Any increase in the drain-to-source voltage above its saturated value, V_{dsat}, simply adds impetus to the attractive force exerted on inversion layer electrons by the lateral electric field promoted by the drain–source voltage.

The problem with the foregoing logic is that the drain current given by Equation 1.107 is premised on Equation 1.105, which in turn invokes the presumption of an electron inversion layer length that is identical to the channel spacing length, L, separating the source region from the drain region. If $V_{ds} = V_{dsat}$ incurs pinch off at the drain site, and hence an inversion layer length equal to the channel length, L, as illustrated in Figures 1.27 and 1.29a, $V_{ds} > V_{dsat}$ necessarily incurs pinch off within the source–drain spacing, as is suggested in Figure 1.29b. Because of the indicated reduction in the effective channel length from L to $(L - \Delta L)$, the integrated form of Equation 1.104 is now

$$\int_0^{L-\Delta L} I_d dx = \int_0^{V_{dsat}} \mu_n C_{ox} W\left[V_{gs} - V_h - \varphi_c(x)\right] d\varphi_c(x). \tag{1.111}$$

The result of this integration exercise is easily demonstrated to be

$$I_d = \frac{K_n}{2}\left(\frac{W}{L - \Delta L}\right)(V_{gs} - V_h)^2 = I_{dsat}\left(\frac{L}{L - \Delta L}\right), \tag{1.112}$$

where the current, I_{dsat}, is given by Equation 1.107 and represents the drain current at the transition boundary between ohmic and saturation operational regimes. For most practical applications of

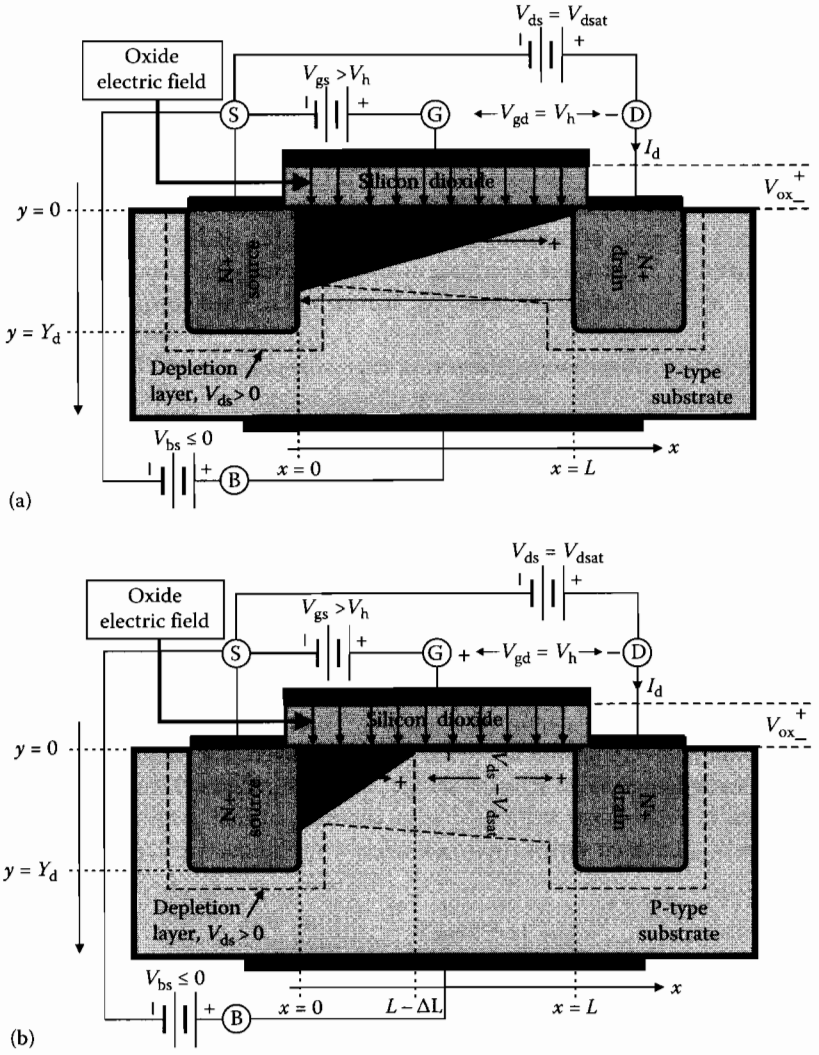

FIGURE 1.29 (a) Cross section of N-channel MOSFET operated in strong inversion and with $V_{ds} = V_{dsat}$. (b) Cross section of the MOSFET in (a) operated with $V_{ds} > V_{dsat}$. The diagrams are not drawn to scale.

MOSFETs [8], the effective reduction, $(L - \Delta L)$, in channel length relates to the drawn channel length, L, as

$$\frac{L}{L - \Delta L} \approx 1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}, \tag{1.113}$$

where V_λ, termed the *channel length modulation voltage*,* is given by the semiempirical expression,

$$V_\lambda = \left(\frac{L}{D_b}\right)\left(\frac{V_j}{V_F}\right)^2 \sqrt{32V_T\left(V_{ds} - V_{dsat} + V_j\right)}. \tag{1.114}$$

* Most HSPICE and other SPICE simulators use a *channel length parameter* to compute the degree to which the drain–source voltage affects the drain saturation current. This channel length parameter, λ, is $\lambda = 1/V_\lambda$.

In Equation 1.114, V_T is the familiar thermal voltage, D_b is the electron Debye length delineated in Equation 1.56, and V_j is the *built-in potential* of the bulk–drain PN junction. Specifically,

$$V_j = V_T \ln\left(\frac{N_A N_D}{N_i^2}\right), \tag{1.115}$$

with N_A, N_D, and N_i respectively denoting the average impurity concentration in the bulk substrate, the average impurity concentration of the drain diffusion (or implant), and the intrinsic carrier concentration of silicon. Equation 1.114 delivers acceptable analytical accuracy for channel lengths, L, that are no smaller than 0.09 μ and drain–source voltages, V_{ds}, that lie within breakdown ratings of the considered transistor.

The drain current in the saturation regime is now expressible as

$$I_d \approx \frac{K_n}{2}\left(\frac{W}{L}\right)(V_{gs} - V_h)^2\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right) = I_{dsat}\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right), \tag{1.116}$$

where it is understood that the gate–source and drain–source voltages, V_{gs} and V_{ds}, respectively, are constrained to satisfy the saturation requirements, $V_{gs} > V_h$ and $V_{ds} \geq (V_{gs} - V_h) = V_{dsat}$. Clearly, the saturation regime drain current is no longer independent of the drain–source voltage. The current is seen to rise with V_{ds} with a slope of I_{dsat}/V_λ. Note, however, that this slope is not constant owing to its square root dependence on V_{ds}. For large V_λ, which is manifested by long transistor channel length, L, this rate of current rise with V_{ds} is modest and indeed, the slope of the current–voltage characteristic curve approaches zero in the limit as V_λ approaches infinity. These observations and Equation 1.116 itself suggest that the drain–source port of a MOSFET does not behave as a constant current source whose value, I_{dsat}, is controlled exclusively by gate–source voltage V_{gs}. Instead, the drain–source port is a practical controlled current source comprised of a constant current generator, albeit controlled by gate–source voltage V_{gs}, in shunt with a resistive branch. To wit, Equation 1.116 can be written as

$$I_d \approx I_{dsat} + \frac{V_{ds} - V_{dsat}}{V_\lambda / I_{dsat}}, \tag{1.117}$$

which suggests the static circuit model provided in Figure 1.30. The subject model is more useful conceptually than computationally since a change made to V_{gs} for the purpose of adjusting the nominal drain current, I_{dsat}, influences the resistance value, V_λ/I_{dsat}, and the voltage offset, V_{dsat}, introduced in the drain–source port.

A complication of the channel length embellishment to the saturation drain current expression is that Equation 1.116 is discontinuous with the ohmic domain drain current expression in Equation 1.105 at the transition boundary between respective operating domains. Simple software fixes in commonly

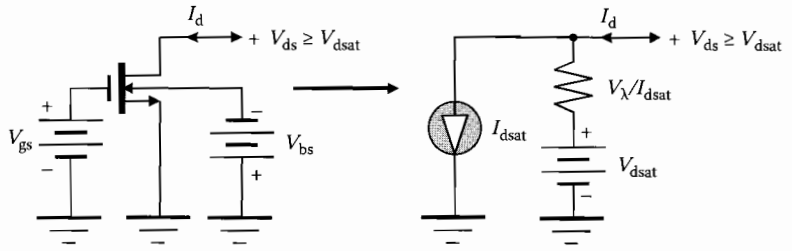

FIGURE 1.30 A large-signal circuit model for an N-channel MOSFET biased to operate in its saturation domain.

available circuit simulators rectify this incongruity. From an analytical perspective, the problem can be tacitly ignored, if V_λ in Equation 1.114 abides by the previously disclosed channel length and voltage restrictions.

For the convenience of the reader, the relevant expressions for the volt–ampere characteristic curves of an N-channel MOSFET are synopsized herewith. In particular,

$$I_d \approx \begin{cases} 0, & V_{gs} < V_h \\ K_n\left(\frac{W}{L}\right)V_{ds}\left(V_{gs} - V_h - \frac{V_{ds}}{2}\right), & V_{gs} \geq V_h;\ V_{ds} < V_{dsat} \\ \frac{K_n}{2}\left(\frac{W}{L}\right)\left(V_{gs} - V_h\right)^2\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right), & V_{gs} \geq V_h;\ V_{ds} \geq V_{dsat} \end{cases}, \tag{1.118}$$

where V_{dsat} is the voltage difference, $(V_{gs} - V_h)$. It is to be understood that the positive reference direction of the drain current in NMOS is a current flowing into the drain, while the positive reference voltage polarities reflect those highlighted in Figure 1.17. Moreover, V_h is recalled as a threshold level dependent on the bulk–source voltage, V_{bs}, in accordance with Equation 1.96. A representative plot of the static volt–ampere characteristics of an NMOS transistor appear in Figure 1.31.

In the interests of clarity and completeness, the PMOS counterpart to Equation 1.118 is

$$I_d \approx \begin{cases} 0, & V_{sg} < V_h \\ K_p\left(\frac{W}{L}\right)V_{sd}\left(V_{sg} - V_h - \frac{V_{sd}}{2}\right), & V_{sg} \geq V_h;\ V_{sd} < V_{dsat} \\ \frac{K_p}{2}\left(\frac{W}{L}\right)\left(V_{sg} - V_h\right)^2\left(1 + \frac{V_{sg} - V_{dsat}}{V_\lambda}\right), & V_{sg} \geq V_h;\ V_{sd} \geq V_{dsat} \end{cases}, \tag{1.119}$$

where, in terms of the source-gate voltage, V_{sg}, V_{dsat} is now given by, $(V_{sg} - V_h)$, the threshold voltage, which is dependent on source-bulk voltage V_{sb} in Equation 1.96, remains a positive number, and

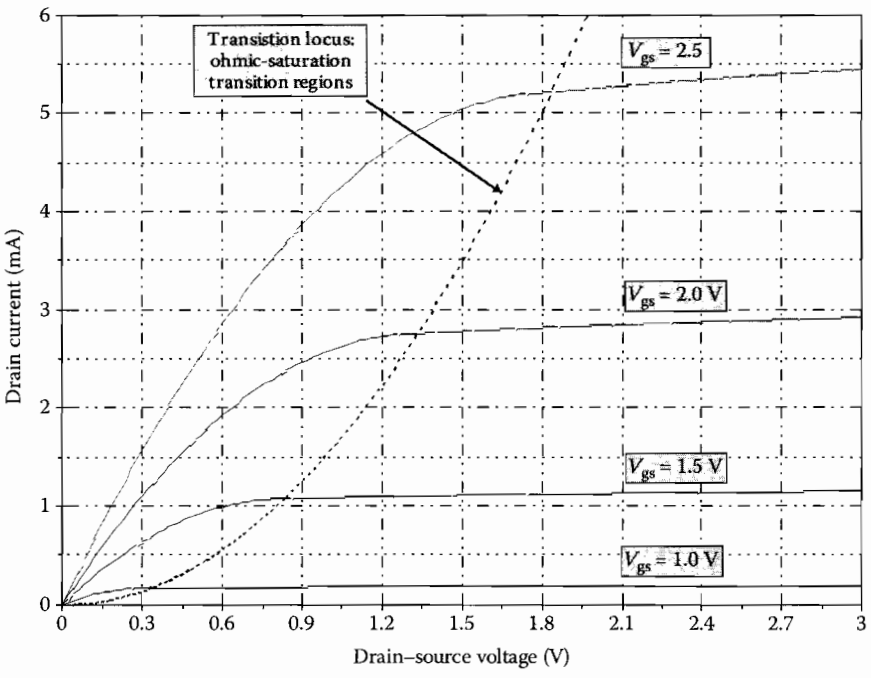

FIGURE 1.31 Common-source volt–ampere characteristic curves for an NMOS transistor.

transconductance parameter K_p is now the product of oxide capacitance density and hole mobility. The positive reference direction of the drain current in PMOS is a current flowing out of the drain, while the positive reference voltage polarities pertain to those delineated in Figure 1.18.

1.2.3.4 Refinements to the Static Model

The static volt–ampere characteristic in Equations 1.118 and 1.119 exhibit observable errors when computed currents are compared to experimental measurements executed on deep submicron MOSFET technology transistors. The principle source of these errors is two types of mobility degradation to which carriers in the source-to-drain channel are subjected. The first form of mobility degradation derives from the large lateral electric fields evidenced when even relatively small drain-to-source voltages are applied across channels whose lengths are smaller than approximately 0.25 μ. The second form of mobility impairment is caused by the strong vertical electric fields established by gate–source voltages applied across thin oxide layers.

1.2.3.4.1 Lateral Electric Fields

The NMOS and PMOS volt–ampere characteristic equations in Equations 1.118 and 1.119 are predicated on the presumption that the drift velocity, say v_c, of carriers propagated through the inverted channel at the oxide–semiconductor interface is proportional to the lateral electric field, $E_x[\varphi_c(x)]$. This field is, of course, established in the channel by applied drain–source voltage, V_{ds}, (in the case of NMOS) or applied source–drain voltage V_{sd} (in the case of PMOS). In particular,

$$v_c = \mu_o |E_x|, \tag{1.120}$$

where μ_o represents either the low field value of the electron mobility, μ_n, in N-channel devices or the low field value of the hole mobility, μ_p, in PMOS. The simpler notation, E_x is adopted in Equation 1.120 to represent the potential-dependent field function, $E_x[\varphi_c(x)]$. The need for the absolute value operation on the right-hand side of Equation 1.120 materializes from the fact that the carrier velocity, which is always a positive metric, is directed against the direction of the channel field in NMOS. In the case of NMOS transistors, carriers drift in the direction of the source-to-the drain, whereas the field is directed from drain-to-source and is therefore negative. For PMOS, no algebraic sign problems are manifested, since carriers drift in the same direction as the lateral field, whence E_x is positive.

The simplicity of Equation 1.120 belies the fact that the carrier drift velocity does not continually increase in proportion to the electric field. In fact, the carrier velocity saturates at a value, say v_{max}, which is of the order of 0.15 μm/ps in silicon, when electric fields are excessive. In recognition of this physical phenomenon, Equation 1.120 is supplanted by the empirical relationship,

$$v_c = \frac{\mu_o |E_x|}{1 + |E_x|/E_c}, \tag{1.121}$$

where

$$E_c = \frac{v_{max}}{\mu_o} \tag{1.122}$$

is termed the *critical electric field*. Typically, E_c is in the range of 3–5 V/μm. A comparison of Equation 1.121 with Equation 1.120 suggests an effective mobility, μ_e, of

$$\mu_e = \frac{\mu_o}{1 + |E_x|/E_c}. \tag{1.123}$$

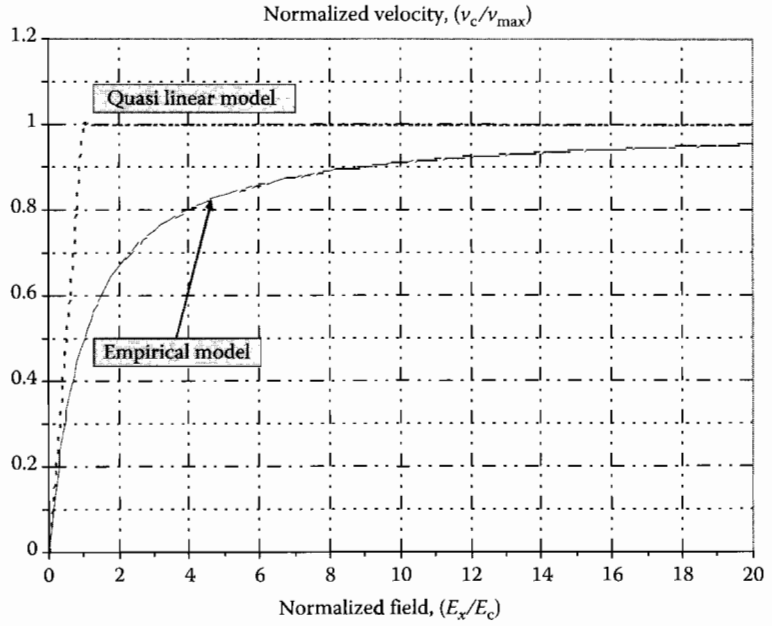

FIGURE 1.32 The dependence of carrier velocity on electric field in a semiconductor. The dashed curve represents the elementary low field approximation to the velocity-field relationship.

The mobility degradation inferred by the last disclosure bodes potentially decreased frequency response attributes of considered transistors, since the less mobile free electrons are in the inverted channel, the longer is the average time required for their transport from the source-to-the drain. Figure 1.32 sketches the velocity-field relationship implied by Equation 1.121. Note in this plot that the linear, or low field, approximation to the velocity characteristic is reasonably accurate up to about only 30% of the saturated limited velocity.

The effect on the ohmic regime drain current of the mobility degradation incurred by strong lateral electric fields can be studied by returning to Equation 1.104 and replacing the electron mobility, μ_n, therein by an adjusted mobility, μ_{ne}, such that

$$\mu_{ne} = \frac{\mu_n}{1 - E_x/E_c} = \frac{\mu_n}{1 + \frac{1}{E_c}\frac{d\varphi_c(x)}{dx}}, \tag{1.124}$$

where Equation 1.100 is applied and μ_n is understood to be the low field value of electron mobility in the inverted source-to-drain channel. Equation 1.104 becomes

$$I_d = \frac{\mu_n C_{ox} W\left[V_{gs} - V_h - \varphi_c(x)\right]}{1 + \frac{1}{E_c}\frac{d\varphi_c(x)}{dx}}\frac{d\varphi_c(x)}{dx}, \tag{1.125}$$

which leads to

$$I_d\left[\int_0^L dx + \frac{1}{E_c}\int_0^{V_{ds}} d\varphi_c(x)\right] = \mu_n C_{ox} W \int_0^{V_{ds}} \left[V_{gs} - V_h - \varphi_c(x)\right] d\varphi_c(x). \tag{1.126}$$

The requisite integrations produce

$$I_d = K_n\left(\frac{W}{L}\right)\left[\frac{V_{ds}\left(V_{gs} - V_h - \frac{V_{ds}}{2}\right)}{1 + \frac{V_{ds}}{V_{le}}}\right], \tag{1.127}$$

where

$$V_{le} = E_c L = \left(\frac{v_{max}}{\mu_n}\right)L \tag{1.128}$$

might be termed the *lateral electric field modulation voltage*. Observe that Equation 1.127 differs from the ohmic region volt-ampere relationship in Equation 1.118 by only the dimensionless factor in the denominator on the right-hand side of Equation 1.127. Appealing to Equation 1.128, this factor is seen to approach one when the channel length, L, is long. Of course, the subject factor also tends toward unity if the drain–source voltage, V_{ds}, is small. The latter point reflects engineering expectations in that small V_{ds} incurs lateral electric fields that are small enough to minimize field-induced mobility degradation.

A complication spawned by Equation 1.127 is that it no longer delivers the simple relationship for the drain saturation voltage witnessed in Equation 1.99. By definition, the drain saturation voltage, V_{dsat}, is the value of the drain–source voltage, V_{ds}, for which the slope of the ohmic regime I_d versus V_{ds} characteristic is zero. An application of this definition to Equation 1.127 leads to the revised drain saturation voltage,

$$V_{dsat} = M_{sat}\left(V_{gs} - V_h\right), \tag{1.129}$$

where, with

$$\alpha \triangleq \frac{V_{gs} - V_h}{V_{le}}, \tag{1.130}$$

$$M_{sat} = \frac{\sqrt{1 + 2\alpha} - 1}{\alpha}. \tag{1.131}$$

It can be demonstrated that $M_{sat} \leq 1$ for $\alpha \geq 0$ and thus, an impact of carrier mobility degradation incurred by strong lateral fields in the inverted channel is a decrease in the low field value of the drain saturation voltage. While mobility degradation is generally an undesirable phenomenon, the drain saturation voltage decrease is actually good news in low-voltage applications that require MOSFETs to function in their saturated regimes.

The drain saturation current corresponding to the revised estimate of the drain saturation voltage can be determined by substituting Equation 1.129 into 1.127. This activity produces the aesthetically pleasing result,

$$I_{dsat} = \frac{K_n}{2}\left(\frac{W}{L}\right)V_{dsat}^2 = \frac{K_n}{2}\left(\frac{W}{L}\right)M_{sat}^2\left(V_{gs} - V_h\right)^2. \tag{1.132}$$

In the limit of large channel lengths, V_{le} in Equation 1.128 is large, thereby rendering parameter α in Equation 1.130 small. But for very small α, M_{sat} in Equation 1.131 approaches unity. It is therefore reassuring that in the limit of large channel lengths, which are incapable of supporting large electric fields in the inverted channel, I_{dsat} in Equation 1.132 collapses to Equation 1.107, a relationship that implicitly

reflects tacit neglect of field-induced carrier mobility degradation. In contrast, very small channel lengths give rise to small V_{le} and large α, whence M_{sat} in Equation 1.131 reduces to

$$M_{sat}\big|_{\text{small L}} = \frac{\sqrt{1+2\alpha}-1}{\alpha}\bigg|_{\text{large }\alpha} \approx \sqrt{\frac{2}{\alpha}}. \tag{1.133}$$

Upon combining the last result with Equation 1.132, the short channel value of I_{dsat} is found to be

$$I_{dsat}\big|_{\text{small L}} \approx WC_{ox}v_{max}(V_{gs}-V_h), \tag{1.134}$$

where Equations 1.128 and 1.106 are exploited. Observe that the resultant drain saturation current is independent of the channel length, L. This independence stems from the fact that in the limit of very small channel lengths, carriers (electrons in the present case of an NMOS transistor) are transported through the inverted channel at their saturated limited, or maximum, velocity. This maximum velocity of carrier propagation renders L inconsequential with respect to the average time of carrier transport from the source region-to-the drain region. But perhaps the most interesting aspect of Equation 1.134 is that the short channel drain saturation current is a linear function of the gate–source voltage, V_{gs}. The linearity posed by Equation 1.134 is an obvious advantage for most analog signal processing applications, but achieving the velocity saturation implicit to this observed linearity may present voltage biasing challenges.

Of course, Equation 1.129 through 1.133 apply to the saturation regime of device operation in that in saturation, the drain current is merely the transistor current, I_{dsat}, evidenced at the boundary of ohmic and saturation regimes, corrected by channel length modulation effects. To wit, short channel phenomena imply that for $V_{gs} \geq V_h$ and $V_{ds} \geq V_{dsat}$,

$$I_d = \frac{K_n}{2}\left(\frac{W}{L}\right)M_{sat}^2(V_{gs}-V_h)^2\left(1+\frac{V_{ds}-V_{dsat}}{V_\lambda}\right), \tag{1.135}$$

where it is essential to remember that the drain saturation voltage, V_{dsat}, is now given by Equation 1.129. It is clear that M_{sat} in Equation 1.129 is properly viewed as a drain saturation voltage correction factor in a short channel (indeed, deep submicron) environment. Because of Equation 1.135, the square of M_{sat} can be accorded the stature of a current correction factor pertinent to short channel drain currents in the saturation regime. The dependence on parameter α of these correction factors is displayed in the plots submitted in Figure 1.33. The indicated correction factors are significant. For example, consider $\alpha = 2$, which might typically represent a gate–source voltage, V_{gs}, that is about a volt over the threshold potential. The curves in the figure at hand suggest an approximate 38% reduction in the drain saturation voltage predicted by the simple long channel model, which corresponds to $\alpha = 0$, as well as about a 62% attenuation of the corresponding drain saturation current.

Although Equations 1.135 and 1.129 are analytically elegant, their utility in a design-oriented environment is questionable in light of the dependence of factor M_{sat} on parameter α set forth by Equation 1.131. In light of this dilemma, an approximate curve fit of both M_{sat} and its square is judicious from an engineering design perspective. A numerical study of Equation 1.131 reveals that the empirical approximation,

$$M_{sat} = \frac{\sqrt{1+2\alpha}-1}{\alpha} \approx 1-\frac{\sqrt{\alpha}}{4}, \tag{1.136}$$

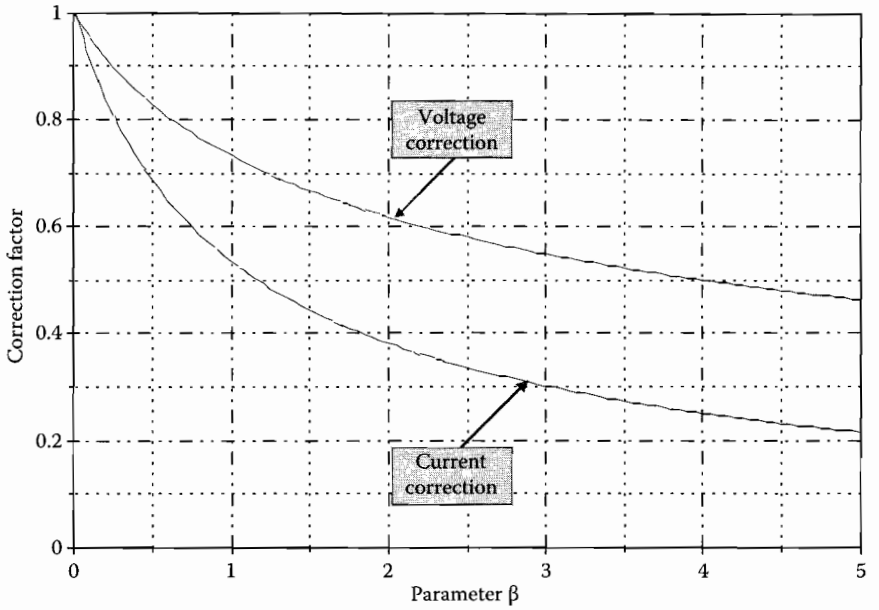

FIGURE 1.33 Voltage and current correction factors precipitated by large lateral electric fields in short channel MOSFETs. The parameter, α, is the effective gate–source voltage, $(V_{gs} - V_h)$, normalized to the lateral electric field modulation voltage, V_{le}.

results in an error of at most 4.8% for $0 \leq \alpha \leq 5$. On the other hand, a similar numerical exercise produces

$$M_{sat}^2 = \left(\frac{\sqrt{1+2\alpha}-1}{\alpha}\right)^2 \approx \frac{1}{1+0.78\alpha} \tag{1.137}$$

to a computational error of at most 5.1% for $0 \leq \alpha \leq 5$. For most design-oriented purposes, Equation 1.129 can therefore be supplanted by

$$\begin{aligned} V_{dsat} = M_{sat}(V_{gs} - V_h) &\approx \left(1 - \frac{\sqrt{\alpha}}{4}\right)(V_{gs} - V_h) \\ &= \left(1 - \frac{1}{4}\sqrt{\frac{V_{gs} - V_h}{V_{le}}}\right)(V_{gs} - V_h), \end{aligned} \tag{1.138}$$

while Equation 1.135 becomes for circuit design applications of MOSFETs operated in saturated regimes,

$$\begin{aligned} I_d &= \frac{K_n}{2}\left(\frac{W}{L}\right)M_{sat}^2(V_{gs} - V_h)^2\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right) \\ &\approx \frac{K_n}{2}\left(\frac{W}{L}\right)(V_{gs} - V_h)^2\left(\frac{1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}}{1 + 0.78\alpha}\right) \\ &= \frac{K_n}{2}\left(\frac{W}{L}\right)(V_{gs} - V_h)^2\left(\frac{1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}}{1 + 0.78\left(\frac{V_{gs} - V_h}{V_{le}}\right)}\right). \end{aligned} \tag{1.139}$$

The academic purist who may understandably balk at the foregoing numerical empiricisms is respectfully reminded that the mobility expression in Equation 1.123 and the "long channel" velocity relationship of Equation 1.120 are hardly grounded in sound physical phenomenology. Moreover, it is interesting to note that of the more than 275 parameters indigenous to the commonly exploited Level 49 HSPICE model of a MOSFET, most are curve fit disclosures that bear no clarion relationship to the physical charge storage and charge transport mechanisms that underpin the volt–ampere characteristics of a MOSFET.

1.2.3.4.2 Vertical Electric Fields

Apart from the carrier mobility degradation incurred by strong lateral fields in the inverted channel of a MOSFET, mobility is impacted by the vertical electric field resulting from the applied effective interface potential, $(V_{gs} - V_h)$, in the case of NMOS or $(V_{sg} - V_h)$ for PMOS devices. In NMOS, increases in V_{gs} strengthens this vertical field so that free electrons transported from the source-to-the drain are encouraged to drift ever closer to the oxide–semiconductor interface. Unfortunately, the interface is far from a perfectly smooth boundary, if for no other reason than routine device processing invariably produces ionic contamination therein. The imperfect boundary causes potentially significant carrier scattering, which in turn results in diminished carrier mobility.

To first order, the mobility attenuation resulting from increased gate overdrive can be addressed analytically by replacing the low field mobility, μ_n (for NMOS), to which K_n in Equation 1.139 is directly proportional, by an effective carrier mobility, μ_{eff}, such that

$$\mu_{eff} = \frac{\mu_n}{1 + \frac{V_{gs} - V_h}{V_{ve}}}. \tag{1.140}$$

In this expression, V_{ve} is the *vertical electric field modulation voltage*, which is nominally directly proportional to the thickness, T_{ox}, of the oxide layer. Of course, an expression analogous to Equation 1.140 prevails for hole mobility in the inverted channel of PMOS transistors. To a very rough approximation,

$$V_{ve} = T_{ox}/15, \tag{1.141}$$

where T_{ox} in units of angstroms returns V_{ve} in units of volts. Because of Equations 1.140 and 1.139 for the saturation domain current becomes

$$\begin{aligned} I_d &= \frac{K_n}{2}\left(\frac{W}{L}\right) M_{sat}^2 \frac{(V_{gs} - V_h)^2}{\left(1 + \frac{V_{gs} - V_h}{V_{ve}}\right)} \left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right) \\ &\approx \frac{K_n}{2}\left(\frac{W}{L}\right) \frac{(V_{gs} - V_h)^2}{\left(1 + \frac{V_{gs} - V_h}{V_{ve}}\right)} \left(\frac{1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}}{1 + 0.78\left(\frac{V_{gs} - V_h}{V_{le}}\right)}\right). \end{aligned} \tag{1.142}$$

An analogous modification, which amounts to an effective reduction of the transconductance parameter, K_n, can be made to the ohmic domain current.

Obviously, Equation 1.142 is inordinately more cumbersome than is the simple, square law, volt–ampere characteristic advanced by Equation 1.110 for device operation in the saturation domain. As a result, the design-oriented determination of a suitable gate–source voltage for a desired drain current and corresponding drain–source voltage can be a daunting challenge. But in addition to the computational problems precipitated merely by algebraic complexity, engineering difficulties are additionally encountered with respect to the accurate numerical delineation of the model metrics, K_n, V_h, V_{ve}, V_λ, and V_{le}. These

latter difficulties derive from the unfortunate fact that the physical device and charge transport properties (saturation velocity, carrier mobility, regional concentrations, etc.) on which these and other model parameters depend are invariably unavailable to the circuit designer. At best, the circuit designer can reasonably expect to have presumably reliable, detailed device model parameters suitable for computer-aided simulation of transistor performance. For example, process foundries routinely supply their customers with device models in the form of Level 49 HSPICE or other computer-based files. Unfortunately, many, if not most, of the hundreds of numerical entries indigenous to these files are themselves nonphysical entities that defy satisfying mathematical relationships to the physical model metrics discussed in earlier paragraphs. These and related other design-oriented problems can prove exasperating. The aforementioned issues are best mitigated by coalescing manual design strategies and calculations with suitable computer-based simulations of device properties and volt–ampere characteristics.

1.2.3.5 Temperature Effects

The operating temperature of the inverted interfacial channel affects the drain current of a transistor in three ways. First, because thermal energy imparted to free carriers increases their scattering, the carrier mobility decreases in response to increased operating temperatures. To first order, the electron mobility, $\mu_n(T)$, at absolute temperature T relates to the mobility, $\mu_n(T_o)$, at a reference temperature, T_o, in accordance with the three-halves power law,

$$\mu_n(T) = \mu_n(T_o)\left(\frac{T_o}{T}\right)^{3/2}. \tag{1.143}$$

Because parameter K_n in Equation 1.142 is directly proportional to carrier mobility, Equation 1.143 implies that the drain current of a MOSFET is characterized by a negative temperature coefficient, that is, the drain current, I_d, decreases with increasing operating temperature.

A second effect of increased thermal energy is a perturbation of threshold voltage. A computation of this perturbation is best initiated by returning to Equation 1.96 to evaluate the derivative of the threshold voltage, V_h, with respect to the Fermi potential, V_F. Recalling Equations 1.96 and 1.97, and noting that the body effect voltage, V_θ, in Equation 1.90 is independent of temperature,

$$\frac{dV_h}{dV_F} = 2 + \sqrt{\frac{V_\theta}{V_F}} + \frac{V_h - V_{ho}}{2V_F} + \left(\frac{V_\theta}{2V_F}\right)\left(\frac{V_{bs}}{V_h - V_{ho} + 2\sqrt{V_\theta V_F}}\right), \tag{1.144}$$

where V_{ho} is recalled as signifying the zero bias ($V_{bs} = 0$) value of the threshold potential. Note that the last two terms on the right-hand side of this expression vanish when a MOSFET is operated with $V_{bs} = 0$. The sensitivity of the threshold voltage with respect to temperature follows as

$$\frac{dV_h}{dT} = \frac{dV_h}{dV_F} \times \frac{dV_F}{dT}. \tag{1.145}$$

The temperature derivative of the Fermi potential derives from Equation 1.55, with the proviso that due account be made of the temperature dependence of the intrinsic carrier concentration, N_i. To this end, a commonly used empiricism is

$$N_i = N_{io} 2^{(T - T_o)/T_n}, \tag{1.146}$$

where T_n is generally taken to be 10°C and, assuming the reference temperature, T_o, is 27°C, N_{io}, the intrinsic carrier concentration at $T = T_o$, is the previously used number, $(1.45)(10^{10})$ atoms/°C. With

$T_n = 10°C$, Equation 1.146 allows N_i to double for each 10°C rise above the reference temperature. Armed with Equations 1.146 and 1.55 produces

$$\frac{dV_F}{dT} = \frac{V_F}{T} - \frac{V_T}{T_n} \ln 2. \tag{1.147}$$

Equations 1.144 and 1.147 combine to yield the final result,

$$\frac{dV_h}{dT} = \left[2 + \sqrt{\frac{V_\theta}{V_F}} + \frac{V_h - V_{ho}}{2V_F} + \left(\frac{V_\theta}{2V_F}\right)\left(\frac{V_{bs}}{V_h - V_{ho} + 2\sqrt{V_\theta V_F}}\right)\right]\left(\frac{V_F}{T} - \frac{V_T}{T_n} \ln 2\right), \tag{1.148}$$

where parameters V_{ho}, V_h, and V_T, are computed at the reference temperature, T_o. The indicated temperature derivative of the threshold voltage is invariably a positive number in the range of 1.5–2.4 mV/°C. Thus, the threshold voltage increases with increasing operating temperatures, thereby leading to a decrease in the drain current. In other words, the temperature dependence of both the carrier mobility and the threshold voltage conduce a drain current exuding a negative temperature coefficient.

The algebraic form of Equation 1.148 is thoroughly depressing and is hardly a relationship stored in the human memories of circuit designers. Fortunately, for MOSFETs featuring thin gate oxides (under 50 Å) and substrate doping concentrations no smaller than 10^{15} atoms/cm^3, the terms in V_θ, $(V_h - V_{ho})$, and V_T/T_n are generally negligible, especially if the bulk–source bias, V_{bs}, is no more negative than 1.5 V. In this event,

$$\frac{dV_h}{dT} \approx \frac{2V_F}{T}, \tag{1.149}$$

which can be shown to be always larger—generally by no more than 5% or 6%—than the result predicted by Equation 1.148. For a substrate doping concentration of $N_A = (5)(10^{15})$ atoms/cm^3, Equation 1.149 predicts a threshold voltage sensitivity at $T = 27°C = 300.16$ K of 2.2 mV/°C.

1.2.4 Transistor Capacitances

At this juncture, the volt-ampere characteristic equations given by Equations 1.118, 1.119, 1.139, and 1.142 pertain to MOSFETs operated exclusively under static or low-frequency signal conditions. Specifically, the drain currents predicted by these relationships are unrealistically cavalier in that they respond instantaneously to applied gate–source, drain–source, and bulk–source excitations. When high-frequency signals are applied, the current responses are slowed by device capacitances arising from the charge storage that prevails in the inverted channel and within the depletion regions formed about the source and drain diffusions or implants. The engineering implications of this inherent inability of drain currents to respond instantly to signal excitations are MOSFET circuits exuding constrained bandwidths, nonzero input/output (I/O) delays and phase shifts, and nonzero rise and fall times in transient responses. In extreme cases, the interaction of these device capacitances with the energy storage elements of the peripheral circuit can produce excessive response peaking in either the frequency or time domains and even outright instability.

1.2.4.1 Depletion Capacitances

The first of the two principle sources of transistor capacitances is the depletion capacitance indigenous to both of the PN junctions formed respectively between the bulk and drain and between the bulk and source. In turn, each of these two transition region capacitances consists of a planar component and a peripheral, or sidewall, component. The planar component embodies the depletion layer established

between the bulk substrate region and the underside of the source and drain regions. On the other hand, the sidewall capacitance embraces the depletion layers in the areas of the source and drain regions that are proximate to the front surface, the back surface, and the side surface area adjacent to the active channel region. For the bulk–drain depletion capacitance, C_{bd},

$$C_{bd} = \frac{A_d C_j}{\left(1 - \frac{V_{bd}}{V_j}\right)^{M_j}} + \frac{P_d C_{jsw}}{\left(1 - \frac{V_{bd}}{V_j}\right)^{M_{jsw}}}, \tag{1.150}$$

where the forms of each of the terms on the right-hand side are observed to mirror the traditional depletion capacitance associated with a back biased PN junction. In Equation 1.150, C_j is the zero bias (meaning, $V_{bd} = 0$), value of the capacitance density, in units of farads/meter^2, associated with the planar component of the bulk–drain capacitance, while C_{jsw} is the zero bias lineal capacitance, in units of farads/meter, of the aforementioned sidewall areas. The planar drain area, A_d, is

$$A_d = WL_{dif}, \tag{1.151}$$

where L_{dif} is recalled in Figures 1.17 and 1.18 to represent the width of the drain region, which is generally identical to the width of the source implant. Generally, the dimension, L_{dif}, must be extracted empirically from measured data but as a rule of thumb, L_{dif} is nominally of the order of twice the channel length, L. Parameter P_d, is the effective length of the perimeter of the sidewall area and is stipulated by

$$P_d = W + 2L_{dif}. \tag{1.152}$$

Voltage V_j in Equation 1.150 is the built-in potential given by Equation 1.115, while M_j and M_{jsw} are the grading coefficients of the planar and sidewall PN junctions, respectively.* Typically $M_j = 0.5$ and $M_{jsw} = 0.33$. An analogous expression, whose terms convey equally analogous engineering interpretations, prevails for the net bulk–source depletion capacitance, C_{bs}. In particular,

$$C_{bs} = \frac{A_s C_j}{\left(1 - \frac{V_{bs}}{V_j}\right)^{M_j}} + \frac{P_s C_{jsw}}{\left(1 - \frac{V_{bs}}{V_j}\right)^{M_{jsw}}}, \tag{1.153}$$

where in general,

$$\left.\begin{aligned} A_s \equiv A_d &= WL_{dif} \approx 2WL \\ P_s \equiv P_d &= W + 2L_{dif} \approx W + 4L \end{aligned}\right\}. \tag{1.154}$$

1.2.4.2 Gate Capacitances

The second source of MOSFET capacitances is the gate capacitance, which itself is comprised of three distinct components. The first of these components appears between the gate and the bulk substrate. As is apparent from Figure 1.23, this particular capacitance has a very small nonzero frequency value in both weak and strong channel inversion modes, which suggests that the channel inversion layer effectively shields the gate from the bulk substrate. Because the gate–bulk capacitance is invariably very small, it bodes little consequence to achievable MOSFET circuit performance and therefore, it is usually ignored tacitly.

* In HSPICE and other forms of SPICE simulators, the built-in potential, V_j, is symbolized by P_B.

The other two components of net gate capacitance are the gate–source capacitance, C_{gs}, and the gate–drain capacitance, C_{gd}. Each of these energy storage elements is a superposition of an intrinsic module, which derives from the gate, gate oxide, and inverted channel, and an extrinsic constituent, which is attributed to gate oxide overlap at the source and drain sites. Since the inversion layer extends from source-to-drain in only the ohmic regime of operation, different values of these two capacitances prevail for ohmic and saturated operation. The maximum possible intrinsic capacitance established between the gate and the inversion layer is clearly WLC_{ox}. In the ohmic operating regime, this maximum capacitance is partitioned equally between the source and the drain to give identical intrinsic gate–source and gate–drain capacitance values; namely, $WLC_{ox}/2$. Accordingly, in the ohmic regime, the effective gate–source capacitance is

$$C_{gs} = \frac{WLC_{ox}}{2} + WC_{gso}, \tag{1.155}$$

where C_{gso} is the capacitance per unit length associated with the oxide–source overlap. Similarly, the effective gate–drain capacitance in the ohmic operating regime is

$$C_{gd} = \frac{WLC_{ox}}{2} + WC_{gdo}, \tag{1.156}$$

where C_{gdo} is the drain overlap capacitance counterpart to the source overlap region. Typically, C_{gso} and C_{gdo} are as small as 0.25 fF/μm in minimal geometry transistors. Thus, for a transistor characterized by $L = 180$ nm, $W/L = 20$, and an oxide thickness of $T_{ox} = 30$ Å, $C_{gs} = C_{gd} = 4.63$ fF. Observe herewith that the net overlap capacitance is $WC_{gso} = WC_{gdo} = 0.9$ fF, which is almost 20% of the total gate–source (or gate–drain) capacitance.

The capacitance situation in saturation is a bit more intricate than that which prevails in the ohmic regime. In saturated domains where $V_{ds} > V_{dsat}$, pinch off occurs within the source-to-drain channel, thereby leaving an effective depletion zone that is free of mobile carriers near the drain site. Accordingly, the drain–source voltage exerts no influence on the channel charge, and the resultant gate–drain capacitance derives exclusively from the oxide overlap with the drain, that is, the gate–drain capacitance, C_{gd}, in saturation is simply

$$C_{gd} = WC_{gdo}. \tag{1.157}$$

In contrast to the charge depletion prevailing in the channel region adjacent to the drain, a large free carrier population is concentrated near the source. Since this concentration is influenced strongly by interface potential, which is determined by the applied gate–source voltage, it is only logical to expect a comparatively substantial intrinsic gate–source capacitance.

An analytical disclosure of the foregoing gate-to-source capacitance commences with a return to Equations 1.102 and 1.103. If these two equations are combined and if Equation 1.63 is recalled,

$$q[Y_s(\varphi_c(x))][N_s(\varphi_c(x))] = C_{ox}\left[V_{gs} - V_h - \varphi_c(x)\right], \tag{1.158}$$

where the left-hand side of this relationship is understood to be the density of mobile charge in the inversion layer. Upon multiplication of both sides of Equation 1.158 by the gate width, W, the resultant left-hand side of the modified expression represents the net mobile charge per unit length of the inversion layer. It follows that the net differential mobile charge (amassed by electrons in NMOS), say $dq_n[\varphi_c(x)]$, contained in a differential channel volume of depth $Y_s[\varphi_c(x)]$, width W, and length extending from x to $(x + dx)$, is

$$dq_n(\varphi_c(x)) = WC_{ox}\left[V_{gs} - V_h - \varphi_c(x)\right]dx. \tag{1.159}$$

Ignoring mobility degradation incurred by lateral electric fields, Equation 1.104 can be used to recast Equation 1.58 in the form

$$\mathrm{d}q_n(\varphi_c(x)) = \frac{\mu_n(WC_{ox})^2}{I_d}\left[V_{gs} - V_h - \varphi_c(x)\right]^2 \mathrm{d}\varphi_c(x). \tag{1.160}$$

Equation 1.59 can be integrated conveniently from $\varphi_c(0)$ to $\varphi_c(V_{dsat})$, where the indicated interfacial potential limits correspond to the boundaries of the channel inversion layer evidenced in saturation. Such an integration of the left-hand side of Equation 1.59 brackets the net mobile charge, say $Q_n(V_{gs})$, observed in saturation for a stipulated gate–source voltage, V_{gs}. In particular,

$$\begin{aligned} Q_n(V_{gs}) &= \int_0^{V_{dsat}} \mathrm{d}q_n(\varphi_c(x)) = \frac{\mu_n(WC_{ox})^2}{I_d} \int_o^{V_{dsat}} \left[V_{gs} - V_h - \varphi_c(x)\right]^2 \mathrm{d}\varphi_c(x) \\ &= \frac{\mu_n(WC_{ox})^2}{3I_d}\left(V_{gs} - V_h\right)^3, \end{aligned} \tag{1.161}$$

where Equation 1.99 is exploited. Using Equation 1.109 to replace the drain current variable, I_d, in this relationship results in

$$Q_n(V_{gs}) = \frac{2}{3} WLC_{ox}(V_{gs} - V_h). \tag{1.162}$$

The saturation region value of the intrinsic gate–source capacitance now follows as

$$\frac{\mathrm{d}Q_n(V_{gs})}{\mathrm{d}V_{gs}} = \frac{2}{3} WLC_{ox}, \tag{1.163}$$

whereupon the saturation region value of the net gate–source capacitance, C_{gs}, inclusive of oxide overlap effects at the source site, is

$$C_{gs} = \frac{2}{3} WLC_{ox} + WC_{gso}. \tag{1.164}$$

1.2.4.3 Large-Signal Model

At this juncture, the large signal, or nonlinear, model of an N-channel MOSFET is the structure advanced in Figure 1.34. Depending on whether the transistor undergoing assessment is operated as an ohmic regime or as a saturated device, the equation for the indicated controlled current source, I_d, derives from expressions formulated in Section 1.2.3.2 or Section 1.2.3.3 or, for that matter, the model refinements addressed in Section 1.2.3.4. The depletion capacitances, C_{bs} and C_{bd}, are not affected by the domain of transistor operation, but the appropriate regional values of the capacitances, C_{gs} and C_{gd}, must be culled from the discourse in Section 1.2.4.2.

The model at hand also incorporates four resistive elements. The resistances, r_d and r_s, are respectively associated with the strongly doped drain and source regions, respectively. These resistances are specified in HSPICE simulation software by a sheet resistance parameter, R_{sh}, and drain and source geometric parameters, N_{rd} and N_{rs}. In particular,

$$\left.\begin{aligned} r_d &= N_{rd}R_{sh} \\ r_s &= N_{rs}R_{sh} \end{aligned}\right\}. \tag{1.165}$$

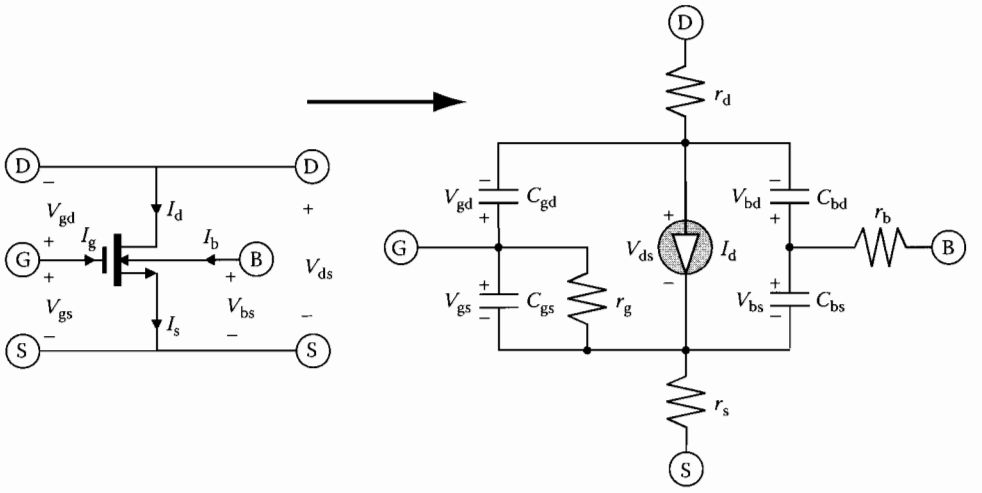

FIGURE 1.34 Large-signal model of an N-channel MOSFET. A topologically identical equivalent circuit prevails for P-channel MOSFETs.

Owing to the high doping concentrations of the drain and the source, which begets a small *sheet resistance* parameter, R_{sh}, resistances r_d and r_s are generally sufficiently small to justify their tacit neglect in most analog circuit applications. In contrast, resistance, r_b, which represents an effective spreading resistance in the bulk substrate, can be as large as the high tens to low hundreds of ohms. Despite its relatively large value, its impact on analog circuit performance is muted by the fact that the bulk rarely conducts significant currents, even at high signal frequencies. However, this resistance does influence the thermal noise characteristics of the drain–source channel.

Like resistance r_b, the gate resistance, r_g, is likewise important from a thermal noise perspective in that it captures the salient effects that thermally agitated mobile charge carriers exert on channel potential. It also looms significant with respect to design problems associated with maximum signal power transfer in radio frequency (RF) circuits [9]. This resistance is computed as [10]

$$r_g = \frac{5}{(\omega C_{gs})^2 R_{ch}}, \tag{1.166}$$

where R_{ch} represents the $V_{ds} = 0$ value of the drain–source channel resistance. Recalling Equation 1.109,

$$R_{ch} = \frac{1}{K_n\left(\frac{W}{L}\right)(V_{gs} - V_h)} = \frac{M_{sat} V_{dsat}}{2I_{dsat}}, \tag{1.167}$$

where Equations 1.129 and 1.132 have been used. Because of the inverse dependence of r_g on the square of radial signal frequency, ω, r_g is infinity for quiescent operating conditions and extremely large for low to even reasonably high frequencies.

1.2.5 Small-Signal Operation

As noted in Section 1.2.1, MOSFETs are the active device of choice in a plethora of high-performance analog integrated circuits. When the fundamental objective of these analog networks is linear I/O signal processing, each MOSFET therein is commonly biased in a saturated regime that ensures, for all applied signals of interest, an instantaneous drain–source voltage, v_{ds}, that is never any smaller than the

instantaneous drain–source saturation voltage, v_{dsat}. To be sure, linear signal processing can also be achieved when transistors operate in their ohmic regimes. But when high performance, in such senses as high gain, wide bandwidth, large dynamic range, and acceptable driving point I/O impedance levels, is a fundamental design objective, saturation is the regime of choice. Accordingly, ohmic linear equivalent circuits of transistors are ignored herewith and left as an investigation exercise for the reader.

A casual inspection of Equation 1.142 suggests that the instantaneous drain current, i_d, flowing in an N-channel MOSFET is a function of three device voltages: the instantaneous gate–source voltage, v_{gs}, the instantaneous drain–source voltage, v_{ds}, and the instantaneous bulk–source voltage, v_{bs} which covertly influences the threshold potential, V_h. An analogous statement applies to P-channel transistors, subject to the current and voltage conventions adopted earlier. Thus, Equation 1.142 can be generalized as

$$i_d \approx \frac{K_n}{2}\left(\frac{W}{L}\right)M_{sat}^2\left(v_{gs}-V_h\right)^2\left(\frac{1+\frac{v_{ds}-v_{dsat}}{V_A}}{1+\frac{v_{gs}-V_h}{V_{ve}}}\right)=f\left(v_{gs},v_{ds},v_{bs}\right). \tag{1.168}$$

Under zero signal conditions, which is tantamount to operating the considered MOSFET at its quiescent operating point, it is understood that Equation 1.168 yields

$$I_d=f\left(V_{gs},V_{ds},V_{bs}\right), \tag{1.169}$$

where the indicated variables in capital letters designate static, or quiescent, device currents and voltages. In other words, the MOSFET described mathematically by Equation 1.169 is in a standby mode that awaits the application of dynamic, invariably time-varying, signals. Prior to signal excitation, the transistor maintains quiescent values of drain current, gate–source voltage, drain–source voltage, and bulk–source voltage that respectively equal I_d, V_{gs}, V_{ds}, and V_{bs}. Signals applied as a current, say I_{ds}, to the drain lead and/or a voltages, say V_1 to the gate–source port, V_2 to the bulk–source port, or V_3 to the drain–source port perturb the quiescent, or *Q-point*, counterparts of these electrical variables to deliver the observable net instantaneous current and voltage responses,

$$\left.\begin{aligned} i_d &= I_d + I_{ds} \\ v_{gs} &= V_{gs} + V_1 \\ v_{bs} &= V_{bs} + V_2 \\ v_{ds} &= V_{ds} + V_3 \end{aligned}\right\}. \tag{1.170}$$

In concert with these relationships, the MOSFET under consideration is said to operate linearly if and only if the signal-induced changes, I_{ds}, V_1, V_2, and V_3, interrelate linearly and if and only if the Q-point currents and voltages are independent of signal strengths. It is crucial to understand that operational linearity in an electronic device does not imply linear relationships among the instantaneous device variables, nor does it imply linearity among the corresponding quiescent values of these variables. Instead, operational linearity implies merely that a selected variable in the selected set of four perturbed variables in Equation 1.170 linearly superimpose with the remaining three electrical signal components.

1.2.5.1 Fundamental Small-Signal Model

Because of the obviously nonlinear nature of Equation 1.168, questions abound as to the plausibility of achieving the aforementioned linearity condition among the electrical perturbations induced by applied signals. Despite its inherently nonlinear nature, Equation 1.168 is a well-behaved functional relationship, which suggests that the desired linearity might be approximated adequately by limiting all signal excursions about respective operating point values to sufficiently small levels. This sufficiently small-signal mandate defines the concept of *small-signal analysis* and produces a *small-signal model* of the

MOSFET. A small-signal analysis reflective of a mathematical exploitation of the corresponding small-signal model are deemed both appropriate and useful if the retention of only the linear terms of the Taylor series expansion of Equation 1.168 about the operating point of the considered device leads to minimal errors in the resultant expression for the signal component of the drain current. Thus,

$$i_d \approx I_d + \left.\frac{\partial i_d}{\partial v_{gs}}\right|_Q (v_{gs} - V_{gs}) + \left.\frac{\partial i_d}{\partial v_{bs}}\right|_Q (v_{bs} - V_{bs}) + \left.\frac{\partial i_d}{\partial v_{ds}}\right|_Q (v_{ds} - V_{ds}), \tag{1.171}$$

where each of the three derivatives on the right-hand side of this relationship are evaluated at the Q-point of the MOSFET, that is, at $i_d = I_d$, $v_{gs} = V_{gs}$, $v_{bs} = V_{bs}$, and $v_{ds} = V_{ds}$. Using Equation 1.170 and noting that each of the three subject derivatives is a constant having units of conductance, Equation 1.171 can be couched in the form

$$I_{ds} \approx g_m V_1 + g_{mb} V_2 + \frac{V_3}{r_o}, \tag{1.172}$$

where

$$\left.\begin{aligned} g_m &\triangleq \left.\frac{\partial i_d}{\partial v_{gs}}\right|_Q \\ g_{mb} &\triangleq \left.\frac{\partial i_d}{\partial v_{bs}}\right|_Q \\ \frac{1}{r_o} &\triangleq \left.\frac{\partial i_d}{\partial v_{ds}}\right|_Q \end{aligned}\right\}. \tag{1.173}$$

Equation 1.71 gives rise to the small-signal, low-frequency equivalent circuit depicted in Figure 1.35a. The subject circuit becomes the small-signal, high-frequency MOSFET model if the four capacitances, C_{gs}, C_{gd}, C_{bd}, and C_{bs}, discussed in Section 1.2.4 are appended as indicated in Figure 1.35b. It is important to underscore the fact that either of the models in Figure 1.35 gives no information about the instantaneous electrical variables of a MOSFET, nor does either model allow for the computation of the quiescent values of these variables. Indeed, the models at hand require a priori knowledge of the Q-point since the small-signal parameters, g_m, g_{mb}, and r_o, depend on the operating point, as is implied by Equation 1.173. Moreover, the four capacitive elements in the model of Figure 1.35b likewise depend on the Q-point at which the considered transistor is biased. In short, the models in Figure 1.35 give first order approximations of the interrelationships among only the small-signal components of the net currents and voltages indigenous to a MOSFET. Although the topologies of both the models drawn in Figure 1.35 pertain to both the ohmic and saturation regimes of N-channel MOSFET operation, the equations to be developed shortly for the low-frequency parameters of these models apply exclusively to the saturation region. Moreover, while Figure 1.35 makes explicit reference to an N-channel transistor, or NMOS, the small-signal, low- and high-frequency equivalent circuits of PMOS units are identical to their NMOS counterparts. This topological identity stems from the fundamental fact that the small-signal models intertwine only signal-induced changes of device currents and voltages about their respective quiescent values. In an attempt to dispel possible confusion, the latter models are offered in Figure 1.36.

In the low-frequency models of either Figure 1.35a or Figure 1.36a the signal component of the bulk current, I_{bs}, flows into an open circuit because the bulk–drain and bulk–source junctions of devices embedded in analog networks are commonly reversed biased. The low-frequency signal component of the gate current, I_{gs}, is very nearly zero because the gate resistance, r_g, is, by Equation 1.166, inversely proportional to the square of the signal frequency. Of course, both of these currents are substantively

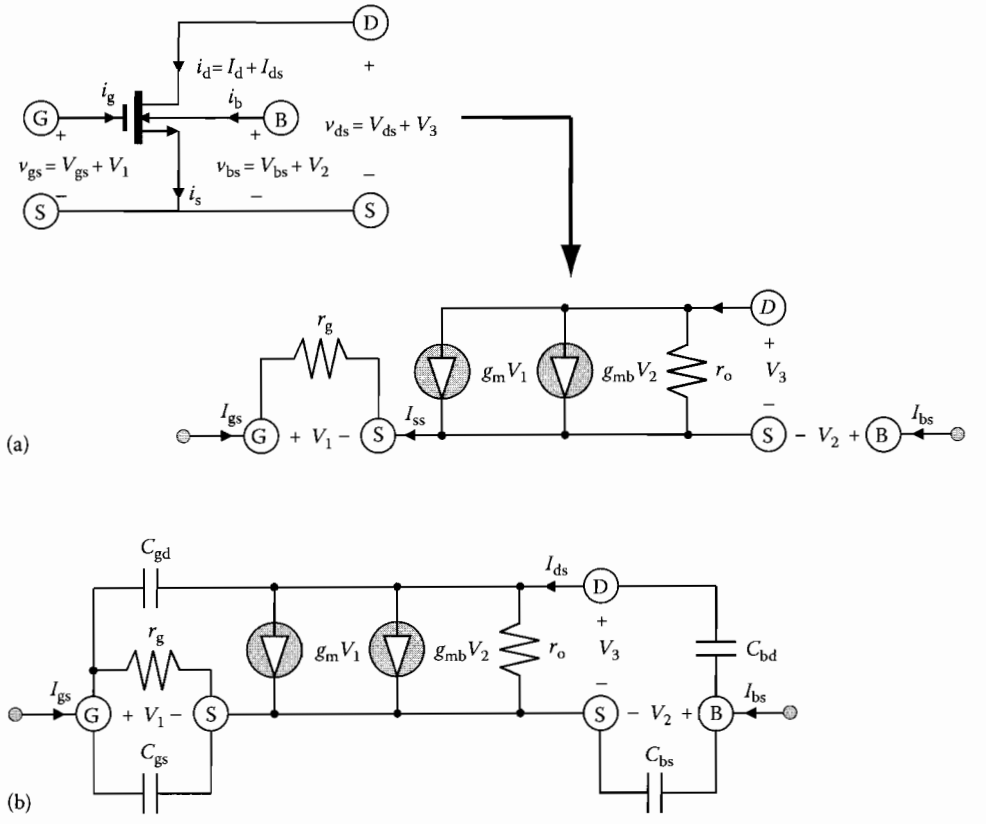

FIGURE 1.35 (a) Small-signal, low-frequency equivalent circuit of an N-channel MOSFET. (b) Small-signal, high-frequency equivalent circuit of an N-channel MOSFET. The topological structures of either model apply to both the ohmic and saturation regimes of operation.

larger at high frequencies where the various capacitances in the models of Figures 1.35b and 1.36b become poor approximations of the open circuits they mirror at low signal frequencies.

The parameter, g_m, is termed the *forward transconductance*. It is a critical analog circuit metric in that it serves as a measure of achievable forward gain. In particular, parameter g_m, when multiplied by the applied gate–source signal voltage, V_1, determines the amount of drain signal current, I_{ds}, manifested by the applied gate–source signal. On the other hand, the *bulk transconductance*, g_{mb}, measures the ability of a MOSFET to transfer applied bulk–source signal, V_2, to the drain signal current response. The controlled current, $g_{mb}V_2$, is negligible when it is much smaller than is its forward transconductance counterpart current, $g_m V_1$. However, it should be noted that depending on the selected quiescent operating point, parameter g_{mb} can be as much as 15%–25% of the forward transconductance, g_m. The current, $g_{mb}V_2$, is entirely inconsequential in numerous analog circuits that configure their utilized MOSFETs in such a way as to operate both the bulk and source terminals at signal ground, which renders $V_2 = 0$. Finally, r_o, the *drain–source channel resistance*, appears as a shunting resistive element across the drain and source terminals. If r_o is infinitely large (which, to be sure, it is not in practical MOSFETs) the drain-to-source small-signal port of a MOSFET behaves as an ideal Norton equivalent current source, that is, the current level determined largely by $g_m V_1$ is unaffected by modulations in the drain–source signal voltage, V_3. It follows that to the extent that the gate–source terminals serve as an input signal port boasting infinitely large impedance and the drain–source terminals function as the output port, the MOSFET emulates an ideal transconductance amplifier if the channel resistance, r_o, is large.

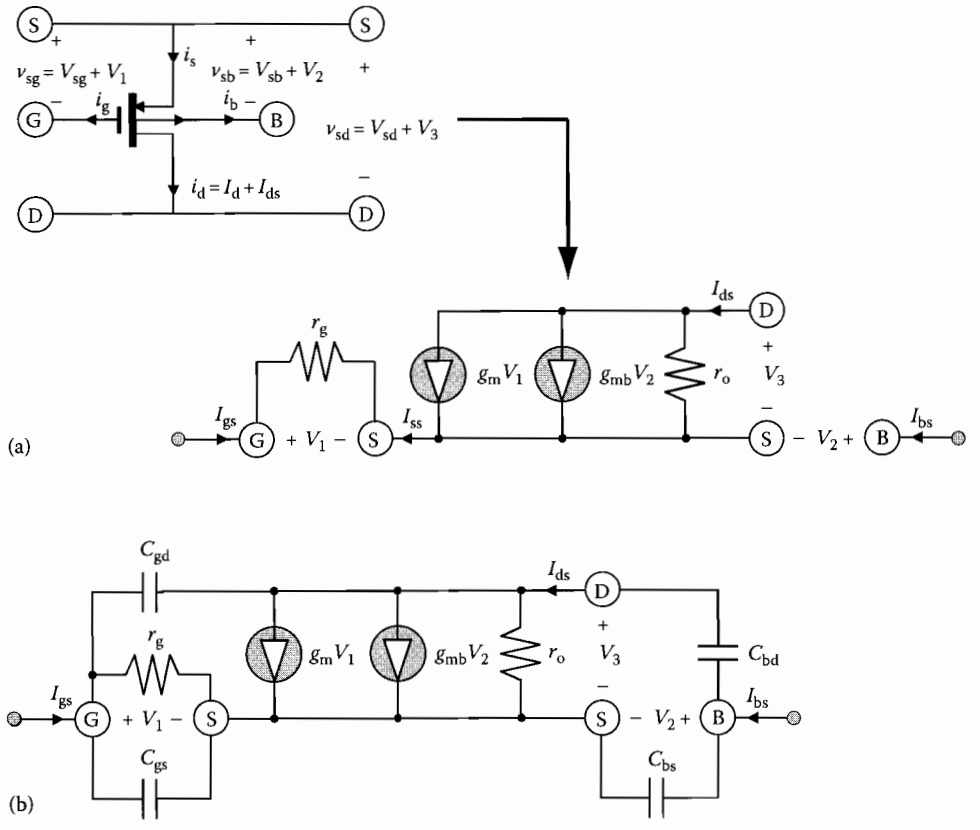

FIGURE 1.36 (a) Small-signal, low-frequency equivalent circuit of a P-channel MOSFET. (b) Small-signal, high-frequency equivalent circuit of a PMOS device.

The determination of the three low-frequency parameters defined in Equation 1.173 requires that the indicated derivatives of the drain current expression in Equation 1.168 be evaluated. This evaluation is an algebraically trying task that borders on a futile engineering enterprise in that many of the physical parameters implicit to Equation 1.168 are rarely disclosed to the circuit designer. It is therefore prudent to condescend to first order approximations of the subject small-signal parameters by replacing Equation 1.168 with the simpler expression,

$$i_d \approx \frac{K_n}{2}\left(\frac{W}{L}\right)(v_{gs} - V_h)^2\left(1 + \frac{v_{ds} - v_{dsat}}{V_\lambda}\right), \tag{1.174}$$

which effectively ignores the influence of both lateral and vertical electric fields in the MOSFET channel. By ignoring the effects of lateral fields, parameter M_{sat} in Equation 1.131 is one, whence the drain saturation voltage in Equation 1.129 is simply the voltage difference, $v_{dsat} = (v_{gs} - V_h)$. Accordingly, Equations 1.173 and 1.174 yield a forward transconductance of

$$g_m \triangleq \left.\frac{\partial i_d}{\partial v_{gs}}\right|_Q \approx \frac{2I_d}{V_{gs} - V_h} - \frac{I_d}{V_\lambda + V_{ds} - V_E}, \tag{1.175}$$

where it is understood that the variables, I_d, V_{gs}, V_h, V_λ, and V_{dsat} reflect the Q-point of the transistor undergoing study. Biasing voltage and standby power constraints ordinarily compel that the transistor be

biased at a drain–source voltage that is only slightly above the drain saturation voltage. Accordingly, V_λ is typically much larger than $(V_{ds} - V_{dsat})$. Moreover, V_λ is generally significantly larger than $V_{dsat}/2$. It follows that the second term on the right-hand side of Equation 1.175 is often negligible, whereupon Equations 1.168 and 1.175 combine for the case of large V_λ to deliver

$$g_m \triangleq \left.\frac{\partial i_d}{\partial v_{gs}}\right|_Q \approx \frac{2I_d}{V_{gs} - V_h} \approx \sqrt{2K_n(W/L)I_d}. \tag{1.176}$$

The result suggests that the forward transconductance of a MOSFET increases with the square root of the product of quiescent drain current and transistor gate aspect ratio. Accordingly, high gain requirements in certain MOSFET amplifiers compel relatively large standby drain currents and/or suitably large gate widths. The former tack conflicts with omnipresent desires for low power operation, while the latter begets increased device capacitances and hence, potentially degraded frequency responses. Observe that while the term in V_λ in Equation 1.175 is usually negligibly small, significant channel length modulation (which translates to small V_λ) is deleterious to high gain objectives.

An evaluation of the bulk transconductance, g_{mb}, requires that Equation 1.96 be considered analytically in conjunction with the threshold voltage term in Equation 1.174. After a bit of messy algebra, it can be shown that

$$g_{mb} \triangleq \left.\frac{\partial i_d}{\partial v_{bs}}\right|_Q \approx \lambda_b g_m, \tag{1.177}$$

where λ_b, which might be termed a *bulk modulation factor*, is

$$\lambda_b = \sqrt{\frac{V_\theta/2}{2V_F - V_{bs}}}. \tag{1.178}$$

Recall a previous assertion to the extent that the bulk transconductance, g_{mb}, may be insignificant in comparison to the small-signal impact of the forward transconductance, g_m. From Equations 1.177 and 1.178, λ_b, and hence g_{mb}, are small if V_θ, the body effect potential defined by Equation 1.90, is small. Since V_θ is proportional to the square of the gate oxide thickness, thin oxide layers conduce small bulk transconductances. It is interesting to note in Equation 1.178 that the small values of the bulk modulation factor that are precipitated by thin oxides are made even smaller by increases in the reverse bias applied between the bulk and source.

The drain–source channel resistance, r_o, in Equation 1.173 is readily confirmed to derive from

$$\frac{1}{r_o} \triangleq \left.\frac{\partial i_d}{\partial v_{ds}}\right|_Q \approx \frac{I_d}{V_\lambda + V_{ds} - V_{dsat}}. \tag{1.179}$$

The result shows that a large channel length modulation voltage, V_λ, gives rise to a large channel resistance, r_o, which in turn implies that the drain–source port of a MOSFET emulates the volt–ampere characteristics of an ideal current source. For conventional values of V_λ, large r_o is seen to require a small drain bias current, I_d.

1.2.5.2 Unity Gain Frequency

The models of Figures 1.35 and 1.36 provide an analytical path for computing a commonly invoked figure of merit for MOSFETs; namely, the *unity gain frequency*, which in radial units is symbolized as ω_T. Although this metric offers a meaningful basis for comparing the high-frequency signal processing

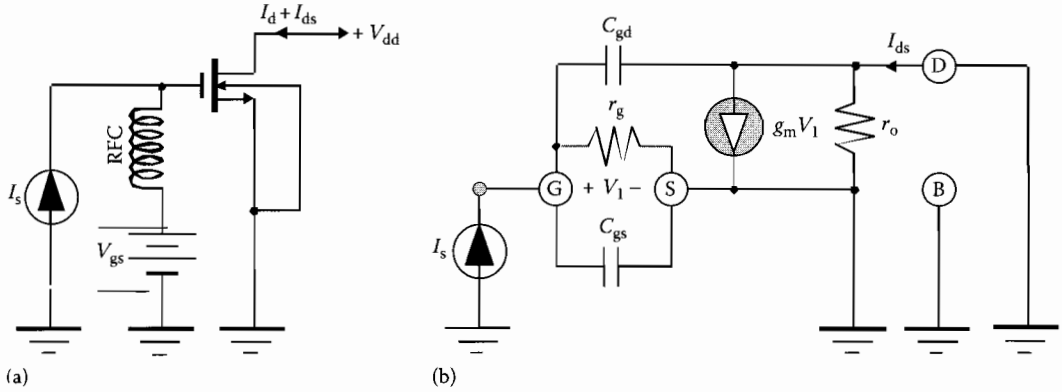

FIGURE 1.37 (a) N-channel common-source MOSFET configured for the evaluation of the unity gain frequency, ω_T, of the transistor. (b) Small-signal, high-frequency equivalent circuit of the network in (a).

capabilities of competing transistors and their associated fabrication processes, its value to bracketing the achievable bandwidths and response speeds of MOSFET circuits is dubious. The latter contention stems from the very definition of the metric. In particular, ω_T is the radial value of signal frequency at which the magnitude of the small-signal, short-circuit current gain of a common-source amplifier degrades to unity. The circuit of relevance is the topology of Figure 1.37a, in which a current signal, I_s is applied to the gate of a MOSFET whose source terminal is incident with signal ground. The radio frequency choke (RFC) provides a conduit for establishing a gate–source bias, V_{gs}, above threshold, while providing a dynamic impedance in series with the gate biasing voltage that is large enough to cajole most of the input signal current to enter the gate terminal.

An input current applied to a gate lead that inherently comprises an open circuit at low frequencies is hardly rational from a circuits perspective. This irrationality is exacerbated by the fact that the drain terminal, where the small-signal current signal response, I_{ds}, to input signal current I_s is extracted, is connected directly to the power supply rail, V_{dd}, thereby rendering the drain terminal short circuited to signal ground (hence the nomenclature, "short circuit" current gain). In other words, the current gain, I_{ds}/I_s, is computed for a common-source amplifier whose gate is driven by signal current and whose drain is short circuited to signal ground, which is hardly a viable analog circuit cell.

Assuming that the transistor at hand operates in saturation, the small-signal equivalent model of the circuit in Figure 1.37a is the structure given in Figure 1.37b. Because the drain, in addition to the bulk and the source, is grounded for signal conditions, the current gain, I_{ds}/I_s, tacitly ignores the high-frequency effects of bulk–drain and bulk–source transistor capacitances. Moreover, the connection of the bulk terminal to the source obviates the need for the bulk transconductance generator, $\lambda_b g_m V_2$, in the model at hand, while short circuiting the drain terminal to the source terminal renders the channel resistance, r_o, inconsequential. Accordingly, an analysis of the structure in Figure 1.37b yields

$$\left.\begin{aligned} I_{ds} &= g_m V_1 - j\omega C_{gd} V_1 \\ I_s &= \frac{V_1}{r_g} + j\omega C_{gs} V_1 + j\omega C_{gd} V_1 \end{aligned}\right\}, \tag{1.180}$$

whence

$$\frac{I_{ds}}{I_s} = \frac{g_m r_g (1 - j\omega C_{gd}/g_m)}{1 + j\omega r_g (C_{gs} + C_{gd})}. \tag{1.181}$$

Since r_g in Equation 1.166 is infinitely large at zero signal frequency, the short circuit gain is seen to be infinity at zero frequency, which reflects engineering intuition in that the gate can conduct no current at zero frequency. In addition, r_g is likely to remain very large in the neighborhood of the 3 dB frequency projected by this gain relationship so that

$$\frac{I_{ds}}{I_s} \approx \frac{g_m}{j\omega(C_{gs} + C_{gd})}. \tag{1.182}$$

Equation 1.182 also invokes the reasonable presumption that the frequency, g_m/C_{gd}, of the right half plane zero evidenced on the right-hand side of Equation 1.181 is significantly larger than the aforementioned 3 dB bandwidth. This presumption is tantamount to neglecting the gate-to-drain feedforward through the gate–drain capacitance, C_{gd}, in comparison to the I/O feedforward promoted by the transistor transconductance, g_m. While this approximation is suspect at very high signal frequencies, the approximations leading to Equation 1.182 allow an extrapolated value of the unity gain frequency of

$$\omega_T = 2\pi f_T = \frac{g_m}{C_{gs} + C_{gd}}. \tag{1.183}$$

Clearly, f_T is a highly optimistic estimate of achievable circuit performance, for it pertains expressly to the special case of a drain that is short circuited to the source terminal, thereby quashing the impact on bandwidth of bulk–drain capacitance and any load capacitance that might be driven by the subject transistor. Using Equations 1.106, 1.156, 1.164, 1.176, and 1.174 with V_λ presumed large, the last result is expressible as

$$\omega_T = 2\pi f_T \approx \frac{3\mu_n(V_{gs} - V_h)}{2L^2\left[1 + \frac{3(C_{gso}+C_{gdo})}{2LC_{ox}}\right]}, \tag{1.184}$$

which reflects a direct dependence of the unity gain frequency on free carrier mobility. The result also infers that to the extent that $2LC_{ox} \gg 3(C_{gso} + C_{gdo})$, which may indeed be an engineering stretch for minimal geometry, deep submicron devices, f_T is inversely proportional to the square of channel length. A significant increase in f_T is therefore portended by even a relatively modest shortening of the channel length. The fact that the unity gain frequency value is common corporate banter in the marketing of state of the art transistors and processes arguably underpins the widespread process-engineering penchant for progressively decreased channel lengths.

With the gate resistance, r_g, presumed very large over the signal frequency range of interest, the resultant short circuit current gain in Equation 1.182 is dependent on only frequency-invariant small-signal transistor parameters. Accordingly, Equation 1.183 allows Equation 1.182 to be generalized as the complex frequency domain expression,

$$\frac{I_{ds}}{I_s} \approx \frac{\omega_T}{s}. \tag{1.185}$$

Moreover, the model in Figure 1.35b reduces to the equivalent circuit shown in Figure 1.38a, where signal current I_1 is identified as the current conducted by the gate–source capacitance, C_{gs}. Since signal voltage V_1 is clearly I_1/sC_{gs}, the voltage controlled current, $g_m V_1$, is

$$g_m V_1 = \frac{g_m I_1}{sC_{gs}} = \frac{\omega_T(C_{gs} + C_{gd})I_1}{sC_{gs}} = \left(\frac{k_g\omega_T}{s}\right)I_1, \tag{1.186}$$

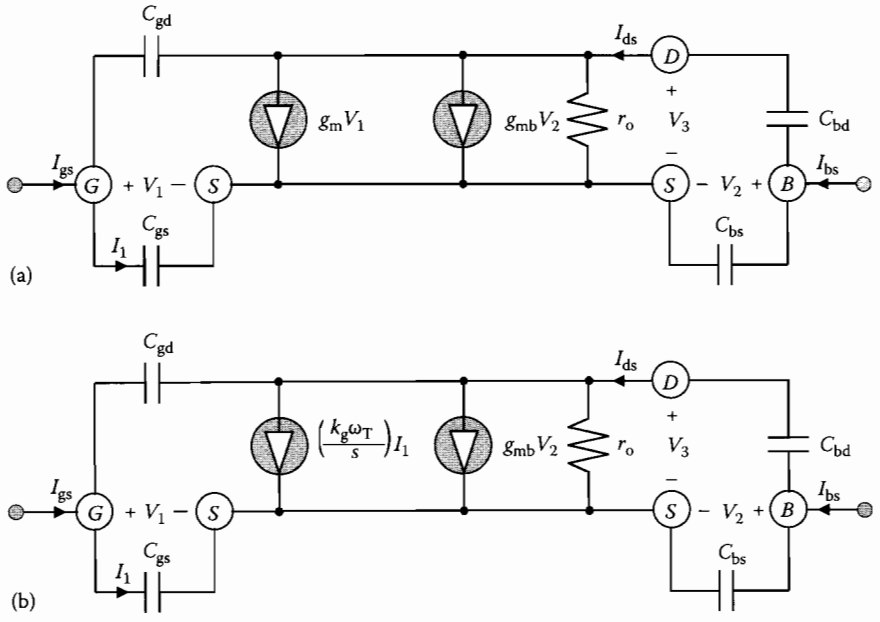

FIGURE 1.38 (a) Small-signal model of Figure 1.35b with gate resistance r_g ignored. (b) Alternative CCCS form of the equivalent circuit in (a).

where

$$k_g = 1 + \frac{C_{gd}}{C_{gs}}. \tag{1.187}$$

Equation 1.186 allows the voltage-controlled current source (VCCS) form of the model shown in Figure 1.38a to be transformed into the equivalent current-controlled structure offered in Figure 1.38b. The latter form proves useful in assessing the performance of amplifiers, such as certain forms of low noise bandpass structures, which utilize source degeneration inductances.

1.2.5.3 Small-Signal Model Development

While the small-signal transistor models shown in Figures 1.35 and 1.36 are topologically correct and conceptually useful from the perspective of linear active network design and first order performance assessment, their engineering utility is limited by two issues. First, the analytical expressions for the parameters embedded in these structures are predicated on a plethora of approximations stemming from the neglect of the effects of lateral and vertical electric fields and simplifications surrounding the charge storage mechanisms of devices and their associated capacitive profiles. These analytically simplifying approximations often place laboratory characterizations of device behavior at odds with physical reality. For example, the substrate doping concentration is not a constant, as is presumed in all foregoing analytical disclosures, but it is less than immediately clear if some sort of weighted average of this dopant level is appropriate for a satisfying voltage–current–charge characterization of a considered transistor. Second, the subject small-signal parameters are dependent on variables, such as carrier mobility, oxide overlap dimensions, doping concentrations, densities of charges trapped in the oxide, regional perimeter dimensions, and the like that are either not released to the circuit designer or are otherwise only vaguely known to the processing foundry.

Because of the foregoing parametric anomalies, reasonably accurate and physically sound assessments of small-signal device and associated circuit performance require that the numerical explication of all relevant small-signal parameters derive from appropriate laboratory measurements conducted on either test device structures or on entire test cells of the circuits undergoing development. A commonly used vehicle toward this characterization end is the *scattering parameters*, or *S-parameters*, measured for a grounded source, grounded gate, or grounded drain interconnection of a subject transistor excited for a suitable range of biasing levels and over an appropriate range of signal frequencies [11]. These parameters are extracted with fixed and known—generally 50 ohm—reference terminations at the input and output ports of the device undergoing test. The measured S-parameters, S_{ij}, are then converted into short circuit admittance (y-) parameters, y_{ij}. The latter two-port parameters are virtually impossible to discern directly in the laboratory because their numerical delineation mandates the imposition of input and output port signal short circuits, which are difficult to sustain over broad frequency passbands. Once the y_{ij} are determined, it is an involved, but nonetheless straightforward, matter to infer realistic values of most of the parameters implicit to the structures of Figures 1.35 and 1.36.

If the process foundry provides a reliable large-signal HSPICE model, such as the fundamentally heuristic Level 49 MOSFET model, of the device under consideration, the short circuit admittance parameters of the subject transistor can be deduced through appropriate small-signal computer-aided simulations. For example, consider the N-channel MOSFET in Figure 1.39, which is shown connected as a grounded source, three-port configuration. The battery voltage, V_{gg}, biases the gate–source terminals at a greater than threshold value of voltage that establishes the desired quiescent drain current, I_d. Of course, zero quiescent gate current flows in the gate lead of the transistor. On the other hand, the battery voltage, V_{dd}, which modestly influences the quiescent drain current, I_d, is chosen to ensure saturation regime operation of the transistor. Finally, the voltage, V_{bb}, biases the bulk substrate terminal, where it is understood that V_{bb} is ordinarily at most zero. In concert with the traditional stipulations of reverse-biased bulk–source and bulk–drain junctions, zero quiescent gate current is presumed to flow into the bulk. The application of any one or more of the three indicated signal voltages, V_{1s}, V_{2s}, and V_{3s}, produces signal current responses in the gate, bulk, and drain of I_{gs}, I_{bs}, and I_{ds}, respectively. Selecting the "AC" simulation option to manifest a strictly linear HSPICE analysis of the aforementioned current signal responses about respective quiescent values allows the applied signal voltages, V_{1s}, V_{2s}, and V_{3s}, to be set conveniently to amplitudes of 1 V.

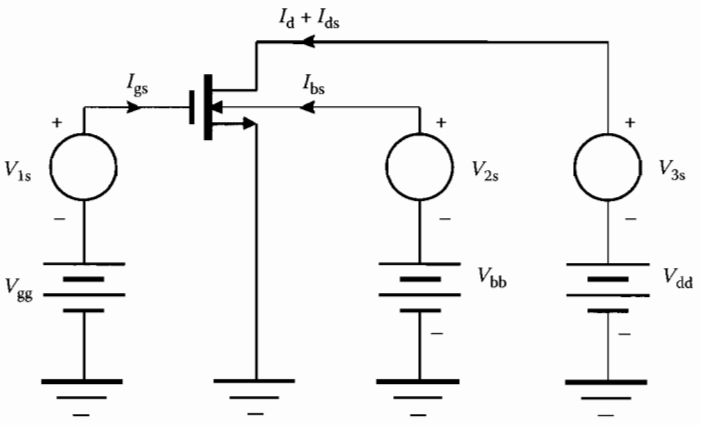

FIGURE 1.39 Common-source test cell of an N-channel transistor configured as a three-port network.

If the applied signal voltages subscribe to the small-signal, linear-operational constraint, superposition theory applies, and the three signal current responses can be described by the linear admittance parameter matrix,

$$\begin{bmatrix} I_{gs} \\ I_{bs} \\ I_{ds} \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{bmatrix} \begin{bmatrix} V_{1s} \\ V_{2s} \\ V_{3s} \end{bmatrix}. \tag{1.188}$$

In Equation 1.188, the short-circuit admittance parameters, y_{ij}, which are invariably complex numbers, are extracted over signal frequency. In particular,

$$\left.\begin{aligned} y_{11} &= I_{gs}/V_{1s}|_{V_{2s}=V_{3s}=0}, \quad y_{12} = I_{gs}/V_{2s}|_{V_{1s}=V_{3s}=0}, \quad y_{13} = I_{gs}/V_{3s}|_{V_{1s}=V_{2s}=0} \\ y_{21} &= I_{bs}/V_{1s}|_{V_{2s}=V_{3s}=0}, \quad y_{22} = I_{bs}/V_{2s}|_{V_{1s}=V_{3s}=0}, \quad y_{23} = I_{bs}/V_{3s}|_{V_{1s}=V_{2s}=0} \\ y_{31} &= I_{ds}/V_{1s}|_{V_{2s}=V_{3s}=0}, \quad y_{32} = I_{ds}/V_{2s}|_{V_{1s}=V_{3s}=0}, \quad y_{33} = I_{ds}/V_{3s}|_{V_{1s}=V_{2s}=0} \end{aligned}\right\}. \tag{1.189}$$

The real and imaginary parts of all nine of these y-parameters can be readily evaluated from a small-signal HSPICE analysis of the structure in Figure 1.39 or alternatively, they can be discerned in terms of scattering parameters gleaned from measurements of a test structure analogous to that of the subject figure.

If the algebraic form of parameter y_{12} in Equation 1.188 is defined as

$$y_{12} \triangleq g_{12} - j\omega C_{12}, \tag{1.190}$$

the first of the equations in Equation 1.188 can be written as

$$I_{gs} = (y_{11} + y_{13} - j\omega C_{12})V_{1s} + g_{12}V_{2s} - y_{13}(V_{1s} - V_{3s}) + j\omega C_{12}(V_{1s} - V_{2s}), \tag{1.191}$$

which can be couched in the form,

$$I_{gs} = \left(\frac{1}{R_i} + j\omega C_i\right)V_{1s} + g_{12}V_{2s} + \left(\frac{1}{R_f} + j\omega C_f\right)(V_{1s} - V_{3s}) + j\omega C_{12}(V_{1s} - V_{2s}). \tag{1.192}$$

The first parenthesized term on the right-hand side of this expression represents a gate-to-ground shunt interconnection of a resistance,

$$R_i = \frac{1}{\mathrm{Re}(y_{11}) + \mathrm{Re}(y_{13})}, \tag{1.193}$$

and a capacitance,

$$C_i = \frac{\mathrm{Im}(y_{11}) + \mathrm{Im}(y_{13})}{\omega} - C_{12}, \tag{1.194}$$

where from Equation 1.190,

$$C_{12} = -\frac{\mathrm{Im}(y_{12})}{\omega}. \tag{1.195}$$

The resistance, R_i, tends to vary as the inverse square of the radial signal frequency. Thus, it is expedient to write R_i as

$$R_i = \frac{K_{Ri}}{\omega^2}, \tag{1.196}$$

where K_{Ri} is a constant boasting the strange dimensions of ohms-$(\mathrm{rad/s})^2$.

The second term in Equation 1.192 represents a VCCS whose bulk-to-gate transconductance is

$$g_{12} = \mathrm{Re}(y_{12}). \tag{1.197}$$

The second parenthesized factor on the right-hand side of the subject equation connotes a gate-to-drain shunt combination of resistance

$$R_f = -\frac{1}{\mathrm{Re}(y_{13})} \tag{1.198}$$

and capacitance

$$C_f = -\frac{\mathrm{Im}(y_{13})}{\omega}. \tag{1.199}$$

As is the case with resistance R_i, R_f also varies as the inverse square of the radial signal frequency. Accordingly,

$$R_f = \frac{K_{Rf}}{\omega^2}. \tag{1.200}$$

Finally, the last term in Equation 1.192 is merely a capacitance, C_{12}, incident between gate and bulk terminals. It is appropriate to interject that over a broad range of signal frequencies that do not exceed the transistor unity gain frequency, f_T, the capacitances, C_i, C_{12}, and C_f, are nearly constants, which suggests that $\mathrm{Im}(y_{11})$, $\mathrm{Im}(y_{12})$, and $\mathrm{Im}(y_{13})$ are nominally linear functions of the radial signal frequency. An analogous statement prevails for all of the other capacitances defined in the forthcoming paragraphs. While resistances R_i and R_f decrease sharply with signal frequency, they are so large (hundreds or even thousands of megohms) that they can usually be neglected in the course of most design-oriented analog circuit analyses.

Letting

$$y_{ij} = g_{ij} - j\omega C_{ij} \tag{1.201}$$

denote the general short circuit admittance parameter, y_{ij}, Equation 1.188 allows the signal current, I_{bs}, conducted by the bulk to be expressed as

$$I_{bs} = (g_{21} - j\omega C_x)V_{1s} + \left(\frac{1}{R_{bb}} + j\omega C_{bb}\right)V_{2s} + (g_{23} - j\omega C_{23})V_{3s} + j\omega C_{12}(V_{2s} - V_{1s}), \tag{1.202}$$

where, recalling Equation 1.201,

$$C_x = C_{21} - C_{12} = \frac{\mathrm{Im}(y_{12}) - \mathrm{Im}(y_{21})}{\omega}. \tag{1.203}$$

and

$$g_{21} = \mathrm{Re}(y_{21}). \tag{1.204}$$

The term, $(g_{21} - j\omega C_x)$ in Equation 1.202 is a transadmittance linking the signal gate voltage to the bulk signal current. The second parenthesized term on the right-hand side of Equation 1.202 reflects a bulk-to-ground parallel combination of a frequency-dependent resistance

$$R_{bb} = \frac{1}{\mathrm{Re}(y_{22})} = \frac{K_{Rbb}}{\omega^2}, \tag{1.205}$$

and capacitance

$$C_{bb} = \frac{\mathrm{Im}(y_{22})}{\omega} - C_{12}. \tag{1.206}$$

A second transadmittance factor, $(g_{23} - j\omega C_{23})$, surfaces to model the coupling of the drain signal voltage, V_{3s}, to the bulk signal current, I_{bs}, where

$$\left.\begin{aligned} g_{23} &= \mathrm{Re}(y_{23}) \\ C_{23} &= -\frac{\mathrm{Im}(y_{23})}{\omega} \end{aligned}\right\}. \tag{1.207}$$

Finally, the last term in Equation 1.200 complements its last term counterpart in Equation 1.192 in that it accounts for the bilateral capacitive coupling prevailing between the drain and bulk terminals.

The only current not yet addressed is the signal drain current, I_{ds}. From Equation 1.188,

$$I_{ds} = (g_m - j\omega C_m)V_{1s} + (g_{mb} - j\omega C_{mb})V_{2s} + \left(\frac{1}{R_o} + j\omega C_o\right)V_{3s} + \left(\frac{1}{R_f} + j\omega C_f\right)(V_{3s} - V_{1s}). \tag{1.208}$$

The factor, $(g_m - j\omega C_m)$, is the forward transadmittance that couples the gate signal voltage to the drain signal current. Its constituent variables are

$$\left.\begin{aligned} g_m &= \mathrm{Re}(y_{31}) - \mathrm{Re}(y_{13}) \\ C_m &= \frac{\mathrm{Im}(y_{13}) - \mathrm{Im}(y_{31})}{\omega} \end{aligned}\right\}. \tag{1.209}$$

On the other hand, $(g_{mb} - j\omega C_{mb})$ is the bulk transadmittance serving to bracket the signal drain current response to the signal bulk voltage, V_{2s}. The variables, g_{mb} and C_{mb}, are

$$\left.\begin{aligned} g_{mb} &= \mathrm{Re}(y_{32}) \\ C_{mb} &= -\frac{\mathrm{Im}(y_{32})}{\omega} \end{aligned}\right\}. \tag{1.210}$$

Following Equation 1.177, the bulk modulation factor, λ_b, can be discerned to be

$$\lambda_b = \frac{g_{mb}}{g_m} = \frac{\mathrm{Re}(y_{32})}{\mathrm{Re}(y_{31}) - \mathrm{Re}(y_{13})}. \tag{1.211}$$

The parenthesized factor of the third term on the right-hand side of Equation 1.206 is the drain-to-ground shunt interconnection of resistance R_o and capacitance C_o, such that

$$\left.\begin{aligned} R_o &= \frac{1}{\mathrm{Re}(y_{33}) + \mathrm{Re}(y_{13})} \\ C_o &= \frac{\mathrm{Im}(y_{33}) + \mathrm{Im}(y_{13})}{\omega} \end{aligned}\right\}. \tag{1.212}$$

The last term in Equation 1.208 reflects the previously introduced, bilateral R_f–C_f coupling between the gate and the drain terminals.

Equation 1.192 for the signal gate current, I_{gs}, Equation 1.202 for the signal bulk current, I_{bs}, and Equation 1.208 for the signal drain current, I_{ds}, can now be exploited to develop the foreboding three-port common-source MOSFET model diagrammed in Figure 1.40a. While the model is intractable for manual circuit analysis and considerably more complicated than its simplified brethren in Figures 1.35 and 1.36, it does serve to bolster circuit design insights. First, and perhaps most obviously, the model at hand illustrates the complex interactions of the bulk with the gate, drain, and source regions of a MOSFET. For example, the bulk signal voltage, V_{2s}, precipitates a real controlled source, $\lambda_b g_m V_{2s}$, in addition to a quadrature controlled source, $j\omega C_{mb} V_{2s}$, at the drain–source port. The first of these sources is the expected effect of bulk-induced modulation of MOSFET threshold voltage, but the latter controlled element is slightly south of transparent. The bulk also gives rise to a VCCS, $g_{12} V_{2s}$, in the gate–source port, which accounts for observable bulk-induced increases in high-frequency gate current. These intricacies, together with the complex transadmittance coupling, $(g_{23} + j\omega C_{23}) V_{3s}$, from the drain-to-the bulk seemingly encourage, whenever possible and prudent, operating the MOSFET with its bulk terminal returned to the transistor source terminal. Under such a topological constraint, the model in Figure 1.40a collapses to the almost shockingly simpler network offered in Figure 1.40b.

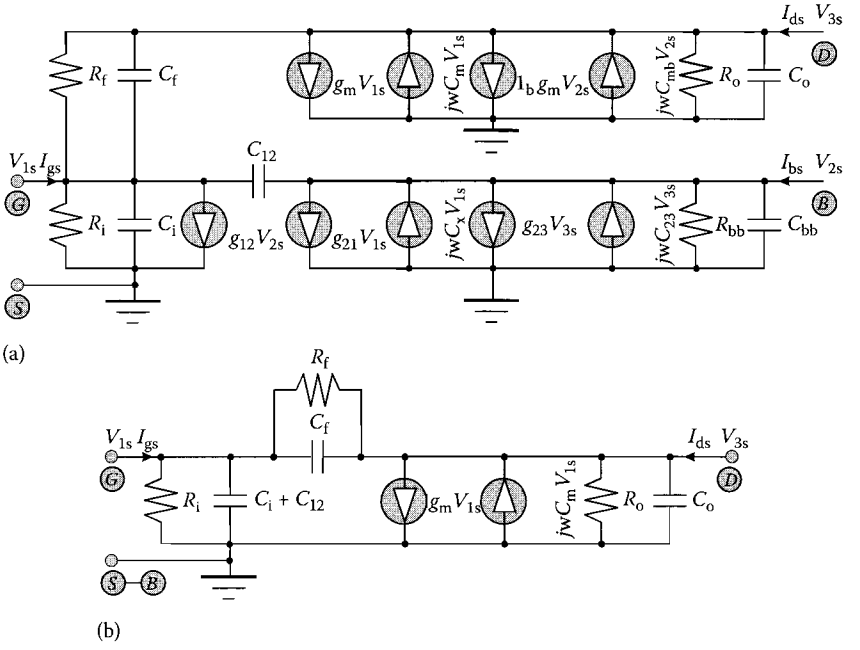

FIGURE 1.40 (a) Small-signal, three-port equivalent circuit for the common-source interconnection of a MOSFET. The three terminal voltages, V_{1s}, V_{2s}, and V_{3s}, denote signal voltages developed with respect to ground at the gate, bulk, and drain terminals, respectively. (b) The equivalent circuit of (a) with the bulk terminal connected directly to the MOSFET source terminal.

A comparison of the model in Figure 1.40b with that of Figure 1.35b suggests that the gate–source resistance, R_i, is effectively the gate resistance, r_g, introduced in Equation 1.166. The resistance, R_f, in Figure 1.40b has no counterpart in Figure 1.35b. Throughout the range of frequencies extending through the unity gain frequency of the considered transistor, both R_i and R_f are so large that they can be ignored for most small-signal analysis ventures, save possibly for a small-signal analysis entailing an assessment of the noise properties of a transistor. The resistance, R_o, is akin to the channel resistance, r_o, in Figure 1.35b. Unlike R_i and R_f, r_o is nominally frequency invariant through the device unity gain frequency metric. With R_i and R_f tacitly ignored and in view of the fact that r_o is independent of frequency, the steady state frequency variable, $j\omega$, in Figure 1.40 can be replaced by the Laplace operator, s, thereby allowing for small-signal step response and other transient investigations of MOSFET amplifiers.

The net capacitance, $(C_i + C_{12})$, in Figure 1.40 is the effective gate–source capacitance, C_{gs}, in Figure 1.35b. Because of the inclusion of capacitance C_{12}, this net gate–source capacitance accounts for gate-to-bulk capacitance, which earlier models presented in this discourse ignore tacitly, primarily because of the high-frequency capacitance characteristics advanced by Figure 1.23. The capacitance, C_f, is the effective gate–drain capacitance, C_{gd}, while capacitance C_o represents the effective bulk–drain capacitance, C_{bd}.

The model in Figure 1.35b highlights a real forward transconductance of g_m, while the models in Figure 1.40 project a complex forward transadmittance, Y_m, of

$$Y_m = g_m - j\omega C_m = g_m e^{j\varphi_m(\omega)} \sqrt{1 + \left(\frac{\omega C_m}{g_m}\right)^2}, \tag{1.213}$$

where

$$\varphi_m(\omega) = -\tan^{-1}\left(\frac{\omega C_m}{g_m}\right) \tag{1.214}$$

denotes an *excess phase angle* associated with the transport of minority carriers in the gate-induced channel extending from the source region-to-the drain region. Equivalently, the angle, $\varphi_m(\omega)$, is associated with an *excess envelope delay*, $T_m(\omega)$, such that

$$T_m(\omega) = -\frac{d\varphi_m(\omega)}{d\omega} = \frac{C_m/g_m}{1 + \left(\frac{\omega C_m}{g_m}\right)^2}, \tag{1.215}$$

whose low-frequency and, in this case, maximum value is obviously (C_m/g_m). Excess delay, for which no account prevails in the simpler models of Figures 1.35 and 1.36, looms potentially critical in feedback circuits in that it acts to degrade the achievable phase margin of the open loop response.

VCCSs having an imaginary transadmittance can be synthesized easily for small-signal, computer-based analyses through the use of a voltage-controlled voltage source (VCVS), a capacitor, and a current-controlled current source (CCCS), as depicted in Figure 1.41a. In this figure, the controlling current, I_i, of the CCCS, αI_i, is

$$I_i = j\omega C\mu V_i, \tag{1.216}$$

whence the indicated controlled current, I_o, is

$$I_o = \alpha I_i = j\omega C\alpha\mu V_i. \tag{1.217}$$

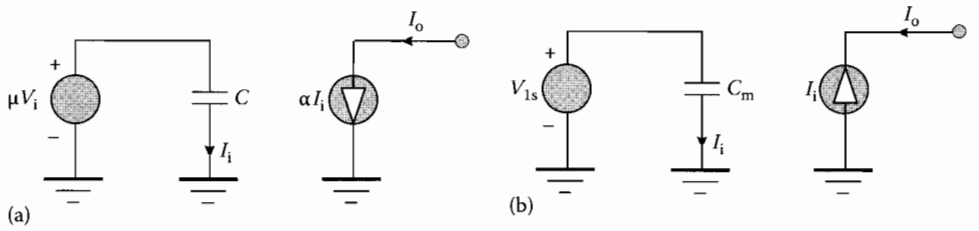

FIGURE 1.41 (a) Synthesis of a VCCS whose transadmittance is imaginary and proportional to radial signal frequency. (b) The synthesis of the controlled current, $-j\omega C_m V_{1s}$, in the models shown in Figure 1.40.

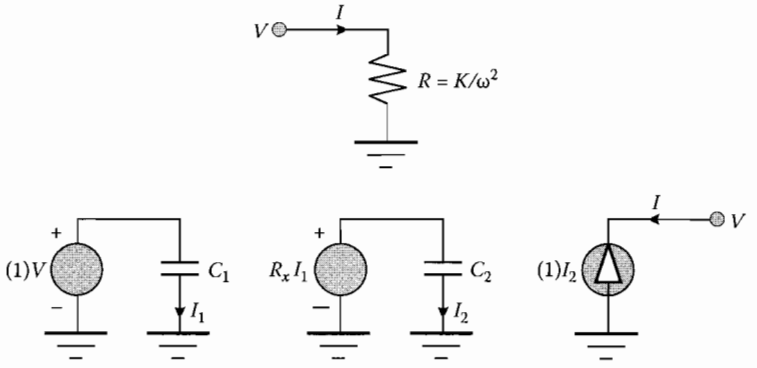

FIGURE 1.42 Synthesis of a branch resistance whose value is inversely proportional to the square of the radial signal frequency. The indicated resistance, R, is synthesized if $R_x = 1/KC_1C_2$.

Thus, for the imaginary component, $j\omega C_m$, of the forward transadmittance, Y_m, in Equation 1.211, $\mu = 1$, $C = C_m$, and $\alpha = -1$ gives the desired controlled current, $-j\omega C_m V_{1s}$, as is abstracted in Figure 1.41b.

Similarly, the frequency variant resistances, R_i, R_f, and R_{bb}, can be synthesized for small-signal, computer-aided analysis purposes using a VCVS, a current-controlled voltage source (CCVS), and a CCCS. This contention is illustrated in Figure 1.42 for the general case of a resistance, R, given by

$$R = -\frac{K}{s^2}, \tag{1.218}$$

which for steady state sinusoidal conditions is the generalized relationship,

$$R = \frac{K}{\omega^2}, \tag{1.219}$$

advanced by Equations 1.196, 1.200, and 1.205. To wit, the controlled current, I_1, generated by the VCVS, $(1)V$, is $I_1 = sC_1V$, while the current, I_2, established in response to the CCVS, (R_xI_1), is $I_2 = sC_2R_xI_1 = s^2C_1C_2R_xV$. It follows that the resistance, R, presented to the port driven by the CCCS, $(1)I_2$, is

$$R = \frac{V}{-I_2} = \frac{V}{-s^2C_1C_2R_xV} = -\frac{1}{s^2C_1C_2R_x}. \tag{1.220}$$

For arbitrary values of capacitances C_1 and C_2, selecting

$$R_x = \frac{1}{KC_1C_2} \tag{1.221}$$

achieves the desired resistance value set forth by Equation 1.218.

Parameterization example:
An N-channel transistor featuring a channel length of 180 nm has the Level 49 HSPICE parameters that appear in Table 1.2. The transistor is implemented with a gate aspect ratio of $W/L = 25$, and is biased at $V_{gs} = 1.1$ V, $V_{ds} = 1$ V and $V_{bs} = 0$ V. The device undergoing study is earmarked for analog small-signal applications that embrace a signal frequency range extending from 100 MHz to 10 GHz. For this frequency passband, determine nominal values of all of the parameters indigenous to the small-signal, common-source model of Figure 1.40b. Also, compute the extrapolated unity gain frequency of the transistor at the given quiescent operating point. Express these results as maximum value, minimum value, average value, and standard deviation (referred to the average value) over a frequency passband extending from 100 MHz to 10 GHz.

TABLE 1.2 Representative Level 49 HSPICE Parameters for an NMOS Transistor in a Fabrication Process Featuring a Nominal Channel Length of 180 nM

Model 180 nM NMOS (Level = 49)

+VERSION = 3.1	TNOM = 27	TOX = 4E-9	XJ = 1E-7
+NCH = 2.3549E17	VTH0 = 0.3627858	K1 = 0.5873035	K2 = 4.793052E-3
+K3 = 1E-3	K3B = 2.2736112	W0 = 1E-7	NLX = 1.675684E-7
+DVT0W = 0	DVT1W = 0	DVT2W = 0	DVT0 = 1.7838401
+DVT1 = 0.5354277	DVT2 = −1.243646E-3	U0 = 263.3294995	UA = −1.359749E-9
+UB = 2.250116E-18	UC = 5.204485E-11	VSAT = 1.083427E5	A0 = 2
+AGS = 0.4289385	B0 = −6.378671E-9	B1 = −1E-7	KETA = −0.0127717
+A1 = 5.347644E-4	A2 = 0.8370202	RDSW = 150	PRWG = 0.5
+PRWB = −0.2	WR = 1	WINT = 1.798714E-9	LINT = 7.631769E-9
+XL = −2E-8	XW = −1E-8	DWG = −3.268901E-9	DWB = 7.685893E-9
+VOFF = −0.0882278	NFACTOR = 2.5	CIT = 0	CDSC = 2.4E-4
+CDSCD = 0	CDSCB = 0	ETA0 = 2.455162E-3	ETAB = 1
+DSUB = 0.0173531	PCLM = 0.7303352	PDIBLC1 = 0.2246297	PDIBLC2 = 2.220529E-3
+PDIBLCB = −0.1	DROUT = 0.7685422	PSCBE1 = 8.697563E9	PSCBE2 = 5E-10
+PVAG = 0	DELTA = 0.01	RSH = 6.7	MOBMOD = 1
+PRT = 0	UTE = −1.5	KT1 = −0.11	KT1L = 0
+KT2 = 0.022	UA1 = 4.31E-9	UB1 = −7.61E-18	UC1 = −5.6E-11
+AT = 3.3E4	WL = 0	WLN = 1	WW = 0
+WWN = 1	WWL = 0	LL = 0	LLN = 1
+LW = 0	LWN = 1	LWL = 0	CAPMOD = 2
+XPART = 0.5	CGDO = 716E-12	CGSO = 716E-12	CGBO = 1E-12
+CJ = 9.725711E-4	PB = 0.7300537	MJ = 0.365507	CJSW = 2.604808E-10
+PBSW = 0.4	MJSW = 0.1	CJSWG = 3.3E-10	PBSWG = 0.4
+MJSWG = 0.1	CF = 0	PVTH0 = 4.289276E-4	PRDSW = −4.2003751
+PK2 = −4.920718E-4	WKETA = 6.938214E-4	LKETA = −0.0118628	PU0 = 24.2772783
+PUA = 9.138642E-11	PUB = 0	PVSAT = 1.680804E3	PETA0 = 2.44792E-6
+PKETA = 4.537962E-5)			

Results:

1. Before proceeding with the simulation, the planar source and drain areas, A_s and A_d, as well as the source and drain peripheral dimensions, P_s and P_d, must be computed through an appeal to Equation 1.154. These are $A_s = A_d = (1.62)(10^{-12})$ m^2 and $P_s = P_d = (5.22)(10^{-6})$ *m*. In arriving at these figures, use is made of the fact that for a channel length of $L = 180$ nm and a gate aspect ratio of $W/L = 25$, the gate width is $W = 4.5$ μm. The parameters, L, W, A_s, A_d, P_s, and P_d are inserted directly on the model line of the HSPICE net list. For example, the model line used in the simulations executed herewith is

$$M2\ 6\ 4\ 0\ 7\ 180\ \text{nM}, \quad L = 180\text{n}, \quad W = 4.5\ \text{u}, \quad A_S = 1.62\ \text{p}, \quad A_D = 1.62\ \text{P},$$
$$P_S = 5.22\ \text{u}, \quad P_D = 5.22\ \text{u}.$$

In this model line, *M2* identifies the transistor undergoing examination, "6" is the number of the drain node, "4" is the gate node number, "0" is the number of the grounded source node, and "7" is the number of the bulk substrate node. The insert, "180 nM," identifies the name of the model used for the subject transistor.

An HSPICE simulation of the simple test cell shown in Figure 1.39 can now be straightforwardly executed. In this test structure, $V_{gg} = 1.1$ V, $V_{dd} = 1$ V, and $V_{bb} = 0$ V combine to set the desired operating point of the transistor. The operating point information disclosed by the static HSPICE simulation is as follows.

ID 8.9407E-04 (drain current is $I_d = 894.1$ μA)
IS -8.9407E-04 (source current flows out of device and is virtually identical to the drain current)
IB -1.0002E-12 (bulk current is about one picoampere and flows out of the device)
IBD -9.9417E-13 (bulk current is sum of the bulk–drain and bulk–source junction currents)
IBS -6.0237E-15 (bulk current is sum of the bulk–drain and bulk–source junction currents)
VGS 1.1000 (desired gate–source quiescent voltage)
VDS 1.0000 (desired drain–source quiescent voltage)
VBS 0.0000 (desired bulk–source quiescent voltage)
VTH 0.5102 (simulated threshold voltage is $V_h = 510.2$ mV)
VDSAT 0.3149 (simulated drain saturation voltage is $V_{dsat} = 314.9$ mV)

It should be noted that the quiescent drain source voltage, $V_{ds} = 1.0$ V, is certainly larger than the simulated drain saturation voltage, $V_{dsat} = 314.9$ mV. Accordingly, the device at hand operates in its saturation regime for suitable small-signal excitations.

2. A small-signal, computer-aided simulation of the test circuit in Figure 1.39 can now be executed at the quiescent operating point established in the preceding step of this exercise. The objective of this simulation is to ascertain the real and imaginary components of each of the nine short circuit admittance parameters, y_{ij}, introduced in Equation 1.188. The model parameters then derive from the pertinent equations given in Section 1.2.5.3.

Gate-to-drain resistance coefficient, K_{Rf}:
Maximum value is $(7.93)(10^{27})$
Minimum value is $(7.92)(10^{27})$
Average value is $(7.93)(10^{27})$
Standard deviation is 0.03%

Gate-to-drain capacitance, C_f:
Maximum value is 3.22 fF
Minimum value is 3.21 fF

Average value is 3.22 fF
Standard deviation is 0.02%

Gate-to-source resistance coefficient, K_{Ri}:
Maximum value is $(4.25)(10^{27})$
Minimum value is $(4.24)(10^{27})$
Average value is $(4.25)(10^{27})$
Standard deviation is 0.03%

Gate-to-source capacitance, C_i:
Maximum value is 7.56 fF
Minimum value is 7.55 fF
Average value is 7.55 fF
Standard deviation is 0.02%

Bulk–gate transconductance, g_{12}:
Maximum value is 1.01 μmho
Minimum value is 0 μmho
Average value is 0.11 μmho
Standard deviation is 191.97%

Gate-to-bulk capacitance, C_{12}:
Maximum value is 0.42 fF
Minimum value is 0.42 fF
Average value is 0.42 fF
Standard deviation is 0%

Forward transconductance, g_m:
Maximum value is 2.03 mmho
Minimum value is 2.03 mmho
Average value is 2.03 mmho
Standard deviation is 0.01%

Transadmittance capacitance, C_m:
Maximum value is 2.00 fF
Minimum value is 1.99 fF
Average value is 2.00 fF
Standard deviation is 0.05%

Gate–bulk transconductance, g_{21}:
Maximum value is 0.82 μmho
Minimum value is 0 μmho
Average value is 0.18 μmho
Standard deviation is 191.97%

Gate–bulk transadmittance capacitance, C_x:
Maximum value is 0.68 fF
Minimum value is 0.67 fF
Average value is 0.67 fF
Standard deviation is 0.03%

Bulk transconductance modulation factor, λ_b:
Maximum value is 0.21
Minimum value is 0.21

Average value is 0.21
Standard deviation is 0.03%

Drain–bulk transconductance, g_{23}:
Maximum value is 0 μmho
Minimum value is −0.41 μmho
Average value is −0.05 μmho
Standard deviation is 191.97%

Bulk transadmittance capacitance, C_{mb}:
Maximum value is 3.19 fF
Minimum value is 3.18 fF
Average value is 3.19 fF
Standard deviation is 0.02%

Drain–bulk transadmittance capacitance, C_{23}:
Maximum value is 2.63 fF
Minimum value is 2.63 fF
Average value is 2.63 fF
Standard deviation is 0.02%

Drain–source channel resistance, R_o:
Maximum value is 10.38 kΩ
Minimum value is 10.33 kΩ
Average value is 10.37 kΩ
Standard deviation is 0.09%

Bulk-to-source resistance coefficient, K_{Rbb}:
Maximum value is $(7.03)(10^{27})$
Minimum value is $(6.78)(10^{27})$
Average value is $(6.99)(10^{27})$
Standard deviation is 0.76%

Drain–source capacitance, C_o:
Maximum value is 2.62 fF
Minimum value is 2.62 fF
Average value is 2.62 fF
Standard deviation is 0.03%

Bulk–source capacitance, C_{bb}:
Maximum value is 6.83 fF
Minimum value is 6.82 fF
Average value is 6.83 fF
Standard deviation is 0.02%

3. Equation 1.183 is the pertinent equation for the computation of the extrapolated unity gain frequency. To this end, the average forward transconductance has been computed to be $g_m = 2.03$ mmho. The effective gate–source capacitance, C_{gs}, is the computed average value, $C_i = 7.55$ fF, which accounts for gate–source overlap and any other second order phenomena embraced by the utilized HSPICE model. On the other hand, the effective average gate–drain capacitance, C_{gd}, is $C_f = 3.22$ fF, which, like C_i, incorporates all pertinent high order device characterization phenomena. Accordingly

$$f_T = \frac{g_m}{2\pi(C_i + C_f)} = 30.0 \text{ GHz}.$$

Comments: With the exception of parameters g_{12}, g_{21}, and g_{23}, the quoted standard deviation numbers indicate an excellent model fit to circuit theoretic issues. These three transconductances can also be made to agree well with theoretical disclosures if they are each allowed to vary as the square of the radial signal frequency. However, their values are so small as to render overt concern of them unproductive.

The computed unity gain frequency, f_T, is within range of the expected frequency performance of representative MOSFETs manufactured in a 180 nM technology process. It is interesting to note, however, that the effective gate–drain capacitance (3.22 fF), which is traditionally ignored in first order, high-frequency circuit analysis ventures, is, in this case, almost 43% of the effective gate–source capacitance (7.55 fF).

1.2.6 Design-Oriented Analysis Strategy

When a MOSFET is exploited for a linear analog signal processing application, an essential early design requirement entails the implementation of suitable biasing. Generally, this biasing must ensure that for all pertinent signal levels, each transistor used to supply gain, impedance conversion, constant current, constant voltage, or other I/O properties operates in its saturated domain where its drain–source voltages, V_{ds}, is at least as large as its drain saturation voltage, V_{dsat}. When $V_{ds} \geq V_{dsat}$, Equation 1.142 is the applicable relationship for ascertaining a gate–source voltage, V_{gs}, commensurate with a target drain current, I_d, conducted at a given or desired value of drain–source voltage.

Unfortunately, academic satisfaction does not often resonate with the engineering reality that underlies predictable, reliable, and reproducible integrated circuit design. For the biasing issue at hand, Equation 1.142 is fraught with numerous shortfalls. Despite its algebraic cumbersomeness, Equation 1.142 is only an approximation of the static volt–ampere characteristics of a MOSFET operated in saturation, owing to a variety of analytical liberties exploited with respect to charge storage, charge transport, carrier mobility, and the other phenomenological issues discussed in preceding sections. Even if Equation 1.142 were an accurate disclosure of the aforementioned static characteristics, challenges surround its utilization because circuit and system designers are rarely privy to the physical and process parameters on which the metrics, K_n, V_h, V_{dsat}, V_{ve}, V_{le}, and V_λ, are dependent. These model variables can be discerned reliably through only laboratory measurement of static device responses or via analyses conducted on appropriate computer-based simulations founded on accurate and reliable transistor models.

On the tacit presumption that the foregoing six model variables can be extracted satisfactorily from measurement and/or simulation, Equation 1.142 might be supplanted by the more familiar, nominally square law relationship,

$$I_d = \frac{K_{ne}}{2}\left(\frac{W}{L}\right)(V_{gs} - V_h)^2\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right), \tag{1.222}$$

where K_{ne} symbolizes the effective transconductance coefficient,

$$K_{ne} = \frac{K_n M_{sat}^2}{1 + \frac{V_{gs} - V_h}{V_{ve}}} \approx \frac{K_n}{\left(1 + \frac{V_{gs} - V_h}{V_{ve}}\right)\left[1 + 0.78\left(\frac{V_{gs} - V_h}{V_{le}}\right)\right]}. \tag{1.223}$$

This effective transconductance coefficient accounts for mobility degradation deriving from strong vertical (gate-to-channel) electric fields through the variable, V_{ve}, as well as mobility degradation caused by lateral (drain-to-source) electric fields, which is monitored by variable V_{le}. While Equation 1.222 suggests a relatively straightforward square law dependence of drain current on the so called *excess*, or *effective, gate-source voltage*, $(V_{gs} - V_h)$, particularly for the commonly encountered situation of $(V_{ds} - V_{dsat}) \ll V_\lambda$, it should be noted that K_{ne} is inversely proportional to a quadratic function of the

excess gate–source voltage. Typically, V_{ve} is of the order of 5- to 20-fold the value of V_{le} and thus, the possibility of simplifying Equation 1.223 to ease computational strain, while preserving computational accuracy, is dubious.

Example:

An N-channel transistor featuring a channel length of 180 nM has the Level 49 HSPICE parameters given in Table 1.2. The transistor is to be biased in saturation at $V_{ds} = 1$ V and $I_d \approx 1$ mA to achieve a small-signal transconductance, g_m, of at least 3 mmhos. Assuming that the bulk terminal is incident with the transistor source terminal, choose a reasonable gate aspect ratio, W/L, determine the required gate–source voltage bias, V_{gs}, and estimate the model parameters implicit to Equation 1.222.

Results:

1. The applicable circuit for computer-aided investigation is offered in Figure 1.43, where the transistor model parameters are those that appear in Table 1.2, and the gate aspect ratio, W/L, is to be determined. The null voltage source in the drain circuit of the device facilitates the extraction of the quiescent drain current, I_d. It is understood that for biasing purposes, the area and perimeter parameters, A_s, A_d, P_s, and P_d, are of no consequence and can therefore be defaulted to any convenient value. Initially, set $V_{gs} = 1$ V and $W/L = 1$ and, of course, $V_{ds} = 1$ V. The HSPICE static simulation reveals $I_d = 46.4$ μA, $V_{dsat} = 262.8$ mV, $V_h = 519.7$ mV, and $g_m = 136.5$ μmho. Since $V_{gs} = 1$ V is certainly larger than $V_h = 519.7$ mV and $V_{ds} = 1$ V $> V_{dsat} = 262.8$ mV, the transistor is clearly turned on and operates in its saturation domain.
2. With $W/L = 1$, the simulated drain current is a factor of 21.55 times smaller than the target current of 1 mA. This observation seemingly suggests the need for increasing the gate aspect ratio from 1 to 21.55, since the drain current is ostensibly proportional to W/L. In truth, the actual drain current is not directly proportional to W/L because of numerous second order effects, including weak dependencies of threshold voltage, drain saturation voltage, and parameter M_{sat} on gate width W. Experience shows that a more viable gate aspect ratio adjustment is about twice that computed or in this case, about 40. With $W/L = 40$ and $V_{gs} = V_{ds} = 1$ V, HSPICE delivers $I_d = 1.08$ mA, $V_{dsat} = 278.2$ mV, $V_h = 510.2$ mV, and $g_m = 3.17$ mmho. The simulated transconductance value satisfies its design target. Although V_{gs} can be decreased modestly to reduce the drain current to 1 mA, this exercise is unnecessary in view of the effects of routinely encountered device processing

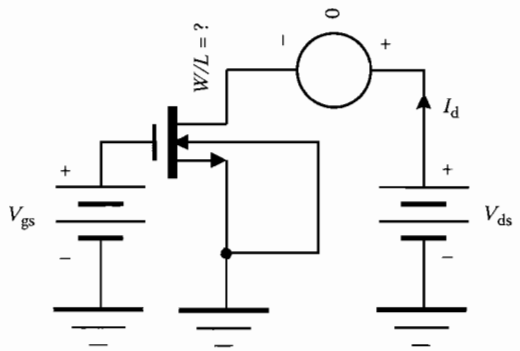

FIGURE 1.43 Circuit structure for MOSFET biasing simulation. The Level 49 HSPICE parameters of the transistor are delineated in Table 1.2.

vagaries and model parameter uncertainties. Thus the design requirement is satisfied for $W/L = 40$ and $V_{gs} = V_{ds} = 1$ V.

3. The model parameterization exercise begins by using Equation 1.141 to compute the voltage, V_{ve}. From Table 1.2, the oxide thickness is $T_{ox} = 4(10^{-9})$ m, which is 40 Å. Accordingly, $V_{ve} = 40/15 = 2.667$ V.
4. The next step in the parameterization process entails operating the transistor undergoing study at a V_{ds} value that equals its saturated value of 278.2 mV. This tack reduces the last parenthesized factor on the right-hand side of Equation 1.222 to unity, thereby simplifying the computation of the effective transconductance parameter, K_{ne}. With $W/L = 40$, $V_{gs} = 1$ V, and $V_{ds} = V_{dsat} = 278.2$ mV, HSPICE produces $I_d = 878.33$ μA and $V_h = 510.0$ mV. Appealing to Equation 1.222, parameter K_{ne} follows forthwith as $K_{ne} = 182.9$ μmho/V.
5. Recalling that $V_{dsat} = 278.2$ mV and $(V_{gs} - V_h) = (1 - 0.510)$ V $= 0.490$ V, Equation 1.129 delivers $M_{sat} = 0.5678$. The previously documented approximate equation, Equation 1.136, relating M_{sat} to variable α can be used to determine the numerical value of α for $V_{gs} = 1$ V and $V_h = 510.0$ mV. Alternatively, Equation 1.131 can be solved for α directly to yield

$$\alpha = \frac{2(1 - M_{sat})}{M_{sat}^2} = 2.682. \tag{E1.1}$$

Using Equation 1.130, parameter V_{le} follows forthwith as $V_{le} = 182.7$ mV.

6. With $K_{ne} = 182.9$ μmho/V, $V_{ve} = 2.667$ V, $M_{sat} = 0.5678$, $V_{gs} = 1$ V, and $V_h = 510.0$ mV, the device transconductance parameter, K_n, follows from Equation 1.223 as $K_n = 671.7$ μmho/V. It is interesting to observe that the effective transconductance factor, K_{ne}, is almost 3.7 times smaller that the "actual" transconductance coefficient, K_n. Experience testifies to the apparent fact that for deep submicron devices, $2.5 \leq K_n/K_{ne} \leq 4$ is typical.
7. In principle, V_{ve}, V_{le}, V_h, V_{dsat}, K_n, and thus K_{ne}, do not vary with changes in the drain–source voltage, V_{ds}. Accordingly, the ratio of the drain current (1.08 mA) for $V_{ds} = 1$ volt to the drain current (878.33 μA) at $V_{ds} = V_{dsat} = 278.2$ mV is solely attributed to the last parenthesized factor on the right-hand side of Equation 1.222, that is,

$$\frac{I_d|_{V_{ds}=1\text{ V}}}{I_d|_{V_{ds}=V_{dsat}}} = \frac{1.08\text{ mA}}{878.33\ \mu\text{A}} = 1.230 = 1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}. \tag{E1.2}$$

It follows that the channel length modulation voltage is $V_\lambda = 3.144$ V.

8. In an attempt to demonstrate the propriety of the foregoing modeling exercise, the forward static transfer characteristic of the subject transistor is modeled in HSPICE for both $V_{ds} = 1$ V and $V_{ds} = 1.5$ V. The simulated results are then compared with calculations deriving from Equations 1.222 and 1.223 using the computed values of V_{ve}, V_{le}, and V_λ and the simulated disclosures for W/L, V_{dsat}, and V_h. Specifically, $V_{ve} = 2.667$ volts, $V_{le} = 182.7$ mV, $V_\lambda = 3.144$ volts, $W/L = 40$, $V_{dsat} = 278.2$ mV, and $V_h = 510.0$ mV.

Figures 1.44 and 1.45 display the results of the foregoing comparative study. In Figure 1.44, the simulated and calculated forward transistor characteristics in the saturation domain are displayed for a drain–source voltage, V_{ds}, of 1.0 V. The calculations corroborate reasonably well with pertinent simulations in that ±15% error is observed for $0.73\text{ V} < V_{gs} < 1.89$ V. It is notable that $V_{gs} = 0.73$ V is only slightly larger than 200 mV above threshold level, while at $V_{gs} = 1.89$ V, the transistor no longer operates in its saturation domain when $V_{ds} = 1$ V. Figure 1.45 confirms better corroboration between calculated and simulated results for $V_{ds} = 1.5$ V. In particular, the computational error is within ±9% for $0.92\text{ V} < V_{gs} < 2$ V and is within ±15% for $0.78\text{ V} < V_{gs} < 2$ V.

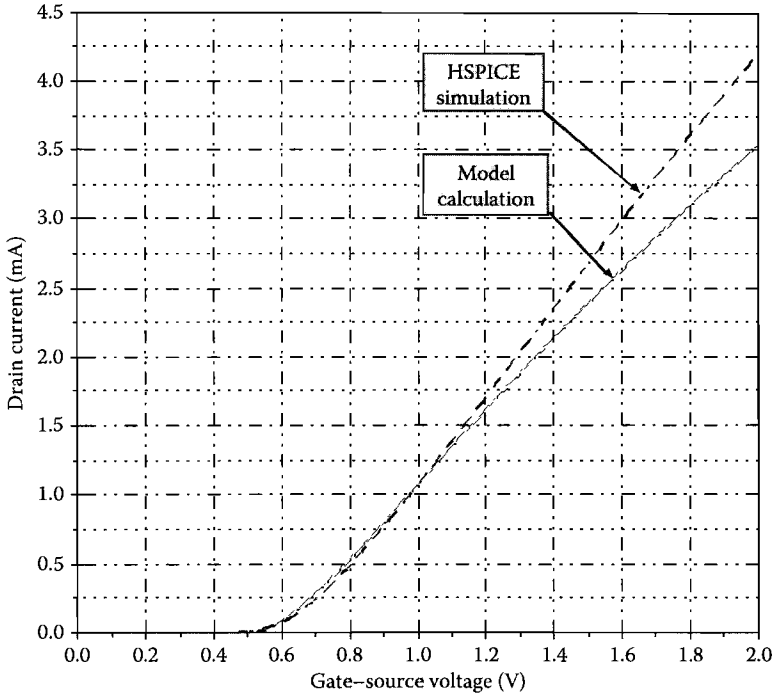

FIGURE 1.44 Simulated and calculated forward static transfer characteristic for the NMOS transistor whose model parameters are delineated in Table 1.2. The transistor is operated at a drain–source voltage, V_{ds}, of 1 V.

Comments: In Step #2 of the foregoing computational procedure, the gate aspect ratio, W/L, is the pivotal metric for achieving the desired transconductance and transistor drain current. If power dissipation is a dominant design concern, W/L can be increased above the value of 40 discerned in this example, with the understanding that the gate–source voltage, V_{gs}, can be reduced commensurately, thereby reducing the static drain current and hence, the power dissipation of the transistor. Of course, the primary penalty of large gate aspect ratio is a possible degradation of high-frequency circuit response since, as is confirmed by Equation 1.154, the capacitance area and peripheral dimensions increase in proportion to the gate width, W.

In Step #3, the metric, V_{ve}, is evaluated in terms of a purely empirical, and indeed crude first order, relationship to the oxide thickness, T_{ox}. A possible way around this dilemma is to compute V_{ve} and all of the other requisite modeling parameters by curve fitting Equation 1.222 to simulated or actually measured static data. While this approach may be academically satisfying, it may be imprudent from a design time perspective. Keep in mind that biasing is not the fundamental performance objective of an analog circuit; rather, biasing is the necessary condition that expedites the desired analog responses.

The drain saturation voltage, V_{dsat}, is obviously a nonlinear function of the excess gate voltage, $(V_{gs}-V_h)$, owing to the parameter, M_{sat}. But in addition, V_{dsat} changes slightly with the applied drain–source voltage, V_{ds}. Indeed, the Level 49 model parameters account for a slight sensitivity of threshold voltage on V_{ds}, which is as anticipated since the interface potential throughout the entire channel varies somewhat as a function of the lateral (drain-to-source) field engendered by V_{ds}.

Finally, it should be noted that the computed value (3.144 V) of the channel length modulation voltage, V_λ, is appreciably smaller than values often propounded in the textbook literature. However, V_λ is indeed a relatively small voltage for deep submicron MOS technology transistors. This anemic voltage is the principle cause of correspondingly small drain–source channel resistances, which renders the

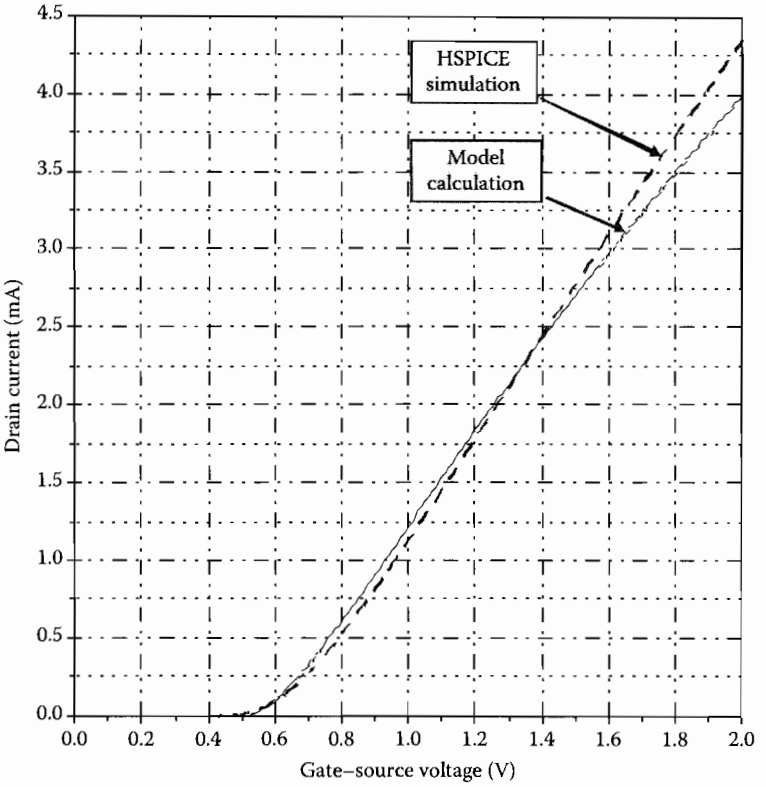

FIGURE 1.45 Simulated and calculated forward static transfer characteristic for the NMOS transistor whose model parameters are delineated in Table 1.2. The transistor is operated at a drain–source voltage, V_{ds}, of 1.5 V.

realization of transconductor amplifiers, as might be used in operational transconductor amplifier-capacitor (OTA-C) filters, a daunting challenge. The desire for accuracy surrounding the enumeration of V_λ is exacerbated by the fact that parameter V_λ is not the constant that is presumed tacitly in the foregoing demonstration. Instead, and as is suggested by Equation 1.114, V_λ is functionally dependent on drain–source voltage, drain saturation voltage, and threshold voltage. If V_λ or the drain–source channel resistance is critical in an analog circuit design endeavor, care must therefore be exercised to ensure that model parameters are extracted in terms of measured or simulated data that largely mirror the desired or expected operating state of the utilized transistor.

References

1. S. W. Sze, *Physics of Semiconductor Devices*. New York: John Wiley & Sons, 1969, pp. 366–379.
2. A. B. Glaser and G. E. Subak-Sharpe, *Integrated Circuit Engineering: Design, Fabrication, and Applications*. Reading, MA: Addison-Wesley Publishing Company, 1977, pp. 80–94.
3. A. S. Grove, *Physics and Technology of Semiconductor Devices*. New York: John Wiley & Sons, 1967, pp. 263–285.
4. A. Goetzberger, Ideal MOS curves for silicon, *Bell System Technical Journal*, 45, 1097, 1966.
5. S. R. Hofstein and G. Warfield, Physical limitation on the frequency response of a semiconductor surface inversion layer, *Solid State Electronics*, 8, 321, 1965.
6. A. S. Grove, E. H. Snow, B. E. Deal, and C. T. Sah, Simple physical model for the space-charge capacitance of metal–oxide–semiconductor structures, *Journal of Applied Physics*, 33, 2458, 1964.

7. R. L. Geiger, P. E. Allen, and N. R. Strader, *VLSI Techniques for Analog and Digital Circuits.* New York: McGraw-Hill Publishing Company, 1990, pp. 174–177.
8. D. Johns and K. Martin, *Analog Integrated Circuit Design.* New York: John Wiley & Sons, Inc, 1997, pp. 24–27.
9. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd Ed. Cambridge, United Kingdom: Cambridge University Press, 2004, Chaps. 11 and 12.
10. A. van der Ziel, *Noise in Solid State Devices and Circuits.* New York: John Wiley & Sons, Inc, 1986.
11. J. Choma and W.-K. Chen, *Feedback Networks: Theory and Applications.* Singapore: World Scientific Press, 2007, Chap. 3.

1.3 JFET, MESFET, and HEMT Technology and Devices

Stephen I. Long

1.3.1 Introduction

Many types of field effect devices are used in analog IC and RFIC design. Section 1.2 described the MOSFET and associated device models. MOSFETs are currently the predominant field effect device used in analog circuit applications due to the pervasive CMOS technology. CMOS fabrication is relatively inexpensive when not scaled below 0.25 μm. However, mask costs for 130 nm and below increase very rapidly, limiting applications to only those requiring extremely high volume. Also, drain breakdown voltage is quite low, of the order of 1 V for 65 nm CMOS. This constrains dynamic range or power output in certain applications.

Other field effect devices are available, but are considered niche market devices in most cases. This would include the legacy silicon JFET technology, still used in conjunction with bipolar transistors for some lower frequency analog applications. Compound semiconductor-based field effect devices (MESFET, HEMT, p-HEMT, m-HEMT) are often the FET of choice for applications requiring very wide bandwidth, extremely low noise, high gain at mm-wave frequencies, and high output powers at frequencies above 2 GHz. Cost of fabrication is frequently less than that of CMOS in smaller volume applications because the mask set costs are typically an order of magnitude less. Also, the compound semiconductor devices are grown on semi-insulating substrates. Passive components such as spiral inductors, MIM capacitors and deposited resistors have less parasitic capacitance and higher Q than is typical for silicon-based RFICs.

In this section, the silicon JFET and the main compound semiconductor HEMT devices will be described. Special emphasis will be placed on the GaN HEMTs whose performance is exceptionally good for microwave and mm-wave power amplifiers.

1.3.2 Silicon JFET Device Operation and Technology

Although the silicon JFET is today a legacy device, it is still used in some bipolar analog ICs to provide an inexpensive BiFET IC technology. Also, the description of its current–voltage characteristic is similar to any FET which uses a pn or Schottky metal–semiconductor junction for the gate electrode. The JFET consists of a conductive channel with source and drain contacts whose conductance is controlled by a gate electrode. The channel can be fabricated in either conductivity type, n or p, and both normally-on (depletion mode) and normally-off (enhancement mode) type devices are possible. The circuit symbols typically used for JFETs are shown in Figure 1.46 along with the bias polarities of active region operation for these four device possibilities. For analog circuit applications, the depletion mode is almost exclusively utilized because it provides a larger range of input voltage and therefore greater dynamic range. In silicon,

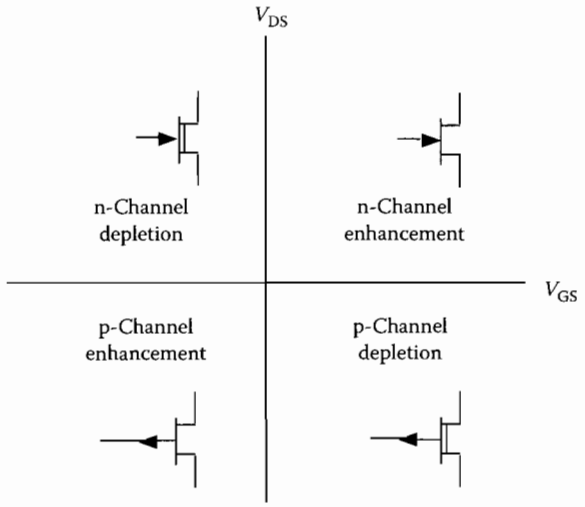

FIGURE 1.46 The circuit symbols typically used for JFETs are shown with the bias polarities for active region operation.

both p- and n-channel JFETs are used, but when compound semiconductor materials such as GaAs or InGaAs are used to build the FET, n-channel devices are used almost exclusively.

When fabricated with silicon, the JFET is used in analog IC processes for its high input impedance, limited by the depletion capacitance and leakage current of a reverse-biased pn junction. When the JFETs are used at the input stage, an op-amp with low input bias current, at least at room temperature, can be built. Fortunately, a p-channel JFET can be fabricated with a standard bipolar process with few additional process steps. This enables inexpensive BiFET processes to be employed for such applications. Unfortunately, the simple process modifications required for integrating JFETs and BJTs are not consistent with the requirements for high-performance devices. Short-gate lengths and high-channel doping levels are generally not possible. So the transconductance per channel width and the gain–bandwidth product of JFETs integrated with a traditional analog BJT process are not very good. The short-circuit current gain–bandwidth product (f_T) is about 50 MHz for an integrated p-channel JFET. The MOSFETs in a BiCMOS process are much better devices, however, a BiCMOS process does not often include both NPN and PNP BJTs needed for high-performance analog circuits.

Discrete silicon JFETs are available with much better performance because they can be fabricated with a process optimized for the JFET. Typical applications are for low-noise amplifiers up to the VHF/UHF range. Noise figures less than 0.1 dB can be obtained at low frequencies with high source impedances and 2 dB at high frequencies at the noise matching input condition with high performance discrete silicon JFETs. The low input gate current, I_G, which can be in the picoamp range, causes the shot noise (proportional to $\sqrt{I_G}$) component to be very low. The input equivalent noise current of the JFET is mainly due to input referred channel (Johnson) noise. This property gives very low noise performance when presented with a high source impedance. In this case, the JFET is often superior to a BJT for noise. For low source impedances, the BJT is generally better.

Compound semiconductor materials such as GaAs and InGaAs are used to fabricate JFET-like devices called metal-semiconductor FET (MESFETs) and high electron mobility transistor (HEMTs). The reason for using these materials is superior performance at high frequencies. These devices are unequaled for gain–bandwidth, ultralow noise, and power amplification at frequencies above 10 GHz and up to 300 GHz. Integrated analog microwave circuits are fabricated with these devices and are commercially available for use in low noise receiver and power amplifier applications. Some representative results will be summarized in Table 1.5.

1.3.2.1 JFET Static *I–V* Characteristics

The JFET differs in structure and in the details of its operation from the MOSFET discussed in Section 1.3.2. Figure 1.47 shows an idealized cross section of a JFET. The channel consists of a doped region, which can be either p- or n-type, with source and drain contacts at each end. The channel is generally isolated from its surrounding substrate material by a reverse biased p–n junction. The depletion regions are bounded in Figure 1.47 by dashed lines and are unshaded. The thin, doped channel region forms a resistor of width W into the page and height d. A gate electrode is located at the center of the channel, defined by a semiconductor region of opposite conductivity type of length L. An n-channel structure is shown here for purposes of illustration. The p-type gate constricts the channel, both through the depth of the diffusion or implant used to produce the gate and through the depletion layer formed at the p–n junction. The height of the channel can be varied by biasing the gate relative to the source (V_{GS}). A reverse bias increases the depletion layer thickness, reducing the channel height and the drain current. If V_{GS} is large enough that the channel is completely depleted, the drain current will become very small. This condition corresponds to the cutoff and subthreshold current regions of operation, and the V_{GS} required to cut-off the channel is called V_P, the pinch-off voltage. V_P corresponds to the threshold voltage that was defined for the MOSFET. Similarly, a forward bias between gate and channel can be used to increase drain current, up to the point where the gate junction begins to conduct. Most JFETs are designed to be depletion-mode (normally on); drain current can flow when $V_{GS}=0$ and they are normally operated with a reverse-biased gate junction. It is also possible, however, to fabricate enhancement-mode JFETs by use of a thinner or more lightly doped channel.

The pinch-off voltage is a sensitive function of the doping and thickness of the channel region. It can be found if the channel-doping profile, $N(x)$, is known through Poisson's equation. For a nonuniform profile,

$$V_P = V_{BI} - \frac{q}{\varepsilon}\int_0^d xN(x)\mathrm{d}x \tag{1.224}$$

For uniform doping, $N(x)=N_D$ and the familiar result in Equation 1.225 shows that the pinch-off voltage depends on the square of the thickness. This result shows that very precise control of profile depth is needed if good matching and reproducibility of pinch-off voltage is to be obtained [7].

$$V_P - V_{BI} - \frac{qN_D d^2}{2\varepsilon} \tag{1.225}$$

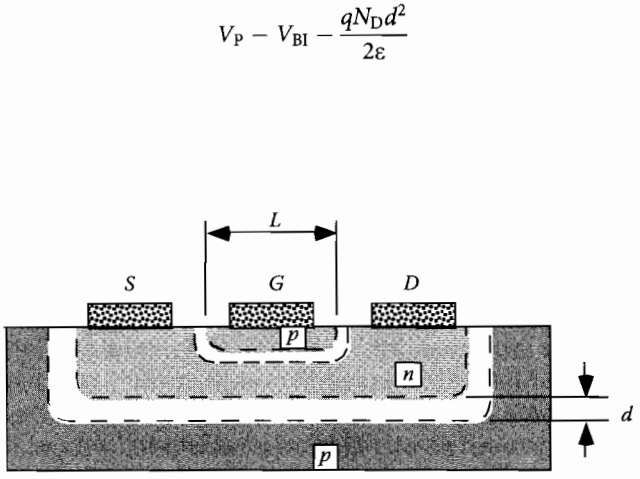

FIGURE 1.47 Idealized cross section of a JFET. The depletion regions are bounded with dashed lines and are unshaded.

1.3.2.2 JFET Operating Regions

The static current–voltage characteristics of the JFET can be categorized by the five regions of operation shown in Figure 1.48 for an n-channel device. The mechanisms that produce these regions can be qualitatively understood by referring to the channel cross sections in Figure 1.49. In these figures, the doped channel region is shaded, and the depletion region is white. First, consider the JFET in Figure 1.49a with small V_{DS} ($\ll V_{GS} - V_P$). This condition corresponds to the ohmic region (sometimes called linear or triode region) where current and voltage are linearly related. At small drain voltages, the depletion layer height is nearly uniform, the electric fields in the channel are too small to saturate the carrier velocity, and thus the channel behaves like a linear resistor. The resistance can be varied by changing V_{GS}. The channel height is reduced by increasing the reverse bias on the gate leading to an increased resistance.

As V_{DS} increases, the depletion layer thickness grows down the length of the channel as shown in Figure 1.49b. This occurs because the drain current causes a voltage increase along the channel as it flows through the channel resistance. Since the depletion layer thickness is governed by the gate-to-channel voltage (V_{GC}), there is an increasing reverse bias that leads to constriction of the channel at the drain end of the gate. Ideally, when $V_{DS} = V_{GS} - V_P$, then $V_{GC} = V_P$, and the channel height will approach zero (pinch-off). The constricted channel will cause the drain current to saturate as shown. Further increases in V_{DS} do not cause the drain current to increase since the channel has already constricted to a minimum height and the additional potential is accommodated by lateral extension of the depletion region at the drain end of the gate. This region of operation is generally described as the pinch-off region (rather than the saturation region in order to avoid confusion with BJT saturation). The height of the channel is not actually zero but is limited by the mobile channel charge, which travels at saturated drift velocity in this high field region.

If $V_{GS} < 0$, then the initial channel height at the source is reduced, I_D is less, and the pinch-off region occurs at a smaller drain voltage $V_{DS} = V_{GS} - V_P$. The saturation of drain current can also occur at smaller V_{DS} if the gate length is very small. In this case, the electric field in the channel is large, and the carrier velocity will saturate before the channel can reach pinch-off. Velocity saturation will also limit drain current.

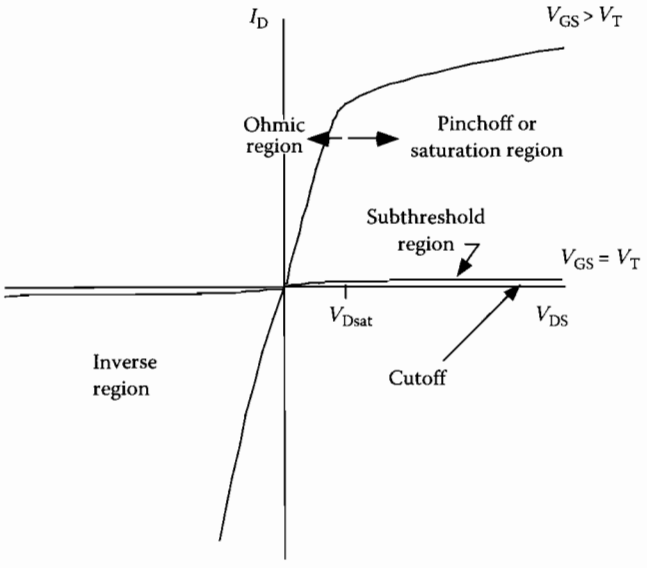

FIGURE 1.48 The static current–voltage characteristics of the JFET can be categorized by five regions of operation. An n-channel device is shown in this illustration.

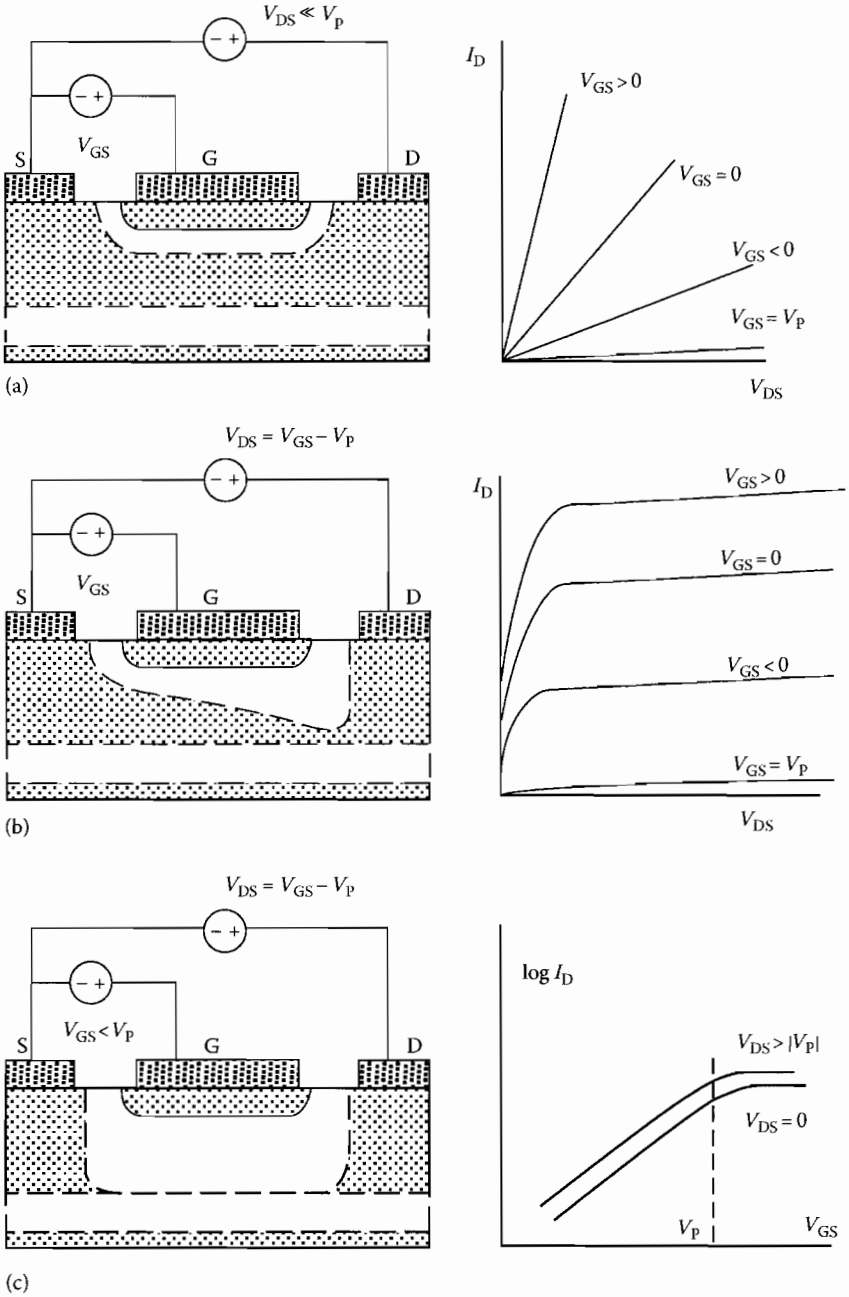

FIGURE 1.49 (a) Ohmic region with small V_{DS} ($\ll V_{GS} - V_P$). (b) When $V_{DS} = V_{GS} - V_P$, the channel height will become narrow at the drain end of the gate. The device enters pinch-off. The constricted channel will cause the drain current to saturate as shown. (c) Cutoff and subthreshold current regions occur when the depletion region extends through the channel.

The subthreshold region of operation, shown in Figure 1.49c is defined when small drain currents continue to flow even though $V_{GS} \leq V_P$. While technically this gate bias should produce cutoff, some small fraction of the electrons from the source region will have sufficient energy to overcome the

potential barrier caused by the gate depletion region and will drift into the drain region and produce a current. Since the energy distribution is exponential with potential, the current flow in this region varies exponentially with V_{GS}.

The inverse region occurs when the polarity of the drain bias is reversed. This region is of little interest for the JFET since gate-to-drain conduction of the gate diode limits the operation to the linear region only.

1.3.2.3 Channel-Length Modulation Effect

A close look at the I–V characteristic in the pinch-off region shows that the incremental conductivity or slope of this region is not equal to zero. There is some finite slope that is not expected from the simple velocity saturation or pinch-off models. Channel length modulation is one explanation for this increase; the position under the gate where pinch-off or velocity-saturation first occurs moves toward the source as V_{DS} increases. This is due to the expansion of the drain side depletion region at large V_{DS}. Figure 1.50 illustrates this point. Here, a channel cross section is shown for $V_{DS} = V_{GS} - V_P$ in Figure 1.50a and for $V_{DS} \gg V_{GS} - V_P$ in Figure 1.50b. While pinch-off always occurs when the gate-to-channel voltage is V_P, the higher drain voltage causes the location of this point ($x = L$) to move closer to the source end of the channel. Since the electric field in this region, E, is roughly proportional to $(V_{GS} - V_P)/L$ where L is now a function of V_{DS} and V_{GS} and the carrier velocity $v = \mu E$ (by assumption), then the current must increase as the channel length decreases due to increasing carrier velocity. If the channel length is short, velocity saturation may cause the drain current to saturate. In this case, the velocity saturation point moves closer to the source as drain voltage is increased. Since the length has decreased, less gate-to-channel voltage is needed to produce the critical field for velocity saturation. Less voltage implies a wider channel opening, hence more current.

1.3.2.4 Temperature Effects

There are two mechanisms that influence the drain current of the JFET when temperature is changed [8,9]. First, the pinch-off voltage becomes more negative (for n-channel) with increase in temperature, therefore requiring lower V_{GS} to cut off the channel or to enter the pinch-off region. Therefore, when the device is operating in the pinch-off region, and $V_{GS} - V_P$ is small, the drain current will increase with temperature. This effect is caused by the decrease in the built-in voltage of the gate-to-channel junction with increasing temperature. Second, the carrier mobility and saturated drift velocity decreases with temperature. This causes a reduction in drain current that is in opposition to the first effect. This effect dominates for large $V_{GS} - V_P$. Therefore, there is a V_{GS} value for which the drain current is exactly compensated by the two effects. This is illustrated qualitatively in Figure 1.51.

The gate current is also affected by temperature, as it is the reverse current of a pn junction. The current increases roughly by a factor of 2 for each 10°C increase in temperature. At high temperatures, the input current of a JFET input stage may become comparable to that of a well-designed BJT input stage of an op-amp, thus losing some of the benefit of the mixed BJT–JFET circuit design.

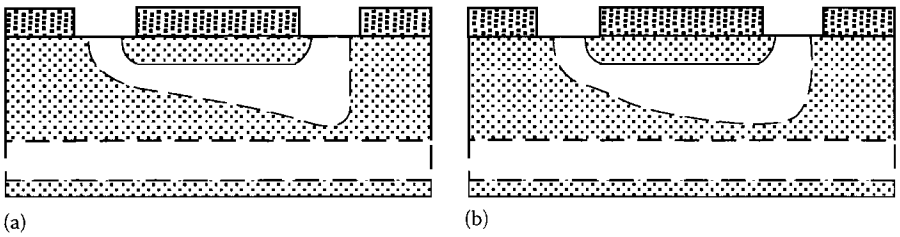

FIGURE 1.50 A channel cross section is shown for $V_{DS} = V_{GS} - V_P$ in (a) and for $V_{DS} \gg V_{GS} - V_P$ in (b). While pinch-off always occurs when the gate-to-channel voltage is V_P, the higher drain voltage causes the location of this point ($x = L$) to move closer to the source end of the channel.

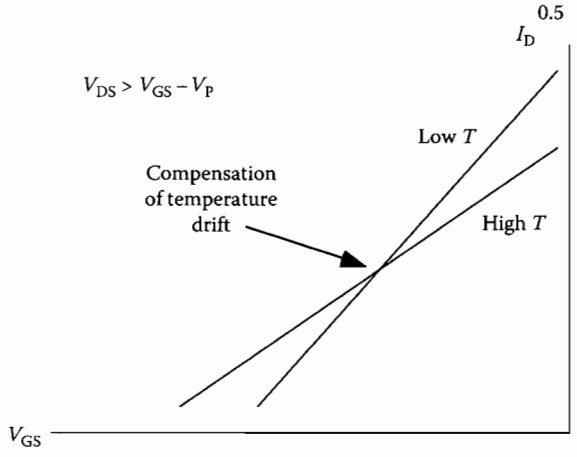

FIGURE 1.51 Effect of temperature on the drain current in the pinch-off region.

1.3.2.5 JFET Models

Most applications of the JFET in analog ICs employ the pinch-off region of operation. It is this region that provides power gain and buffer (source follower) capability for the device, so the models for the JFET presented below will concentrate on this region. It will also be assumed that the gate–source junction will not be biased into forward conduction. Although forward conduction is simple to model using the ideal diode equation within the FET equivalent circuit models, this bias condition is not useful for the principal analog circuit applications of the JFET and will also be avoided in the discussion that follows.

1.3.2.5.1 Large-Signal Model: Drain Current Equations

Equations modeling the large signal JFET $I_D - V_{GS}$ characteristic can be derived for the two extreme cases of FET operation in the pinch-off region. A gradually decreasing channel height and mobility limited drift velocity in the channel are appropriate assumptions for very long gate length FETs. A fixed channel height at pinch-off with velocity saturation limited drift velocity are more suitable for short gate lengths.

The square-law transfer characteristic [10] given by Equation 1.226 provides a good

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_P}\right)^2 (1 + \lambda V_{DS}) \tag{1.226}$$

approximation to measured device characteristics in the case of long gate length (>5 μm) or very low electric fields in the channel $(V_{GS} - V_P)/L < E_{sat}$. In both cases, the channel height varies slowly and the velocity remains proportional to mobility. E_{sat} is the critical field for saturation of drift velocity, about 3.5 kV/cm for GaAs and 20 kV/cm for Si. I_{DSS} is defined as the drain current in the pinch-off region when $V_{GS} = 0$. The first two terms of the equation are useful for approximate calculation of DC biasing. The third term models the finite drain conductance caused by the channel length modulation effect. The parameter λ in this term is derived from the intercept of the drain current when extrapolated back to zero as shown in Figure 1.52.

Equation 1.226 is also used to represent the pinch-off region in the SPICE JFET model. It is parameterized in a slightly different form as shown below in Equation 1.227.

$$I_D = \beta(V_{GS,i} - V_{T0})^2(1 + \lambda V_{DS}) \tag{1.227}$$

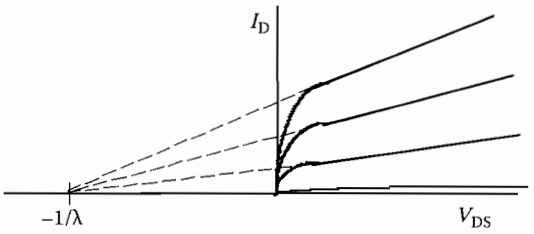

FIGURE 1.52 The channel length modulation parameter λ is defined by the extrapolation of the drain current in saturation to $I_D = 0$.

These equations are the same if $V_{T0} = V_P$, and

$$\beta = \frac{I_{DSS}}{V_P^2} \tag{1.228}$$

and

$$\begin{aligned} V_{GS,i} &= V_{GS} - I_D R_S \\ V_{DS,i} &= V_{DS} - I_E(R_S + R_D) \end{aligned} \tag{1.229}$$

The pinch-off region is defined for $V_{DS,i} \geq V_{GS,i} - V_{T0}$ as is usual for the gradual channel approximation. R_S and R_D are the parasitic source and drain resistances associated with the contacts and the part of the channel that is outside of the gate junction. These resistances will reduce the internal device voltages below the applied terminal voltages as shown in Equations 1.229.

For shorter gate length devices, improved models have been proposed and implemented in SPICE3 and some of the many commercial SPICE products, often in the MESFET model. The Statz model [11] is frequently used for this purpose. This model modifies the drain current dependence on V_{GS} by adding a velocity saturation model parameter b in the denominator as shown in Equation 1.230.

$$I_D = \left[\frac{\beta(V_{GS,i} - V_{T0})^2}{1 + b(V_{GS,i} - V_{T0})}\right](1 + \lambda V_{DS,i}) \tag{1.230}$$

This added term allows the drain current to be nearly square law in V_{GS} for small $V_{GS} - V_{T0}$, but it becomes almost linear when V_{GS} is large, effectively emulating the rapid rise in transconductance followed by saturation that is typical in short channel devices. Although the specific behavior of the drain current is sensitive to the vertical doping profile in the channel), Equation 1.230 is flexible enough to accommodate most short channel device characteristics with uniform or nonuniform channel doping. Another feature of short gate length FETs that this model predicts adequately is a saturation of I_D at $V_{DS,i} < V_{GS,i} - V_{T0}$. This early transition into the pinch-off region is also a consequence of velocity saturation and is widely observed.

1.3.2.5.2 Small-Signal Model

The small-signal model for the JFET in the pinch-off region is shown in Figure 1.53. The voltage dependent current source models the transconductance g_m as a constant which can be derived from the drain current equations above from

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{1.231}$$

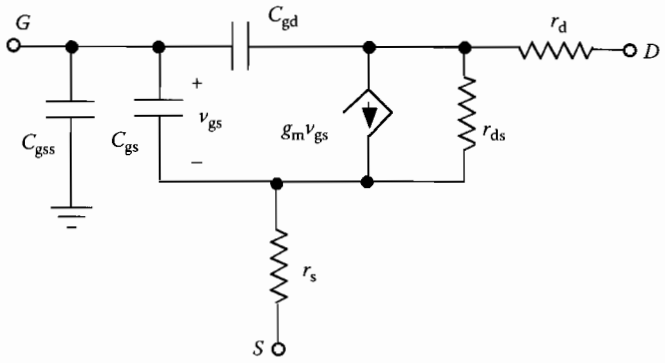

FIGURE 1.53 Small-signal model for the JFET in the pinch-off region.

The square-law current model (Equation 1.226) predicts a linearly increasing g_m with V_{GS}

$$g_m = -\frac{2I_{DSS}}{V_P}\left(1 - \frac{V_{GS}}{V_P}\right) \tag{1.232}$$

whereas a model which includes some velocity saturation effects such as Equation 1.230 would predict a saturation in g_m.

The small-signal output resistance, r_o, models the channel length modulation effect. This is also derived from the drain current equations through

$$r_o^{-1} = \frac{\partial I_D}{\partial V_{DS}} \tag{1.233}$$

For both models, r_o is determined by

$$r_o = \frac{1}{I_D \lambda} \tag{1.234}$$

The small-signal capacitors representing the nonlinear, voltage-dependent C_{gs}, C_{gd}, and C_{gss} are also shown in Figure 1.53. Parasitic source and drain resistances, R_S and R_D can also be included, as shown. If they are not included in the small-signal model, the effect of these parasitics can sometimes be produced in the intrinsic FET model by reducing the intrinsic g_m of the device.

The short-circuit current gain–bandwidth product, f_T, defined in Equation 1.235 is a high-frequency figure of merit for transistors. It is inversely proportional to the transit time τ of the channel charge, and it is increased by reducing the gate length. Reduced L also reduces the gate capacitance and increases transconductance. The material also affects f_T as higher drift velocity leads to higher g_m.

$$f_T = \frac{g_m}{2\pi(C_{gss} + C_{gs} + C_{gd})} = \frac{1}{\tau} \tag{1.235}$$

1.3.2.6 Silicon JFET Technologies

The IC fabrication technology used to make JFETs depends primarily on the material. Discrete Si JFETs are available that provide f_T above 500 MHz and very low input rms noise currents through optimizing the channel design and minimizing parasitic capacitances, resistances, and gate diode leakage currents.

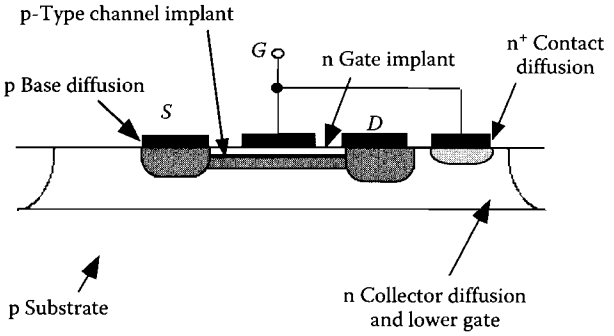

FIGURE 1.54 Cross section of an ion implanted silicon JFET (not to scale).

However, a silicon IC process is rarely designed to optimize the performance of the JFET; rather, the JFET is made to accommodate an existing bipolar process with as few modifications as possible [10]. Then, the extra circuit design flexibility and performance benefits of a relatively inexpensive mixed FET/BJT process (often called BiFET) can be obtained with small incremental cost.

In principle, it would be possible to build p-channel Si JFETs in a standard analog BJT process without additional mask steps if the base diffusion had suitable doping and thickness to give a useful pinch-off voltage when overlaid with the emitter diffusion. Unfortunately, this is usually not the case, since the emitter diffusion is too shallow, and the pinch-off voltage resulting from this approach would be too high (positive in the case of the p-channel device). Therefore, the channel of the JFET must be made thinner either through the use of an additional diffusion or by providing the channel and gate with ion implantations.

In analog ICs applications, silicon JFETs are passengers on a bipolar process; they must be compatible with the BJT process that they inhabit. Most flexibility in the JFET design is achieved using the ion implantation method. Figure 1.54 illustrates the cross section of an ion implanted JFET. In order to gain good control of the pinch-off voltage and transconductance, both the channel and the gate are formed by ion implantation. In addition, the forced compatibility with the BJT process requires use of the collector layer under the channel. This forms a lower gate electrode which is less heavily doped than the channel. Therefore, the depletion region at this interface extends primarily into the collector region, and the lower gate is less effective in contributing to the total transconductance of the JFET. It does add the parasitic capacitance C_{gss} to the device at the collector to substrate junction, limiting frequency response. In addition, the predeposition of channel and gate charge is much more repeatable with ion implantation than with earlier double diffusion methods, so device matching and reproducibility of pinch-off voltage is greatly improved. The f_T will be improved by the larger g_m per unit width and the slightly reduced gate capacitances, and the drain breakdown voltage will be increased as is often needed for an analog IC process. However, low-channel doping is not a good recipe for a high-frequency transistor with short gate length, so the f_T of these devices is still only 50 MHz or so.

1.3.3 Compound Semiconductor FET Technologies

An introduction to compound semiconductor materials will be presented in this section to establish the underlying rationale for using these materials for extremely high-performance MMIC and RFIC applications. The transport properties of typical III–V materials are compared with silicon and SiGe alloys.

There is no denying that silicon is the workhorse of the semiconductor industry. Large, high-quality substrates are relatively inexpensive, a highly stable oxide can be grown with low interface state density, and a highly advanced processing technology has enabled extremely large circuit density and extremely

fine lines to be achieved with low parasitic capacitances. Its greatest weakness for electronic device applications is the relatively low electron velocity and mobility. These intrinsic properties lead to higher transit times and access resistances, respectively, a limitation on high frequency device performance. The deeply scaled submicron technology has compensated for these deficiencies to some degree by aggressive reduction in gate length or base width. Also, p-SiGe has higher hole mobility than p-Si, so access resistance can be improved. And, using the strain induced by local depositions of SiGe in MOSFETs increases electron and hole mobilities. As good as Si IC technology is, there exist compound semiconductor materials whose intrinsic electron velocity and mobility are greatly superior to Si and so can potentially offer higher frequency, higher speed or higher power performance.

The III–V FET and bipolar device technology can provide the highest frequency and lowest noise circuit applications. Its main limitation is density. Device footprints are often significantly larger than those of similar Si devices. Thus, the high intrinsic performance of these devices is achieved in circuits of relatively low complexity.

1.3.3.1 Defining III–V Compound Semiconductors

The compound semiconductor family, as traditionally defined, is composed of the group III and group V elements shown in Table 1.3 [12]. Each semiconductor is formed from at least one group III and one group V element.

The main motivation for using the III–V compound semiconductors for device applications is found in their electronic properties when compared with those of the dominant semiconductor material, silicon. Figure 1.55 is a plot of steady-state electron velocity of several n-type semiconductors versus electric

TABLE 1.3 The Group II–VI Elements

II	III	IV	V	VI
Be	B	C	N	O
Mg	Al	Si	P	S
Zn	Ga	Ge	As	Se
Cd	In	Sn	Sb	Te

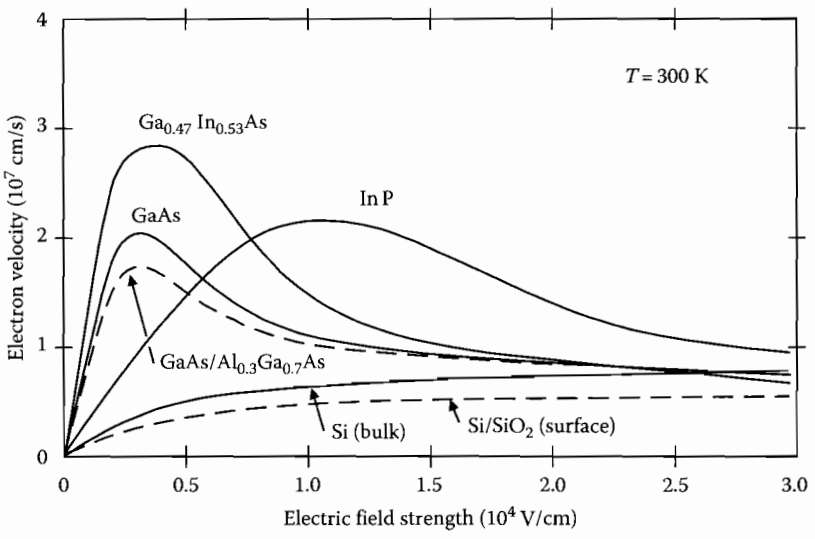

FIGURE 1.55 Electron velocity versus electric field for several n-type semiconductors.

field [12]. From this graph, we see that at low electric fields the slope of the III–V semiconductor curves (mobility) is higher than that of silicon. High mobility means that the semiconductor resistivity will be less for III–V n-type materials, and it therefore will be easier to achieve lower access resistance. Access resistance is the series resistance between the device contacts and the internal active region. An example would be the base resistance of a bipolar transistor or source resistance of a FET. Lower resistance will reduce some of the fundamental device time constants that often dominate device high frequency performance. Figure 1.55 also shows that the peak electron velocity is higher for the III–V's, and the peak velocity can be achieved at much lower electric fields. High velocity reduces transit time, the time required for a charge carrier to travel from its source to its destination, and improves device high-frequency performance. Achieving this high velocity at lower electric fields means that the devices will reach their peak performance at lower voltages, useful for low power, high-speed applications. Higher velocity of electrons also increases the current density of a device since current is the product of charge and velocity. Mobility and peak velocities of several semiconductors are compared in Table 1.4 [12].

The higher velocities are a consequence of the band structure of III–V materials. Since Si is an indirect bandgap material, conduction electrons reside in a high effective mass conduction band (CB). Mobility is dominated by the high effective mass. At high electric fields, the optical phonon generation process limits the maximum achievable electron drift velocity. GaAs, on the other hand, is direct gap, the electron mobility is high because of the lower energy, low effective mass CB where conduction electrons are confined at low fields. However, the average electron velocity will be reduced at higher electric fields due to scattering into the higher mass CB. This produces a saturated drift velocity less than the peak drift velocity, typical of the direct-gap III–V's.

To obtain significant transit velocity improvement over silicon, one must use a ternary III–III–V semiconductor such as InGaAs. The high effective mass CB is separated by 50% of the bandgap for InGaAs, whereas for GaAs it was only 20%. Thus, the peak velocity in InGaAs can be much higher than GaAs because more energy can be transferred to the conduction electrons before they begin scattering to the high mass CB. This results in higher peak velocity, 2.7×10^7 cm/s vs. 2×10^7 cm/s for GaAs.

Also shown in Table 1.4, p-type III–V semiconductors have rather poor hole mobility when compared with elemental semiconductor materials such as silicon or germanium. Holes also reach their peak velocities at much higher electric fields than electrons. Consequently, there has been very little use of p-channel III–V FET devices. The only reason to use compound semiconductor FETs is their superb high-frequency performance. The p-channel devices cannot provide this.

TABLE 1.4 Electronic Properties of Compound Semiconductors Compared with Si and Ge

Semiconductor	E_G (eV)	ε_r	Electron Mobility (cm^2/V-s)	Hole Mobility (cm^2/V-s)	Peak Electron Velocity (cm/s)
Si (bulk)	1.12	11.7	1,450	450	NA
Ge	0.66	15.8	3,900	1,900	NA
InP	1.35 D	12.4	4,600	150	2.1×10^7
GaAs	1.42 D	13.1	8,500	400	2×10^7
$Ga_{0.47}In_{0.53}As$	0.78 D	13.9	11,000	200	2.7×10^7
InAs	0.35 D	14.6	22,600	460	4×10^7
$Al_{0.3}Ga_{0.7}As$	1.80 D	12.2	1,000	100	—
AlAs	2.17	10.1	280	—	—
$Al_{0.48}In_{0.52}As$	1.92 D	12.3	800	100	—
GaN	3.39D	9.0	1,500	30	$2.5–2.7 \times 10^7$
SiC (4H)	3.26	9.8	500		2.2×10^7

Note: In bandgap energy column the symbol "D" indicates direct bandgap, otherwise it is indirect bandgap.

1.3.3.2 Heterojunctions

Heterojunctions provide an additional degree of freedom that is widely used to improve performance of compound semiconductor FET devices. The heterojunction formed by an atomically abrupt transition between AlGaAs and GaAs, shown in the energy band diagram of Figure 1.56 [12], creates discontinuities in the valence and CBs. The CB energy discontinuity is labeled ΔE_C and the valence band discontinuity, ΔE_V. Their sum equals the energy bandgap difference between the two materials. The potential energy steps caused by these discontinuities are used as barriers to electrons or holes. The relative sizes of these potential barriers depend on the composition of the semiconductor materials on each side of the heterojunction. In this example, an electron barrier in the CB is used to confine carriers into a narrow potential energy well with triangular shape. Quantum well structures such as these are used to improve device performance through two-dimensional charge transport channels, similar to the role played by the inversion layer in MOS devices. The structure and operation of heterojunctions in FETs will be described in Section 1.3.3.

The overall principle of the use of heterojunctions is summarized in a *Central Design Principle*:

> Heterostructures use energy gap variations in addition to electric fields as forces acting on holes and electrons to control their distribution and flow [13,14].

The energy barriers can control motion of charge both across the heterojunction and in the plane of the heterojunction. In addition, heterojunctions are most widely used in light emitting devices since the compositional differences also lead to either stepped or graded index of refraction, which can be used to confine, refract, and reflect light. The barriers also control the transport of holes and electrons in the light generating regions.

Figure 1.57 shows a plot of bandgap versus lattice constant for many of the III–V semiconductors [12]. Consider GaAs as an example. GaAs and AlAs have the same lattice constant (approximately 0.56 nm) but different band gaps (1.4 and 2.2 eV, respectively). An alloy semiconductor, AlGaAs, can be grown epitaxially on a GaAs substrate wafer using standard growth techniques. The composition can be selected by the Al to Ga ratio giving a bandgap that can be chosen across the entire range from GaAs to AlAs. Since both lattice constants are essentially the same, very low lattice mismatch can be achieved for any composition of $Al_xGa_{1-x}As$. Lattice matching permits low defect density, high quality materials to be grown that have good electronic and optical properties.

It quickly becomes apparent from Figure 1.57, however, that a requirement for lattice matching to the substrate greatly restricts the combinations of materials available to the device designer. For electron devices, the low mismatch GaAs/AlAs alloys, GaSb/AlSb alloys, and ternary combinations $GaAs/Ga_{0.49}In_{0.51}P$ and $InP/In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ alone are available. Efforts to utilize

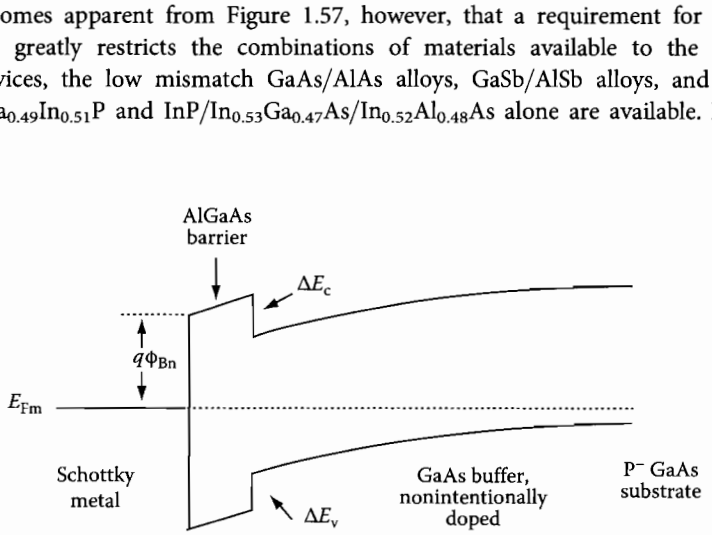

FIGURE 1.56 Energy band diagram of an abrupt heterojunction. Typical AlGaAs/GaAs HEMT band diagram.

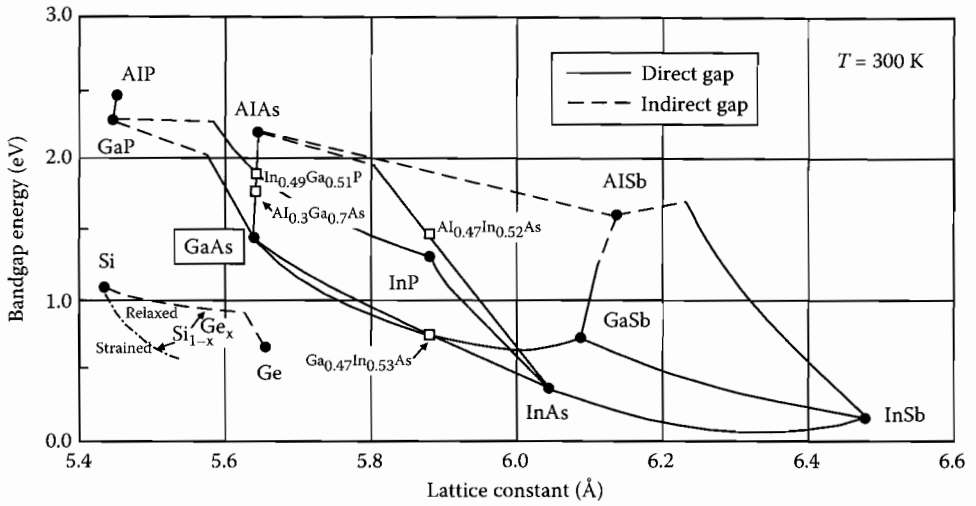

FIGURE 1.57 Energy bandgap versus lattice constant for compound semiconductor materials.

combinations such as GaP on Si or GaAs on Ge that lattice match have been generally unsuccessful because of problems with interface structure, polarization, and autodoping.

For several years, lattice matching was considered to be a necessary condition if mobility-damaging defects were to be avoided. This barrier was later broken when it was discovered that high quality semiconductor materials could still be obtained although lattice-mismatched if the thickness of the mismatched layer is sufficiently small [15,16]. This technique, called pseudomorphic growth, opened another dimension in III–V device technology, and allowed device structures to be optimized over a wider range of bandgap for better electron or hole dynamics and optical properties.

Two of the pseudomorphic systems that have been very successful in high performance millimeter-wave FETs are the InAlAs/InGaAs/GaAs and InAlAs/InGaAs/InP systems. The $In_xGa_{1-x}As$ layer is responsible for the high electron mobility and velocity which both improve as the In concentration x is increased. Up to $x = 0.25$ for GaAs substrates and $x = 0.80$ for InP substrates have been demonstrated and result in great performance enhancements when compared with lattice-matched combinations. [6]

InP substrates, however, are more expensive, smaller, and more easily broken than GaAs. And, the 3.8% lattice mismatch would seem to be too great for direct epitaxy of $In_{0.53}Ga_{0.47}As$ on GaAs substrates. It has been demonstrated, however, that good quality devices can be obtained using the metamorphic growth technique. A thick InP transition layer or a graded InGaP layer is grown directly upon a GaAs substrate. The defects caused by the lattice mismatch are largely contained in this layer, and low defect layers can be obtained when grown upon this transitional buffer layer [17,18].

1.3.3.3 Compound Semiconductor HEMT Devices

High performance GaAs MESFET* and HEMT† devices are constructed with a metal-to-semiconductor junction gate instead of a diffused or implanted pn junction gate. The metal gate forms a Schottky barrier diode directly on an n-type channel or on a wider bandgap barrier layer. In the case of the MESFET as shown in Figure 1.58 [19], the gate, directly on the n-type doped channel, forms a depletion layer which allows the channel height to be varied in the same manner as the JFET. No gate dielectric or p-type diffusion is necessary. Often the gate is deposited in a recess, etched below the surface of the channel. This allows for thicker and sometimes more highly doped regions at source and drain to be used to

* Metal–semiconductor FET.

† High electron mobility transistor.

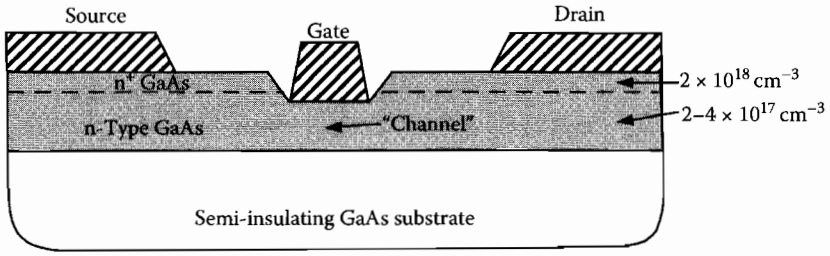

FIGURE 1.58 Cross section of recessed gate GaAs MESFET. (From Estreich, D., in *The VLSI Handbook*, CRC Press, Boca Raton, FL, 2006.)

reduce parasitic resistances. When gate cross sections are very small, for example less than 150 nm, the gate metal is taller than the width, and a thicker, wider region is often deposited on the top to reduce gate access resistance.

With the HEMT device, the gate potential modulates the height of a triangular potential well (Figure 1.56) thereby varying the channel charge available for source–drain conduction. The channel layer is confined by the triangular potential well formed at the interface between the higher bandgap barrier (InAlAs or AlGaAs) and channel (InGaAs or GaAs) as illustrated in the device cross sectional drawing in Figure 1.59 [19]. In some devices, the back side of the channel is also confined by a wide gap barrier. The confinement provided by these energy barriers provides large channel electron sheet concentrations, improving g_m and current density. The active region of the HEMT is formed by epitaxial growth of the channel and barrier region with molecular beam epitaxy.

In Figure 1.59, the device is also shown with a recessed gate. This type of structure enables the use of more highly doped, lower bandgap material on the surface to reduce parasitic source and drain resistances.

These compound semiconductor FETs are used as the primary active device in analog microwave and mm-wave monolithic integrated circuits (MMICs or RFICs). Extremely low noise figure and wide bandwidth have been obtained by the use of HEMT, p-HEMT (pseudomorphic HEMT), and m-HEMT (metamorphic HEMT) devices. These devices achieve their improved performance mainly through the high mobility, undoped InGaAs channel material. The electron velocity vs. electric field of $In_{0.53}Ga_{0.47}As$ is compared with GaAs and Si in Figure 1.55 where it can be seen that higher drift velocity is obtained in $In_{0.53}Ga_{0.47}As$ [20] than either GaAs [21] or Si [22]. The higher the In concentration in the InGaAs, the higher the mobility and velocity and the lower the noise.

Finally, the gate barrier heterojunction also enables good Schottky gate characteristics to be obtained even though the channel material itself has a low bandgap and would otherwise provide a poor barrier height if the metal were in direct contact.

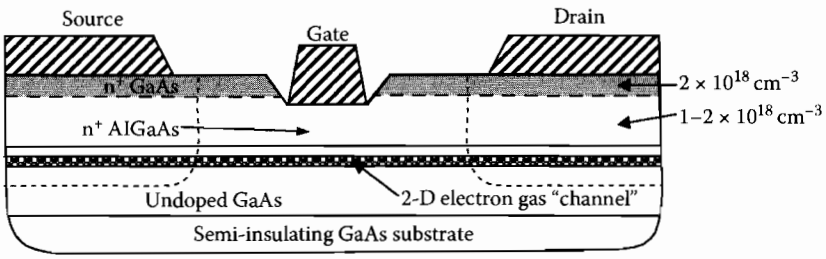

FIGURE 1.59 Cross section of recessed gate AlGaAs/GaAs HEMT device structure. (From Estreich, D., in *The VLSI Handbook*, CRC Press, Boca Raton, FL, 2006.)

TABLE 1.5 Microwave and mm-Wave Performance Comparison between Compound Semiconductor FETs

Frequency (GHz)	Device	Gain (dB)	Noise Figure (dB)	Reference
4–9	Dual Gate	21	<1.75	[1]
10–20	100 nm	17	<2.75	
20–40	GaAs pHEMT	20	<2.5	
23	130 nm	43	1.9	[2]
18–40	InP HEMT	>40	—	
0.5–80	100 nm InP HEMT	>17	<2.5	[3]
70–105	50 nm	20	2.4	[4]
220	m-HEMT	21	9	
192–235	50 nm m-HEMT	>15	—	[5]
270	35 nm	11.6	—	[6]
300	InP p-HEMT	6		

Excellent performance of HEMT, p-HEMT, and m-HEMT MMICs at microwave and millimeter wave frequencies has been reported. Table 1.5 presents a summary of some representative MMIC amplifiers where both narrowband and wideband amplifiers are reported. Gate lengths down to 35 nm have been successfully used for mm-wave and sub-mm-wave amplifiers.

GaAs HEMT devices also exhibit higher breakdown voltages (often 15 to 20 V) that make them suitable for power amplifier applications. In a recent article, a wideband distributed GaAs HEMT amplifier was reported that provided over 4 W of output power from 4 to 18 GHz at a 5 V drain bias with a power added efficiency of 23% [23]. Narrowband HEMT amplifiers can provide much higher efficiency and power.

1.3.3.4 Wide Bandgap Compound Semiconductors

In recent years there has been increasing interest in the wide bandgap compound semiconductors, SiC and GaN (and associated alloys of Al/In/Ga with N). The applications have been primarily for microwave power applications because the wider bandgap increases breakdown voltage while the band structure allows for high electron peak velocities in both materials [24,25]. Table 1.6 compares the fundamental physical properties of the wide-gap compound semiconductors with GaAs and Si [12]. It should be noted that there is not uniform agreement on the wide-gap parameter values from one reference to the next, but the numbers presented are representative of the current literature. As seen in Table 1.6, thermal conductivity is very high for both SiC and GaN, allowing for effective removal of heat from power devices. In fact, at room temperature, SiC has a higher thermal conductivity than any metal.

Figure 1.60 compares the electron velocity of GaN and SiC with GaAs and silicon [12,26]. The peak velocity of GaN is reached at electric fields above 150 kV/cm. Both SiC and GaN retain their good

TABLE 1.6 High Electron Mobility GaN

Material	Bandgap (eV)	Mobility (cm^2/V-s)	E_c (V/cm)	Saturation Drift Velocity (cm/s)	Thermal Conductivity (W/cm-K)
n-SiC (4H)	3.26	500	2.2×10^6	2×10^7	3.0–3.8
n-GaN	3.39	1500	3×10^6	1.5×10^7	2.2
n-GaAs	1.4	5000	3×10^5	0.6×10^7	0.45
n-Si	1.1	1300	2.5×10^5	1×10^7	1.45

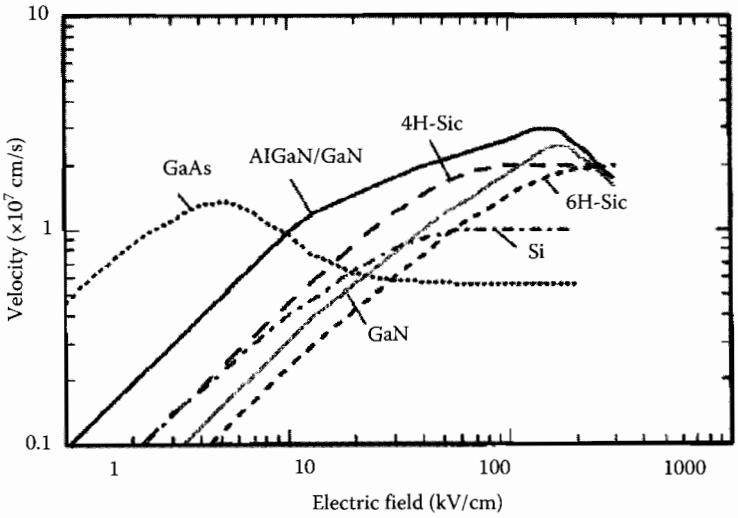

FIGURE 1.60 Electron velocity versus electric field of GaN and SiC compared with Silicon and GaAs. (From Trew, R.J., *Proc. IEEE*, 90, 1032, 2002. With permission.)

transport properties for high power applications. Table 1.6 shows that GaN has high electron mobility as well, which helps to reduce parasitic source resistance. Hole mobility for wide bandgap compound semiconductors is quite low, however, generally less than 50 $cm^2/V-s$.

There is significant lattice mismatch between a GaN channel in a heterojunction FET and the AlGaN barrier layer. However, the strain caused by this mismatch produces polarization and piezoelectric effects that induce large sheet charge densities, above $10^{13}\,cm^{-2}$ in the channel, beneficial for high current density operation of these devices [26]. This level of charge is about five times higher than what can be induced in GaAs channels in the AlGaAs/GaAs heterostructure.

Figures of merit are often employed when comparing materials for microwave power amplifier applications. Johnson's FOM [27]

$$\text{JFOM} = \frac{E_c v_{sat}}{2\pi} \tag{1.236}$$

has units of power–frequency. E_c is the maximum or critical electric field for breakdown and v_{sat} is the saturated drift velocity at high electric fields. This expresses the electronic merits of the material but neglects to consider thermal conductivity, also of importance for power electronics. Nevertheless, based on the electronic properties alone, GaN and SiC have a JFOM approximately 18 times greater than Si or GaAs. If the superior thermal conductivity is also considered, it becomes clear that these materials are extremely well suited for microwave power.

1.3.3.5 GaN HEMT Field Effect Transistors

The unusual electronic and thermal properties of the wide bandgap materials such as GaN are very attractive for applications requiring transistors with both high breakdown voltage and high frequency performance. This range of applications is focused especially on microwave power transistors. To understand why, consider first what is most desirable in a high power transistor's drain current–voltage characteristics. Figure 1.61 shows an idealized representation of this characteristic. To obtain the maximum output sinusoidal voltage and current amplitudes, one would like a device with high breakdown voltage, V_{break}, low "knee" voltage, V_{knee}, and high maximum current, I_{max}. The voltage swing is

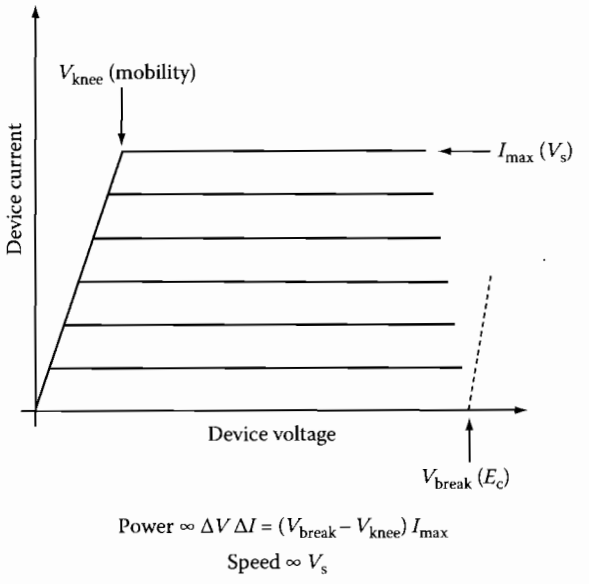

FIGURE 1.61 Idealized GaN HEMT drain current–drain voltage characteristic.

determined by $V_{break} - V_{knee}$, and the current swing by I_{max}. This combination provides high power density (W/mm) in a device, therefore requiring smaller device area. Power is proportional to

$$P_{OUT} \propto (V_{break} - V_{knee}) \times I_{max}.$$

DC to RF conversion efficiency, equally important in a power device, is proportional to

$$\text{Efficiency} \propto (1 - V_{knee}/V_{break}).$$

Thus, a large breakdown voltage is helpful for both power and efficiency.

But, Si LDMOS devices also have high breakdown voltage. So, why is GaN better? First, the current density is far higher in GaN. The high current density in the GaN HEMT is a result of the exceptionally high sheet charge, n_S, in the channel, typically 1×10^{13} cm^{-2} or higher. The high sheet charge density is a result of the static polarization that occurs at the interface between the GaN channel and the AlGaN barrier. To satisfy the charge balance at the interface, a high density of negative charge is required. A cross section of a GaN HEMT device is shown in Figure 1.62.

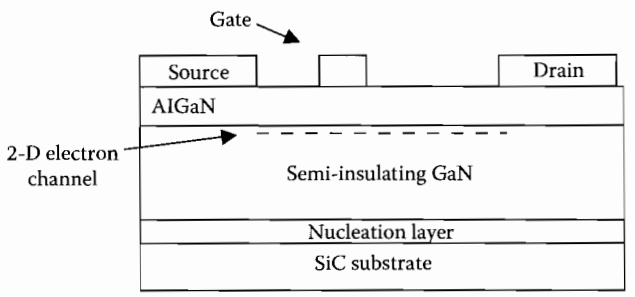

FIGURE 1.62 Cross section of GaN HEMT device structure.

Second, the high current and power density means that smaller device areas can be used to meet a particular power requirement. Smaller area translates into higher impedances because capacitances are proportional to device area. This simplifies the matching at the input and output, and therefore can lead to wider bandwidth and lower losses. The GaN devices are grown on a semi-insulating SiC substrate, thus the drain–source capacitance, C_{ds}, is small: typically about 0.25 pF/mm of device width. A 25 W device would require about 5 mm of channel width giving C_{ds} of about 1.25 pF. The Si device drain capacitance is generally quite high by comparison. For example, a 25 W LDMOS device designed for operation at 500 MHz has a C_{ds} of about 30 pF. This greatly limits the application of such devices for higher frequency or switching mode PA applications.

The current density and high frequency performance are also aided by the high saturated electron velocity in GaN. Refer once again to Figure 1.60 where electron velocity at high electric field in GaN is compared with other semiconductor materials. The electron transit time across the channel is inversely proportional to the carrier velocity; the current directly proportional.

The GaN HEMT gate structure often includes a field plate at the drain end of the gate as shown in Figure 1.63. The field plate distributes the electric field in the gate–drain region over a wider area leading to reduction in peak electric field and higher breakdown voltage. In addition, surface trap charge is less affected by potential variation on the surface, thus there is less depletion of channel charge by these traps leading to reduction in the low frequency dispersion effects.

Recent GaN HEMT microwave power amplifier performance highlights are presented in Table 1.7. Power outputs of several hundreds of watts have been obtained under low duty cycle pulsed operating conditions where thermal effects are less serious. CW output powers in the same range have been obtained by power combining of two or more amplifiers. Operation at 35 GHz was reported on MMIC PAs using devices with reduced gate length (0.15 μm) and smaller drain voltages.

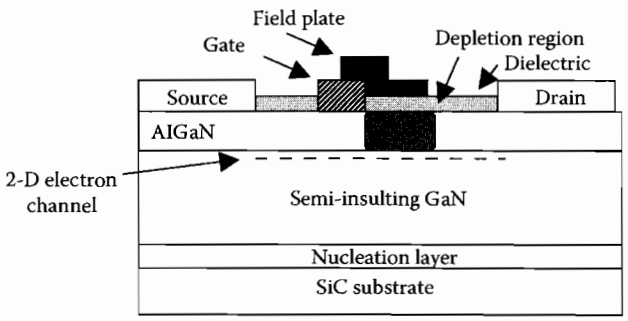

FIGURE 1.63 Cross section of GaN HEMT device structure with gate connected field plate.

TABLE 1.7 Recent GaN/HEMT Microwave Power Amplifier Performance Highlights

F (GHz)	P_{OUT} (W)	PAE (%)	V_{DC} (V)	W/mm	Reference
1.5	500[a]	49	65	13.9[a]	[28]
2.14	750[a]	—	50	7.8[a]	[29]
2.14	370	—	45	3.8	[30]
6	130[a]	45	50[a]	5.4[a]	[31]
9.5	80	34	30	3.5	[32]
35	4[b]	23	24	3.3	[33]
35	0.9[b]	51	20	4.5	[34]

[a] Pulsed CW unless otherwise noted.
[b] MMIC.

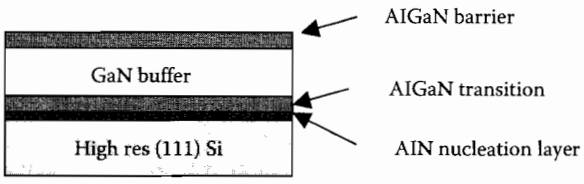

FIGURE 1.64 GaN on Si m-HEMT substrate.

GaN has also been successfully grown on high resistivity silicon substrates as illustrated by Figure 1.64. A thick nucleation layer leads to a metamorphic structure. Good power amplifier GaN on Si HEMT devices have been demonstrated to provide comparable performance in power and efficiency to those grown on Si [35,36]. The thermal conductivity of silicon is considerably less than that of SiC, however, so one would not expect the thermal resistance of the GaN on Si HEMT to be as low as the former.

1.3.4 Conclusion

While the mainstream semiconductor device and circuit technology is defined by silicon and its related materials, the superior electron transport properties of compound semiconductor materials offer unique advantages in applications requiring the highest frequency, speed and power or lowest noise figure. The range of device structural possibilities with compound semiconductor heterojunctions is far greater than what can be realized without this option, and this has led to high-performance FET structures HEMT, p-HEMT, and m-HEMT, with excellent bandwidth, noise figure, and power.

References

1. Deal, W. R. et al., Design and analysis of broadband dual-gate balanced low-noise amplifiers, *IEEE J. Solid State Circuits*, 42(10), 2107–2115, Oct. 2007.
2. Matsuda, S. et al., Very compact high-gain broadband low-noise amplifier in InP HEMT technology *IEEE Trans. Microwave Theory Tech.*, 54(12) part 2, 4565–4571, Dec. 2006.
3. Grundbacher, R., et al., High performance millimeter wave 0.1/spl mu/m InP HEMT MMIC LNAs fabricated on 100 mm wafers, *International Conference on Indium Phosphide and Related Materials 16th IPRM*, May 31–June 4; Kagoshima, Japan, pp. 284–287, 2004.
4. Schlechtweg, M. and A. Tessmann, From 100 GHz to terahertz electronics—activities in Europe *IEEE Compound Semiconductor IC Symposium*, pp. 8–11, San Antonio, TX, Nov. 2006.
5. Leuther, A., et al., 50 nm MHEMT technology for G- and H-Band MMICs, *International Conference on Indium Phosphide and Related Materials, 19th IPRM*, 14–18 May, Matsue Japan, pp. 24–27, 2007.
6. Deal, W. R., et al., Demonstration of a 270-GHz MMIC amplifier using 35-nm InP HEMT technology, *IEEE Microwave Wireless Components Lett.*, 17(5), 391–393, May 2007.
7. Sze, S. M., *Physics of Semiconductor Devices*. Wiley-Interscience, New York, 1981.
8. Taylor, G. W., H. M. Darley, et al., A device model for an ion-implanted MESFET. *IEEE Trans. Electron. Dev.*, ED-26, 172–179, 1979.
9. Lee, S. J. and C. P. Lee, Temperature effect on low threshold voltage ion-implanted GaAs MESFETs *Electron. Lett.*, 17(20), 760–761, 1981.
10. Gray, P. R. and R. Meyer, *Analysis and Design of Analog Integrated Circuits*. 4th ed., John Wiley New York, 2004.
11. Statz, H., P. Newman, et al., GaAs FET device and circuit simulation in SPICE, *IEEE Trans. Electron Dev.*, 34(2), 160–169, 1987.
12. Long, S. I., Compound semiconductor materials, Chap. 71 in *The VLSI Handbook*, 2nd edn. Ed. W.-K Chen, CRC Press, Boca Raton, FL, 2006.

13. Kroemer, H., Heterostructures for everything: Device principles of the 1980's? *Japanese J. Appl. Phys.*, 20, 9, 1981.
14. Kroemer, H., Heterostructure bipolar transistors and integrated circuits, *Proc. IEEE*, 70, 13, 1982.
15. Matthews, J. W. and A. E. Blakeslee, Defects in epitaxial multilayers, III. Preparation of almost perfect layers, *J. Crystal Growth*, 32, 265, 1976.
16. Matthews, J. W. and A. E. Blakeslee, Coherent strain in epitaxially grown films, *J. Crystal Growth*, 27, 118, 1974.
17. Hoke, W. E., et al., Properties of metamorphic materials and device molecular beam epitaxy, *2002 International Conference on Molecular Beam Epitaxy*, 15–20 September, San Francisco, CA, pp. 69–70, 2002.
18. Schlechtweg, M., et al., Millimeter-wave and mixed-signal integrated circuits based on advanced metamorphic HEMT technology, *16th International Conference on Indium Phosphide and Related Materials*, May 31–June 4; Kagoshima, Japan, pp. 609–614, 2004.
19. Estreich, D., Compound semiconductor devices for analog and digital circuits, Chap. 72 in *The VLSI Handbook*, 2nd edn., Ed. W.-K Chen, CRC Press, Boca Raton, FL, 2006.
20. Littlejohn, M. A., et al., High-field transport in InGaAs and related heterostructures, in *Properties of Lattice-Matched and Strained Indium Gallium Arsenide*, Inspec—IEE, London, pp. 107–116, 1993.
21. Pozela, F. and A. Reklaitis, Electron transport properties in GaAs at high electric fields, *Solid-State Elect.*, 23, 927–933, 1980.
22. Jacobini, C., et al., A review of some charge transport properties of Silicon, *Solid-State Elect.*, 20, 77, 1977.
23. Meharry, D., et al., Multi-watt wideband MMICs in GaN and GaAs, *IEEE/MTT-S International Microwave Symposium*, June 3–June 8; Honolulu; Hawaii, pp. 631–634, 2007.
24. Gelmont, B., K. Kim, and M. Shur, Monte-Carlo simulation of electron transport in gallium nitride, *J. Appl. Phys.*, 74, 1818–1821, 1993.
25. F. Schwierz, An electron mobility model for wurtzite GaN, *Solid State Elect.*, 49, 889–895, 2005.
26. Trew, R. J., SiC and GaN transistors: Is there one winner for microwave power applications? *Proc. IEEE, Special Issue on Wide Bandgap Semiconductors*, 90, 1032–1047, June 2002
27. Johnson, E. O., Physical limitations on frequency and power parameters of transistors, *IRE International Convention Record*, Vol. 13; Part 5, pp. 27–34, March 1965.
28. Maekawa, A., et al., A 500W Push-Pull AlGaN/GaN HEMT amplifier for L band high power applications, *IEEE/MTT-S International Microwave Symposium*, IEEE Mircrowave Theory and Techniques Society, San Francisco, CA, June 11–16, pp. 722–725, 2006.
29. Wakejima, A., et al., Pulsed 0.75 kW output single-ended GaN-FET amplifier for L/S band applications, *Electron. Lett.*, 42, 1349–1350, Nov. 2006.
30. Wakejima, A., et al., 370 W output power GaN-FET amplifier for W-CDMA cellular base stations, *Electron. Lett.*, 41, 1371–1372, Dec. 2005.
31. Yamanaka, K., et al., C-band GaN HEMT power amplifier with 220W output power, *2007 IEEE International Microwave Symposium*, June 8, Honolulu, Hawaii, pp. 1251–1254, 2007.
32. Takagi, K., et al., Xband AlGaN/GaN HEMT with over 80 W output power, *IEEE Compound Semiconductor IC Symposium (CSICS)*, Nov 12–15, San Antonio, TX, pp. 265–268, 2006.
33. Darwish, A., et al., 4 W Ka-band AlGaN/GaN power amplifier MMIC, *IEEE/MTT-S International Microwave Symposium*, IEEE Microwave Theory and Techniques Society, San Francisco, CA, June 11–16, pp. 730–733, 2006.
34. Kao, M.-Y., et al., AlGaN/GaN HEMTs with PAE of 53% at 35 GHz for HPA and multi-function MMIC applications, *2007 IEEE International Microwave Symposium*, Honolulu, Hawaii, June 3–8, pp. 627–629, 2007.
35. Johnson, J. W., Piner, E. L., Vescan, A., Therrien, R., Rajagopal, P., Roberts, J. C., Brown, J. D., Singhal, S., and Linthieum, K. J., 12 W/mm AlGaN/GaN HFETs on silicon substrates, *IEEE Elect. Dev. Lett.*, 25(7), 459–461, July 2004.
36. Nagy, W. et al., 150 W GaN on Si RF power transistor, *IEEE/MIT-S International Microwave Symposium*, Microwave Theory and Techniques Society, Long Beach, CA, 4pp, June 2005.

1.4 Passive Components

Nhat M. Nguyen

1.4.1 Resistors

Resistors available in monolithic form are classified in general as semiconductor resistors and thin-film resistors. Semiconductor structures include diffused, pinched, epitaxial, and ion-implanted resistors. Commonly used thin-film resistors include tantalum, nickel–chromium (Ni–Cr), cermet (Cr–SiO), and tin oxide (SnO_2). Diffused, pinched, and epitaxial resistors can be fabricated along with other circuit elements without any additional processing steps. Ion-implanted and thin-film resistors require additional processing steps for monolithic integration but offer lower temperature coefficient, smaller absolute value variation, and superior high-frequency performance.

Resistor calculation. The simplified structure of a uniformly doped resistor of length L, width W, and thickness T is shown in Figure 1.65. The resistance is

$$R = \frac{1}{\sigma}\frac{L}{WT} = \left(\frac{\rho}{T}\right)\frac{L}{W} = R_n\frac{L}{W} \tag{1.237}$$

where

σ and ρ are conductivity and resistivity of the sample, respectively
R_N is referred to as the *sheet resistance*

From the theory of semiconductor physics, the conductivity of a semiconductor sample is

$$\sigma = q\left(\mu_n n + \mu_p p\right) \tag{1.238}$$

where

q is the electron charge (1.6×10^{-19} C)
$\mu_n(\text{cm}^2/\text{V}\cdot\text{s})$ is the electron mobility
$\mu_p(\text{cm}^2/\text{V}\cdot\text{s})$ is the hole mobility
$n(\text{cm}^{-3})$ is the electron concentration
$p(\text{cm}^{-3})$ is the hole concentration
$\sigma(\Omega/\text{cm})^{-1}$ is the electrical conductivity

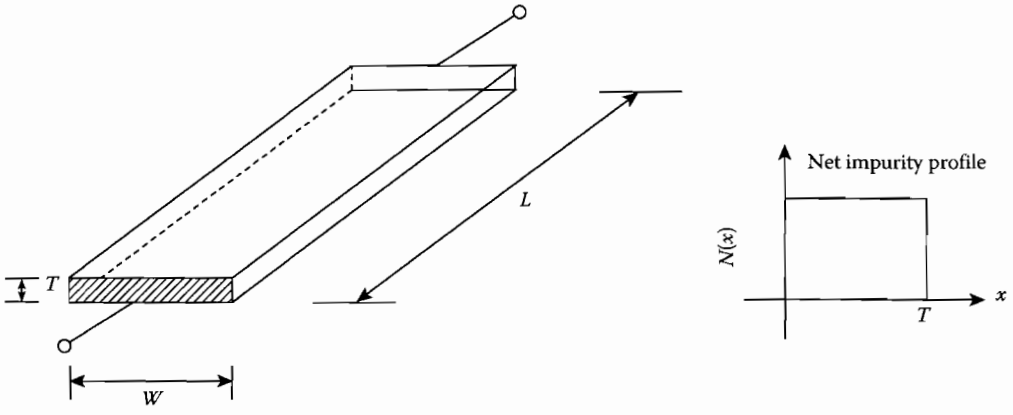

FIGURE 1.65 Simplified structure of a uniformly doped resistor.

For an n-type doped sample with a concentration $N_D(cm^{-3})$ of donor impurity atoms, the electron concentration n is approximately equal to N_D. Given the mass-action law $np = n_i^2$, the conductivity of an n-type doped sample is approximated by

$$\sigma = q\left(\mu_n N_D + \mu_p \frac{n_i^2}{N_D}\right) \approx q\mu_n N_D \tag{1.239}$$

where $n_i(cm^{-3})$ is the *intrinsic* concentration. For a p-type doped sample, the conductivity is

$$\sigma = q\left(\mu_n \frac{n_i^2}{N_A} + \mu_p N_A\right) \approx q\mu_p N_A \tag{1.240}$$

where $N_A(cm^{-3})$ is the concentration of p-type donor impurity atoms. The sheet resistance of an n-type uniformly doped resistor is thus

$$R_n = \left(\frac{1}{q\mu_n N_D T}\right) \tag{1.241}$$

For an n-type nonuniformly doped resistor as shown in Figure 1.66, where n-type impurity atoms are introduced into the p-type region by means of a high-temperature diffusion process, the sheet resistance [7] is

$$R_n = \left[\int_0^{x_j} q\mu_n N_D(x)dx\right]^{-1} \tag{1.242}$$

where x_j is the distance from the surface to the edge of the junction depletion layer.

Measured values of electron mobility and hole mobility in silicon material as a function of impurity concentration are shown in Figure 1.67 [4]. The resistivity ρ (Ω-cm) of n-type and p-type silicon as a function of impurity concentration is shown in Figure 1.68 [12].

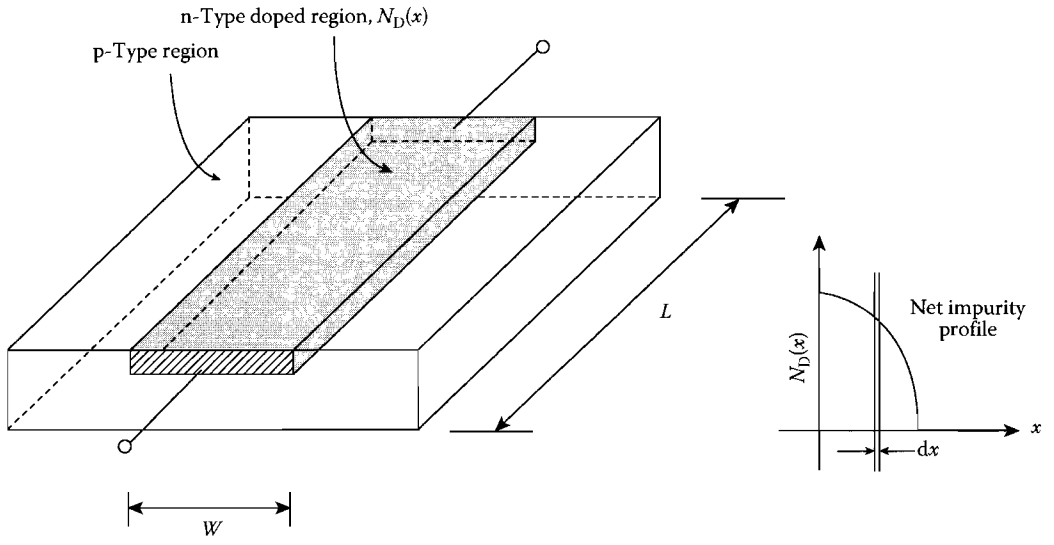

FIGURE 1.66 Simplified structure of an n-type nonuniformly doped resistor.

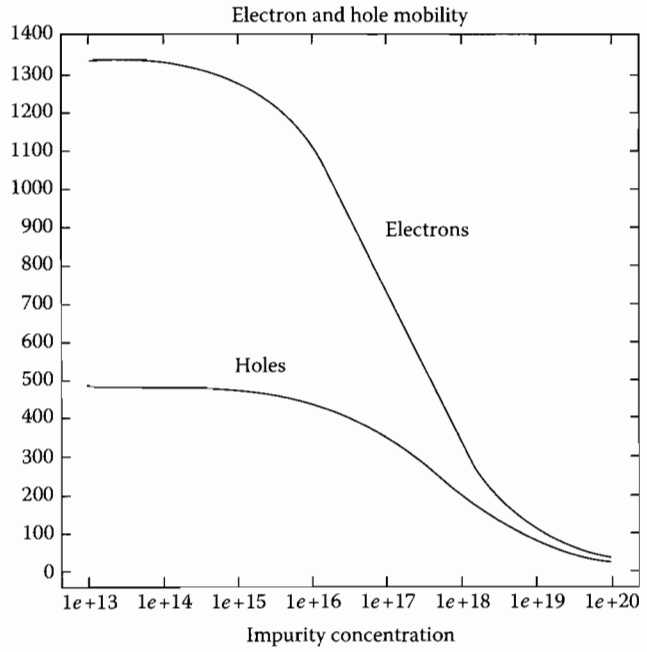

FIGURE 1.67 Electron and hole mobility vs. impurity concentration in silicon.

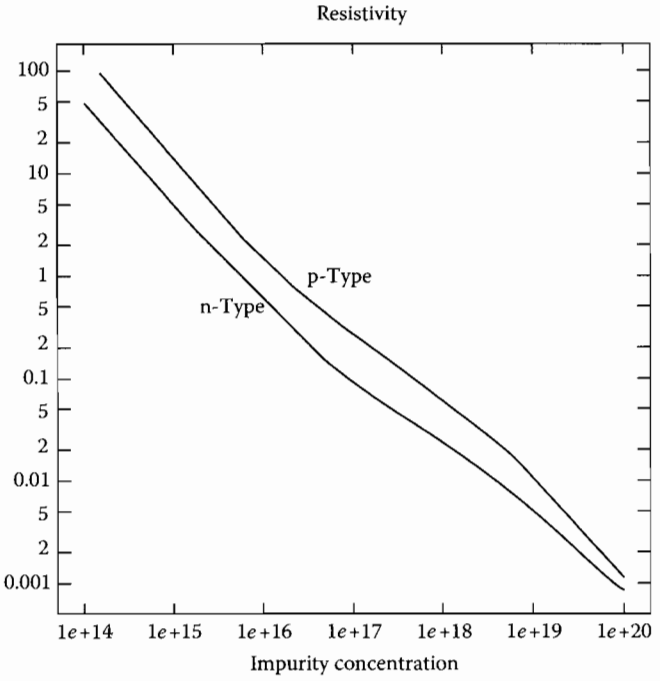

FIGURE 1.68 Resistivity of p-type and n-type silicon vs. impurity concentration.

The sheet resistance depends also on temperature since both electron mobility and hole mobility vary with temperature [17]. This effect is accounted for by utilizing a temperature coefficient quantity that measures the sheet resistance variation as a function of temperature. A mathematical model of the temperature effect is

$$R_v(T) = R_v(T_o)[(T - T_o)\text{TC}] \tag{1.243}$$

where

T_o is the room temperature
"TC" is the temperature coefficient

1.4.1.1 Diffused Resistors

In metal-oxide-semiconductor (MOS) technology, the diffused layer forming the source and drain of the MOS transistors can be used to form a diffused resistor. In silicon bipolar technology, the available diffused layers are base diffusion, emitter diffusion, active base region, and epitaxial layer.

Base-diffused resistors. The structure of a typical base diffused resistor is shown in Figure 1.69, where the substrate material is assumed of p-type silicon material. The diffused resistor is formed by using the p-type base diffusion of the npn transistors. The resistor contacts are formed by etching selected windows of the SiO_2 passivation layer and depositing thin films of conductive metallic material. The isolation region can be formed with either a p-type doped junction or a trench filled with SiO_2 dielectric material. The pn junction formed by the p-type resistor and the n-type epitaxial (epi) layer must be reverse biased

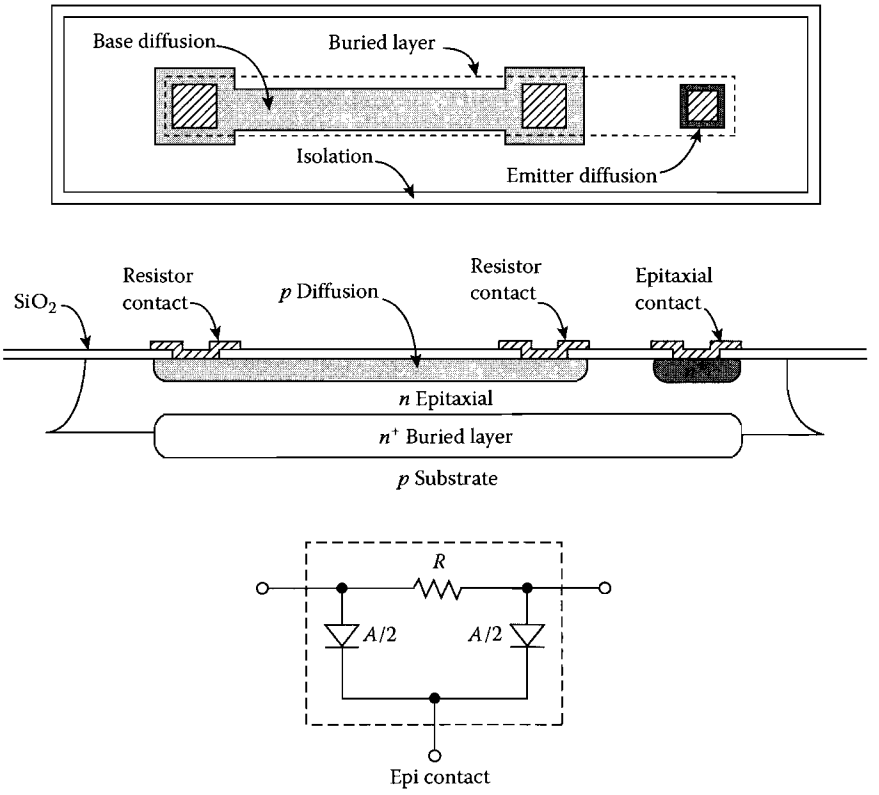

FIGURE 1.69 p-Type base-diffused resistor.

in order to eliminate the undesired dc current path through the pn junction. The impedance associated with a forward-biased pn junction is low and thus would also cause significant ac signal loss. To ensure this reverse bias constraint the epi region must be connected to a potential that is more positive than either end of the resistor contacts. Connecting the epi region to a relatively higher potential also eliminates the conductive action due to the parasitic pnp transistor formed by the p-type resistor, the n-type epi region, and the p-type substrate. When the base-diffused resistor is fabricated along with other circuit elements to form an integrated circuit (IC), the epitaxial contact is normally connected to the most positive supply of the circuit.

The resistance of a diffused resistor is given by Equation 1.237, where the diffused sheet resistance is between 100 and 200 Ω/N. Due to the lateral diffusion of impurity atoms, the effective cross-sectional area of the resistor is larger than the width determined by photomasking. This lateral or side diffusion effect can be accounted for by replacing the resistor width W by an effective width W_{eff}, where $W_{\text{eff}} \geq W$. The resistance from the two resistor contacts must also be accounted for, especially for small values of L/W [3]. Base-diffused resistors have a typical temperature coefficient between +1500 and +2000 ppm/°C.

The maximum allowable voltage for the base-diffused resistor of Figure 1.69 is limited by the breakdown voltage between the p-type base diffusion and the n-type epi. This voltage equals the breakdown voltage BV_{CBO} of the collector–base junction of the npn transistor and typically causes an *avalanche breakdown* mechanism across the base–epi junction. As the applied voltage approaches the breakdown voltage, a large leakage current flows from the epi region to the base region and can cause excessive heat dissipation.

For analog IC applications where good matching tolerance between adjacent resistors is required, the resistor width should be made as large as possible. Base-diffused resistors with 50 μm resistor widths can achieve a matching tolerance of ±0.2%. The minimum resistor width is limited by photolithographic consideration with typical values between 3 and 5 μm. Also, in order to avoid the self-heating problem of the resistor it is important to ensure a minimum resistor width for a given dc current level, with a typical value of about 3 μm for every 1 mA of current.

With respect to high-frequency performance, the reverse-biased pn junction between the p-type base diffusion and the n-type epi contributes a distributed depletion capacitance which in turn causes an impedance roll-off at 20 dB/decade. This capacitance depends on the voltage applied across the junction and the junction impurity-atom dopings. For most applications the electrical lumped model as shown in Figure 1.69 is adequate for characterizing this capacitive effect where the effective pn junction area is divided equally between the two diodes. Figure 1.70 shows a normalized impedance response as a function of the RC distributed stage. The frequency at which impedance value is reduced by 3 dB is given by

$$f_{-3\text{ dB}} = \begin{cases} \left(\dfrac{1}{2\pi}\right)\dfrac{2.0}{RC} & N = 1 \text{ (Circuit model of Figure 1.62)} \\ \left(\dfrac{1}{2\pi}\right)\dfrac{2.32}{RC} & N = 2 \\ \left(\dfrac{1}{2\pi}\right)\dfrac{2.42}{RC} & N = 3 \\ \left(\dfrac{1}{2\pi}\right)\dfrac{2.48}{RC} & N = 4 \end{cases} \tag{1.244}$$

Emitter-diffused resistors. Emitter-diffused resistors are formed by using the heavily doped n^+ emitter diffusion layer of the npn transistors. Due to the high doping concentration, the sheet resistance can be as low as 2 to 10 Ω/N with a typical absolute value tolerance of ±20%.

Figure 1.71 shows an emitter-diffused resistor structure where an n^+ diffusion layer is formed directly on top of the n-type epitaxial region and the ohmic contacts are composed of conductive

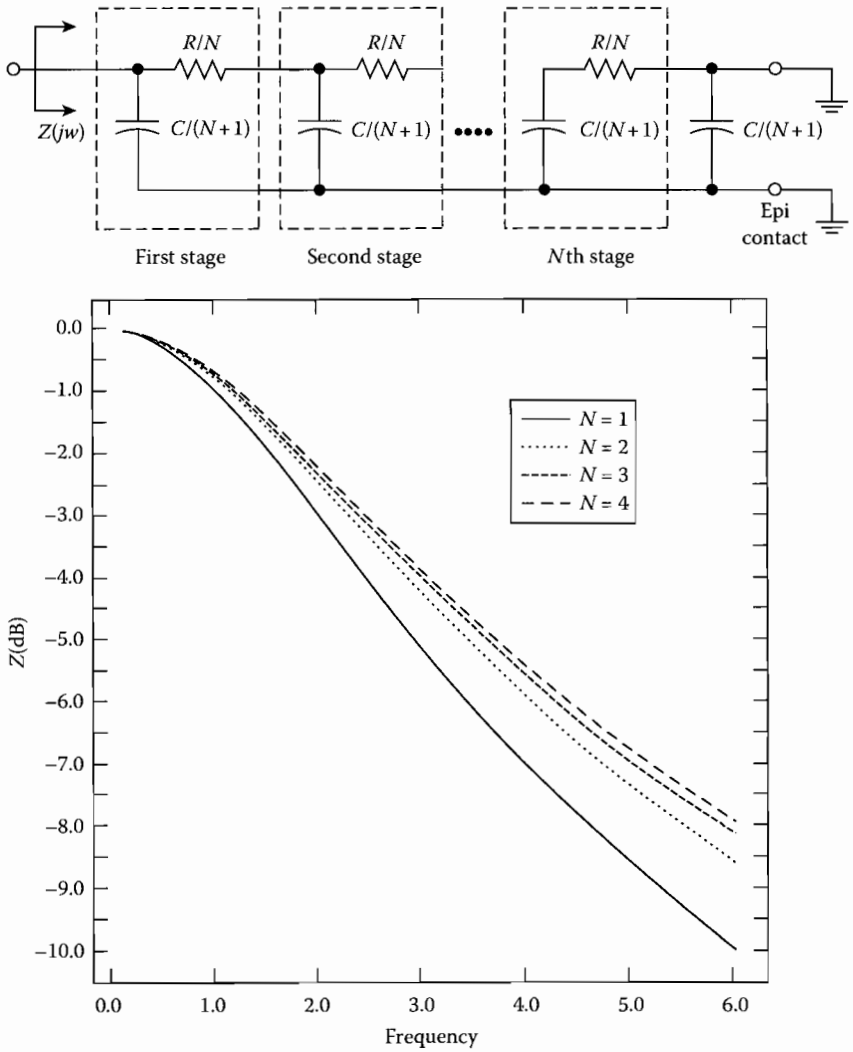

FIGURE 1.70 Normalized frequency response of a diffused resistor for $N = 1, 2, 3, 4$. The epi contact and one end of the resistor are grounded.

metal thin films. Since the resistor body and the epi layer are both n-type doped, they are electrically connected in parallel but the epi layer is of much higher resistivity due to its lower concentration doping, and thus the effective sheet resistance of the resistor structure is determined solely by the n^+ diffusion layer. The pn junction formed between the p-type substrate and the n-type epi region must always be reverse biased, which is accomplished by connecting the substrate to a most negative potential. Because of the common n-type epi layer, each resistor structure of Figure 1.71 requires a separate isolation region.

Figure 1.72 shows another emitter diffused resistor structure where the n^+ diffusion layer is situated within a p-type diffused well. Several such resistors can be fabricated in the same p-type well or in the same isolation region because the resistors are all electrically isolated. The p-type well and the n^+ diffusion region form a pn junction that must always be reverse biased for electrical isolation. In order to eliminate the conductive action due to the parasitic npn transistor formed by the n-type resistor body,

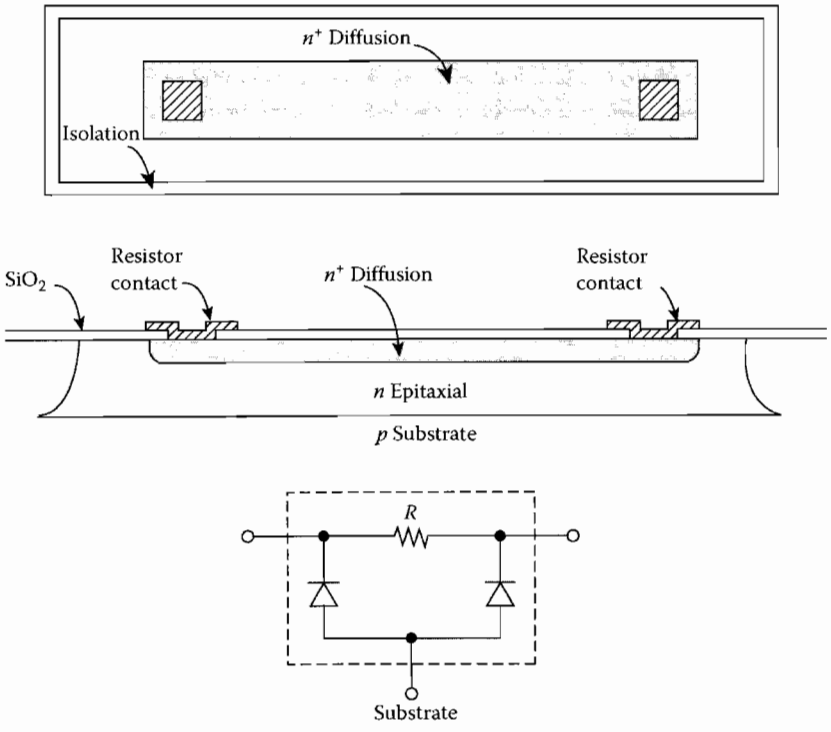

FIGURE 1.71 n-Type emitter-diffused resistor I.

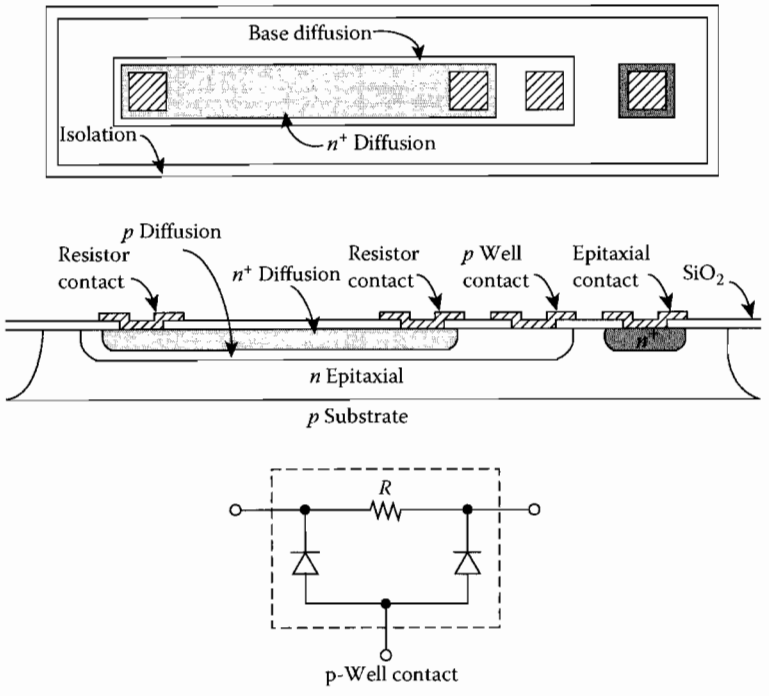

FIGURE 1.72 n-Type emitter-diffused resistor II.

the p-type well, and the n-type epi, the junction potential across the well contact and the epi contact must be either short-circuited or reverse-biased. The maximum voltage that can be applied across the emitter-diffused resistor of Figure 1.72 is limited by the breakdown voltage between the n^+ diffusion and the p-type well. This voltage equals the breakdown voltage BV_{EBO} of the emitter–base junction of the npn transistor, with typical values between 6 and 8 V.

1.4.1.2 Pinched Resistors

The active base region for the npn transistor can be used to construct pinched resistors with typical sheet resistance range from 2 to 10 KΩ/N. These high values can be achieved due to a thin cross-sectional area through which the resistor current traverses. The structure of a p-type base-pinched resistor is shown in Figure 1.73, where the p-type resistor body is "pinched" between the n^+ diffusion layer and the n-type epitaxial layer. The n^+ diffusion layer overlaps the p-type diffusion layer and is therefore electrically connected to the n-type epi. In many aspects the base-pinched resistor behaves like a p-channel JFET, in which the active base region functions as the p-channel, the two resistor contacts assume the drain and source, and the n^+ diffusion and the epi constitute the n-type gate. When the pn junction formed between the active base and the surrounding n^+ diffusion and n-epi is subject to a reverse bias potential, the carrier-free depletion region increases and extends into the active base region, effectively reducing the resistor cross section and consequently increasing the sheet resistance. Since the carrier-free depletion region varies with reverse bias potential, the pinched resistance is voltage controlled and is nonlinear.

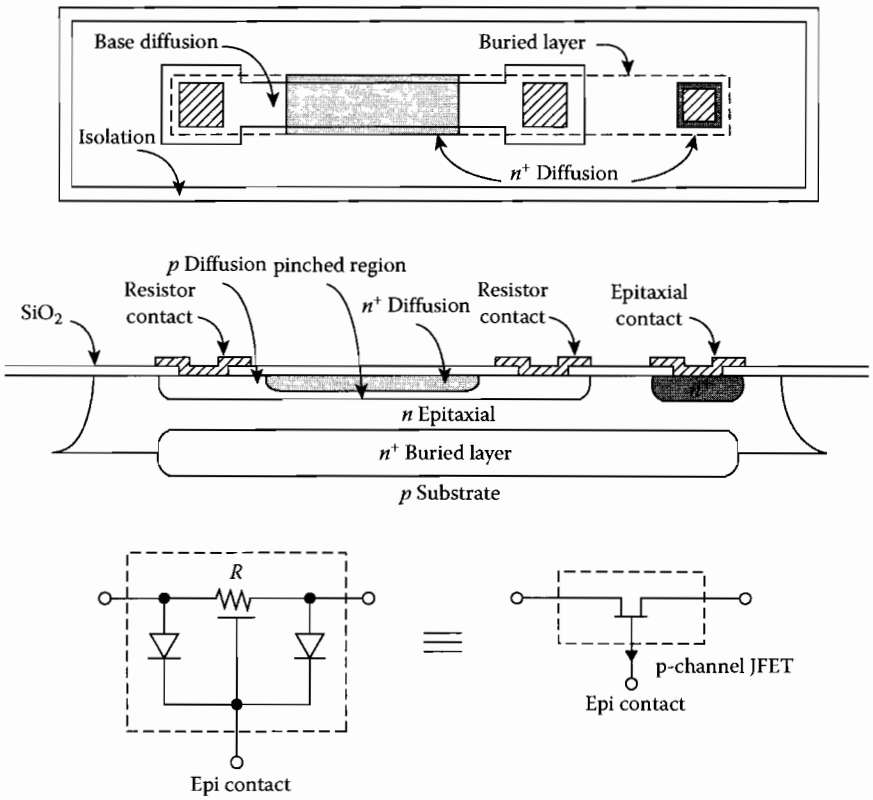

FIGURE 1.73 p-Type base-pinched resistor.

Absolute values for the base-pinched resistors can vary as much as ±50% due to large process variation in the fabrication of the active base region. The maximum voltage that can be applied across the base-pinched resistor of Figure 1.73 is restricted by the breakdown voltage between the n^+ diffusion layer and the p-type base diffusion. The breakdown voltage has a typical value around 6 V.

1.4.1.3 Epitaxial Resistors

Large values of sheet resistance can be obtained either by reducing the effective cross-sectional area of the resistor structure or by using a low doping concentration that forms the resistor body. The first technique is used to realize the pinched resistor while the second is used to realize the epitaxial resistor. Figure 1.74 shows an epitaxial resistor structure where the resistor is formed with a lightly doped epitaxial layer. For an epi thickness of 10 μm and a doping concentration of 10^{15} donor atoms/cm^3, this structure achieves a resistivity of 5 Ω-cm and an effective sheet resistance of 5 KΩ/N. The temperature coefficient of the epitaxial resistor is relatively high with typical values around +3000 ppm/°C. This large temperature variation is a direct consequence of the hole and electron mobilities undergoing more drastic variations against temperature at particularly low doping concentrations [13]. The maximum voltage that can be applied across the epitaxial resistor is significantly higher than that for the pinched resistor. This voltage

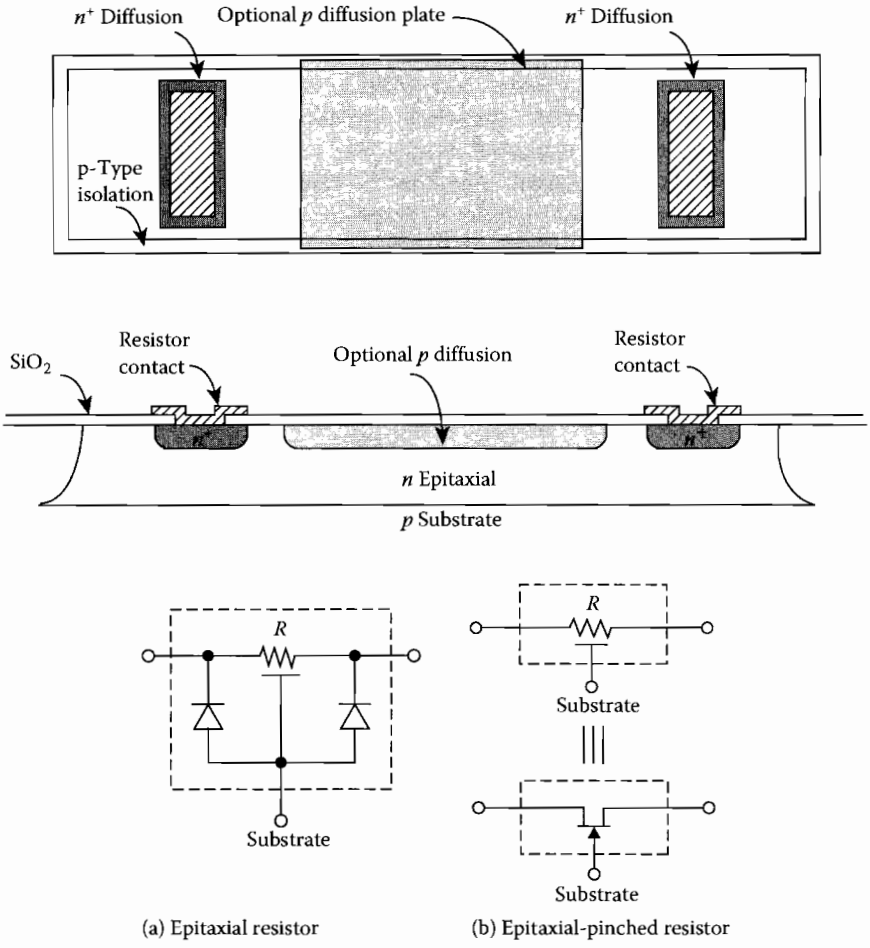

FIGURE 1.74 n-Type epitaxial and epitaxial-pinched resistors.

is set by the breakdown voltage between the n-type epi and the p-type substrate which varies inversely with the doping concentration of this pn junction.

Epitaxial-pinched resistors. By putting a p-type diffusion plate on top of the epitaxial resistor of Figure 1.74, even larger sheet resistance value can be obtained. The p-type diffusion plate overlaps the epi region and is electrically connected to the substrate through the p-type isolation. The epi layer is thus pinched between the p-type diffusion plate and the p-type substrate. When the n-type epi and the surrounding p-type regions is subject to a reverse bias potential, the junction depletion width extends into the epi region and effectively reduces the cross-sectional area. Typical sheet resistance values are between 4 and 5 KΩ/N. The epitaxial-pinched resistor behaves like an n-channel JFET, in which the effective channel width is controlled by the substrate voltage.

1.4.1.4 Ion-Implanted Resistors

Ion implantation is an alternative technique beside diffusion for inserting impurity atoms into a silicon wafer [17]. Commonly used impurities for implantation are the p-type boron atoms. The desired impurity atoms are first ionized and then accelerated to a high energy by an electric field. When a beam of these high-energy ions is directed at the wafer, the ions penetrate into exposed regions of the wafer surface. The penetration depth depends on the velocity at contact and is typically between 0.1 and 0.8 μm. The exposed regions of the wafer surface are defined by selectively etching a thick thermally grown SiO_2 layer that covers the wafer and functions as a barrier against the implanted ions. Unique characteristics of the ion-implantation technique include a precise control of the impurity concentration, uniformly implanted layers of impurity atoms, and no lateral diffusion. The structure of a p-type ion-implanted resistor is shown in Figure 1.75, where the p-type diffused regions at the contacts are used to achieve good ohmic contacts to the implanted resistor. The pn junction formed between the p-type implanted region and the n-type epitaxial layer must be reverse biased for electrical isolation. By connecting the epi region to a potential relatively more positive than the substrate potential, the conductive action due to the parasitic pnp transistor formed by the p-type implanted, the n-type epi, and the p-type substrate is also eliminated. Ion-implanted resistors exhibit relatively tight absolute value tolerance and excellent matching. Absolute value tolerance down to ±3% and matching tolerance of ±2% are typical performance.

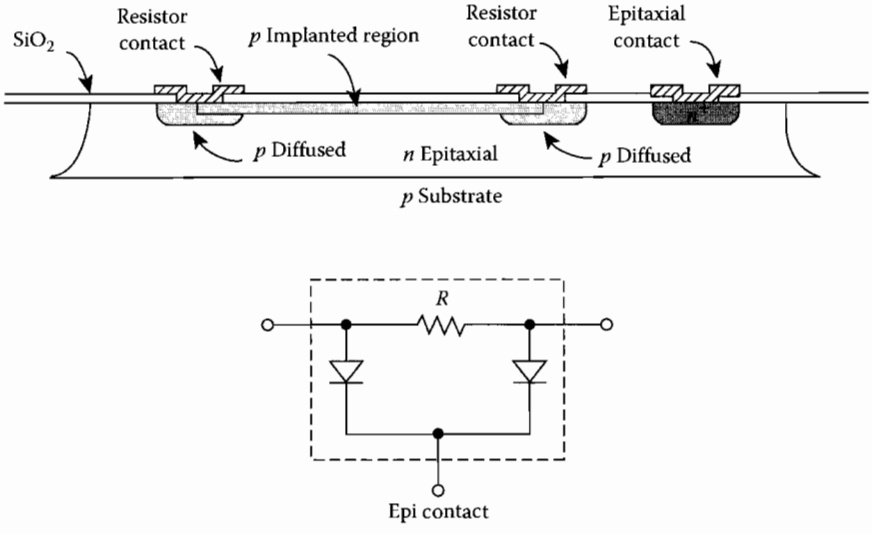

FIGURE 1.75 p-Type ion-implanted resistor.

TABLE 1.8 Typical Properties of Semiconductor Resistors

Resistor Type	Sheet ρ(Ω/N)	Absolute Tolerance (%)	Matching Tolerance (%)	Temperature Coefficient (ppm/°C)
Base-diffused	100–200	±20	±2 (5 μm wide) ±0.2 (50 μm wide)	+1500 to +2000 —
Emitter-diffused	2–10	±20	±2	+600
Base-pinched	2–10 K	±50	±10	+2500
Epitaxial	2–5 K	±30	±5	+3000
Epitaxial-pinched	4–10 K	±50	±7	+3000
Ion-implanted	100–1000	±3	±2 (5 μm wide) ±0.15 (50 μm wide)	Controllable to ±100

Source: Gray, P.R. and Meyer, R.G., *Analysis and Design of Analog Integrated Circuits*, Wiley, New York, 1984, p. 119.

Table 1.8 provides a summary of the typical characteristics for the diffused, pinched, epitaxial, and ion-implanted resistors.

1.4.1.5 Thin-Film Resistors

Compared with diffused resistors, thin-film resistors offer advantages of a lower temperature coefficient, a smaller absolute value variation, and an excellent high-frequency characteristic. Commonly used resistive thin films are tantalum, Ni–Cr, Cr–SiO, and SnO_2. A typical thin-film resistor structure is shown in Figure 1.76, where a thin-film resistive layer is deposited on top of a thermally grown SiO_2 layer and a thin-film conductive metal layer is used to form the resistor contacts. The oxide layer functions as an insulating layer for the resistor. Various CVD techniques can be used to form the thin films [8]. The oxide passivation layer deposited on top of the resistive film and the conductive film protects the device surface from contamination. The electrical lumped model as shown in Figure 1.76 is adequate to characterize the high-frequency performance of the resistor. The parallel-plate capacitance formed

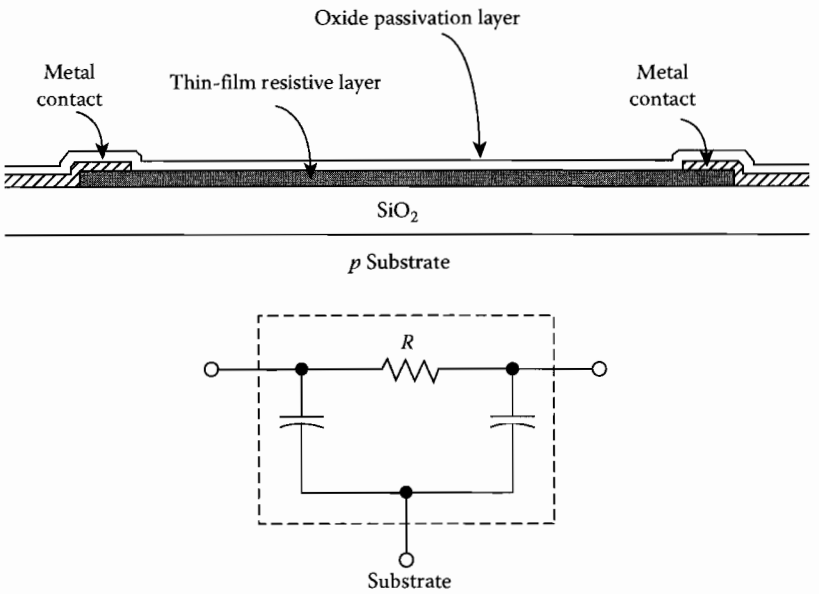

FIGURE 1.76 Thin-film resistor.

TABLE 1.9 Typical Characteristic of Thin-Film Resistors

Resistor Type	Sheet ρ(Ω/N)	Absolute Tolerance (%)	Matching Tolerance (%)	Temperature Coefficient (ppm/°C)
Ni–Cr	40–400	±5	±1	±100
Ta	10–1000	±5	±1	±100
SnO_2	80–4000	±8	±2	0–1500
Cr–SiO	30–2500	±10	±2	±50 to ±150

Source: Grebene, A.B., *Bipolar and MOS Analog Integrated Circuit Design*, Wiley, New York, 1984, p. 155.

between the thin-film resistive and the substrate is divided equally between the two capacitors. Table 1.9 provides a summary of the characteristics for some commonly used thin-film resistors.

1.4.2 Capacitors

Monolithic capacitors are widely used in analog and digital ICs for functions such as circuit stability, bandwidth enhancement, ac signal coupling, impedance matching, and charge storage cells. Capacitor structures available in monolithic form include pn junction, MOS, and polysilicon capacitors. pn junctions under reverse-biased conditions exhibit a nonlinear voltage-dependent capacitance. MOS and polysilicon capacitors, on the other hand, closely resemble the linear parallel-plate capacitor structure as shown in Figure 1.77. If the insulator thickness T of the parallel-plate structure is small compared with the plate width W and length L, the electric field between the plates is uniform (fringing field neglected). Under this condition the capacitance can be calculated by

$$C = \frac{\kappa\varepsilon_0}{T} WL \tag{1.245}$$

where κ is the relative dielectric constant of the insulating material and ε_0 is the permittivity constant in vacuum (8.854×10^{-14} F/cm).

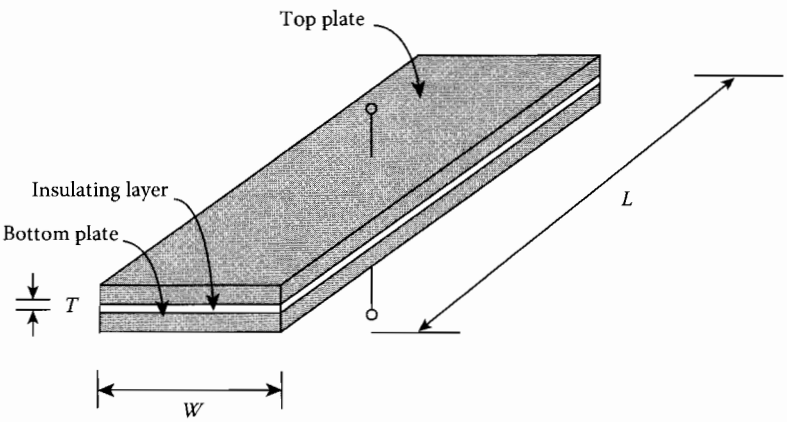

FIGURE 1.77 Structure of a parallel-plate capacitor.

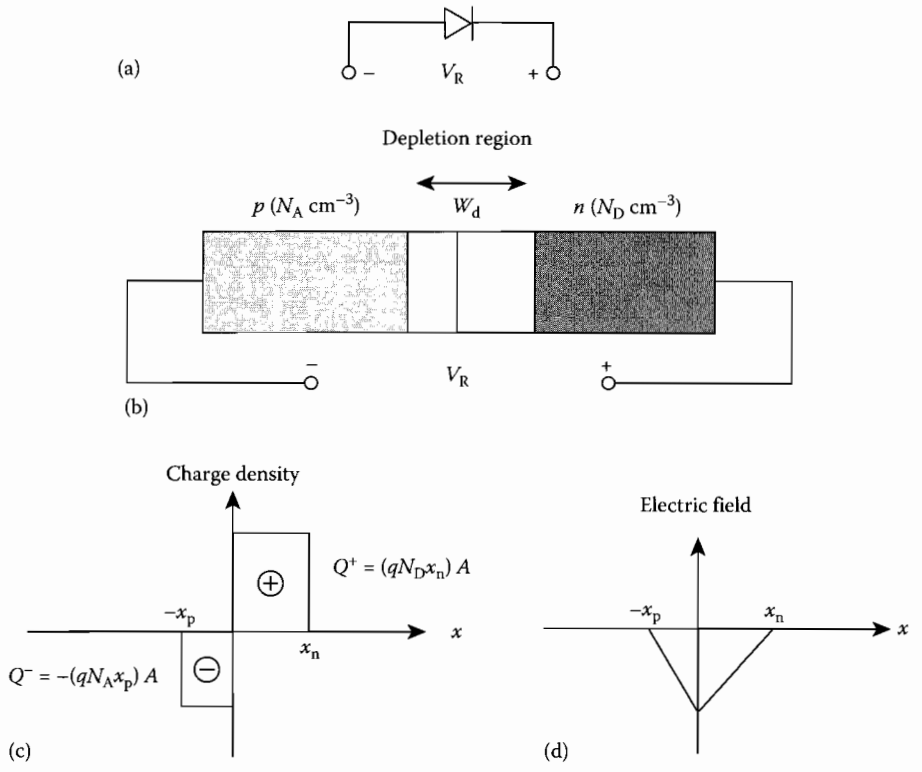

FIGURE 1.78 Abrupt p–n junction: (a) p–n junction symbol; (b) depletion region; (c) charge density within the depletion region; and (d) electric field.

1.4.2.1 Junction Capacitors

The structure of an *abrupt* pn junction is shown in Figure 1.78, where the doping is assumed uniform throughout the region on both sides. The acceptor impurity concentration of the p region is N_A atoms/cm^3 and the donor impurity concentration of the n region is N_D atoms/cm^3. When the two regions are brought in contact, mobile holes from the p region diffuse across the junction to the n region and mobile electrons diffuse from the n to the p region. This diffusion process creates a *depletion* region that is essentially free of mobile carriers (depletion approximation) and contains only fixed acceptor and donor ions. Ionized acceptor atoms are negatively charged and ionized donor atoms are positively charged. In equilibrium the diffusion process is balanced out by a drift process that arises from a *built-in* voltage ψ_o across the junction. This voltage is positive from the n region relative to the p region and is given by [17]

$$\psi_o = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} \tag{1.246}$$

where

k is the Boltzmann constant (1.38×10^{-23} V · C/K)
T is the temperature in Kelvin (K)
q is the electron charge (1.60×10^{-19} C)
n_i(cm^{-3}) is the *intrinsic* carrier concentration in a pure semiconductor sample

For silicon at 300 K, $n_i \approx 1.5 \times 10^{10}$ cm^{-3}.

When the pn junction is subject to an applied reverse bias voltage V_R, the drift process is augmented by the external electric field and more mobile electrons and holes are pulled away from the junction. Because of this effect, the depletion width W_d and consequently the charge Q on each side of the junction vary with the applied voltage. A junction capacitor can thus be defined to correlate this charge–voltage relationship. The Poisson's equation relating the junction voltage $\phi(x)$ to the electric field $\xi(x)$ and the total charge Q is

$$\begin{aligned}\frac{d^2\varphi(x)}{dx^2} &= -\frac{d\xi(x)}{dx} = -\frac{q}{\varepsilon_S}(p - n + N_D - N_A) \\ &\approx \begin{cases} \dfrac{qN_A}{\varepsilon_S} & -x_p < x < 0 \\ -\dfrac{qN_D}{\varepsilon_S} & 0 < x < x_n \end{cases}\end{aligned} \tag{1.247}$$

where ε_S ($11.8\varepsilon_0 = 1.04 \times 10^{-12}$ F/cm) is the permittivity of the silicon material. The first integral of Equation 1.247 yields the electric field as

$$\xi(x) = \begin{cases} -\dfrac{qN_A}{\varepsilon_S}(x + x_p) & -x_p < x < 0 \\ -\dfrac{qN_D}{\varepsilon_S}(x_n + x) & 0 < x < x_n \end{cases} \tag{1.248}$$

The electric field is shown in Figure 1.78, where the maximum field strength occurs at the junction edge. This value is given by

$$|\xi_{max}| = \frac{qN_A}{\varepsilon_S}x_p = \frac{qN_D}{\varepsilon_S}x_n$$

The partial depletion width x_p on the p region and the partial depletion width x_n on the n region can then be related to the depletion width W_d as

$$\begin{aligned} x_p + x_n &= W_d \\ x_p &= \frac{N_D}{N_A + N_D}W_d \\ x_n &= \frac{N_A}{N_A + N_D}W_d \end{aligned}$$

Taking the second integral of Equation 1.247 yields the junction voltage

$$\varphi(x) = \begin{cases} \dfrac{qN_A}{\varepsilon_S}\left(\dfrac{x_p^2}{2} + x_p x + \dfrac{x^2}{2}\right) & -x_p < x < 0 \\ \dfrac{qN_D}{\varepsilon_S}\left(\dfrac{x_n x_p}{2} + x_n x - \dfrac{x^2}{2}\right) & 0 < x < x_n \end{cases} \tag{1.249}$$

where the voltage at x_p is arbitrarily assigned to be zero. The total voltage $\psi_0 + V_R$ can be expressed as

$$\psi_o + V_R = \varphi(x_n) = \frac{qN_D}{2\varepsilon_S}\left(1 + \frac{N_D}{N_A}\right)x_n^2$$

Finally, the depletion width W_d and the total charge Q in terms of the total voltage across the junction can be derived to be

$$W_d = \left[\frac{2\varepsilon_S}{q}(\psi_o + V_R)\left(\frac{1}{N_A} + \frac{1}{N_D}\right)\right]^{1/2}$$
$$|Q| = A(qN_A x_p) = A(qN_D x_n) = A\left[2q\varepsilon_S(\psi_o + V_R)\left(\frac{1}{N_A} + \frac{1}{N_D}\right)^{-1}\right]^{1/2} \tag{1.250}$$

The junction capacitance is thus

$$C_j = \left|\frac{dQ}{dV_R}\right| = A\left[\frac{q\varepsilon_S}{2}\left(\frac{1}{\psi_o + V_R}\right)\left(\frac{1}{N_A} + \frac{1}{N_D}\right)^{-1}\right]^{1/2}$$
$$= \frac{C_{jo}}{\left(1 + \frac{V_R}{\psi_o}\right)^{1/2}} \tag{1.251}$$

where

A is the effective cross-sectional junction area

C_{jo} is the value of C_j for $V_R = 0$

If the doping concentration in one side of the pn junction is much higher than that in the other, the depletion width and the junction capacitance can be simplified to

$$W_d = \left[\frac{2\varepsilon_S}{qN_L}(\psi_o + V_R)\right]^{1/2} \tag{1.252}$$

$$C_j = A\left[\frac{\varepsilon_S qN_L}{2}\left(\frac{1}{\psi_o + V_R}\right)\right]^{1/2} \tag{1.253}$$

where N_L is the concentration of the lightly doped side. Figure 1.79 displays the junction capacitance per unit area as a function of the total voltage $\psi_o + V_R$ and the concentration on the lightly doped side of the junction [3].

In silicon bipolar technology the base–emitter, the base–collector, and the collector–substrate junctions under reverse bias are often utilized for realizing a junction capacitance. The collector–substrate junction has only a limited use since it can only function as a shunt capacitor due to the substrate being connected to an ac ground.

Base–collector junction capacitor. A typical base–collector capacitor structure is shown in Figure 1.80 together with an equivalent lumped circuit model. A heavily doped n^+ buried layer is used to minimize the series resistance R_C. For the base–collector junction to operate in reverse bias, the n-type collector must be connected to a voltage relatively higher than the voltage at the p-type base. The junction breakdown voltage is determined by BV_{CBO} of the npn transistor, which has a typical value between 25 and 50 V.

Base–emitter junction capacitor. Figure 1.81 shows a typical base–emitter capacitor structure where the parasitic junctions D_{BC} and D_{SC} must always be in reverse bias. The base–emitter junction achieves the highest capacitance per unit area among the base–collector, base–emitter, and collector–substrate junctions due to the relatively higher doping concentrations in the base and emitter regions.

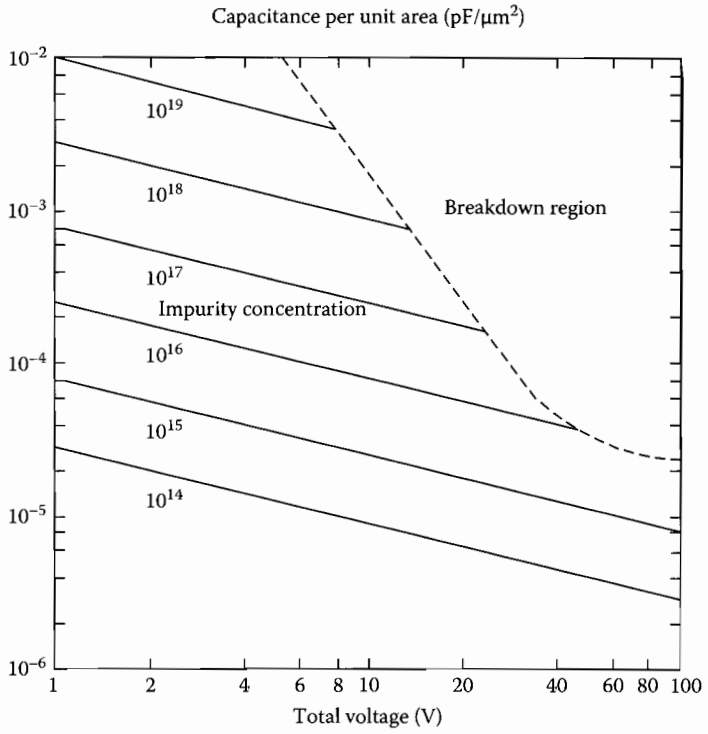

FIGURE 1.79 Junction capacitance as a function of the total voltage and the concentration on the lightly doped side.

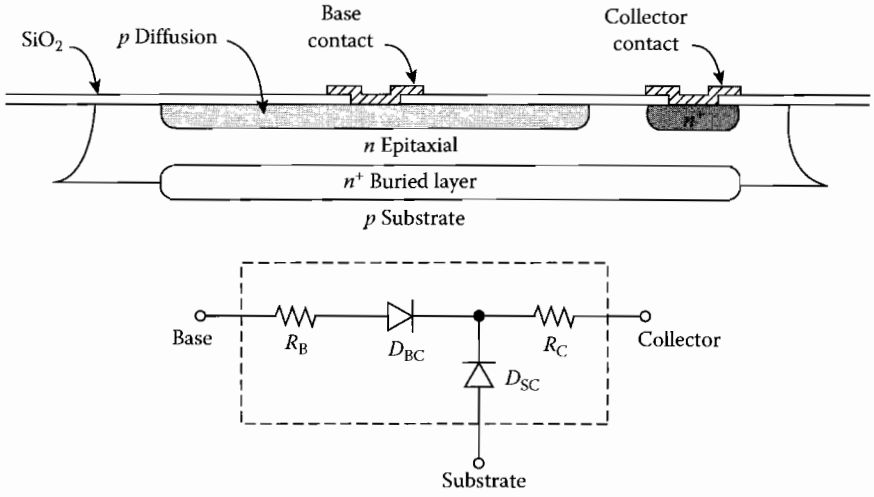

FIGURE 1.80 Base–collector junction capacitor.

For the base–emitter junction to operate in reverse bias, the n-type emitter must be connected to a voltage relatively higher than the voltage at the p-type base. The breakdown voltage of the base–emitter junction is relatively low, determined by the BV_{EBO} of the npn transistor, which has a typical value of about 6 V.

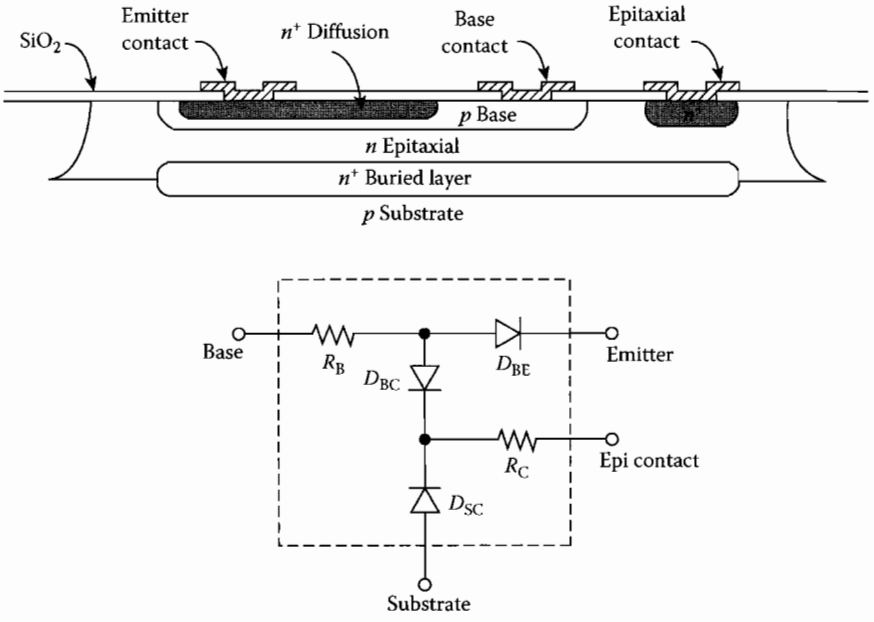

FIGURE 1.81 Base-emitter junction capacitor.

1.4.2.2 MOS Capacitors

MOS capacitors are preferable and commonly used in ICs since they are linear and not confined to a reverse-biased operating condition as in the junction capacitors. The structure of a MOS capacitor is shown in Figure 1.82, where by means of a local oxidation process a thin oxide layer is thermally grown on top of a heavily doped n^+ diffusion layer. The oxide layer has a typical thickness between 500 and 1500 Å (Å $= 10^{-10}$ m $= 10^{-4}$ μm) and functions as the insulating layer of the parallel-plate capacitor. The top plate is formed by overlapping the thin oxide area with a deposited layer of conductive metal. The bottom-plate diffusion layer is heavily doped for two reasons: to minimize the bottom-plate resistance and to minimize the depletion width at the oxide-semiconductor interface when the capacitor operates in the *depletion* and *inversion* modes [17]. By keeping the depletion width small, the effective capacitance is dominated by the parallel-plate oxide capacitance. The MOS capacitance is thus given by

$$C = \frac{\kappa_{ox}\varepsilon_0}{T} A \tag{1.254}$$

where

κ_{ox} is the relative dielectric constant of SiO_2 (2.7 to 4.2)
ε_0 is the permittivity constant
T is the oxide thickness
A is the area defined by the thin oxide layer

In practice, a thin layer of silicon nitride (Si_3N_4) is often deposited on the thin oxide layer and is used to minimize the charges inadvertently introduced in the oxide layer during oxidation and subsequent processing steps. These *oxide charges* are trapped within the oxide and can cause detrimental effect to the capacitor characteristic [17]. The silicon nitride assimilates an additional insulating layer and effectively

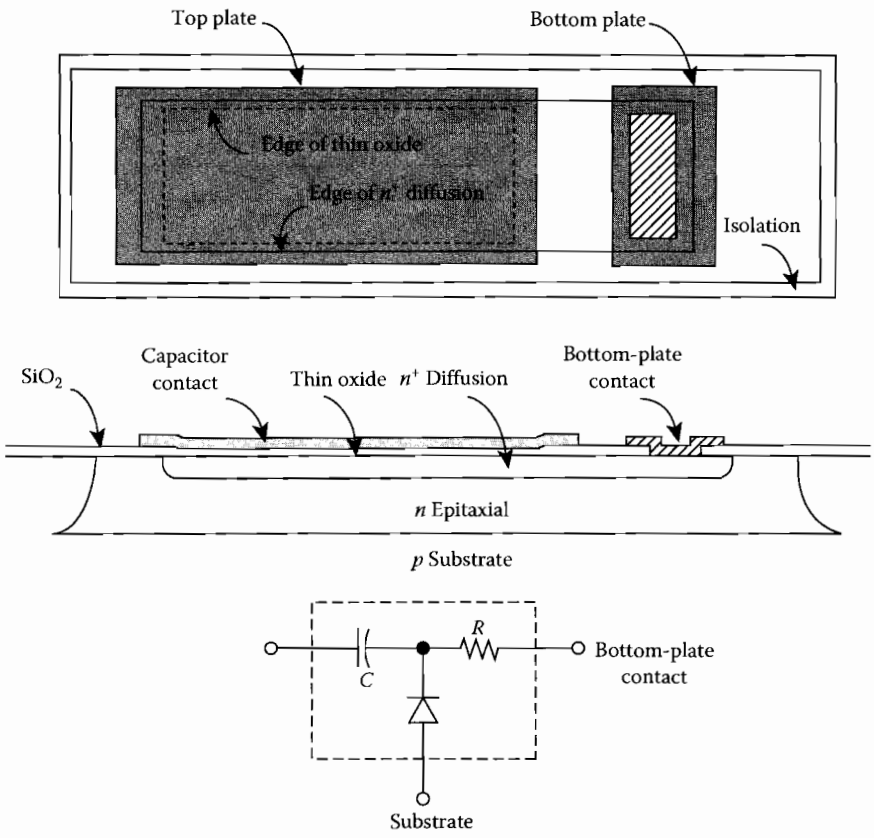

FIGURE 1.82 MOS capacitor.

creates an additional capacitor in series with the oxide capacitor. The capacitance for such a structure can be determined by an application of *Gauss's law*. It is given by

$$C = \frac{\varepsilon_o}{\left(\frac{T_{ni}}{\kappa_{ni}}\right) + \left(\frac{T_{ox}}{\kappa_{ox}}\right)} A \tag{1.255}$$

where

T_{ni} and T_{ox} are the thickness of the silicon nitride and oxide layers, respectively

κ_{ni} (2.7 to 4.2) and κ_{ni} (3.5 to 9) are the relative dielectric constant of oxide and silicon nitride, respectively

In the equivalent circuit model of Figure 1.82, the parasitic junction between the p-type substrate and the n-type bottom plate must always be reverse biased. The bottom-plate contact must be connected to a voltage relatively higher than the substrate voltage.

1.4.2.3 Polysilicon Capacitors

Polysilicon capacitors are conveniently available in MOSFET technology, where the gate of the MOSFET transistor is made of polysilicon material. Polysilicon capacitors also assimilate the parallel-plate capacitor. Figure 1.83 shows a typical structure of a polysilicon capacitor, where a thin oxide is deposited on top of a polysilicon layer and serves as an insulating layer between the top-plate metal layer and the

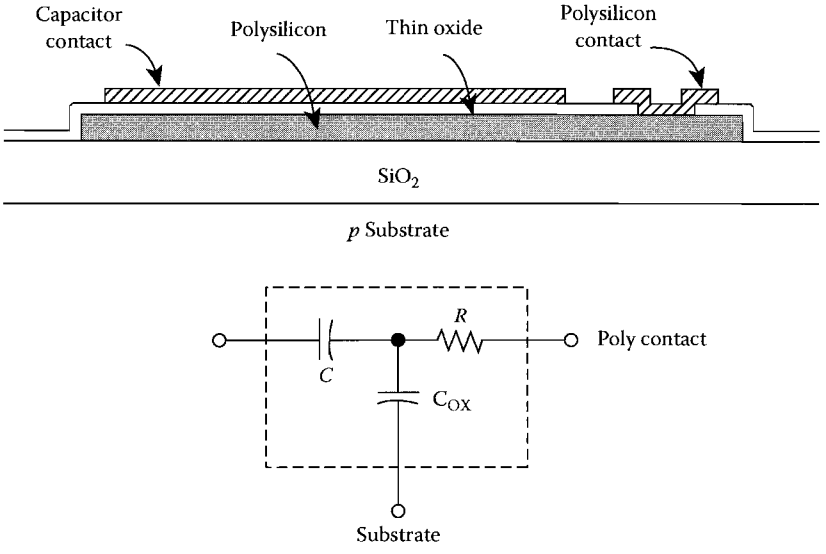

FIGURE 1.83 Polysilicon capacitor.

bottom-plate polysilicon layer. The polysilicon region is isolated from the substrate by a thick oxide layer that forms a parasitic parallel-plate capacitance between the polysilicon layer and the substrate. This parasitic capacitance must be accounted for in the equivalent circuit model. The capacitance of the polysilicon capacitor is determined by either Equation 1.254 or 1.255 depending on whether a thin silicon nitride is used in conjunction with the thin oxide.

1.4.3 Inductors

Planar inductors have been implemented using a variety of substrates such as standard PC boards, ceramic and sapphire hybrids, monolithic GaAs [24], and more recently monolithic silicon [18]. In the early development of silicon technology, planar inductors were investigated [26], but the prevailing lithographic limitations and relatively large inductance requirements (for low-frequency applications) resulted in excessive silicon area and poor performance. Reflected losses from the conductive silicon substrate were a major contribution to low inductor *Q*. Recent advances in silicon IC processing technology have achieved fabrication of metal width and metal spacing in the low micrometer range and thus allow many more inductor turns per unit area. Also, modern oxide-isolated processes with multilayer metal options allow thick oxides to help isolate the inductor from the silicon substrate. Practical applications of monolithic inductors in low-noise amplifiers, impedance matching amplifiers, filters and microwave oscillators in silicon technologies have been successfully demonstrated [19,20].

Monolithic inductors are especially useful in high-frequency applications where inductors of a few nano-Henrys of inductance are sufficient. Inductor structures in monolithic form include strip, loop, and spiral inductors. Rectangular and circular spiral inductors are by far the most commonly used structures.

1.4.3.1 Rectangular Spiral Inductors

The structure of a rectangular spiral inductor is shown in Figure 1.84, where the spiral loops are formed with the top metal layer M_2 and the connector bridge is formed with the bottom metal layer M_1. Using the top metal layer to form the spiral loops has the advantage of minimizing the parasitic metal-to-substrate

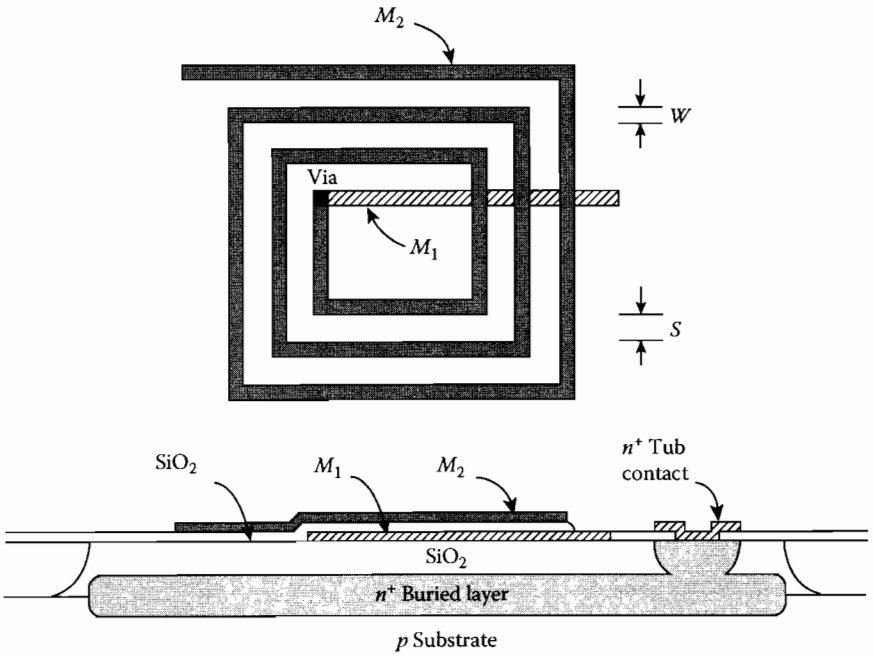

FIGURE 1.84 Rectangular spiral inductor.

capacitance. The metal width is denoted by W and the metal spacing is denoted by S. The total inductance is given by

$$L_T = \sum_{i=1}^{4N} L_S(i) + 2 \cdot \sum_{i=1}^{4N-1} \sum_{j=i+1}^{4N} L_M(ij) \tag{1.256}$$

where

N is the number of turns
$L_S(i)$ is the *self inductance* of the rectangular metal segment i
$L_M(ij)$ is the *mutual inductance* between metal segments i and j

The self-inductance is due to the magnetic flux surrounding each metal segment. The mutual inductance is due to the magnetic flux coupling around every two parallel metal segments and has a positive value if the currents applied to the metal conductors flow in the same direction and a negative value otherwise. Perpendicular metal segments have negligible mutual inductance.

The self-inductance and mutual inductance for straight rectangular conductors can be determined by the *geometric mean distance* method [10], in which the conductors are replaced by equivalent straight filaments whose basic inductive characteristics are well known.

Self-inductance. The self-inductance for the rectangular conductor of Figure 1.85 depends on the conductor length L, the conductor width W, and the conductor thickness T. The static self-inductance is given by [9,10].

$$L_S = 2L\left[\ln\left(\frac{2L}{GMD}\right) - 1.25 + \left(\frac{AMD}{L}\right) + \left(\frac{\mu_r}{4}\right)\zeta\right] (nH) \tag{1.257}$$

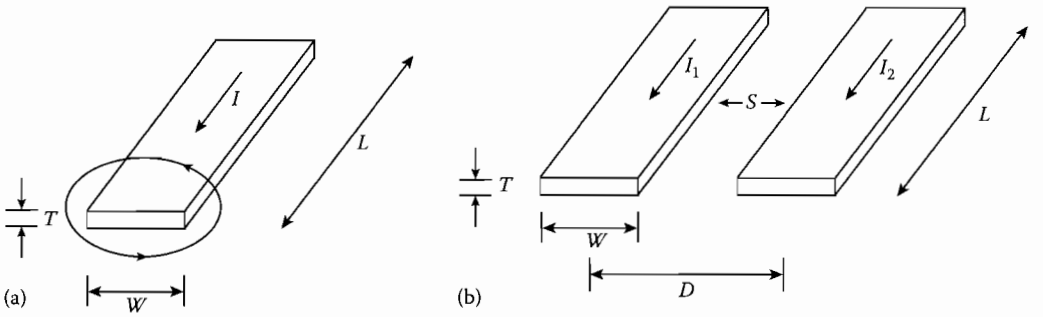

FIGURE 1.85 Calculation of (a) self-inductance and (b) mutual inductance for parallel rectangular conductors.

where

μ_r is the relative permeability constant of the conductor
GMD is the geometric mean distance
AMD is the arithmetic mean distance
ζ is a frequency-dependent parameter that equals 1 for direct and low-frequency alternating currents and approaches 0 for very high-frequency alternating currents

The AMD and GMD for the rectangular conductor of Figure 1.85 are

$$\text{AMD} = \left(\frac{W+T}{3}\right)$$
$$\text{GMD} = \begin{cases} 0.22313 \cdot (W+T) & T \to 0 \\ 0.22360 \cdot (W+T) & T = \text{W}/2 \\ 0.223525 \cdot (W+T) & T \to \text{W} \end{cases} \tag{1.258}$$

The rectangular dimensions L, W, and T are normalized to the centimeter in the preceding expressions.

Mutual inductance. The mutual inductance for the two parallel rectangular conductors of Figure 1.85 depends on the conductor length L, the conductor width W, and the conductor thickness T, and the distance D separating the conductor centers. The static mutual inductance is [10]

$$L_M = 2L\alpha(\text{nH}) \tag{1.259}$$

where

$$\alpha = \ln\left[\left(\frac{L}{\text{GMD}}\right) + \left[1 + \left(\frac{L}{\text{GMD}}\right)^2\right]^{1/2}\right] - \left[1 + \left(\frac{\text{GMD}}{L}\right)^2\right]^{1/2} + \left(\frac{\text{GMD}}{L}\right)$$

and

$$\text{GMD} = \exp(\ln D - \beta) \tag{1.260}$$

$$\beta = \begin{cases} \frac{1}{12}\left(\frac{D}{W}\right)^{-2} + \frac{1}{60}\left(\frac{D}{W}\right)^{-4} + \frac{1}{168}\left(\frac{D}{W}\right)^{-6} + \frac{1}{360}\left(\frac{D}{W}\right)^{-8} \\ + \frac{1}{660}\left(\frac{D}{W}\right)^{-10} + \cdots \\ 0.1137 \quad \text{for} \quad D = W \end{cases}$$

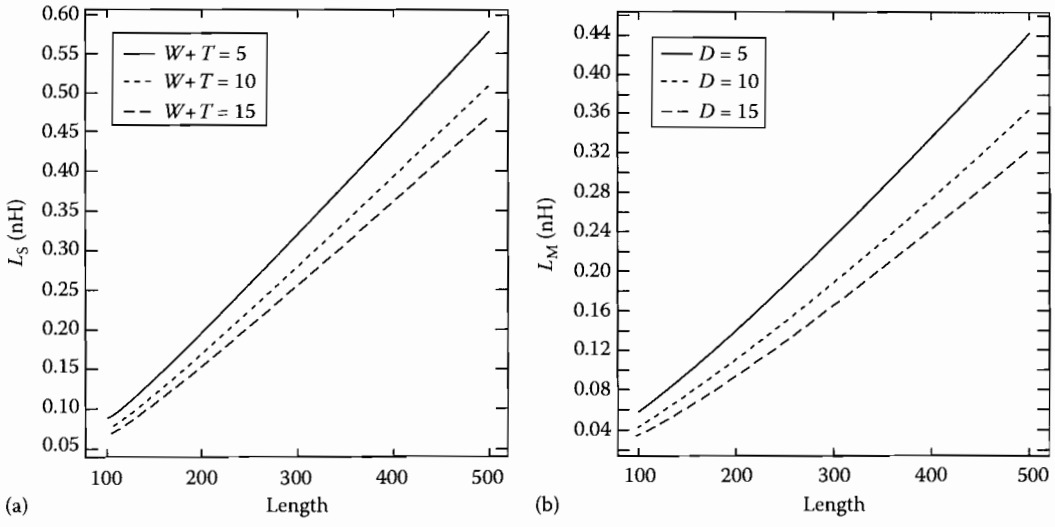

FIGURE 1.86 (a) Self-inductance as a function of width, thickness, and length for rectangular conductors. (b) Mutual inductance as a function of distance and length for rectangular conductors ($W=5$, $T=0$).

The GMD closed-form expression Equation 1.260 is valid for rectangular conductors with small thickness-to-width ratios T/W. As the thickness T approaches the width W, the GMD approaches the distance D and the GMD is no longer represented by the above closed-form expression. Figure 1.86 shows plots of the self inductance and the mutual inductance as expressed in Equations 1.257 and 1.259, respectively. The conductor dimensions are given in μm (μm $= 10^{-4}$ cm).

For the inductor structure of Figure 1.84 it is important to emphasize that since the spiral loops are of nonmagnetic metal material, the total inductance depends only on the geometry of the conductors and not on the current strength. At high-frequencies, especially those above the *self-resonant* frequency of the inductor, the *skin effect* due to current crowding toward the surface and the propagation delay as the current traverses the spiral must be fully accounted for [16,22]. The ground-plane effect due to the inductor image must also be considered regardless of the operation frequency.

An equivalent lumped model for the rectangular spiral inductor of Figure 1.84 is shown in Figure 1.87. This model consists of the total inductance L_T, the accumulated metal resistance R_S, the coupling

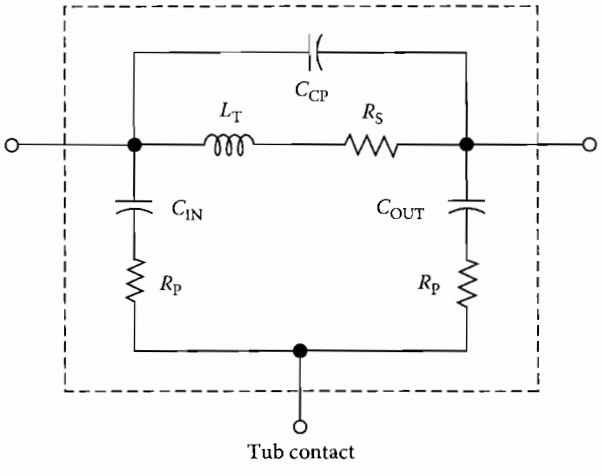

FIGURE 1.87 Electrical model for the spiral inductor.

capacitance C_{CP} between metal segments due to the electric fields in both the oxide region and the air region, the parasitic capacitances C_{IN} and C_{OUT} from the metal layers to the buried layer [2,11,15], and the buried-layer resistance R_P. Since the spiral structure of Figure 1.84 is not symmetrical, the parasitic capacitors C_{IN} and C_{OUT} are not the same, though the difference is relatively small. The self-resonant frequency can be approximated using the circuit model of Figure 1.87 with one side of the inductor being grounded. For simplicity, let $C_{IN} = C_{OUT} + C_P$ and neglect the relatively small coupling capacitor C_{CP}, the self-resonant frequency is given by

$$f_R = \frac{1}{2\pi} \frac{1}{\sqrt{L_T C_P}} \left[\frac{1 - R_S^2 \left(\frac{C_P}{L_T} \right)}{1 - R_P^2 \left(\frac{C_P}{L_T} \right)} \right]^{1/2} \tag{1.261}$$

Transformer structures. Transformers are often used in high-performance analog ICs that require conversions between single-ended signals and differential signals. In monolithic technology, transformers can be fabricated using the basic structure of the rectangular spiral inductor. Figure 1.88 shows a planar interdigitated spiral transformer that requires only two metal layers M_1 and M_2. The structure of Figure 1.89, on the other hand, requires three layers of metal for which the top metal layer M_3 is used for the upper spiral, the middle metal layer M_2 is used for the lower spiral, and the bottom metal layer M_1 is used for the two connector bridges. This structure can achieve a higher inductance per unit than that of Figure 1.88 due to a stronger magnetic coupling between the upper spiral and the lower spiral through a relatively thin oxide layer separating metal layers M_2 and M_3. An equivalent lumped model is

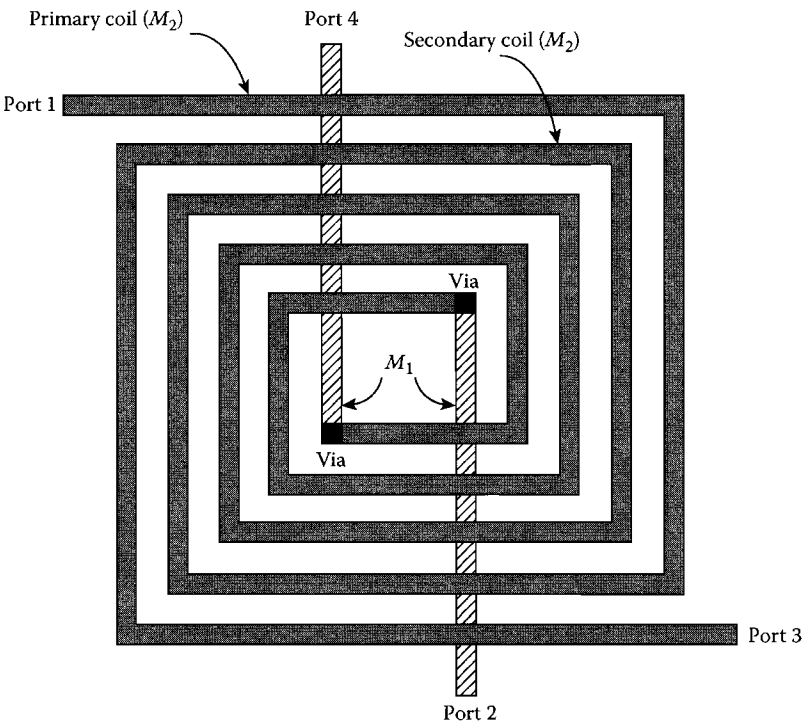

FIGURE 1.88 Rectangular spiral transformer I.

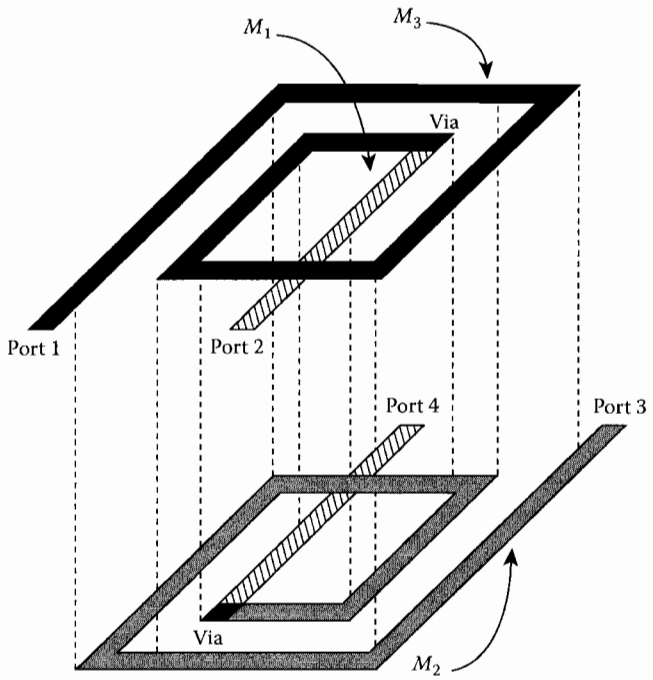

FIGURE 1.89 Rectangular spiral transformer II.

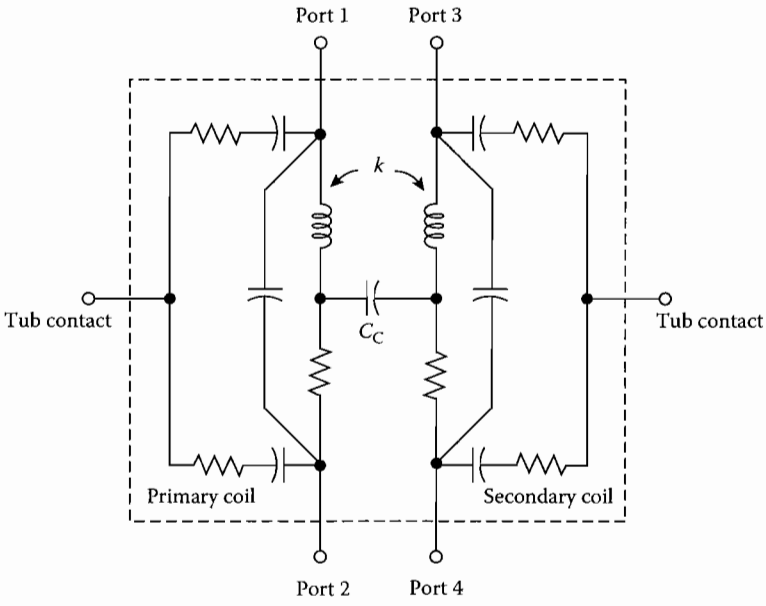

FIGURE 1.90 Electrical model for the spiral transformer.

shown in Figure 1.90. In addition to all the circuit elements of the two individual spiral inductors, there are also a magnetic coupling factor k and a coupling capacitance C_C between the primary and secondary coils.

1.4.3.2 Circular Spiral Inductors

The structure of a concentric circular spiral inductor is shown in Figure 1.91 where the circular loops share the same center point. The top metal layer M_2 is used for the circular conductors and the bottom metal layer M_1 is used for the connector bridge. The metal width is denoted by W and the spacing between two adjacent loops is denoted by S. The total inductance is given by

$$L_T = \sum_{i=1}^{N} L_S(i) + 2 \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} L_M(ij) \tag{1.262}$$

where

N is the number of circular turns
$L_S(i)$ is the *self-inductance* of the circular conductor i
$L_M(ij)$ is the *mutual inductance* between conductors i and j

Self-inductance. Consider the single circular conductor of Figure 1.92a that has a radius R and a width W. A current I applied to this conductor produces a magnetic flux encircled by the loop and another magnetic flux inside the conductor itself. The inductance associated with the former and the latter magnetic flux component is referred to as the *external* self-inductance and the *internal* self-inductance, respectively. The external self-inductance characterizing the change in the encircled magnetic flux to the change in current is [25].

$$L_S = \mu(2R - \delta)\left[\left(1 - \frac{k^2}{2}\right)K(k) - E(k)\right] \text{(nH)} \tag{1.263}$$

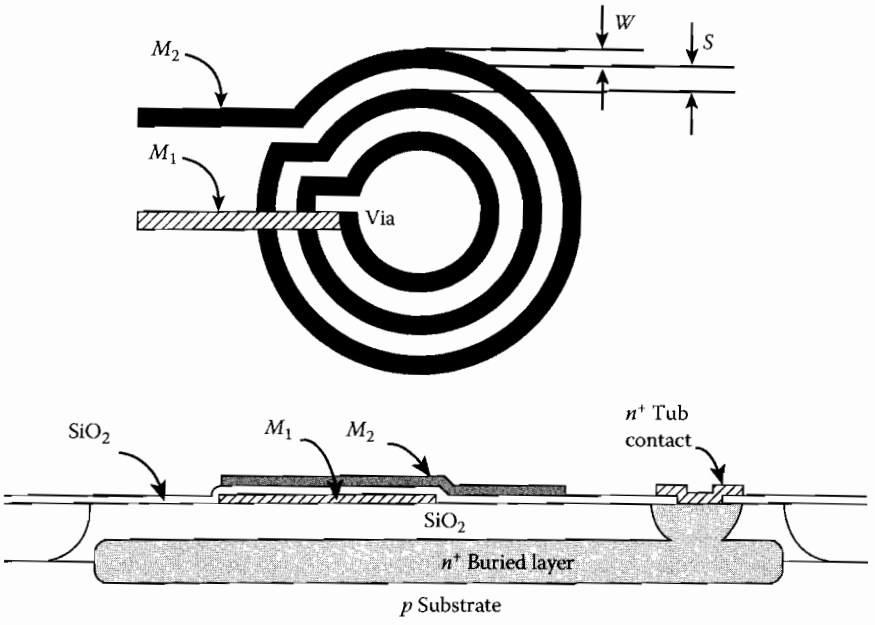

FIGURE 1.91 Concentric circular spiral inductor.

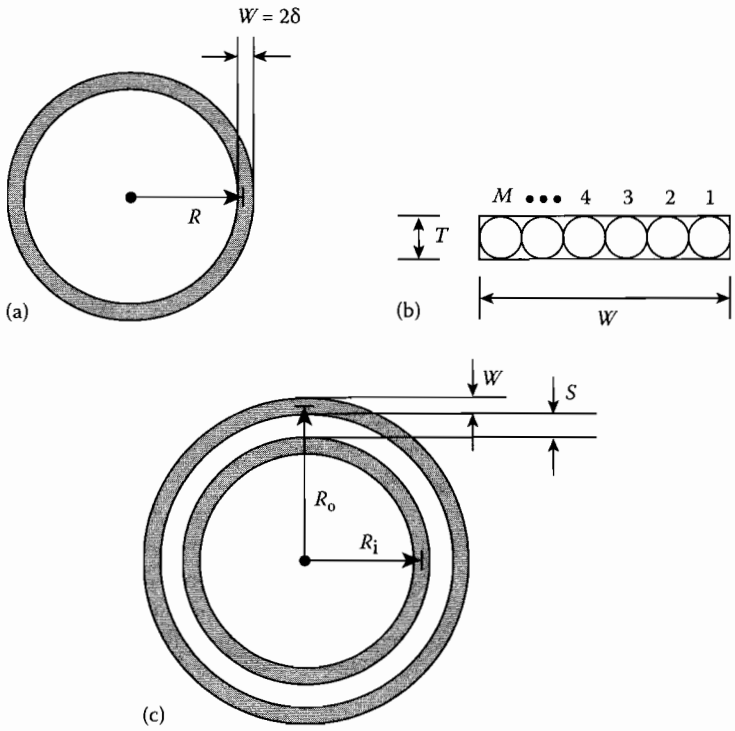

FIGURE 1.92 Calculation of self-inductance and mutual inductance for circular conductors. (a) External self-inductance; (b) internal self-inductance; and (c) mutual inductance.

where

$$k^2 = \frac{4R(R-\delta)}{(2R-\delta)^2} \tag{1.264}$$

and μ is the permeability of the conductor (equals 4π nH/cm for nonmagnetic conductors), and δ is one-half the conductor width W. $K(k)$ and $E(k)$ are the *complete elliptic integrals* of the first and second kind, respectively, and are given by

$$K(k) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1-k^2\sin^2\phi}}$$

$$E(k) = \int_0^{\pi/2} \sqrt{1-k^2\sin^2\phi}\, d\phi$$

The internal self-inductance is determined based on the concept of magnetic field energy. As shown in Figure 1.92b, the flat circular conductor is first approximated by an M number of round circular conductors [14] that are electrically in parallel and each conductor has a diameter equal to the thickness T of the flat conductor. The internal self-inductance of each round conductor is then determined as [25].

$$L = \frac{\mu}{8\pi}(\text{nH/cm})$$

The internal self-inductance of the flat conductor thus equals the parallel combination of these M components

$$L_S \approx \frac{\mu}{4}\left\{\sum_{i=1}^{M}[R - \delta + T(i - 0.5)]^{-1}\right\}^{-1}(\text{nH}) \tag{1.265}$$

where $R - \delta + T(i - 0.5)$ is the effective radius from the center of the loop to the center of the round conductor i. The typical contribution from the internal self-inductance of Equation 1.265 is less than 5% the contribution from the external self-inductance of Equation 1.263.

Mutual inductance. The mutual inductance of the two circular loops of Figure 1.92c depends on the inner radius R_i and the outer radius R_o. For any two adjacent loops of the circular spiral inductor, the outer radius is related to the inner radius by the simple relation $R_o = R_i + (W + S)$. The mutual inductance is determined based on the *Neumann's* line integral given as follows:

$$L_M = \frac{\mu}{4\pi}\int_C\int_C \frac{\mathbf{dl}_1 \cdot \mathbf{dl}_2}{D}$$

where

$\mathbf{dl}_1 \cdot \mathbf{dl}_2$ represents the dot product of the differential lengths
D is the distance separating the differential $\mathbf{l}_1$ vector and $\mathbf{l}_2$ vector

The static mutual inductance [25] is

$$L_M = \mu\sqrt{R_i R_o}\left[\left(\frac{2}{k} - k\right)K(k) - \frac{2}{k}E(k)\right](\text{nH}) \tag{1.266}$$

where

$$k^2 = \frac{4R_iR_o}{(R_i + R_o)^2} \tag{1.267}$$

Figure 1.93 shows plots of the external self-inductance and the mutual inductance as expressed in Equations 1.263 and 1.266, respectively. The conductor dimensions are given in μm.

As in the rectangular spiral inductor, the ground-plane effect and the retardation effect of the circular spiral inductor must be fully accounted for. The circuit model of Figure 1.87 can be used to characterize the electrical behavior of the circular inductor.

A comparison between the rectangular spiral of Figure 1.84 and the circular spiral of Figure 1.91 is shown in Figure 1.94, where the total inductance L_T is plotted against the turn number N. Both inductors have the same innermost dimension, the same conductor width, space, and thickness. The dimensions are given in μm, and the ground-plane effect and the retardation effect are not considered. For a given turn number, the rectangular spiral yields a higher inductance per semiconductor area than the circular spiral. Figure 1.95 shows a plot of the inductor Q vs. the total inductance of the same spiral inductors under consideration. Due to a higher inductance per length ratio, the Q of the circular inductor is higher than that of the rectangular inductor, about 10% for high inductance values.

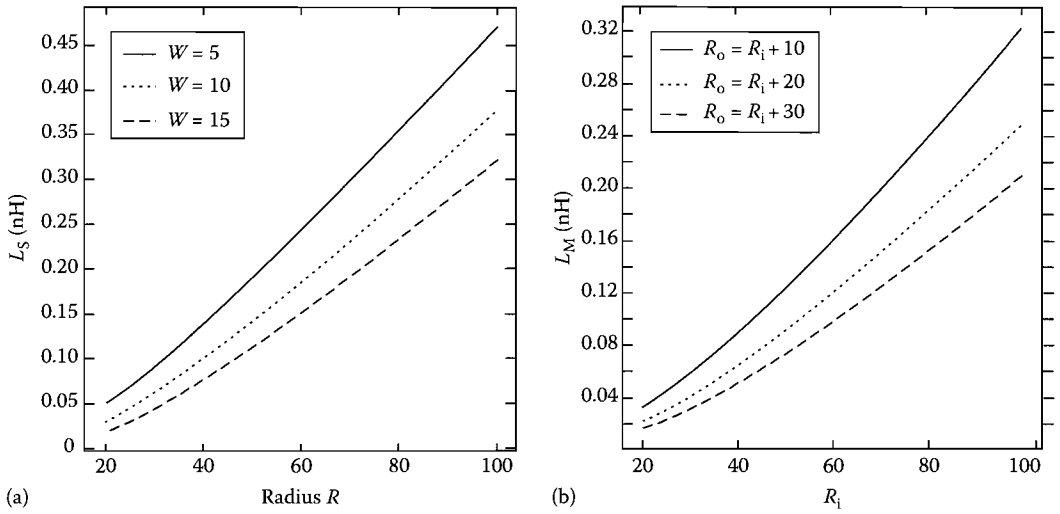

FIGURE 1.93 (a) External self-inductance as a function of radius and width for circular conductors. (b) Mutual inductance as a function of radii R_i and R_o for circular conductors.

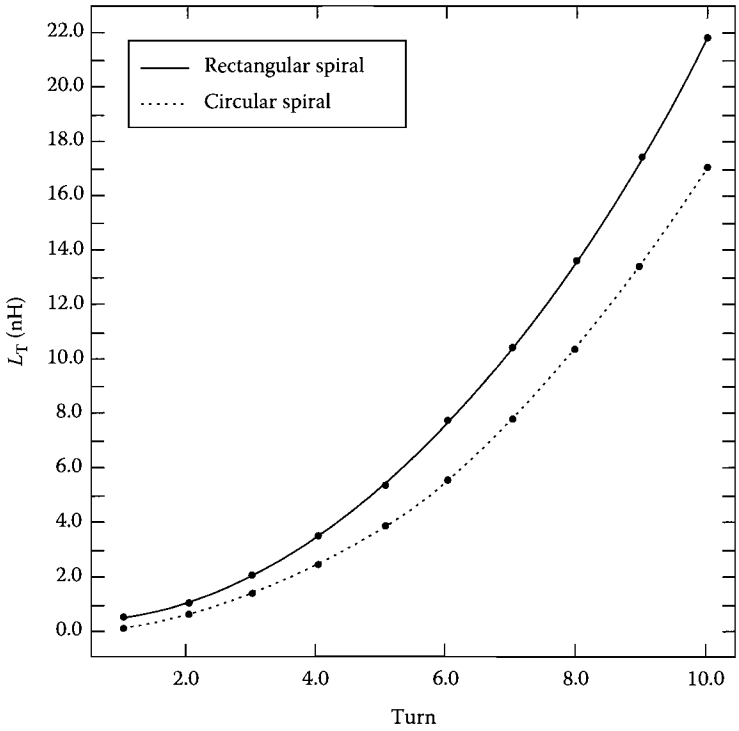

FIGURE 1.94 Total static inductance vs. turn number for the rectangular and circular inductors. The ground-plane effect is neglected. Innermost center dimension is 88 by 88, $W = 6$, $S = 3$, $T = 1.2$.

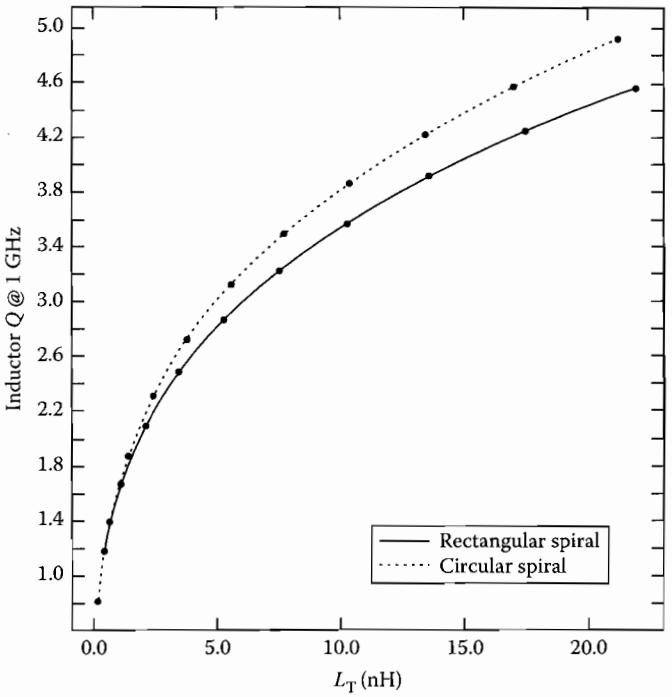

FIGURE 1.95 Inductor Q vs. total inductance for the rectangular and circular inductors. Metal sheet resistance is 25 mΩ/N. Innermost dimension is 88, $W = 6$, $S = 3$, $T = 1.2$.

References

1. I. Bahl and P. Bhartia, *Microwave Solid State Circuit Design*, New York: Wiley, 1988.
2. T. G. Bryant and J. A. Weiss, Parameters of microstrip transmission lines and of coupled pairs of microstrip lines, *IEEE Trans. Microwave Theory Tech.*, MTT-16, 1021–1027, 1968.
3. H. R. Camenzind, *Electronic Integrated Systems Design*, New York: Van Nostrand Reinhold, 1972.
4. E. M. Conwell, Properties of silicon and germanium, *Proc. IRE*, 46, 1281–1300, 1958.
5. R. Garg and I. J. Bahl, Characteristics of coupled microstrip lines, *IEEE Trans. Microwave Theory Tech.*, MTT-27, 700–705, 1979.
6. F. R. Gleason, Thin-film microelectronic inductors, in *Proc. Nat. Electron. Conf.*, 1964, pp. 197–198.
7. P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 2nd ed., New York: Wiley, 1984.
8. A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design*, New York: Wiley, 1984.
9. H. M. Greenhouse, Design of planar rectangular microelectronic inductor, *IEEE Trans. Parts, Hybrids, Packaging*, PHP-10, 101–109, 1974.
10. F. W. Grover, *Inductance Calculations*, New York: Van Nostrand, 1946.
11. E. Hammerstad and O. Jensen, Accurate models for microstrip computer-aided design, *IEEE MTT-S Dig.*, 80, 407–409, 1980.
12. J. C. Irwin, Resistivity of bulk silicon and of diffused layers in silicon, *Bell Syst. Tech. J.*, 41, 387–410, 1962.
13. C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta, A review of some charge transport properties of silicon, *Solid State Electron.*, 20, 77–89, 1977.

14. R. L. Kemke and G. A. Burdick, Spiral inductors for hybrid and microwave applications, in *Proc. Electron. Components Conf.*, 1974, pp. 152–161.
15. M. Kirschning and R. H. Jansen, Accurate wide-range design equations for the frequency-dependent characteristics of parallel-coupling microstrip lines, *IEEE Trans. Microwave Theory Tech.*, MTT-32, 83–90, 1984.
16. D. Krafcsik and D. Dawson, A close-form expression for representing the distributed nature of the spiral inductor, *IEEE MTT-S Dig.*, 86, 87–92, 1986.
17. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed., New York: Wiley, 1986.
18. N. M. Nguyen and R. G. Meyer, Si IC-compatible inductors and LC passive filters, *IEEE J. Solid-State Circuits*, 25, 1028–1031, 1990.
19. N. M. Nguyen and R. G. Meyer, A Si bipolar monolithic RF bandpass amplifier, *IEEE J. Solid-State Circuits*, 27, 123–127, 1992.
20. N. M. Nguyen and R. G. Meyer, A 1.8-GHz monolithic LC voltage-controlled oscillator, *IEEE J. Solid-State Circuits*, 27, 444–450, 1992.
21. N. M. Nguyen and R. G. Meyer, Start-up and frequency stability in high-frequency oscillators, *IEEE J. Solid-State Circuits*, 27, 810–820, 1992.
22. M. Parisot, Y. Archambault, D. Pavlidis, and J. Magarshack, Highly accurate design of spiral inductors for MMIC's with small size and high cut-off frequency characteristics, *IEEE MTT-S Dig.*, 84, 106–110, 1984.
23. E. Pettenpaul, H. Kapusta, A. Weisgerber, H. Mampe, J. Luginsland, and I. Wolff, CAD models of lumped elements on GaAs up to 18 GHz, *IEEE Trans. Microwave Theory Tech.*, 36, 294–304, 1988.
24. R. A. Pucel, Design considerations for monolithic microwave circuits, *IEEE Trans. Microwave Theory Tech.*, MTT-29, 513–534, 1981.
25. S. Ramon, J. R. Whinnery, and T. V. Duzer, *Fields and Waves in Communication Electronics*, 2nd ed., New York: Wiley, 1984.
26. R. M. Warner, Jr., and J. N. Fordemwalt, *Integrated Circuits*, New York: McGraw-Hill, 1965.

1.5 Chip Parasitics in Analog Integrated Circuits

Martin A. Brooke

The parasitic elements in electronic devices and interconnect limit the performance of all ICs. No amount of improvement in device performance or circuit design can completely eliminate these effects. Thus, as circuit speeds increase, unaccounted for interconnect parasitics become a more and more common cause of analog IC design failure. Hence, the causes, characterization, and modeling of significant interconnect parasitics are essential knowledge for good analog IC design [1–4].

1.5.1 Interconnect Parasitics

The parasitics due to the wiring used to connect devices together on chip produce a host of problems. Unanticipated feedback through parasitic capacitances can cause unwanted oscillation. Mismatch due to differences in interconnect resistance contribute to unwanted offset voltages. For very-high-speed ICs, the inductance of interconnects is both a useful tool and a potential cause of yield problems.

Even the interactions between interconnect lines are both important and very difficult to model. So too are the distributed interactions of resistance, capacitance, and (in high-speed circuits) inductance that produce transmission line effects.

1.5.1.1 Parasitic Capacitance

Distributed capacitance of IC lines is perhaps the most important of all IC parasitics. It can lower the bandwidth of amplifiers, alter the frequency response of filters, and cause oscillations.

Physics. Every piece of IC interconnect has capacitance to the substrate. In the case of silicon circuitry, the substrate is conductive and connected to an ac ground, thus there is a capacitance to ground from every circuit node due to the interconnect. Figure 1.96 illustrates this substrate capacitance interconnect parasitic. The capacitance value will depend on the total area of the interconnect, and on the length of edge associated with the interconnect. This edge effect is due to the nonuniformity of the electric field at the interconnect edges. The nonuniformity of the electric field at edges is such that the capacitance value is larger for a given area of interconnect near the edge than elsewhere.

In addition to the substrate capacitance, all adjacent pieces of an interconnect will have capacitance between them. This capacitance is classified into two forms, overlap capacitance, and parallel line capacitance (also known as proximity capacitance). Overlap capacitance occurs when two pieces of interconnect cross each other, while parallel line capacitance occurs when two interconnect traces run close to each other for some distance.

When two lines cross each other, the properties of the overlapping region will determine that size of the overlap capacitance. The electric field through a cross section of two overlapping lines is illustrated in Figure 1.97. The electric field becomes nonuniform near the edges of the overlapping region, producing an edge-dependent capacitance term. The capacitance per unit area at the edge is always greater than elsewhere and, if the overlapping regions are small, the edge capacitance effect can be significant.

The size of parallel line capacitance depends on the distance for which the two lines run side by side and on the separation of the lines. Since parallel line capacitance occurs only at the edges of an interconnect, the electric field that produces it is very nonuniform. This particular nonuniformity, as illustrated in Figure 1.98, makes the capacitance much smaller for a given area of interconnect than either overlap or substrate capacitance. Thus, two lines must run parallel for some distance for this capacitance to be important. The nonuniformity of the electric field makes the dependence of the capacitance on line separation highly nonlinear, as a result the capacitance value decreases much more rapidly with separation than it would if it depended linearly on the line separation.

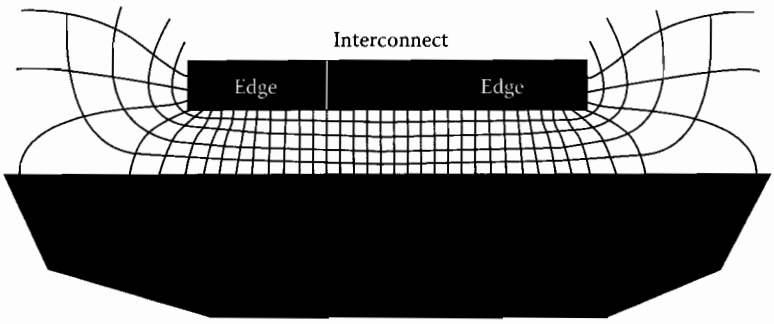

FIGURE 1.96 Substrate capacitance. The electric field distorts at the edges, making the capacitance larger there than elsewhere.

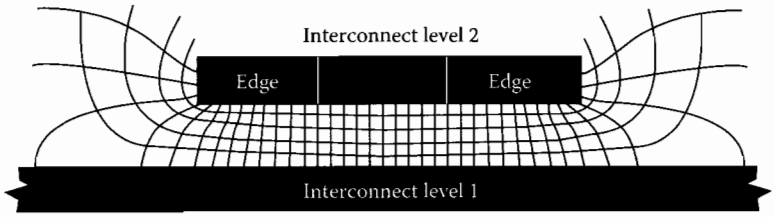

FIGURE 1.97 Overlap capacitance. The bottom interconnect level will have edges into and out of the page with distorted electric field similar to that shown for the top level of interconnect.

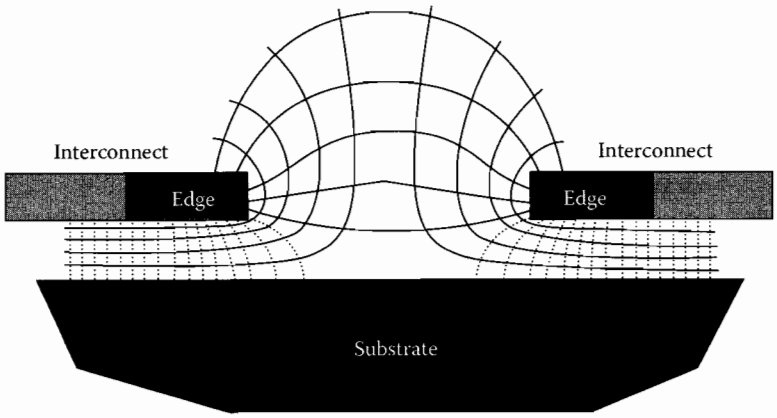

FIGURE 1.98 Parallel line capacitance. Only the solid field lines actually produce line-to-line capacitance, the dashed lines form substrate capacitance.

Modeling. In the absence of significant interconnect resistance effects, all of the parasitic capacitances can be modeled with enough accuracy for most analog circuit design applications by dissecting the interconnect into pieces with similar capacitance characteristics and adding up the capacitance of each piece to obtain a single capacitance term. For example, the dissected view of a piece of interconnect with substrate capacitance is shown in Figure 1.99. The interconnect has been dissected into squares that fall into three

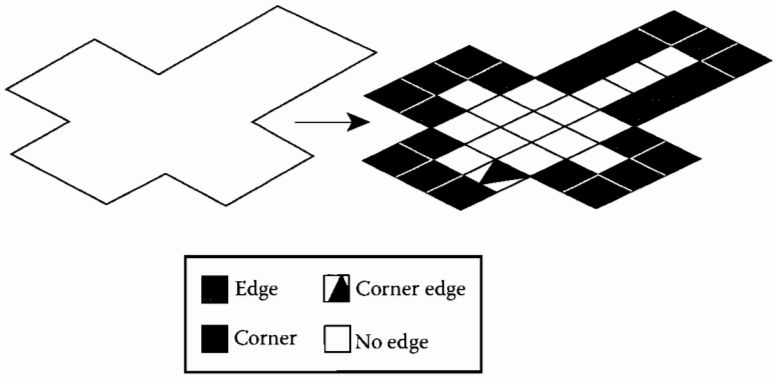

FIGURE 1.99 Determining substrate capacitance. The capacitance of each square in the dissected interconnect segment is summed.

classes: two types of edges, and one center type. The capacitance to the substrate for each of these squares in parallel and thus the total capacitance of the interconnect segment is simply the sum of the capacitance of each square. If the substrate capacitance contribution of each square has been previously measured or calculated, the calculation of the total interconnect segment substrate capacitance involves summing each type of squares capacitance multiplied by the number of squares of that type in the segment.

The accuracy of this modeling technique depends solely on the accuracy of the models used for each type of square. For example, in Figure 1.99, the accuracy could be improved by adding one more type of edge square to those that are modeled. One of these squares has been shaded differently in the figure and is called the corner edge square.

For the nonedge pieces of the interconnect the capacitance is approximately a parallel-plate capacitance and can be computed from Equation 1.268.

$$C = \frac{A \cdot \varepsilon_r \cdot \varepsilon_0}{t} \tag{1.268}$$

where

A is the area of the square or piece of interconnect
t is the thickness of the insulation layer beneath the interconnect
ε_r is the relative dielectric constant of the insulation material
ε_0 is the dielectric constant of free space

For silicon ICs insulated with silicon dioxide the parameters are given in Table 1.10.

The capacitance of edge interconnect pieces will always be larger than nonedge pieces. The amount by which the edge capacitance increases will depend on the ratio of the size of the piece of interconnect and the thickness of the insulation layer beneath the interconnect. If the interconnect width is significantly larger than the thickness of the insulation then edge effects are probably small and can be ignored. However, when thin lines are used in ICs the edge effects are usually significant. The factor by which the edge capacitance can increase over the parallel-plate approximation can easily be as high as 1.5 for thin lines.

The modeling of overlap capacitance is handled in the same fashion as substrate capacitance. The region where interconnect lines overlap is dissected into edges and nonedges and the value of capacitance for each type of square summed up to give a total capacitance between the two circuit nodes associated with each piece of interconnect that overlaps. The area of overlap between the two layers of interconnect can be used as A in Equation 1.268, while that separation between the layers can be used as t. The strong distortion of the electric fields will increase the actual value above this idealized computed value by a factor that depends on the thickness of the lines. This factor can be as high as 2 for thin lines.

Parallel line capacitance can also be handled in a manner similar to that used for substrate and overlap capacitance. However, we must now locate *pairs* of edge squares, one from each of the adjacent interconnect lines. In Figure 1.100, one possible pairing of the squares from adjacent pieces of interconnect is shown. The capacitance for each type of pair of squares is added together, weighted by the number of pairs of each type to get a single capacitance that connects the circuit nodes associated with each interconnect line.

The effect on the capacitance of the spacing between pairs must be either measured or computed for each possible spacing, and type of pair of squares. One approach to this is to use a table of measured or

TABLE 1.10 Parameters for Calculation of Substrate Capacitance in Silicon ICs Insulated with Silicon Dioxide

Parameter	Value
ε_r	3.9
ε_0	$8.854 \cdot 10^{-12}$ F/m
t	$1\text{–}5 \cdot 10^{-6}$ m

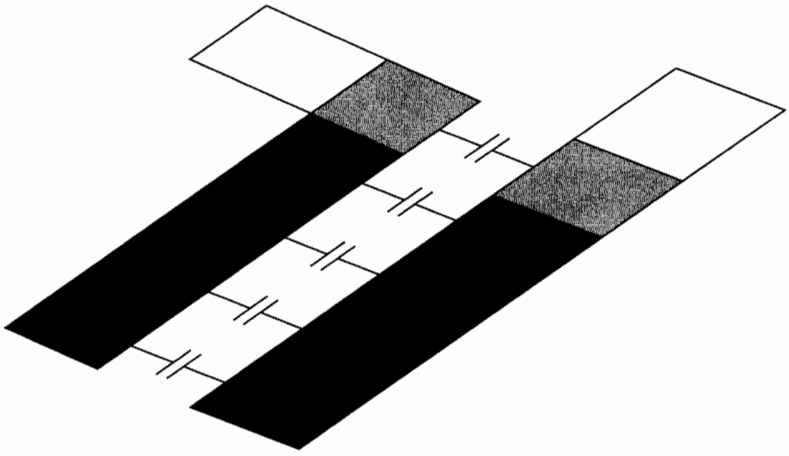

FIGURE 1.100 Determining parallel line capacitance. The differently shaded pairs of squares are different types and will each have a different capacitance between them.

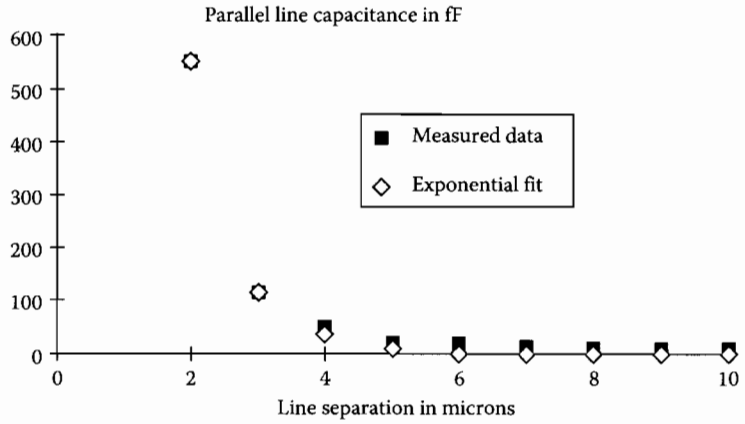

FIGURE 1.101 Parallel line capacitance measured from a silicon IC. The diamonds are an exponential fit to the data (using Equation 1.269). The fit is excellent at short separations when the capacitance is largest.

computed capacitances and separation distances. The measured parallel line capacitance between silicon IC lines for a variety of separations is presented in Figure 1.101. From the figure, we see that the capacitance value decreases exponentially with line separation. Thus an exponential fit to measured or simulated data is good choice for computing the capacitance [7,8].

Equation 1.269 can be used to predict the parallel line capacitance C for each type of pair of edge squares. L is the length of the edge of the squares, and the parameters C_c and S_d are computed or fit to measured coupling capacitance data like that in Figure 1.101.

$$C = C_c \cdot L \cdot e^{-(s/S_d)} \tag{1.269}$$

Effects on circuits. The effects that parasitic capacitances are likely to produce in circuits range from parametric variations, such as reduced bandwidth, to catastrophic failures, such as amplifier oscillation. Each type of parasitic capacitance produces a characteristic set of problems. Being aware of these typical problems will ease diagnosis of actual, or potential, parasitic capacitance problems.

Substrate capacitance usually causes lower than expected bandwidth in amplifiers and lowering of the poles in filters. The capacitance is always to ac ground and thus increases device and circuit capacitances to ground. Thus, circuit nodes that have a dominant effect on amplifier bandwidth, or filter poles, should be designed to have as little substrate capacitance as possible. Another, more subtle, parametric variation that can be caused by substrate capacitance is frequency-dependent mismatch. For example, if the parasitic capacitance to ground is different between the two inputs of a differential amplifier, then, for fast transient signals, the amplifier will appear unbalanced. This could limit the accuracy high-speed comparators, and is sometimes difficult to diagnose since the error only occurs at high speeds.

Overlap and parallel line capacitance can cause unwanted ac connections to be added to a circuit. These connections will produce crosstalk effects and can result in unstable amplifiers. The output interconnect and input interconnect of high-gain or high-frequency amplifiers must thus be kept far apart at all times. Care must be taken to watch for series capacitances of this type. For example, if the output and input interconnect of an amplifier both cross the power supply interconnect, unwanted feedback can result if the power supply line is not well ac grounded. This is a very common cause of IC amplifier oscillation. Because of the potential for crosstalk between parallel or crossing lines, great care should also be taken to keep weak (high-impedance) signal lines away from strong (low-impedance) signal lines.

1.5.1.2 Parasitic Resistance

For analog IC designers, the second most important interconnect parasitic is resistance. This unexpected resistance can cause both parametric problems, such as increased offset voltages, and catastrophic problems such as amplifier oscillation (for example, poorly sized power supply lines can cause resistive positive feedback paths in high gain amplifiers called "ground loops"). To make matters worse, the resistivity of IC interconnect has been steadily increasing as the line widths of circuits have decreased.

Physics. Except for superconductors, all conductors have resistance. A length of interconnect used in an IC is no exception. The resistance of a straight section of interconnect is easily found by obtaining the resistance per square for the particular interconnect layer concerned, and then adding up the resistance of each of the series of squares that makes up the section. This procedure is illustrated in Figure 1.102.

For more complicated interconnect shapes the problem of determining the resistance between two points in the interconnect is also more complex. The simplest approach is to cut the interconnect up into rectangles and assume each rectangle has a resistance equal to the resistance per square of the interconnect material times the number of full and partial squares that will fit along the direction of current flow in the rectangle [5]. This scheme works whenever the direction of current flow is clear; however, for corners and intersections of interconnect the current flow is in fact quite complex. Figure 1.103 shows the kind of current flow that can occur in an interconnect section with complex geometry.

Modeling. To account for the effects of complex current flows the resistance of complex interconnect geometries must be determined by measurement or simulation. One simple empirical approach is to cut out sections of resistive material in the same shape as the interconnect shape to be modeled, and then

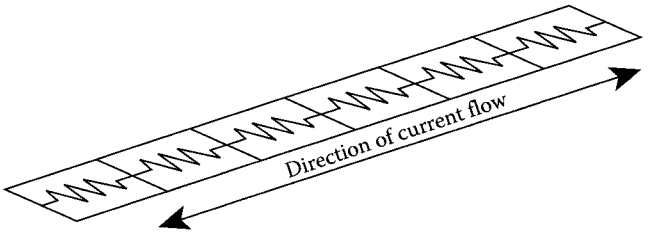

FIGURE 1.102 Determining the resistance of a length of interconnect. Each square has the same resistance regardless of size.

FIGURE 1.103 Current flow in a complex interconnect geometry.

measure the resistance. The resistance for other materials can be found by multiplying by the ratio of the respective resistances per square of the two materials.

Once the resistance has been found for a particular geometry it can be used for any linear scaling of that geometry. For most types of IC interconnect all complex geometries can be broken up into relatively few important subgeometries. If tables of the resistance of these subgeometries for various dimensions and connection patterns are obtained, the resistance of quite complex shapes can be accurately calculated by connecting the resistance of each subgeometry together and calculating the resistance of the connected resistances. This calculation can usually be performed quickly by replacing series and parallel connected resistor pairs with their equivalents. The process of breaking a complex geometry into subgeometries, constructing the equivalent connected resistance, and forming a single resistance for an interconnect section is illustrated in Figure 1.104.

Effects on circuits. The resistance of interconnect can have both parametric and catastrophic effects on circuit performance. Even small differences in the resistance on either input side of a differential amplifier can lead to increased offset voltage. Thus, when designing differential circuits care must be taken to make the interconnect identical on both input sides, as this ensures that the same resistance is present in both circuits.

The resistance of power supply interconnect can lead to both parametric errors in the voltages supplied and catastrophic failure due to oscillation. If power supply voltages are assumed to be identical in two parts of a circuit and, due to interconnect resistance, there is a voltage drop from one point to the next, designs that rely on the voltages being the same may fail. In high-gain and feedback circuits the resistance of the ground and power supply lines may become an unintentional positive feedback resistance which could lead to oscillation. Thus output and input stages for high-gain amplifiers will usually require separate ground and power supply interconnects. This ensures that no parasitic resistance is in a feedback path.

When using resistors provided in an IC process, the extra resistance provided by the interconnect may cause inaccuracies in resistor values. This would be most critical for small resistance values. The only solution in this case is to accurately compute the interconnect resistance. Since most resistance layers

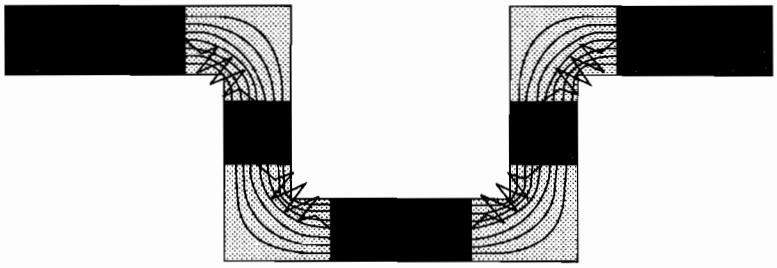

FIGURE 1.104 The process of breaking a complex geometry into subgeometries, constructing the equivalent connected resistance, and forming a single resistance for an interconnect section. In this example, only two subgeometries are used: a corner subgeometry and a basic rectangular subgeometry.

provided in analog IC processes are just a form of high resistivity interconnect, the methods described here for accurately computing the resistance of interconnect are also useful for predicting the resistance of resistors to be fabricated.

1.5.1.3 Parasitic Inductance

In high-speed ICs the inductance of long lines of interconnect becomes significant. In IC technologies that have an insulating substrate, such as gallium arsenide (GaAs) and silicon on insulator (SOI), reasonably high-performance inductive devices can be made from interconnect. In technologies with conductive substrates, resistive losses in the substrate restrict the application of interconnect inductance. High-frequency circuits are often tuned using interconnect inductance and capacitance (*LC*) to form a narrow bandpass filter or tank circuit, and *LC* transmission lines, or stubs, made from interconnect are useful for impedance matching. There is much similarity between this use of parasitic inductance and the design of microstripline-printed circuit boards. The major difference being that inductance does not become significant in IC interconnect until frequencies in the gigahertz are reached.

In order to make a good interconnect inductance, there are two requirements. First, there must not be any resistive material within range of the magnetic field of the inductance. If this occurs then induced currents flowing in the resistive material will make the inductor have high series resistance (low *Q* factor). This would make narrow bandwidth bandpass filters difficult to make using the inductance, and make transmission lines made from the interconnect lossy. The solution is to have an insulating substrate, or to remove the substrate from beneath the inductor.

The second requirement for large inductance is to form a coil or other device to concentrate the magnetic field lines. Within the confines of current IC manufacturing, spiral inductors, like that illustrated in Figure 1.105 are the most common method used to obtain useful inductances.

1.5.1.4 Transmission Line Behavior

Two types of transmission line behavior are important in ICs, *RC* transmission lines and *LC*/*RLC* transmission lines. For gigahertz operation inductive transmission lines are important. These can be lossy *RLC* transmission lines if a conductive substrate such as silicon is used, or nearly lossless *LC* transmission lines if an insulating substrate such as GaAs is used. The design of inductive transmission lines is very similar to designing microstripline-printed circuit boards. At lower frequencies of

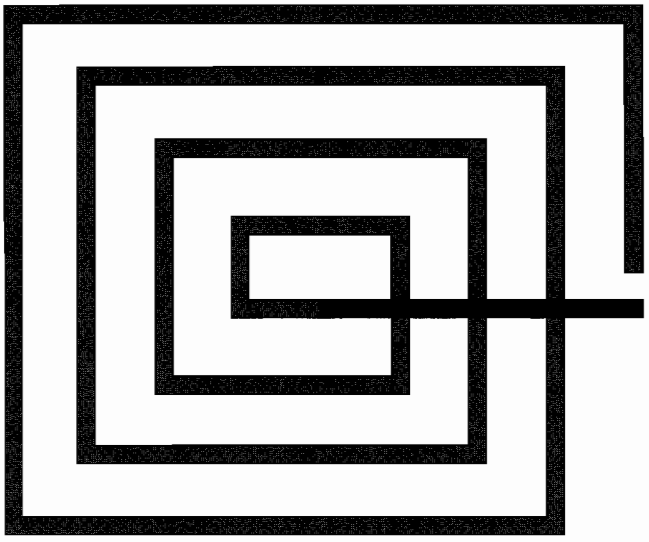

FIGURE 1.105 Spiral inductance used in insulated substrate ICs for gigahertz frequency operation.

10–1000 MHz resistive capacitive (*RC*) transmission lines are important for long low resistivity interconnect lines or short high resistivity lines.

RC transmission lines are of concern to analog circuit designers working in silicon ICs. When used correctly, an interconnect can behave as though it were purely capacitive in nature. However, when a higher resistivity interconnect layer, such as polysilicon or diffusion is used, the distributed resistance and capacitance can start to produce transmission line effects at relatively short distances. Similarly, for very long signal distribution lines or power supply lines, if they are not correctly sized, transmission line behavior ensues.

Physics. One method for modeling distributed transmission line interconnect effects is lumped equivalent modeling [6]. This method is useful for obtaining approximate models of complex geometries quickly, and is the basis of accurate numerical finite element simulation techniques. For analog circuit designers the conversion of interconnect layout sections into lumped equivalent models also provides an intuitive tool to understanding distributed transmission line interconnect behavior.

To be able to model a length of interconnect as a lumped *RC* equivalent, the error between the impedance of the interconnect when correctly treated as a transmission line, and when replaced with the lumped equivalent, must be kept low. If this error is *e*, then it can be shown that the maximum length of interconnect that can be modeled as a simple *RC T* or Π network is given in Equation 1.270. In the equation, *R* is the resistance per square of the particular type of interconnect used, *C* is the capacitance per unit area, and ω is the frequency of operation in radians per second.

$$D < \sqrt{\frac{3 \cdot e}{\omega \cdot R \cdot C}} \tag{1.270}$$

This length can be quite short. Consider the case of a polysilicon interconnect line in a 1.2 μm CMOS process that has a resistance per square of 40 Ω a capacitance per unit are of 0.1 fF/μm^2. For an error *e* of 10% the maximum line length of minimum width line that can be treated as a lumped T or Π network for various frequencies is given in Table 1.11. Longer interconnect lines than this must be cut up into lengths less than or equal to the length given by Equation 1.270.

Modeling. The accurate modeling of distributed transmission line effects in ICs is best performed with lumped equivalent circuits. These circuits can be accurately extracted by dissecting the interconnect geometry into lengths that are, at most, as long as the length given by Equation 1.270. These lengths are then modeled by either a *T* or Π lumped equivalent *RC* network. The extraction of the resistance and capacitance for these short interconnect sections can now follow the same procedures as were described in Sections 1.5.1.2 and 1.5.1.1. The resulting *RC* network is then an accurate transmission line model of the interconnect. Figure 1.106 shows an example of this process.

Effects on circuits. Several parametric and catastrophic problems can arise due to unmodeled transmission line behavior. Signal propagation delays in transmission lines are longer than predicted by a single lumped capacitance and resistance model of interconnect. Thus, ignoring the effects of transmission lines can result in slower circuits than expected. If the design of resistors for feedback networks

TABLE 1.11 The Maximum Length of Minimum Width Polysilicon Line That Can Be Modeled with a Single Lumped *RC T* or π Network and Remain 10% Accurate

Frequency (MHz)	Length (μm)
10	1262
100	399
1000	126

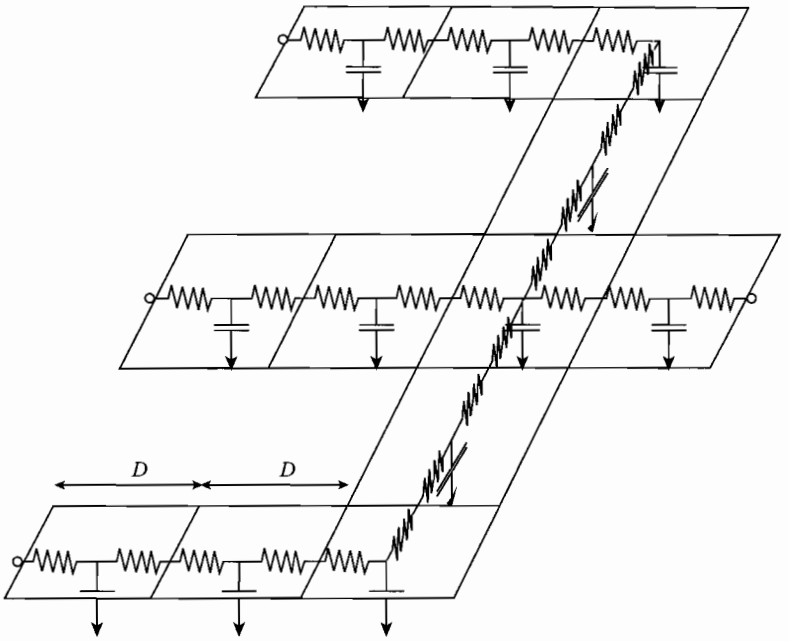

FIGURE 1.106 The extraction of an accurate *RC* transmission line model for resistive interconnect. The maximum allowable length *D* is computed from Equation 1.270.

results in long lengths of the resistive interconnect used to make the resistors, these resistors may in fact be *RC* transmission lines. The extra delay produced by the transmission line may well cause oscillation of the feedback loops using these resistors. The need for decoupling capacitors in digital and analog circuit power supplies is due to the *RC* transmission line behavior of the power supply interconnect. Correct modeling of the *RC* properties of the power distribution interconnect is needed to see whether fast power supply current surges will cause serious changes in the supply voltage or not.

1.5.1.5 Nonlinear Interconnect Parasitics

A number of types of interconnect can have nonlinear parasitics. These nonlinear effects are a challenge to model accurately because the effect can change with the operating conditions of the circuit. A conservative approach is to model the effects as constant at the worst likely value they can attain. This is adequate for predicting parameters, like circuit bandwidth, that need only exceed a specification value. If the specifications call for accurate prediction of parasitics then large nonlinear parasitics are generally undesirable and should be avoided.

Most nonlinear interconnect parasitics are associated with depletion or inversion of the semiconductor substrate. A diffusion interconnect is insulated from conducting substrates such as silicon by a reversed biased diode. This diode's depletion region width varies with the interconnect voltage and results in a voltage-dependent capacitance to the substrate. For example, the diffusion interconnect in Figure 1.107 has voltage-dependent capacitance to the substrate due to a depletion region. The capacitance value depends on the depletion region thickness, which depends on the voltage difference between the interconnect and the substrate.

$$C = C_0 \cdot \left(1 - \left(\frac{V_s}{\varphi_B}\right)\right)^M \tag{1.271}$$

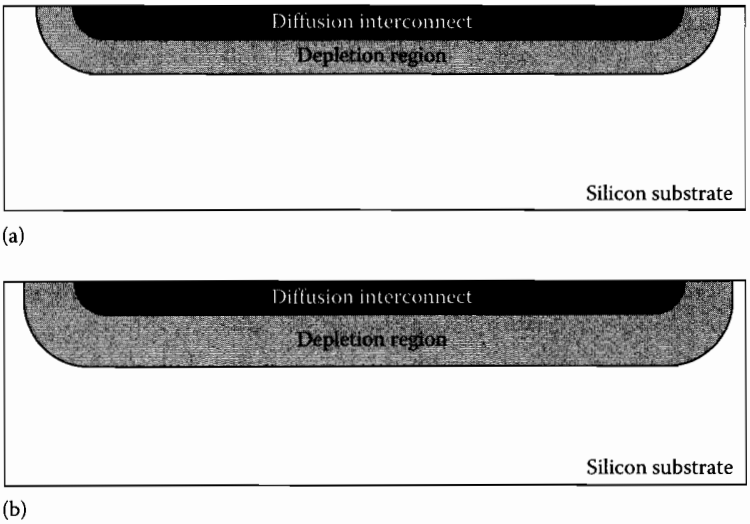

FIGURE 1.107 Diffusion interconnect has a voltage-dependent capacitance produced by the depletion region between the interconnect and the substrate. At low voltage difference between the interconnect and substrate (a), the capacitance is large. However, the capacitance decreases for larger voltage differences (b).

The typical equation for depletion capacitance is given in Equation 1.271. In this equation V_S is the voltage from the interconnect to the substrate, ϕ_B is the built-in potential of the semiconductor junction, M is the grading coefficient of the junction, while C_0 is the zero-bias capacitance of the junction. Since the capacitance is less than C_0 for reverse bias and the junction would not insulate for forward bias, we can assume that the capacitance is always less than C_0 and use C_0 as a conservative estimate of C. Because of the uncertainty in the exact structure of most semiconductor junctions φ_B and M are usually fit to measured capacitance versus voltage (CV) data.

Another common nonlinear parasitic occurs when metal interconnect placed over a conducting semiconductor substrate creates inversions at the semiconductor surface. This inversion layer increases the substrate capacitance of the interconnect and is voltage-dependent. To prevent this most silicon-IC manufacturers place an inversion-preventing implant on the surface of the substrate. The depletion between the substrate and n-type or p-type wells diffused into the substrate also creates a voltage-dependent capacitance. Thus use of the well as a high resistivity interconnect for making high value resistors will require consideration of a nonlinear capacitance to the substrate.

1.5.2 Pad and Packaging Parasitics

All signals and supply voltages that exit an IC must travel across the packaging interconnections. Just like the on-chip interconnect, the packaging interconnect has parasitic resistance, capacitance, and inductance. However, some of the packaging materials are significantly different in properties and dimension to those used in the IC, thus there are major differences in the importance of the various types of parasitics. Figure 1.108 is a typical packaged IC. The chief components of the packaging are the pads on the chip, the wire or bump bond used to connect the pad to the package, and then the package interconnect.

The pads used to attach wire bonds or bump bonds to ICs are often the largest features on an IC. The typical pad is 100 μm on a side and has a capacitance of 100 fF. In addition, protection diodes are often used on pads that will add a small nonlinear component to the pad capacitance.

The wire bonds that attach the pads to the package are typically very low resistivity and have negligible capacitance. Their major contribution to package parasitics is inductance. Typically, the

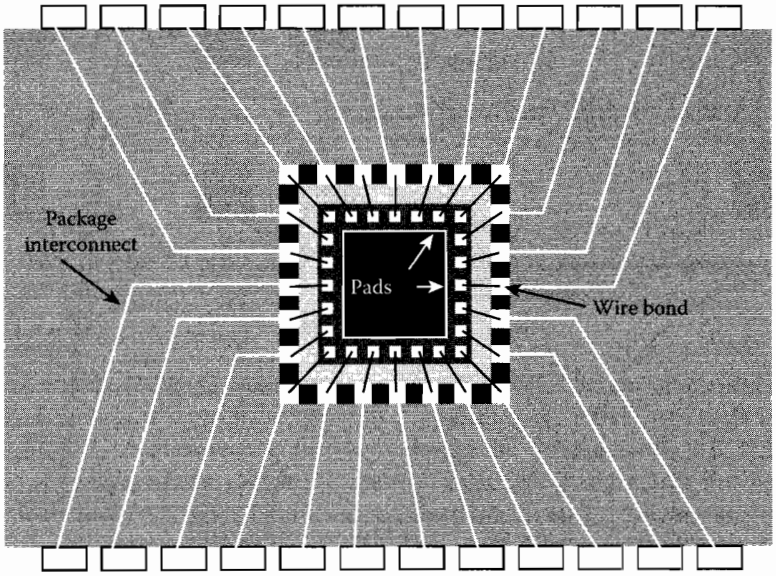

FIGURE 1.108 A packaged IC. The main sites of parasitics are the pad, bond, and package interconnect.

package interconnect inductance is greater than the wire bond inductance; however, when wire bonds are used to connect two ICs directly together, then the wire bond inductance is significant.

Often, the dominant component of package parasitics comes from the packaging interconnect itself. Depending on the package, there is inductance, capacitance to ground, and parallel line capacitance produced by this interconnect. Carefully made high-frequency packages do not exhibit much parallel line capacitance (at the expense of much capacitance to ground due to shielding), but in low-frequency packages with many connections this can become a problem.

Typical inductance and capacitance values for a high-speed package capable of output bandwidths around 5 GHz are incorporated into a circuit model for the package parasitics in Figure 1.109. When simulated with a variety of circuit source resistances (*RS*) this circuit reaches maximum bandwidth without peaking when the output resistance is 4 Ω. At lower output resistance, Figures 1.110 and 1.111 show that considerable peaking in the output frequency response occurs.

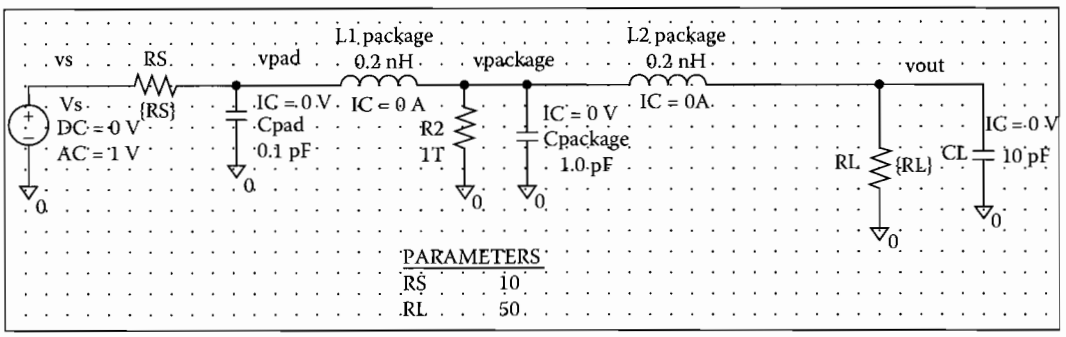

FIGURE 1.109 The circuit model of a high-frequency package output and associated parasitics. Cpad is the pad capacitance. L1package, Cpackage, R2, and L2package model the package interconnect. *RS* is the source resistance of the circuit. *RL* and *CL* are the external load.

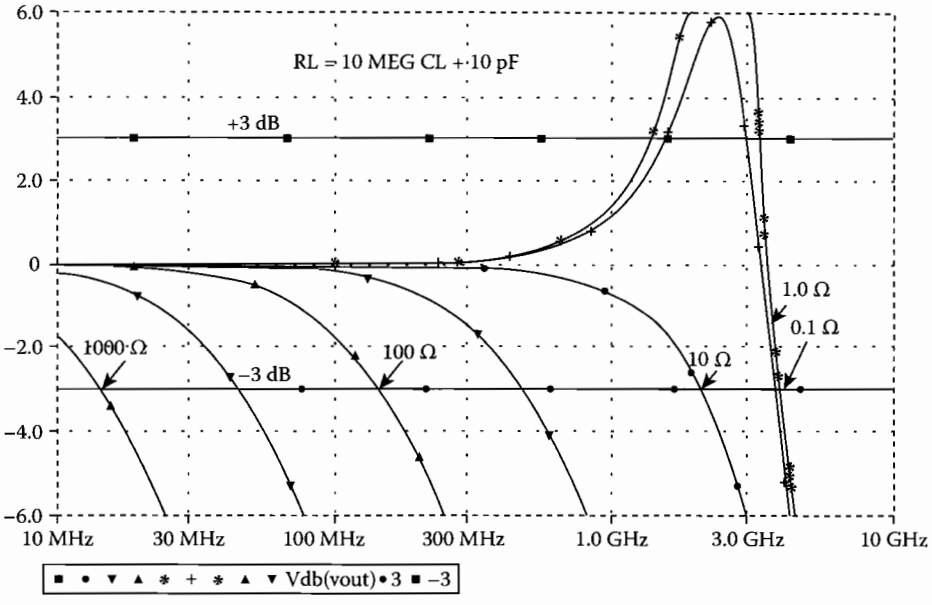

FIGURE 1.110 The PSPICE ac simulation of circuit in Figure 1.109 when the load resistance *RL* is 10 MΩ. This shows how the package inductance causes peaking for sufficiently low output resistance. In this case, peaking occurs for *RS* below 4 Ω and at about 2 GHz.

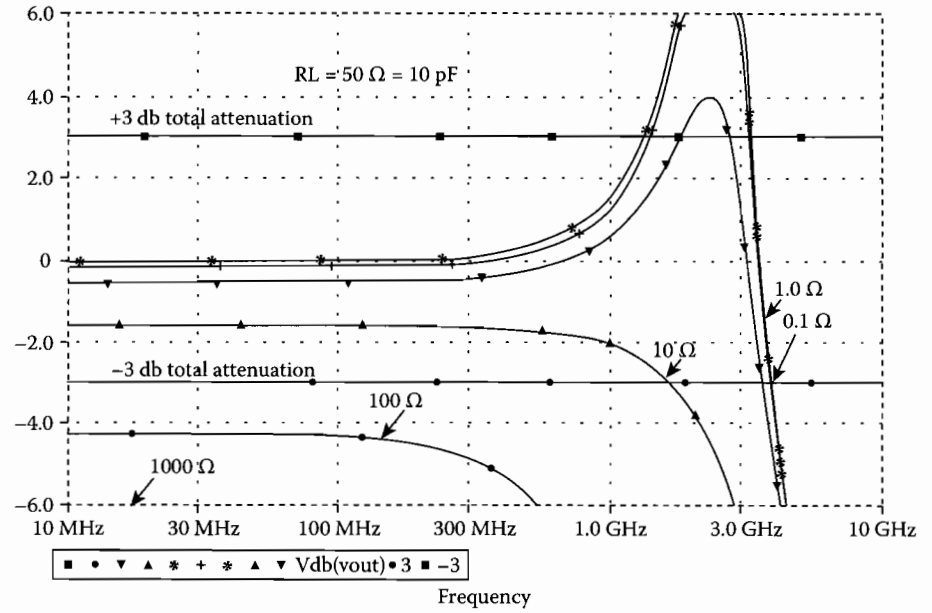

FIGURE 1.111 The PSPICE ac simulation of circuit in Figure 1.109 when the load is 50 Ω. The package inductance still causes peaking for *RS* below 4 Ω.

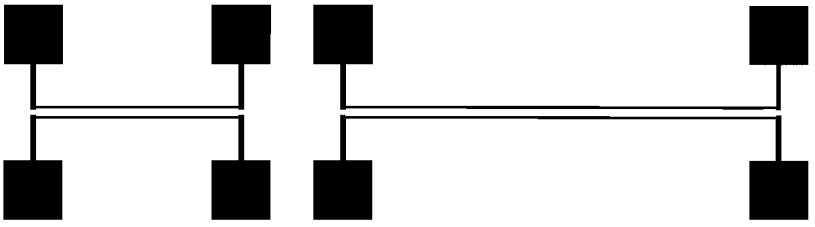

FIGURE 1.112 Test structures for measuring parallel line capacitance.

1.5.3 Parasitic Measurement

The major concern when measuring parasitics is to extract the individual parasitic values independently from measured data. This is normally achieved by exaggerating the effect that causes each individual parasitic in a special test structure, and then reproducing the structure with two or more different dimensions that will affect only the parasitic of interest. In this fashion, the effects of the other parasitics are minimized and can be subtracted from the desired parasitic in each measurement.

$$C_P = \frac{C_1 - C_2}{L_1 - L_2} \tag{1.272}$$

For example, to measure parallel line capacitance, the test structures in Figure 1.112 would be fabricated. These structures vary only in the length of the parallel lines. This means that if other parasitic capacitance ends up between the two signal lines used to measure the parasitic, then it will be a constant capacitance that can be subtracted from both measurements. The parallel line capacitance will vary in proportion to the variation of length between the two test structures. Thus the parallel line capacitance per unit length can be found from Equation 1.272. In this equation C_P is the parallel line capacitance per unit length, C_1 and C_2 are the capacitances measured from each test structure, and L_1 and L_2 are the length of the two parallel interconnect segments.

References

1. D. L. Carter and D. F. Guise, Effects of interconnections on submicron chip performance, *VLSI Design*, 4, 63–68, 1984.
2. H. B. Lunden, Detailed extraction of distributed networks and parasitics in IC designs, in *Proc. Euro. Conf. Circuit Theory, Design*, 1989, pp. 84–88.
3. R. A. Sainati and T. J. Moravec, Estimating high-speed circuit interconnect performance, *IEEE Trans. Circuits Syst.*, 36, 533–541, April 1989.
4. D. S. Gao, A. T. Yang, and S. M. Kang, Modeling and simulation of interconnection delays and crosstalks in high-speed integrated circuits, *IEEE Trans, Circuits Syst.*, 37, 1–8, Jan. 1990.
5. M. Horowitz and R. W. Dutton, Resistance extraction from mask layout data, *IEEE Trans. Computer-Aided Design Integrat. Circuits Syst.*, 7, 1029–1037, Oct. 1988.
6. R. J. Antinone and G. W. Brown, The modeling of resistive interconnections for integrated circuits, *IEEE J. Solid-State Circuits*, SC-18, 200–203, April 1983.
7. A. E. Ruehli and P. A. Brennan, Capacitance models for integrated circuit, metallization wires, *IEEE J. Solid-State Circuits*, SC-10, 530–536, Dec. 1975.
8. S. Mori, I. Suwa, and J. Wilmore, Hierarchical capacitance extraction in an IC artwork verification system, in *Proc. IEEE Int. Conf. Computer-Aided Design*, 1984, pp. 266–268.