# 1

# MOSFET Device Physics and Operation

## 1.1 INTRODUCTION

A field effect transistor (FET) operates as a conducting semiconductor channel with two ohmic contacts – the *source* and the *drain* – where the number of charge carriers in the channel is controlled by a third contact – the *gate*. In the vertical direction, the gate-channel-substrate structure (gate junction) can be regarded as an orthogonal two-terminal device, which is either a MOS structure or a reverse-biased rectifying device that controls the mobile charge in the channel by capacitive coupling (field effect). Examples of FETs based on these principles are metal-oxide-semiconductor FET (MOSFET), junction FET (JFET), metal-semiconductor FET (MESFET), and heterostructure FET (HFETs). In all cases, the stationary gate-channel impedance is very large at normal operating conditions. The basic FET structure is shown schematically in Figure 1.1.

The most important FET is the MOSFET. In a silicon MOSFET, the gate contact is separated from the channel by an insulating silicon dioxide ($SiO_2$) layer. The charge carriers of the conducting channel constitute an inversion charge, that is, electrons in the case of a $p$-type substrate ($n$-channel device) or holes in the case of an $n$-type substrate ($p$-channel device), induced in the semiconductor at the silicon-insulator interface by the voltage applied to the gate electrode. The electrons enter and exit the channel at $n^+$ source and drain contacts in the case of an $n$-channel MOSFET, and at $p^+$ contacts in the case of a $p$-channel MOSFET.

MOSFETs are used both as discrete devices and as active elements in digital and analog monolithic integrated circuits (ICs). In recent years, the device feature size of such circuits has been scaled down into the deep submicrometer range. Presently, the 0.13-μm technology node for complementary MOSFET (CMOS) is used for very large scale ICs (VLSIs) and, within a few years, sub-0.1-μm technology will be available, with a commensurate increase in speed and in integration scale. Hundreds of millions of transistors on a single chip are used in microprocessors and in memory ICs today.

CMOS technology combines both $n$-channel and $p$-channel MOSFETs to provide very low power consumption along with high speed. New silicon-on-insulator (SOI) technology may help achieve three-dimensional integration, that is, packing of devices into many

*Device Modeling for Analog and RF CMOS Circuit Design*. T. Ytterdal, Y. Cheng and T. A. Fjeldly
© 2003 John Wiley & Sons, Ltd ISBN: 0-471-49869-6

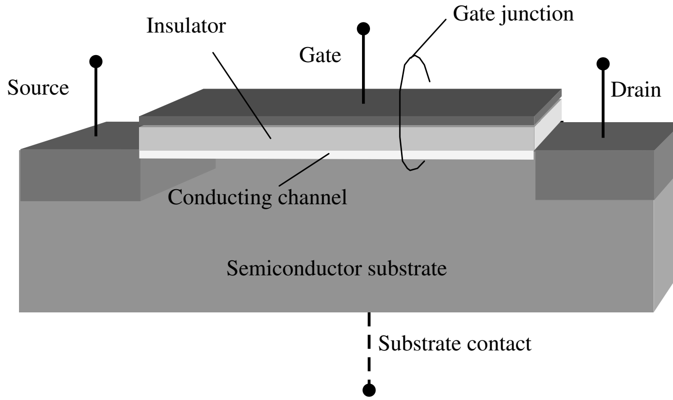


**Figure 1.1** Schematic illustration of a generic field effect transistor. This device can be viewed as a combination of two orthogonal two-terminal devices

layers, with a dramatic increase in integration density. New improved device structures and the combination of bipolar and field effect technologies (BiCMOS) may lead to further advances, yet unforeseen. One of the rapidly growing areas of CMOS is in analog circuits, spanning a variety of applications from audio circuits operating at the kilohertz (kHz) range to modern wireless applications operating at gigahertz (GHz) frequencies.

## 1.2 THE MOS CAPACITOR

To understand the MOSFET, we first have to analyze the MOS capacitor, which constitutes the important gate-channel-substrate structure of the MOSFET. The MOS capacitor is a two-terminal semiconductor device of practical interest in its own right. As indicated in Figure 1.2, it consists of a metal contact separated from the semiconductor by a dielectric insulator. An additional ohmic contact is provided at the semiconductor substrate. Almost universally, the MOS structure utilizes doped silicon as the substrate and its native oxide, silicon dioxide, as the insulator. In the silicon–silicon dioxide system, the density of surface states at the oxide–semiconductor interface is very low compared to the typical channel carrier density in a MOSFET. Also, the insulating quality of the oxide is quite good.

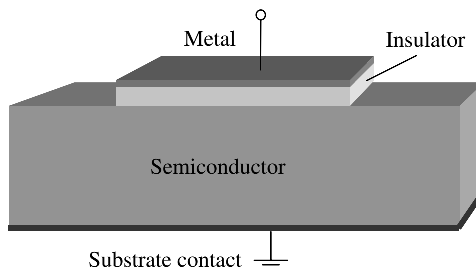


**Figure 1.2** Schematic view of a MOS capacitor

We assume that the insulator layer has infinite resistance, preventing any charge carrier transport across the dielectric layer when a bias voltage is applied between the metal and the semiconductor. Instead, the applied voltage will induce charges and counter charges in the metal and in the interface layer of the semiconductor, similar to what we expect in the metal plates of a conventional parallel plate capacitor. However, in the MOS capacitor we may use the applied voltage to control the type of interface charge we induce in the semiconductor – majority carriers, minority carriers, and depletion charge.

Indeed, the ability to induce and modulate a conducting sheet of minority carriers at the semiconductor–oxide interface is the basis for the operation of the MOSFET.

### 1.2.1 Interface Charge

The induced interface charge in the MOS capacitor is closely linked to the shape of the electron energy bands of the semiconductor near the interface. At zero applied voltage, the bending of the energy bands is ideally determined by the difference in the work functions of the metal and the semiconductor. This band bending changes with the applied bias and the bands become flat when we apply the so-called flat-band voltage given by

$$V_{FB} = (\Phi_m - \Phi_s)/q = (\Phi_m - X_s - E_c + E_F)/q, \tag{1.1}$$

where $\Phi_m$ and $\Phi_s$ are the work functions of the metal and the semiconductor, respectively, $X_s$ is the electron affinity for the semiconductor, $E_c$ is the energy of the conduction band edge, and $E_F$ is the Fermi level at zero applied voltage. The various energies involved are indicated in Figure 1.3, where we show typical band diagrams of a MOS capacitor at zero bias, and with the voltage $V = V_{FB}$ applied to the metal contact relative to the semiconductor–oxide interface. (Note that in real devices, the flat-band voltage may be

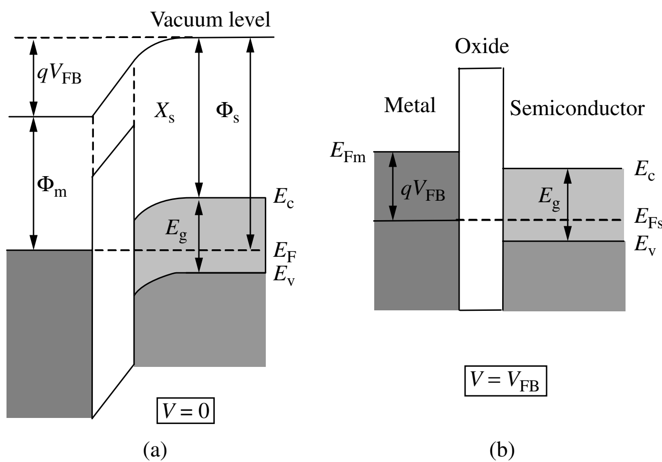


**Figure 1.3** Band diagrams of MOS capacitor (a) at zero bias and (b) with an applied voltage equal to the flat-band voltage. The flat-band voltage is negative in this example

affected by surface states at the semiconductor–oxide interface and by fixed charges in the insulator layer.)

At stationary conditions, no net current flows in the direction perpendicular to the interface owing to the very high resistance of the insulator layer (however, this does not apply to very thin oxides of a few nanometers, where tunneling becomes important, see Section 1.5). Hence, the Fermi level will remain constant inside the semiconductor, independent of the biasing conditions. However, between the semiconductor and the metal contact, the Fermi level is shifted by $E_{Fm} - E_{Fs} = qV$ (see Figure 1.3(b)). Hence, we have a quasi-equilibrium situation in which the semiconductor can be treated as if in thermal equilibrium.

A MOS structure with a $p$-type semiconductor will enter the *accumulation* regime of operation when the voltage applied between the metal and the semiconductor is more negative than the flat-band voltage ($V_{FB} < 0$ in Figure 1.3). In the opposite case, when $V > V_{FB}$, the semiconductor–oxide interface first becomes depleted of holes and we enter the so-called *depletion* regime. By increasing the applied voltage, the band bending becomes so large that the energy difference between the Fermi level and the bottom of the conduction band at the insulator–semiconductor interface becomes smaller than that between the Fermi level and the top of the valence band. This is the case indicated for $V = 0\,\mathrm{V}$ in Figure 1.3(a). Carrier statistics tells us that the electron concentration then will exceed the hole concentration near the interface and we enter the *inversion* regime. At still larger applied voltage, we finally arrive at a situation in which the electron volume concentration at the interface exceeds the doping density in the semiconductor. This is the strong inversion case in which we have a significant conducting sheet of inversion charge at the interface.

The symbol $\psi$ is used to signify the potential in the semiconductor measured relative to the potential at a position $x$ deep inside the semiconductor. Note that $\psi$ becomes positive when the bands bend down, as in the example of a $p$-type semiconductor shown in Figure 1.4. From equilibrium electron statistics, we find that the intrinsic Fermi level $E_i$ in the bulk corresponds to an energy separation $q\varphi_b$ from the actual Fermi level $E_F$ of the doped semiconductor,

$$\varphi_b = V_{th} \ln\left(\frac{N_a}{n_i}\right), \tag{1.2}$$

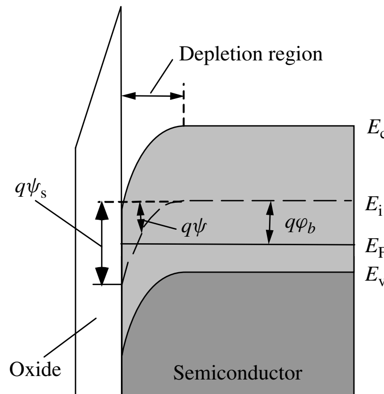


**Figure 1.4** Band diagram for MOS capacitor in weak inversion ($\varphi_b < \psi_s < 2\varphi_b$)

where $V_{th}$ is the thermal voltage, $N_a$ is the shallow acceptor density in the $p$-type semiconductor and $n_i$ is the intrinsic carrier density of silicon. According to the usual definition, strong inversion is reached when the total band bending equals $2q\varphi_b$, corresponding to the surface potential $\psi_s = 2\varphi_b$. Values of the surface potential such that $0 < \psi_s < 2\varphi_b$ correspond to the depletion and the weak inversion regimes, $\psi_s = 0$ is the flat-band condition, and $\psi_s < 0$ corresponds to the accumulation mode.

The surface concentrations of holes and electrons are expressed in terms of the surface potential as follows using equilibrium statistics,

$$p_s = N_a \exp(-\psi_s / V_{th}), \tag{1.3}$$

$$n_s = n_i^2 / p_s = n_{po} \exp(\psi_s / V_{th}), \tag{1.4}$$

where $n_{po} = n_i^2/N_a$ is the equilibrium concentration of the minority carriers (electrons) in the bulk.

The potential distribution $\psi(x)$ in the semiconductor can be determined from a solution of the one-dimensional Poisson's equation:

$$\frac{d^2\psi(x)}{dx^2} = -\frac{\rho(x)}{\varepsilon_s}, \tag{1.5}$$

where $\varepsilon_s$ is the semiconductor permittivity, and the space charge density $\rho(x)$ is given by

$$\rho(x) = q(p - n - N_a). \tag{1.6}$$

The position-dependent hole and electron concentrations may be expressed as

$$p = N_a \exp(-\psi / V_{th}), \tag{1.7}$$

$$n = n_{po} \exp(\psi / V_{th}). \tag{1.8}$$

Note that deep inside the semiconductor, we have $\psi(\infty) = 0$.

In general, the above equations do not have an analytical solution for $\psi(x)$. However, the following expression can be derived for the electric field $F_s$ at the insulator–semiconductor interface, in terms of the surface potential (see, e.g., Fjeldly *et al.* 1998),

$$F_s = \sqrt{2}\frac{V_{th}}{L_{Dp}} f\left(\frac{\psi_s}{V_{th}}\right), \tag{1.9}$$

where the function $f$ is defined by

$$f(u) = \pm\sqrt{[\exp(-u) + u - 1] + \frac{n_{po}}{N_a}[\exp(u) - u - 1]}, \tag{1.10}$$

and

$$L_{Dp} = \sqrt{\frac{\varepsilon_s V_{th}}{q N_a}} \tag{1.11}$$

is called the *Debye length*. In (1.10), a positive sign should be chosen for a positive $\psi_s$ and a negative sign corresponds to a negative $\psi_s$.

Using Gauss' law, we can relate the total charge $Q_s$ per unit area (carrier charge and depletion charge) in the semiconductor to the surface electric field by

$$Q_s = -\varepsilon_s F_s. \tag{1.12}$$

At the flat-band condition ($V = V_{FB}$), the surface charge is equal to zero. In accumulation ($V < V_{FB}$), the surface charge is positive, and in depletion and inversion ($V > V_{FB}$), the surface charge is negative. In accumulation (when $|\psi_s|$ exceeds a few times $V_{th}$) and in strong inversion, the mobile sheet charge density is proportional to $\exp[|\psi_s|/(2V_{th})]$. In depletion and weak inversion, the depletion charge is dominant and its sheet density varies as $\psi_s^{1/2}$. Figure 1.5 shows $|Q_s|$ versus $\psi_s$ for $p$-type silicon with a doping density of $10^{16}/\text{cm}^3$.

In order to relate the semiconductor surface potential to the applied voltage $V$, we have to investigate how this voltage is divided between the insulator and the semiconductor. Using the condition of continuity of the electric flux density at the semiconductor–insulator interface, we find

$$\varepsilon_s F_s = \varepsilon_i F_i, \tag{1.13}$$

where $\varepsilon_i$ is the permittivity of the oxide layer and $F_i$ is the constant electric field in the insulator (assuming no space charge). Hence, with an insulator thickness $d_i$, the voltage drop across the insulator becomes $F_i d_i$. Accounting for the flat-band voltage, the applied voltage can be written as

$$V = V_{FB} + \psi_s + \varepsilon_s F_s / c_i, \tag{1.14}$$

where $c_i = \varepsilon_i / d_i$ is the insulator capacitance per unit area.

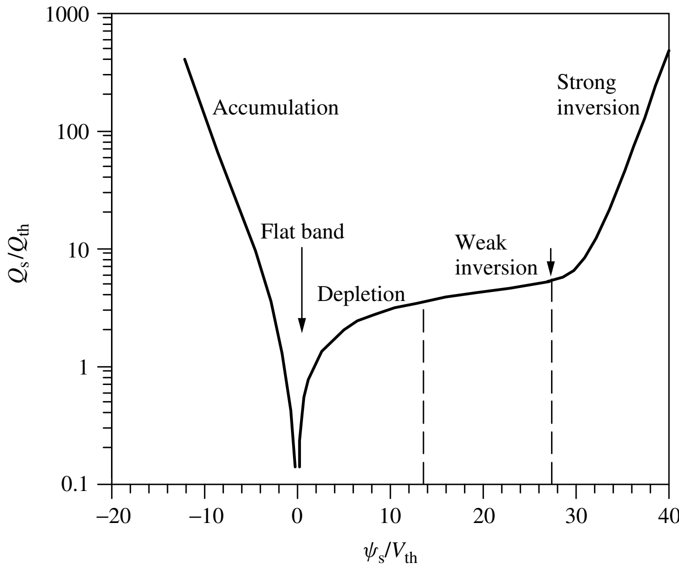


**Figure 1.5** Normalized total semiconductor charge per unit area versus normalized surface potential for $p$-type Si with $N_a = 10^{16}/\text{cm}^3$. $Q_{th} = (2\varepsilon_s q N_a V_{th})^{1/2} \approx 9.3 \times 10^{-9}\ \text{C/cm}^2$ and $V_{th} \approx 0.026\ \text{V}$ at $T = 300\ \text{K}$. The arrows indicate flat-band condition and onset of strong inversion

### 1.2.2 Threshold Voltage

The threshold voltage $V = V_T$, corresponding to the onset of the strong inversion, is one of the most important parameters characterizing metal-insulator-semiconductor devices. As discussed above, strong inversion occurs when the surface potential $\psi_s$ becomes equal to $2\varphi_b$. For this surface potential, the charge of the free carriers induced at the insulator–semiconductor interface is still small compared to the charge in the depletion layer, which is given by

$$Q_{dT} = -qN_a d_{dT} = -\sqrt{4\varepsilon_s q N_a \varphi_b}, \tag{1.15}$$

where $d_{dT} = (4\varepsilon_s \varphi_b / qN_a)^{1/2}$ is the width of the depletion layer at threshold. Accordingly, the electric field at the semiconductor–insulator interface becomes

$$F_{sT} = -Q_{dT}/\varepsilon_s = \sqrt{4qN_a\varphi_b/\varepsilon_s}. \tag{1.16}$$

Hence, substituting the threshold values of $\psi_s$ and $F_s$ in (1.14), we obtain the following expression for the threshold voltage:

$$V_T = V_{FB} + 2\varphi_b + \sqrt{4\varepsilon_s q N_a \varphi_b}/c_i. \tag{1.17}$$

Figure 1.6 shows typical calculated dependencies of $V_T$ on doping level and dielectric thickness.

For the MOS structure shown in Figure 1.2, the application of a bulk bias $V_B$ is simply equivalent to changing the applied voltage from $V$ to $V - V_B$. Hence, the threshold

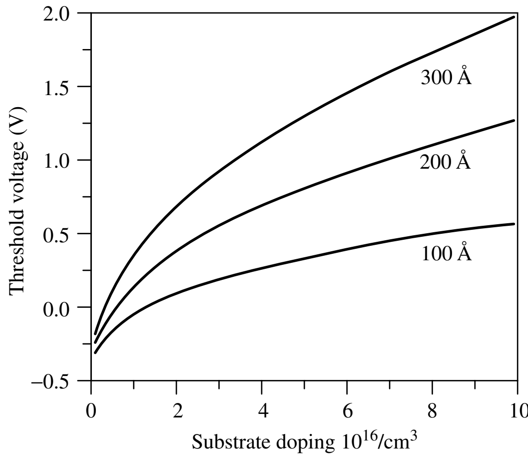


**Figure 1.6** Dependence of MOS threshold voltage on the substrate doping level for different thicknesses of the dielectric layer. Parameters used in calculation: energy gap, 1.12 eV; effective density of states in the conduction band, $3.22 \times 10^{25}/\text{m}^3$; effective density of states in the valence band, $1.83 \times 10^{25}/\text{m}^3$; semiconductor permittivity, $1.05 \times 10^{-10}$ F/m; insulator permittivity, $3.45 \times 10^{-11}$ F/m; flat-band voltage, −1 V; temperature: 300 K. Reproduced from Lee K., Shur M., Fjeldly T. A., and Ytterdal T. (1993) *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, NJ

referred to the ground potential is simply shifted by $V_B$. However, the situation will be different in a MOSFET where the conducting layer of mobile electrons may be maintained at some constant potential. Assuming that the inversion layer is grounded, $V_B$ biases the effective junction between the inversion layer and the substrate, changing the amount of charge in the depletion layer. In this case, the threshold voltage becomes

$$V_T = V_{FB} + 2\varphi_b + \sqrt{2\varepsilon_s q N_a (2\varphi_b - V_B)}/c_i. \tag{1.18}$$

Note that the threshold voltage may also be affected by so-called fast surface states at the semiconductor–oxide interface and by fixed charges in the insulator layer. However, this is not a significant concern with modern day fabrication technology.

As discussed above, the threshold voltage separates the subthreshold regime, where the mobile carrier charge increases exponentially with increasing applied voltage, from the above-threshold regime, where the mobile carrier charge is linearly dependent on the applied voltage. However, there is no clear point of transition between the two regimes, so different definitions and experimental techniques have been used to determine $V_T$. Sometimes (1.17) and (1.18) are taken to indicate the onset of so-called moderate inversion, while the onset of strong inversion is defined to be a few thermal voltages higher.

### 1.2.3 MOS Capacitance

In a MOS capacitor, the metal contact and the neutral region in the doped semiconductor substrate are separated by the insulator layer, the channel, and the depletion region. Hence, the capacitance $C_{mos}$ of the MOS structure can be represented as a series connection of the insulator capacitance $C_i = S\varepsilon_i/d_i$, where $S$ is the area of the MOS capacitor, and the capacitance of the active semiconductor layer $C_s$,

$$C_{mos} = \frac{C_i C_s}{C_i + C_s}. \tag{1.19}$$

The semiconductor capacitance can be calculated as

$$C_s = S\left|\frac{dQ_s}{d\psi_s}\right|, \tag{1.20}$$

where $Q_s$ is the total charge density per unit area in the semiconductor and $\psi_s$ is the surface potential. Using (1.9) to (1.12) for $Q_s$ and performing the differentiation, we obtain

$$C_s = \frac{C_{so}}{\sqrt{2} f(\psi_s/V_{th})}\left\{1 - \exp\left(-\frac{\psi_s}{V_{th}}\right) + \frac{n_{po}}{N_a}\left[\exp\left(\frac{\psi_s}{V_{th}}\right) - 1\right]\right\}. \tag{1.21}$$

Here, $C_{so} = S\varepsilon_s/L_{Dp}$ is the semiconductor capacitance at the flat-band condition (i.e., for $\psi_s = 0$) and $L_{Dp}$ is the Debye length given by (1.11). Equation (1.14) describes the relationship between the surface potential and the applied bias.

The semiconductor capacitance can formally be represented as the sum of two capacitances – a depletion layer capacitance $C_d$ and a free carrier capacitance $C_{fc}$. $C_{fc}$ together with a series resistance $R_{GR}$ describes the delay caused by the generation/recombination

mechanisms in the buildup and removal of inversion charge in response to changes in the bias voltage (see following text). The depletion layer capacitance is given by

$$C_{\mathrm{d}} = S\varepsilon_{\mathrm{s}}/d_{\mathrm{d}}, \tag{1.22}$$

where

$$d_{\mathrm{d}} = \sqrt{\frac{2\varepsilon_{\mathrm{s}}\psi_{\mathrm{s}}}{qN_{\mathrm{a}}}} \tag{1.23}$$

is the depletion layer width. In strong inversion, a change in the applied voltage will primarily affect the minority carrier charge at the interface, owing to the strong dependence of this charge on the surface potential. This means that the depletion width reaches a maximum value with no significant further increase in the depletion charge. This maximum depletion width $d_{\mathrm{dT}}$ can be determined from (1.23) by applying the threshold condition, $\psi_{\mathrm{s}} = 2\varphi_b$. The corresponding minimum value of the depletion capacitance is $C_{\mathrm{dT}} = S\varepsilon_{\mathrm{s}}/d_{\mathrm{dT}}$.

The free carrier contribution to the semiconductor capacitance can be formally expressed as

$$C_{\mathrm{fc}} = C_{\mathrm{s}} - C_{\mathrm{d}}. \tag{1.24}$$

As indicated, the variation in the minority carrier charge at the interface comes from the processes of generation and recombination mechanisms, with the creation and removal of electron–hole pairs. Once an electron–hole pair is generated, the majority carrier (a hole in $p$-type material and an electron in $n$-type material) is swept from the space charge region into the substrate by the electric field of this region. The minority carrier is swept in the opposite direction toward the semiconductor–insulator interface. The variation in minority carrier charge at the semiconductor–insulator interface therefore proceeds at a rate limited by the time constants associated with the generation/recombination processes. This finite rate represents a delay, which may be represented electrically in terms of an $RC$ product consisting of the capacitance $C_{\mathrm{fc}}$ and the resistance $R_{\mathrm{GR}}$, as reflected in the equivalent circuit of the MOS structure shown in Figure 1.7. The capacitance $C_{\mathrm{fc}}$ becomes important in the inversion regime, especially in strong inversion where the mobile charge is important. The resistance $R_{\mathrm{s}}$ in the equivalent circuit is the series resistance of the neutral semiconductor layer and the contacts.

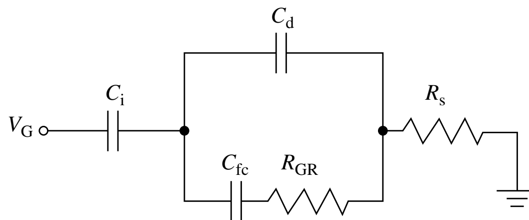


**Figure 1.7** Equivalent circuit of the MOS capacitor. Reproduced from Shur M. (1990) *Physics of Semiconductor Devices*, Prentice Hall, Englewood Cliffs, NJ

This equivalent circuit is clearly frequency-dependent. In the low-frequency limit, we can neglect the effects of $R_{GR}$ and $R_s$ to obtain (using $C_s = C_d + C_{fc}$)

$$C_{mos}^{o} = \frac{C_s C_i}{C_s + C_i}. \quad (1.25)$$

In strong inversion, we have $C_s \gg C_i$, which gives

$$C_{mos}^{o} \approx C_i \quad (1.26)$$

at low frequencies.

In the high-frequency limit, the time constant of the generation/recombination mechanism will be much longer than the signal period ($R_{GR}C_{fc} \gg 1/f$) and $C_d$ effectively shunts the lower branch of the parallel section of the equivalent in Figure 1.7. Hence, the high-frequency, strong inversion capacitance of the equivalent circuit becomes

$$C_{mos}^{\infty} = \frac{C_{dT} C_i}{C_{dT} + C_i}. \quad (1.27)$$

The calculated dependence of $C_{mos}$ on the applied voltage for different frequencies is shown in Figure 1.8. For applied voltages well below threshold, the device is in accumulation and $C_{mos}$ equals $C_i$. As the voltage approaches threshold, the semiconductor passes the flat-band condition where $C_{mos}$ has the value $C_{FB}$, and then enters the depletion and the weak inversion regimes where the depletion width increases and the capacitance value drops steadily until it reaches the minimum value at threshold given by (1.27). The calculated curves clearly demonstrate how the MOS capacitance in the strong inversion regime depends on the frequency, with a value of $C_{mos}^{\infty}$ at high frequencies to $C_i$ at low frequencies.

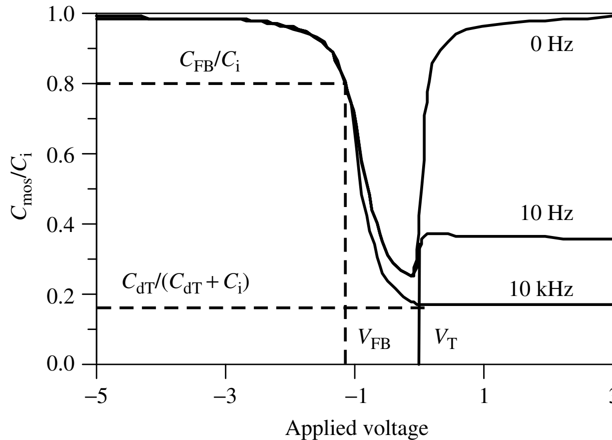


**Figure 1.8** Calculated dependence of $C_{mos}$ on the applied voltage for different frequencies. Parameters used: insulator thickness, $2 \times 10^{-8}$ m; semiconductor doping density, $10^{15}/cm^3$; generation time, $10^{-8}$ s. Reproduced from Shur M. (1990) *Physics of Semiconductor Devices*, Prentice Hall, Englewood Cliffs, NJ

We note that in a MOSFET, where the highly doped source and drain regions act as reservoirs of minority carriers for the inversion layer, the time constant $R_{GR}C_{fc}$ must be substituted by a much smaller time constant corresponding to the time needed for transporting carriers from these reservoirs in and out of the MOSFET gate area. Consequently, high-frequency strong inversion MOSFET gate-channel $C-V$ characteristics will resemble the zero frequency MOS characteristic.

Since the low-frequency MOS capacitance in the strong inversion is close to $C_i$, the induced inversion charge per unit area can be approximated by

$$qn_s \approx c_i(V - V_T). \tag{1.28}$$

This equation serves as the basis of a simple charge control model (SCCM) allowing us to calculate MOSFET current–voltage characteristics in strong inversion.

From measured MOS $C-V$ characteristics, we can easily determine important parameters of the MOS structure, including the gate insulator thickness, the semiconductor substrate doping density, and the flat-band voltage. The maximum measured capacitance $C_{max}$ (capacitance $C_i$ in Figure 1.7) yields the insulator thickness

$$d_i \approx S\varepsilon_i/C_{max}. \tag{1.29}$$

The minimum measured capacitance $C_{min}$ (at high frequency) allows us to find the doping concentration in the semiconductor substrate. First, we determine the depletion capacitance in the strong inversion regime using (1.27),

$$1/C_{min} = 1/C_{dT} + 1/C_i. \tag{1.30}$$

From $C_{dT}$ we obtain the thickness of the depletion region at threshold as

$$d_{dT} = S\varepsilon_s/C_{dT}. \tag{1.31}$$

Then we calculate the doping density $N_a$ using (1.23) with $\psi_s = 2\varphi_b$ and (1.2) for $\varphi_b$. This results in the following transcendental equation for $N_a$:

$$N_a = \frac{4\varepsilon_s V_{th}}{qd_{dT}^2} \ln\left(\frac{N_a}{n_i}\right). \tag{1.32}$$

This equation can easily be solved by iteration or by approximate analytical techniques. Once $d_i$ and $N_a$ have been obtained, the device capacitance $C_{FB}$ under flat-band conditions can be determined using $C_s = C_{so}$ ((1.21) at flat-band condition) in combination with (1.19):

$$C_{FB} = \frac{C_{so}C_i}{C_{so} + C_i} = \frac{S\varepsilon_s\varepsilon_i}{\varepsilon_s d_i + \varepsilon_i L_{Dp}}. \tag{1.33}$$

The flat-band voltage $V_{FB}$ is simply equal to the applied voltage corresponding to this value of the device capacitance.

We note that the above characterization technique applies to ideal MOS structures. Different nonideal effects, such as geometrical effects, nonuniform doping in the substrate,

interface states, and mobile charges in the oxide may influence the $C-V$ characteristics of the MOS capacitor.

### 1.2.4 MOS Charge Control Model

Well above threshold, the charge density of the mobile carriers in the inversion layer can be calculated using the parallel plate charge control model of (1.28). This model gives an adequate description for the strong inversion regime of the MOS capacitor, but fails for applied voltages near and below threshold (i.e., in the weak inversion and depletion regimes). Several expressions have been proposed for a unified charge control model (UCCM) that covers all the regimes of operation, including the following (see Byun *et al.* 1990):

$$V - V_T = q(n_s - n_o)/c_a + \eta V_{th} \ln\left(\frac{n_s}{n_o}\right), \tag{1.34}$$

where $c_a \approx c_i$ is approximately the insulator capacitance per unit area (with a small correction for the finite vertical extent of the inversion channel, see Lee *et al.* (1993)), $n_o = n_s(V = V_T)$ is the density of minority carriers per unit area at threshold, and $\eta$ is the so-called subthreshold ideality factor, also known as the subthreshold swing parameter. The ideality factor accounts for the subthreshold division of the applied voltage between the gate insulator and the depletion layer, and $1/\eta$ represents the fraction of this voltage that contributes to the interface potential. A simplified analysis gives

$$\eta = 1 + C_d/C_i, \tag{1.35}$$

$$n_o = \eta V_{th} c_a/2q. \tag{1.36}$$

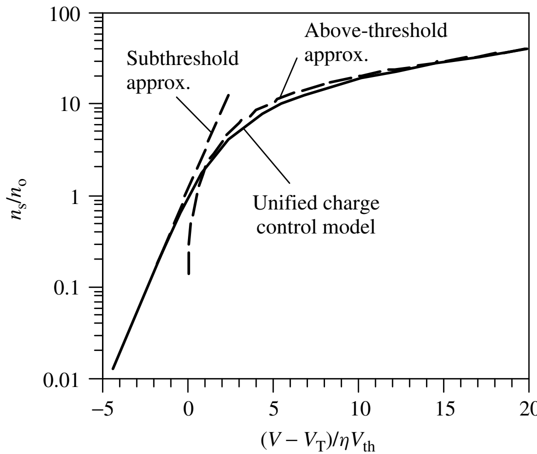


**Figure 1.9** Comparison of various charge control expression for the MOS capacitor. Equation (1.38) is a close approximation to (1.34), while the above- and below-threshold approximations are given by (1.28) and (1.37), respectively. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

In the subthreshold regime, (1.34) approaches the limit

$$n_s = n_o \exp\left(\frac{V - V_T}{\eta V_{th}}\right). \tag{1.37}$$

We note that (1.34) does not have an exact analytical solution for the inversion charge in terms of the applied voltage. However, for many purposes, the following approximate solution may be suitable:

$$n_s = 2n_o \ln\left[1 + \frac{1}{2}\exp\left(\frac{V - V_T}{\eta V_{th}}\right)\right]. \tag{1.38}$$

This expression reproduces the correct limiting behavior both in strong inversion and in the subthreshold regime, although it deviates slightly from (1.34) near threshold. The various charge control expressions of the MOS capacitor are compared in Figure 1.9.

## 1.3 BASIC MOSFET OPERATION

In the MOSFET, an inversion layer at the semiconductor–oxide interface acts as a conducting channel. For example, in an $n$-channel MOSFET, the substrate is $p$-type silicon and the inversion charge consists of electrons that form a conducting channel between the $n^+$ ohmic source and the drain contacts. At DC conditions, the depletion regions and the neutral substrate provide isolation between devices fabricated on the same substrate. A schematic view of the $n$-channel MOSFET is shown in Figure 1.10.

As described above for the MOS capacitor, inversion charge can be induced in the channel by applying a suitable gate voltage relative to other terminals. The onset of strong inversion is defined in terms of a threshold voltage $V_T$ being applied to the gate electrode relative to the other terminals. In order to assure that the induced inversion channel extends all the way from source to drain, it is essential that the MOSFET gate structure either overlaps slightly or aligns with the edges of these contacts (the latter is achieved by a self-aligned process). Self-alignment is preferable since it minimizes the parasitic gate-source and gate-drain capacitances.

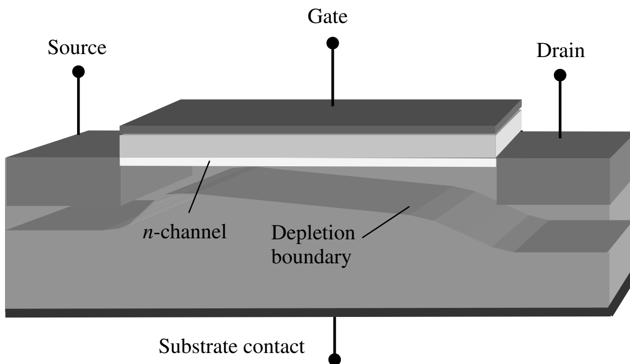


**Figure 1.10** Schematic view of an $n$-channel MOSFET with conducting channel and depletion region

When a drain-source bias $V_{DS}$ is applied to an $n$-channel MOSFET in the above-threshold conducting state, electrons move in the channel inversion layer from source to drain. A change in the gate-source voltage $V_{GS}$ alters the electron sheet density in the channel, modulating the channel conductance and the device current. For $V_{GS} > V_T$ in an $n$-channel device, the application of a positive $V_{DS}$ gives a steady voltage increase from source to drain along the channel that causes a corresponding reduction in the local gate-channel bias $V_{GX}$ (here $X$ signifies a position $x$ within the channel). This reduction is greatest near drain where $V_{GX}$ equals the gate-drain bias $V_{GD}$.

Somewhat simplistically, we may say that when $V_{GD} = V_T$, the channel reaches threshold at the drain and the density of inversion charge vanishes at this point. This is the so-called pinch-off condition, which leads to a saturation of the drain current $I_{ds}$. The corresponding drain-source voltage, $V_{DS} = V_{SAT}$, is called the *saturation voltage*. Since $V_{GD} = V_{GS} - V_{DS}$, we find that $V_{SAT} = V_{GS} - V_T$. (This is actually a result of the SCCM, which is discussed in more detail in Section 1.4.1.)

When $V_{DS} > V_{SAT}$, the pinched-off region near drain expands only slightly in the direction of the source, leaving the remaining inversion channel intact. The point of transition between the two regions, $x = x_p$, is characterized by $V_{XS}\ (x_p) \approx V_{SAT}$, where $V_{XS}(x_p)$ is the channel voltage relative to source at the transition point. Hence, the drain current in saturation remains approximately constant, given by the voltage drop $V_{SAT}$ across the part of the channel that remain in inversion. The voltage $V_{DS} - V_{SAT}$ across the pinched-off region creates a strong electric field, which efficiently transports the electrons from the strongly inverted region to the drain.

Typical current–voltage characteristics of a long-channel MOSFET, where pinch-off is the predominant saturation mechanism, are shown in Figure 1.11. However, with shorter MOSFET gate lengths, typically in the submicrometer range, velocity saturation will occur in the channel near drain at lower $V_{DS}$ than that causing pinch-off. This leads to more evenly spaced saturation characteristics than those shown in this figure, more in

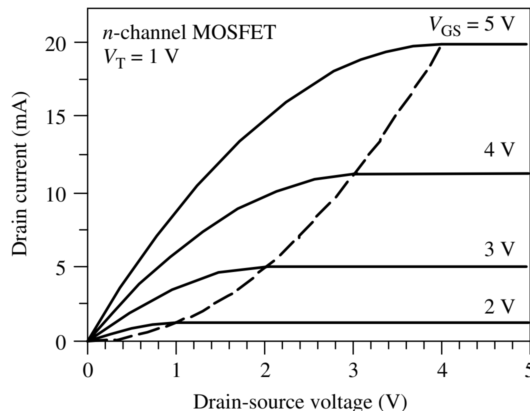


**Figure 1.11** Current–voltage characteristics of an $n$-channel MOSFET with current saturation caused by pinch-off (long-channel case). The intersections with the dotted line indicate the onset of saturation for each characteristic. The threshold voltage is assumed to be $V_T = 1$ V. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

agreement with those observed for modern devices. Also, phenomena such as a finite channel conductance in saturation, a drain bias–induced shift in the threshold voltage, and an increased subthreshold current are important consequences of shorter gate lengths (see Section 1.5).

## 1.4 BASIC MOSFET MODELING

Analytical or semianalytical MOSFET models are usually based on the so-called gradual channel approximation (GCA). Contrary to the situation in the ideal two-terminal MOS device, where the charge density profile is determined from a one-dimensional Poisson's equation (see Section 1.2), the MOSFET generally poses a two-dimensional electrostatic problem. The reason is that the geometric effects and the application of a drain-source bias create a lateral electric field component in the channel, perpendicular to the vertical field associated with the ideal gate structure. The GCA states that, under certain conditions, the electrostatic problem of the gate region can be expressed in terms of two coupled one-dimensional equations – a Poisson's equation for determining the vertical charge density profile under the gate and a charge transport equation for the channel. This allows us to determine self-consistently both the channel potential and the charge profile at any position along the gate. A direct inspection of the two-dimensional Poisson's equation for the channel region shows that the GCA is valid if we can assume that the electric field gradient in the lateral direction of the channel is much less than that in the vertical direction perpendicular to the channel (Lee *et al.* 1993).

Typically, we find that the GCA is valid for long-channel MOSFETs, where the ratio between the gate length and the vertical distance of the space charge region from the gate electrode, the so-called aspect ratio, is large. However, if the MOSFET is biased in saturation, the GCA always becomes invalid near drain as a result of the large lateral field gradient that develops in this region. In Figure 1.12, this is schematically illustrated for a MOSFET in saturation.

Next, we will discuss three relatively simple MOSFET models, the simple charge control model, the Meyer model, and the velocity saturation model. These models, with extensions, can be identified with the models denoted as MOSFET Level 1, Level 2, and Level 3 in SPICE.

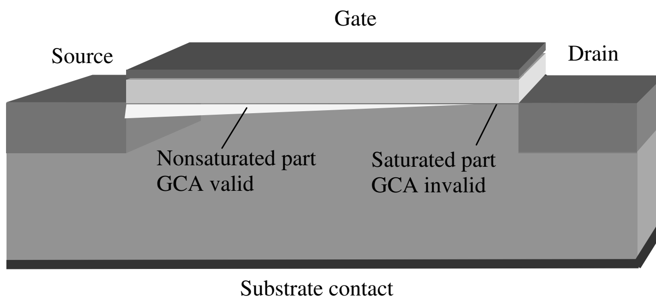


**Figure 1.12** Schematic representation of a MOSFET in saturation, where the channel is divided into a nonsaturated region where the GCA is valid and a saturated region where the GCA is invalid

We should note that the analysis that follows is based on idealized device structures. Especially in modern MOSFET/CMOS technology, optimized for high-speed and low-power applications, the devices are more complex. Additional oxide and doping regions are used for the purpose of controlling the threshold voltage and to avoid deleterious effects of high electric fields and so-called short- and narrow-channel phenomena associated with the steady downscaling device dimensions. These effects will be discussed more in Section 1.5 and in later chapters.

### 1.4.1 Simple Charge Control Model

Consider an $n$-channel MOSFET operating in the above-threshold regime, with a gate voltage that is sufficiently high to cause inversion in the entire length of the channel at zero drain-source bias. We assume a long-channel device, implying that GCA is applicable and that the carrier mobility can be taken to be constant (no velocity saturation). As a first approximation, we can describe the mobile inversion charge by a simple extension of the parallel plate expression (1.28), taking into account the potential variation $V(x)$ along the channel, that is,

$$qn_{\mathrm{s}}(x) \approx c_{\mathrm{i}}[V_{\mathrm{GT}} - V(x)], \tag{1.39}$$

where $V_{\mathrm{GT}} \equiv V_{\mathrm{GS}} - V_{\mathrm{T}}$. This simple charge control expression implies that the variation of the depletion layer charge along the channel, which depends on $V(x)$, is negligible. Furthermore, since the expression relies on GCA, it is only applicable for the nonsaturated part of the channel. Saturation sets in when the conducting channel is pinched-off at the drain side, that is, for $n_{\mathrm{s}}(x = L) \geq 0$. Using the pinch-off condition and $V(x = L) = V_{\mathrm{DS}}$ in (1.39), we obtain the following expression for the saturation drain voltage in the SCCM:

$$V_{\mathrm{SAT}} = V_{\mathrm{GT}}. \tag{1.40}$$

The threshold voltage in this model is given by (1.18), where we have accounted for the substrate bias $V_{\mathrm{BS}}$ relative to the source. We note that this expression is only valid for negative or slightly positive values of $V_{\mathrm{BS}}$, when the junction between the source contact and the $p$-substrate is either reverse-biased or slightly forward-biased. For high $V_{\mathrm{BS}}$, a significant leakage current will take place.

Figure 1.13 shows an example of calculated dependences of the threshold voltage on substrate bias for different values of gate insulator thickness. As can be seen from this figure and from (1.18), the threshold voltage decreases with decreasing insulator thickness and is quite sensitive to the substrate bias. This so-called body effect is essential for device characterization and in threshold voltage engineering. For real devices, it is important to be able to carefully adjust the threshold voltage to match specific application requirements.

Equation (1.18) also shows that $V_{\mathrm{T}}$ can be adjusted by changing the doping or by using different gate metals (including heavily doped polysilicon). As discussed in Section 1.2, the gate metal affects the flat-band voltage through the work-function difference between the metal and the semiconductor. Threshold voltage adjustment by means of doping is often performed with an additional ion implantation through the gate oxide.

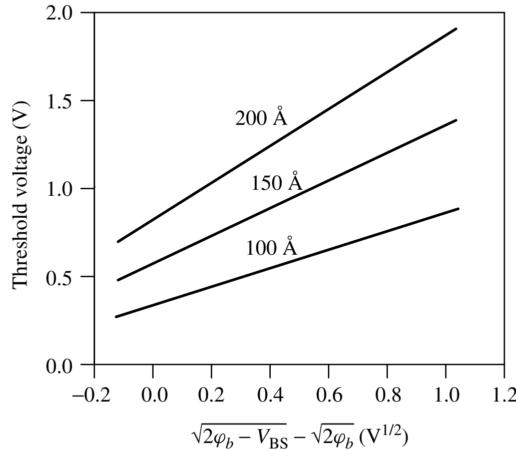


**Figure 1.13** Body plot, the dependence of the threshold voltage on substrate bias in MOSFETs with different insulator thicknesses. Parameters used in the calculation: flat-band voltage −1 V, substrate doping density $10^{22}/\mathrm{m}^3$, temperature 300 K. The slope of the plots are given in terms of the body-effect parameter $\gamma = (2\varepsilon_s q N_a)^{1/2}/c_i$. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

Assuming a constant electron mobility $\mu_n$, the electron velocity can be written as $v_n = -\mu_n \, \mathrm{d}V/\mathrm{d}x$. Neglecting the diffusion current, which is important only near threshold and in the subthreshold regime, the absolute value of the drain current can be written as

$$I_{\mathrm{ds}} = W \mu_n q n_s F, \tag{1.41}$$

where $F = |\mathrm{d}V/\mathrm{d}x|$ is the magnitude of the electric field in the channel and $W$ is the channel width. Integrating this expression over the gate length and using the fact that $I_{\mathrm{ds}}$ is independent of position $x$, we obtain the following expression for the current–voltage characteristics:

$$I_{\mathrm{ds}} = \frac{W \mu_n c_i}{L} \times \begin{cases} (V_{\mathrm{GT}} - V_{\mathrm{DS}}/2)V_{\mathrm{DS}}, & \text{for } V_{\mathrm{DS}} \le V_{\mathrm{SAT}} = V_{\mathrm{GT}} \\ V_{\mathrm{GT}}^2/2, & \text{for } V_{\mathrm{DS}} > V_{\mathrm{SAT}} \end{cases}. \tag{1.42}$$

As implied above, the pinch-off condition implies a vanishing carrier concentration at the drain side of the channel. Hence, at a first glance, one might think that the drain current should also vanish. However, instead the saturation drain current $I_{\mathrm{dsat}}$ is determined by the resistance of nonsaturated part of the channel and the current across it. In fact, this channel resistance changes very little when $V_{\mathrm{DS}}$ increases beyond $V_{\mathrm{SAT}}$, since the pinch-off point $x_p$ moves only slightly away from the drain, leaving the nonsaturated part of the channel almost intact. Moreover, the voltage at the pinch-off point will always be approximately $V_{\mathrm{SAT}}$ since the threshold condition at $x_p$ is determined by $V_G - V(x_p) = V_T$, or $V(x_p) = V_{\mathrm{GT}} = V_{\mathrm{SAT}}$. Hence, since the resistance of the nonsaturated part is constant and the voltage across it is constant, $I_{\mathrm{dsat}}$ will also remain constant. Therefore, the saturation current $I_{\mathrm{SAT}}$ is determined by substituting $V_{\mathrm{DS}} = V_{\mathrm{SAT}}$ from (1.40) into the nonsaturation expression in (1.42). In reality, of course, the electron concentration never vanishes, nor

does the electric field become infinite. This is simply a consequence of the breakdown of GCA near drain in saturation, pointing to the need for a more accurate and detailed analysis of the saturation regime.

The MOSFET current–voltage characteristics shown in Figure 1.11 were calculated using this simple charge control model.

Important device parameters are the channel conductance,

$$g_d = \frac{\partial I_d}{\partial V_{DS}}\bigg|_{V_{GS}} = \begin{cases} \beta(V_{GT} - V_{DS}), & \text{for } V_{DS} \le V_{SAT} \\ 0, & \text{for } V_{DS} > V_{SAT} \end{cases}, \tag{1.43}$$

and the transconductance,

$$g_m = \frac{\partial I_d}{\partial V_{GS}}\bigg|_{V_{DS}} = \begin{cases} \beta V_{DS}, & \text{for } V_{DS} \le V_{SAT} \\ \beta V_{GT}, & \text{for } V_{DS} > V_{SAT} \end{cases}, \tag{1.44}$$

where $\beta = W\mu_n c_i/L$ is called the *transconductance parameter*. As can be seen from these expressions, high values of channel conductance and transconductance are obtained for large electron mobilities, large gate insulator capacitances (i.e., thin gate insulator layers), and large gate width to length ratios.

The SCCM was developed at a time when the MOSFET gate lengths were typically tens of micrometers long, justifying some of the above approximations. With today's deep submicron technology, however, the SCCM is clearly not applicable. We therefore introduce two additional models that include significant improvements. In the first of these, the Meyer model, the lateral variation of the depletion charge in the channel is taken into account. In the second, the velocity saturation model (VSM), we introduce the effects of saturation in the carrier velocity. The former is important at realistic levels of substrate doping, and the latter is important because of the high electric fields generated in short-channel devices. Additional effects of small dimensions and high electric fields will be discussed in Section 1.5.

### 1.4.2 The Meyer Model

The total induced charge $q_s$ per unit area in the semiconductor of an $n$-channel MOSFET, including both inversion and depletion charges, can be expressed in terms of Gauss's law as follows, assuming that the source and the semiconductor substrate are both connected to ground (see Section 1.2),

$$q_s = -c_i[V_{GS} - V_{FB} - 2\varphi_b - V(x)]. \tag{1.45}$$

Here, the content of the bracket expresses the voltage drop across the insulator layer. The induced sheet charge density includes both the inversion charge density $q_i = -qn_s$ and the depletion charge density $q_d$, that is, $q_s = q_i + q_d$. Using (1.15) and including the added channel-substrate bias caused by the channel voltage, the depletion charge per unit area can be expressed as

$$q_d = -qN_a d_d = -\sqrt{2\varepsilon_s q N_a[2\varphi_b + V(x)]}, \tag{1.46}$$

where $d_d$ is the local depletion layer width at position $x$. Hence, the inversion sheet charge density becomes

$$q_i = -qn_s = -c_i[V_{GS} - V_{FB} - 2\varphi_b - V(x)] + \sqrt{2\varepsilon_s q N_a[2\varphi_b + V(x)]}. \quad (1.47)$$

A constant electron mobility is also assumed in the Meyer model. Hence, the nonsaturated drain current can again be obtained by substituting the expression for $n_s$ in

$$I_{ds} = W\mu_n q n_s(x) F(x) \quad (1.48)$$

to give (Meyer 1971)

$$\begin{aligned} I_{ds} = \frac{W\mu_n c_i}{L} \Bigg\{ & \left( V_{GS} - V_{FB} - 2\varphi_b - \frac{V_{DS}}{2} \right) V_{DS} \\ & - \frac{2\sqrt{2\varepsilon_s q N_a}}{3c_i} [(V_{DS} + 2\varphi_b)^{3/2} - (2\varphi_b)^{3/2}] \Bigg\}. \end{aligned} \quad (1.49)$$

The saturation voltage is obtained using the pinch-off condition $n_s = 0$,

$$V_{SAT} = V_{GS} - 2\varphi_b - V_{FB} + \frac{\varepsilon_s q N_a}{c_i^2} \left[ 1 - \sqrt{1 + \frac{2c_i^2(V_{GS} - V_{FB})}{\varepsilon_s q N_a}} \right]. \quad (1.50)$$

At low doping levels, we see that $V_{SAT}$ approaches $V_{GT}$, which is the result found for the simple charge control model.

### 1.4.3 Velocity Saturation Model

The linear velocity-field relationship (constant mobility) used in the above MOSFET models works reasonably well for long-channel devices. However, the implicit notion of a diverging carrier velocity as we approach pinch-off is, of course, unphysical. Instead, current saturation is better described in terms of a saturation of the carrier drift velocity when the electric field near drain becomes sufficiently high. The following two-piece model is a simple, first approximation to a realistic velocity-field relationship:

$$v(F) = \begin{cases} \mu_n F & \text{for } F < F_s \\ v_s & \text{for } F \geq F_s \end{cases}, \quad (1.51)$$

where $F = |dV(x)/dx|$ is the magnitude of lateral electrical field in the channel, $v_s$ is the saturation velocity, and $F_s = v_s/\mu_n$ is the saturation field. In this description, current saturation in FETs occurs when the field at the drain side of the gate reaches the saturation field. A somewhat more precise expression, which is particularly useful for $n$-channel MOSFETs, is the so-called Sodini model (Sodini *et al.* 1984),

$$v(F) = \begin{cases} \dfrac{\mu_n F}{1 + F/2F_s} & \text{for } F < 2F_s \\ v_s & \text{for } F \geq 2F_s \end{cases}. \quad (1.52)$$

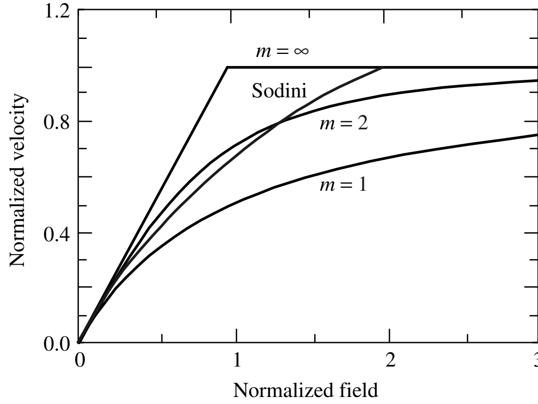


**Figure 1.14** Velocity-field relationships for charge carriers in silicon MOSFETs. The electric field and the velocity are normalized to $F_s$ and $v_s$, respectively. Two of the curves are calculated from (1.53) using $m = 1$ for holes and $m = 2$ for electrons. The curve marked $m = \infty$ corresponds to the linear two-piece model in (1.51). The Sodini model (1.52) is also shown

Even more realistic velocity-field relationships for MOSFETs are obtained from

$$v(F) = \frac{\mu F}{[1 + (F/F_s)^m]^{1/m}}, \tag{1.53}$$

where $m = 2$ and $m = 1$ are reasonable choices for $n$-channel and $p$-channel MOSFETs, respectively. The two-piece model in (1.51) corresponds to $m = \infty$ in (1.53). Figure 1.14 shows different velocity-field models for electrons and holes in silicon MOSFETs.

Using the simple velocity-field relationship of (1.51), current–voltage characteristics can easily be derived from either the SCCM or the Meyer model, since the form of the nonsaturated parts of the characteristics will be the same as before (see (1.42) and (1.49)). However, the saturation voltage will now be identical to the drain-source voltage that initiates velocity saturation at the drain side of the channel. In terms of (1.51), this occurs when $F(L) = F_s$. Hence, using this condition in combination with the SCCM, we obtain the following expressions for the drain current and the saturation voltage:

$$I_{ds} = \frac{W\mu_n c_i}{L} \times \begin{cases} V_{GT}V_{DS} - V_{DS}^2/2, & \text{for } V_{DS} \le V_{SAT} \\ (V_{GT} - V_{SAT})V_L, & \text{for } V_{DS} > V_{SAT} \end{cases}, \tag{1.54}$$

$$V_{SAT} = V_{GT} - V_L\left[\sqrt{1 + (V_{GT}/V_L)^2} - 1\right], \tag{1.55}$$

where $V_L = F_s L = L v_s/\mu_n$. The Meyer VSM leads to a much more complicated relationship for $V_{SAT}$.

For large values of $V_L$ such that $V_L \gg V_{GT}$, the square root terms in (1.55) may be expanded into a Taylor series, yielding the previous long-channel result for the SCCM without velocity saturation. Assuming, as an example, that $V_{GT} = 3$ V, $\mu_n = 0.08$ m$^2$/Vs, and $v_s = 1 \times 10^5$ m/s, we find that velocity saturation effects may be neglected for $L \gg 2.4$ µm. Hence, velocity saturation is certainly important in modern MOSFETs with gate lengths typically in the deep submicrometer range.

In the opposite limiting case, when $V_L \ll V_{GT}$, we obtain $V_{SAT} \approx V_L$ and $I_{dsat} \approx \beta V_L V_{GT}$. Since $I_{dsat}$ is proportional to $V_{GT}^2$ in long-channel devices and proportional to $V_{GT}$ in short-channel devices, we can use this difference to identify the presence of short-channel effects on the basis of measured device characteristics.

### 1.4.4 Capacitance Models

For the simulation of dynamic events in MOSFET circuits, we also have to account for variations in the stored charges of the devices. In a MOSFET, we have stored charges in the gate electrode, in the conducting channel, and in the depletion layers. Somewhat simplified, the variation in the stored charges can be expressed through different capacitance elements, as indicated in Figure 1.15.

We distinguish between the so-called parasitic capacitive elements and the capacitive elements of the intrinsic transistor. The parasitics include the overlap capacitances between the gate electrode and the highly doped source and drain regions ($C_{os}$ and $C_{od}$), the junction capacitances between the substrate and the source and drain regions ($C_{js}$ and $C_{jd}$), and the capacitances between the metal electrodes of the source, the drain, and the gate.

The semiconductor charges of the intrinsic gate region of the MOSFET are divided between the mobile inversion charge and the depletion charge, as indicated in Figure 1.15. In addition, these charges are nonuniformly distributed along the channel when drain-source bias is applied. Hence, the capacitive coupling between the gate electrode and the semiconductor is also distributed, making the channel act as an $RC$ transmission line. In practice, however, because of the short gate lengths and limited bandwidths of FETs, the distributed capacitance of the intrinsic device is usually very well represented in terms of a lumped capacitance model, with capacitive elements between the various intrinsic device terminals.

An accurate modeling of the intrinsic device capacitances still requires an analysis of how the inversion charge and the depletion charge are distributed between source, drain, and substrate for different terminal bias voltages. As discussed by Ward and Dutton (1978), such an analysis leads to a set of charge-conserving and nonreciprocal capacitances between the different intrinsic terminals (nonreciprocity means $C_{ij} \neq C_{ji}$, where $i$ and $j$ denote source, drain, gate, or substrate).

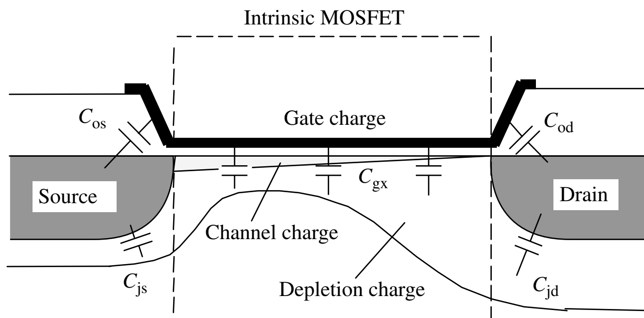


**Figure 1.15** Intrinsic and parasitic capacitive elements of the MOSFET. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

In a simplified and straightforward analysis by Meyer (1971) based on the SCCM, a set of reciprocal capacitances ($C_{ij} = C_{ji}$) were obtained as derivatives of the total gate charge with respect to the various terminal voltages. Although charge conservation is not strictly enforced in this case, since the Meyer capacitances represent only a subset of the Ward–Dutton capacitances, the resulting errors in circuit simulations are usually small, except in some cases of transient analyzes of certain demanding circuits. Here, we first consider Meyer's capacitance model for the long-channel case, but return with modifications of this model and comments on charge-conserving capacitance models in Section 1.5.3.

In Meyer's capacitance model, the distributed intrinsic MOSFET capacitance can be split into the following three lumped capacitances between the intrinsic terminals:

$$C_{\mathrm{GS}} = \left.\frac{\partial Q_{\mathrm{G}}}{\partial V_{\mathrm{GS}}}\right|_{V_{\mathrm{GD}},V_{\mathrm{GB}}}, \quad C_{\mathrm{GD}} = \left.\frac{\partial Q_{\mathrm{G}}}{\partial V_{\mathrm{GD}}}\right|_{V_{\mathrm{GS}},V_{\mathrm{GB}}}, \quad C_{\mathrm{GB}} = \left.\frac{\partial Q_{\mathrm{G}}}{\partial V_{\mathrm{GB}}}\right|_{V_{\mathrm{GS}},V_{\mathrm{GD}}}, \tag{1.56}$$

where $Q_{\mathrm{G}}$ is the total intrinsic gate charge. The intrinsic MOSFET equivalent circuit corresponding to this model is shown in Figure 1.16.

In general, the gate charge reflects both the inversion charge and the depletion charge and can therefore be written as $Q_{\mathrm{G}} = Q_{\mathrm{Gi}} + Q_{\mathrm{Gd}}$. However, in the SCCM for the drain current, the depletion charge is ignored in strong inversion, except for its influence on the threshold voltage (see (1.18)). Likewise, in the Meyer capacitance model, the gate-source capacitance $C_{\mathrm{GS}}$ and the gate-drain capacitance $C_{\mathrm{GD}}$ can be assumed to be dominated by the inversion charge. Here, we include gate-substrate capacitance $C_{\mathrm{GB}}$ in the subthreshold regime, where the depletion charge is dominant.

The contribution of the inversion charge to the gate charge is determined by integrating the sheet charge density given by (1.39), over the gate area, that is,

$$Q_{\mathrm{Gi}} = W c_{\mathrm{i}} \int_0^L [V_{\mathrm{GT}} - V(x)]\,\mathrm{d}x. \tag{1.57}$$

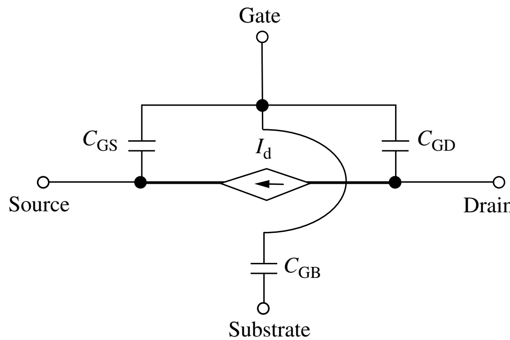


**Figure 1.16** Large-signal equivalent circuit of intrinsic MOSFET based on Meyer's capacitance model. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

From (1.41), we notice that $dx = W\mu_n c_i(V_{GT} - V)\,dV/I_{ds}$, which allows us to make a change of integration variable from $x$ to $V$ in (1.57). Hence, we obtain for the nonsaturated regime

$$Q_{Gi} = \frac{W\mu_n C_i^2}{LI_{ds}}\int_0^{V_{DS}}(V_{GT} - V)^2\,dV = \frac{2}{3}C_i\frac{(V_{GS} - V_T)^3 - (V_{GD} - V_T)^3}{(V_{GS} - V_T)^2 - (V_{GD} - V_T)^2}, \qquad (1.58)$$

where $C_i$ is the total gate oxide capacitance and where we expressed $I_{ds}$ using (1.42) and replaced $V_{DS}$ by $V_{GS} - V_{GD}$ everywhere.

Using the above relationships, the following strong inversion, long-channel Meyer capacitances are obtained:

$$C_{GS} = \frac{2}{3}C_i\left[1 - \left(\frac{V_{GT} - V_{DS}}{2V_{GT} - V_{DS}}\right)^2\right], \qquad (1.59)$$

$$C_{GD} = \frac{2}{3}C_i\left[1 - \left(\frac{V_{GT}}{2V_{GT} - V_{DS}}\right)^2\right], \qquad (1.60)$$

$$C_{GB} = 0. \qquad (1.61)$$

We recall that $V_{SAT} = V_{GT}$ is the saturation voltage in the SCCM. The capacitances at saturation are found by replacing $V_{DS} = V_{SAT}$ in the above expressions, that is,

$$C_{GSs} = \frac{2}{3}C_i, \quad C_{GDs} = C_{GBs} = 0. \qquad (1.62)$$

This result indicates that in saturation, a small change in the applied drain-source voltage does not contribute to the gate or the channel charge, since the channel is pinched off. Instead, the entire channel charge is assigned to the source terminal, giving a maximum value of the capacitance $C_{GS}$. Normalized dependencies of the Meyer capacitances $C_{GS}$ and $C_{GD}$ on bias conditions are shown in Figure 1.17.

In the subthreshold regime, the inversion charge becomes negligible compared to the depletion charge, and the MOSFET gate-substrate capacitance will be the same as that of a MOS capacitor in depletion, with a series connection of the gate oxide capacitance $C_i$ and the depletion capacitance $C_d$ (see (1.19) to (1.23)). According to the discussion in Section 1.2, the applied gate-substrate voltage $V_{GB}$ can be subdivided as follows:

$$V_{GB} = V_{FB} + \psi_s - q_{dep}/C_i, \qquad (1.63)$$

where $V_{FB}$ is the flat-band voltage, $\psi_s$ is the potential across the semiconductor depletion layer (i.e., the surface potential relative to the substrate interior), and $-q_{dep}/c_i$ is the voltage drop across the oxide. In the depletion approximation, the depletion charge per unit area $q_{dep}$ is related to $\psi_s$ by $q_{dep} = -\gamma c_i \psi_s^{1/2}$ where $\gamma = (2\varepsilon_s q N_a)^{1/2}/c_i$ is the body-effect parameter. Using this relationship to substitute for $\psi_s$ in (1.63), we find

$$Q_{Gd} = -WLq_{dep} = \gamma C_i\left(\sqrt{\gamma^2/4 + V_{GB} - V_{FB}} - \gamma/2\right), \qquad (1.64)$$

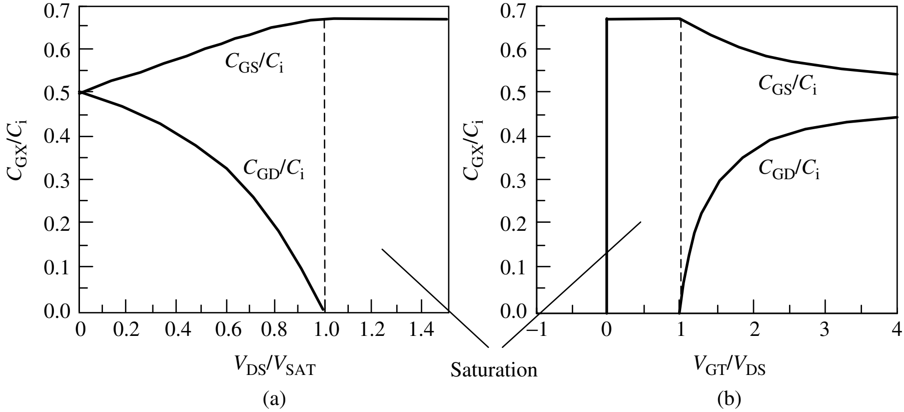


**Figure 1.17** Normalized strong inversion Meyer capacitances according to (1.59) to (1.62) versus (a) drain-source bias and (b) gate-source bias. Note that $V_{SAT} = V_{GT}$ in this model. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

from which we obtain the following subthreshold capacitances:

$$C_{GB} = \frac{C_i}{\sqrt{1 + 4(V_{GB} - V_{FB})/\gamma^2}}, \qquad C_{GS} = C_{GD} = 0. \tag{1.65}$$

We note that (1.65) gives $C_{GB} = C_i$ at the flat-band condition, which is different from the flat-band capacitance of (1.33). This discrepancy arises from neglecting the effects of the free carriers in the subthreshold regime in the present simplified treatment. For the same reason, we observe the presence of discontinuities in the Meyer capacitances at threshold. Discontinuities in the derivatives of the Meyer capacitances occur at the onset of saturation as a result of additional approximations. Such discontinuities should be avoided in the device models since they give rise to increased simulation time and conversion problems in circuit simulators. These issues will be discussed further in Section 1.5.

In the MOSFET VSM, the above-threshold capacitance expressions derived on the basis of the SCCM are still valid in the nonsaturated regime $V_{DS} \le V_{SAT}$. The capacitance values at the saturation point are found by replacing $V_{DS}$ in (1.59) and (1.60) by $V_{SAT}$ from (1.55), yielding

$$C_{GSs} = \frac{2}{3}C_i\left[1 - \left(\frac{V_{SAT}}{2V_L}\right)^2\right], \tag{1.66}$$

$$C_{GDs} = \frac{2}{3}C_i\left[1 - \left(1 - \frac{V_{SAT}}{2V_L}\right)^2\right]. \tag{1.67}$$

However, well into saturation, the intrinsic gate charge will change very little with increasing $V_{DS}$, similar to what takes place in the case of saturation by pinch-off (see preceding text). Hence, the real capacitances have to approach the same limiting values in saturation as the Meyer capacitances, that is, $C_{GS}/C_i \to 2/3$ and $C_{GD}/C_i \to 0$. In

fact, since the behavior of $C_{GS}$ and $C_{GD}$ in the VSM and in the SCCM coincide for $V_{DS} < V_{SAT}$ and have the same asymptotic values in saturation, the Meyer capacitance model offers a reasonable approximation for the MOSFET capacitances also in short-channel devices. This suggests a separate "saturation" voltage for the capacitances close to the long-channel pinch-off voltage ($\approx V_{GT}$), which is larger than $V_{SAT}$ associated with the onset of velocity saturation.

### 1.4.5 Comparison of Basic MOSFET Models

The *I–V* characteristics shown in Figure 1.18 were calculated using the three basic MOSFET models discussed above – the simple charge control model (SCCM), the Meyer *I–V* model (MM), and the velocity saturation model (VSM). The same set of MOSFET parameters were used in all cases. We note that all models coincide at small drain-source voltages. However, in saturation, SCCM always gives the highest current. This is a direct consequence of omitting velocity saturation and spatial variation in the depletion charge in SCCM, resulting in an overestimation of both carrier velocity and inversion charge. The characteristics for VSM and MM clearly demonstrate how inclusion of velocity saturation and distribution of depletion charge, respectively, affect the saturation current.

The intrinsic capacitances according to Section 1.4.4 are shown in Figure 1.19. Meyer's capacitance model can be used in conjunction with all the MOSFET models illustrated in Figure 1.18 (SCCM, MM and VSM). In the present device example, we note that velocity saturation and depletion charge may be quite important. Therefore, we emphasize that SCCM is usually applicable only for long-channel, low-doped devices, while MM applies to long-channel devices with an arbitrary doping level. VSM gives a reasonable description of short-channel devices, although important short-channel effects such as channel-length modulation and drain-induced barrier lowering (DIBL) are still unaccounted for in these

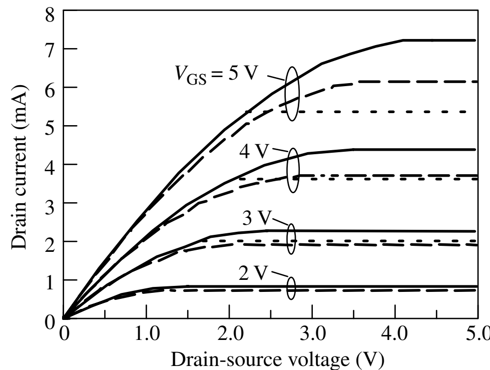


**Figure 1.18** Comparison of *I–V* characteristics obtained for a given set of MOSFET parameters using the three basic MOSFET models: simple charge control model (solid curves), Meyer's *I–V* model (dashed curves), and velocity saturation model (dotted curves). The MOSFET device parameters are $L = 2\,\mu\text{m}$, $W = 20\,\mu\text{m}$, $d_i = 300\,\text{Å}$; $\mu_n = 0.06\,\text{m}^2/\text{Vs}$, $v_s = 10^5$ m/s; $N_a = 10^{22}/\text{m}^3$, $V_T = 0.43$ V; $V_{FB} = -0.75$ V; $\varepsilon_i = 3.45 \times 10^{-11}$ F/m; $\varepsilon_s = 1.05 \times 10^{-10}$ F/m; $n_i = 1.05 \times 10^{16}/\text{m}^3$. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

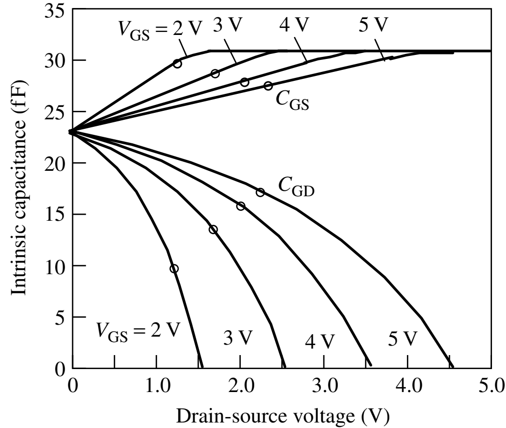


**Figure 1.19** Intrinsic MOSFET *C–V* characteristics for the same devices as in Figure 1.18, obtained from the Meyer capacitance model. The circles indicate the onset of saturation according to (1.66) and (1.67). Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

models. Likewise, we have ignored certain high-field effects (avalanche breakdown), and advanced MOSFET designs. Some of these issues will be discussed in Section 1.5 and in later chapters of this book.

## 1.4.6 Basic Small-signal Model

So far, we have considered large-signal MOSFET models, which are suitable for digital electronics and for determining the operating point in small-signal applications. The small-signal regime is, of course, a very important mode of operation of MOSFETs as well as for other active devices. Typically, the AC signal amplitudes are so small relative to the DC values of the operating point that a linear relationship can be assumed between an incoming signal and its response. Normally, if sufficiently accurate large-signal models are available, the AC designers will use such large-signal models also for small-signal applications, since this mode is readily available in circuit simulators such as SPICE. However, in cases when suitable large-signal models are unavailable or when simple hand calculations are needed, it is convenient to use a dedicated small-signal MOSFET model based on a linearized network.

Figure 1.20 shows an intrinsic, common-source, small-signal model for MOSFETs. The model is generalized to include inputs at both the gate and the substrate terminal, and the response is observed at the drain (Fonstad 1994). The network elements are obtained as first derivatives of current–voltage and charge–voltage characteristics, resulting in fixed small-signal conductances, transconductances, and capacitances for a given operating point.

To build a more complete model, some of the extrinsic parasitics may be added, including the gate overlap capacitances and the source and drain junction capacitances, shown in Figure 1.15, and the source and drain series resistances. At very high frequencies, in the radio frequency (RF) range, the junction capacitances become very important since

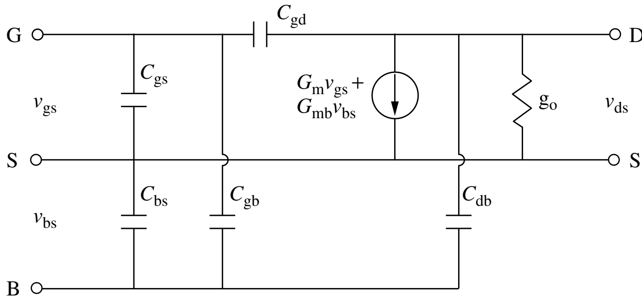


**Figure 1.20** Basic small-signal equivalent circuit of an intrinsic, common-source MOSFET. Reproduced from Fonstad C. G. (1994) *Microelectronic Devices and Circuits*, McGraw-Hill, New York

they couple efficiently to the MOSFET substrate. Other important parasitics in this range are the gate resistance and the series inductances associated with the conducting paths. RF CMOS modeling will be discussed in more detail in Chapter 3 of this book.

## 1.5 ADVANCED MOSFET MODELING

The rapid evolution of semiconductor electronics technology is fueled by a never-ending demand for better performance, combined with a fierce global competition. For silicon CMOS technology, this evolution is often measured in generations of three years – the time it takes for manufactured memory capacity on a chip to be increased by a factor of 4 and for logic circuit density to increase by a factor of between 2 and 3. Technologically, this long-term trend is made possible by a steady downscaling of CMOS feature size by about a factor of 2 per two generations.

At present, CMOS in high volume manufacturing has progressed to the 130-nm technology node. The technology node, used as a measure of the technology scaling, typically signifies the half-pitch size of the first-level interconnect in dynamic RAM (DRAM) technology, while the smallest features, the MOSFET gate lengths, are presently at 65 nm. Following the evolutionary trend, the technology node is expected to decrease below 100 nm within a few years, as indicated in Figure 1.21. Simultaneously, the performance of CMOS ICs rises steeply, packing several 100 million transistors on a chip and operating with clock rates well into the gigahertz range.

Very important issues in this development are the increasing levels of complexity of the fabrication process and the many subtle mechanisms that govern the properties of deep submicrometer FETs. These mechanisms, dictated by the device physics, have to be described and implemented into process modeling and circuit design tools, to empower the circuit designers with abilities to fully utilize the potential of existing and future technologies.

The downscaling of FETs tends to augment important nonideal phenomena, most of which have to be incorporated into any viable device model for use in circuit simulation and device design. These include the so-called short-channel effects, which tend to weaken the gate control over the channel charge. Among the manifestations of short-channel phenomena are serious leakage currents associated with punch-through and threshold voltage shifts resulting from increasing influence of the source and drain contacts over the intrinsic

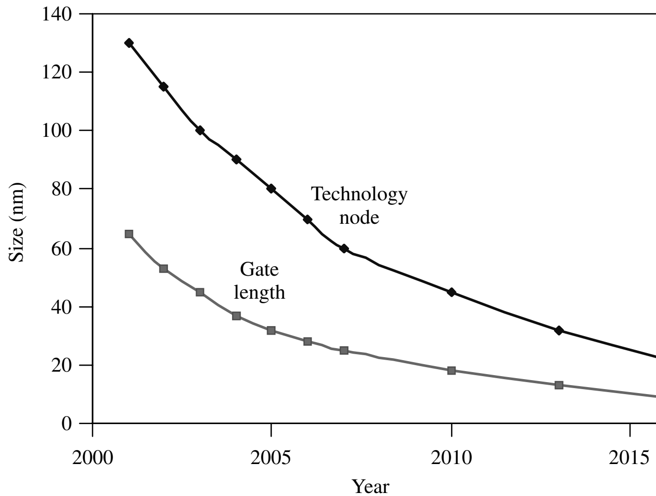


**Figure 1.21** Projected CMOS scaling according to International Technology Roadmap for Semiconductors. Reproduced from *ITRS – International Technology Roadmap for Semiconductor*, Semiconductor Industry Assoc., Austin, TX (2001)

channel and depletion charges. The drain-source bias induces an additional lowering of the injection barrier near the source, giving rise to further shifts in the threshold voltage. The latter also causes an increased output conductance in saturation. The loss of gate control may be interpreted as resulting from an improper collective scaling of dimensions, doping levels, and voltages in the device, since an ideal scaling scheme is difficult to enforce in practice.

Gate leakage is another deleterious effect that occurs in radically downscaled devices with gate oxide thicknesses of one to two nanometers. This leakage is the result of quantum-mechanical tunneling, an effect that actually poses a fundamental limitation for further MOSFET scaling within the next few decades.

In addition to these "new" phenomena, well-known effects from earlier FET generations become magnified at short gate lengths owing to enhanced electric fields associated with improper scaling of voltages. Examples are channel-length modulation (CLM), bias dependence of the field effect mobility, and phenomena related to hot electron–induced impact ionization near drain.

The above mechanisms also have important consequences for the modeling of MOSFETs. All the presently accepted MOSFET models used by industry, including the latest BSIM models (Berkeley short-channel IGFET models), are, in effect, based on the GCA. As discussed in Section 1.4, the GCA allows a separation of the model development into two coupled equations, one describing the local vertical field and charge distribution by means of a one-dimensional Poisson's equation and another describing the lateral charge transport in the channel. In improperly scaled devices, this description becomes seriously flawed since the electrostatic problem of the gate region truly becomes a two-dimensional one, with lateral and vertical fields and field gradients of similar magnitudes. The consequence is that the GCA-based models have to be augmented by numerous empirical and semiempirical "fixes" to maintain the required accuracy. This has resulted in a plethora of device parameters, counting in the hundreds for the latest BSIM models.

### 1.5.1 Modeling Approach

For any FET, the threshold gate voltage $V_T$ is a key parameter. It separates the *on*- (above-threshold) and the *off*- (subthreshold) states of operation. As indicated in Figure 1.22, the average potential energy of the channel electrons in the *off*-state is high relative to those of the source, creating an effective barrier against electron transport from source to drain. In the *on*-state, this barrier is significantly lowered, promoting a high population of free electrons in the channel region. For long-channel devices, with gate lengths of several micrometers and with high power supply voltages, the behavior in the transition region near threshold is not important in digital applications. However, for MOSFETs with deep submicrometer feature size and reduced power supply voltages (such as in low-power operation), the transition region becomes increasingly important, and the distinction between *on*- and *off*-states becomes blurred. Accordingly, a precise modeling of all regimes of device operation, including the near-threshold regime, is needed for short-channel devices, both for digital and high-frequency analog applications.

In the basic MOSFET models considered in Section 1.4, the subthreshold regime is simply considered an *off*-state of the device, ideally blocking all drain current (although the SPICE implementations of some of these models include descriptions of this regime). In practice, however, there will always be some leakage current in the *off*-state owing to a finite amount of mobile charge in the channel and a finite rate of carrier injection from the source to the channel.

This effect is enhanced in modern day downscaled MOSFETs owing to short-channel phenomena such as drain-induced barrier lowering. DIBL is a mechanism whereby the application of a drain-source bias causes a lowering of the source-channel junction barrier. In a long-channel device biased in the subthreshold regime, the applied drain-source voltage drop will be confined to the channel-drain depletion zone. The remaining part of the channel is essentially at a constant potential (flat energy bands), where diffusion is the primary mode of charge transport. However, in a short-channel device the effect of the applied drain-source voltage will be distributed over the length of the channel, giving rise to a shift of the conduction band edge near the source end of the channel, as illustrated in Figure 1.23. Such a shift represents an effective lowering of the injection barrier between the source and the channel. Since the dominant injection mechanism is thermionic emission, this barrier lowering translates into a significant increase of the injected current. This phenomenon can be described in terms of a shift in the threshold voltage (see, e.g., Fjeldly and Shur 1993). Well above threshold, the injection barrier is much reduced, and the DIBL effect eventually disappears.

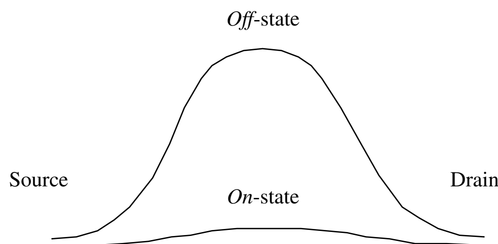


**Figure 1.22** Schematic conduction band profile through the channel region of a short-channel MOSFET in the *on*-state and the *off*-state

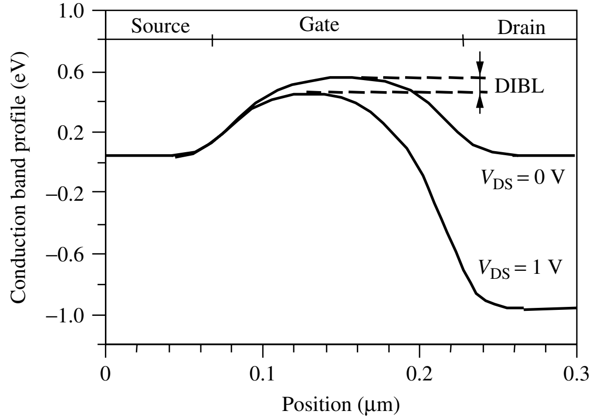


**Figure 1.23** Conduction band profile at the semiconductor–oxide interface of a short $n$-channel MOSFET with and without drain bias. The figure indicates the origin of DIBL. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

The magnitude of the subthreshold current is obviously very important since it has consequences for the power supply voltages and the logic levels needed to achieve a satisfactory *off*-state in digital operations. Hence, it affects the power dissipation in logic circuits. Likewise, the holding time in dynamic memory circuits is affected by the level of subthreshold current.

To correctly model the subthreshold operation of MOSFETs, we need a charge control model for this regime. Also, to avoid convergence problems when using the model in circuit simulators, it is preferable to use a UCCM that covers both the above- and below-threshold regimes with continuous expressions. One such model is a generalization of the UCCM that was introduced in Section 1.2.4 for the purpose of accurately describing the inversion charge density in MOS structures (Lee *et al.* 1993),

$$V_{GT} - \alpha V_F(x) = \eta V_{th} \ln\left(\frac{n_s(x)}{n_o}\right) + a[n_s(x) - n_o]. \tag{1.68}$$

Here, $V_F$ is the quasi-Fermi potential in the channel measured relative to the Fermi potential at the source and $\alpha$ is a constant with a value close to unity called the *bulk effect parameter*. We note that in strong inversion, $V_F(x)$ can be replaced by the channel potential $V(x)$ and the linear term in $n_s(x)$ will dominate on the right-hand side, signifying that charge transport in the channel will be drift current. Below threshold, the logarithmic term dominates on the right-hand side and the charge transport is primarily by diffusion.

Although (1.68) does not have an analytical solution with respect to $n_s$, we can use a generalized version of the approximate analytical expression introduced for the MOS capacitor in (1.38),

$$n_s = 2n_o \ln\left[1 + \frac{1}{2}\exp\left(\frac{V_{GT} - \alpha V_F}{\eta V_{th}}\right)\right] \tag{1.69}$$

This and related models have since been successfully applied to various FETs including MOSFETs, MESFETs, HFETs, poly-Si thin film transistors (TFTs), and a-Si TFTs (see

Fjeldly *et al.* 1998). The unified MOSFET model based on the UCCM expression in (1.68) is discussed in Chapter 8 (see also Shur *et al.* 1992). More elaborate MOSFET models such as the BSIM4 and EKV models are discussed in Chapters 6 and 7, respectively. They include a number of advanced features relating to small feature sizes and scaling of device dimensions.

BSIM4 is presently the most advanced MOSFET model supplied with Berkeley SPICE, and has been adopted in most commercial simulators. Although the BSIM models are characterized by a large number of SPICE parameters (in the hundreds), they have gained a wide popularity for use in professional circuit simulation and design, and have been accepted as an industry standard in the United States.

In Section 1.5.2, we consider more closely some of the advanced features included in modern MOSFET models, such as velocity saturation, gate bias–dependent mobility, impact ionization, drain and source series resistances (extrinsic modeling), channel-length modulation, and DIBL. In Section 1.5.3, we continue the discussion of the MOSFET capacitances from Section 1.4.4 and present a unified and charge-conserving description of the intrinsic capacitance–voltage characteristics.

## 1.5.2 Nonideal Effects

### *1.5.2.1 High-field effects*

*Channel-length modulation* When the drain-source bias of a FET approaches the drain saturation voltage, a region of high electric field forms near the drain and the electron velocity in this region saturates (in long devices, we instead have pinch-off where $n_s$ becomes very small near drain). In saturation, the length $\Delta L$ of the high-field region expands in the direction of the source with increasing $V_{DS}$, and the MOSFET behaves as if the effective channel length has been reduced by $\Delta L$. This phenomenon is called channel-length modulation (CLM). The following simplified expression links $V_{DS}$ to the length of the saturated region (see Lee *et al.* 1993):

$$V_{DS} = V_p + V_\alpha \left[\exp\left(\frac{\Delta L}{l}\right) - 1\right] \tag{1.70}$$

where $V_p$, $V_\alpha$, and $l$ are parameters related to the electron saturation velocity, the field effect mobility, and the drain conductance in the saturation regime. In fact, $V_p$ is the potential at the point of saturation in the channel, which is usually approximated by the saturation voltage $V_{SAT}$. Good agreement has been obtained between the potential profile described by (1.70) and that obtained from a two-dimensional simulation for the saturated region of an $n$-channel MOSFET.

The CLM effect manifests itself as a finite output conductance in saturation, which tends to remain constant over a wide range of drain biases. The output conductance also increases steadily with increasing gate bias. This observation suggests an even simpler model than that in (1.70) for describing CLM, where the basic expression for the drain current is simply multiplied by the first-order term $(1 + \lambda V_{DS})$. In this case, the CLM parameter $\lambda$ can easily be extracted from the output conductance in the saturation regime, well above threshold. This first-order approximation is implemented in several

FET models used in circuit simulators, while expressions similar to (1.70) are used in the BSIM models.

*Hot-carrier effects* Hot-carrier effects are among the main concerns when shrinking FET dimensions into the deep submicrometer regime. Reducing the channel length while retaining high power supply levels, known as constant voltage scaling, results in increased electric field strengths in the channel, causing acceleration and heating of the charge carriers.

Some of the manifestations of hot electrons on device operation are breakdown and substrate current caused by impact ionization, creation of interface states, gate current resulting from hot-electron emission across the interface barrier, oxide charges owing to tunneling of charge carriers into oxide states, and photocurrents caused by electron–hole recombination with emission of photons (see following text).

The substrate current resulting from electron–hole pair generation may overload substrate-bias generators, introduce snapback breakdown, cause CMOS latch-up, and generate a significant increase in the subthreshold drain current. A complete model for the substrate current is too complex for use in circuit level simulation. Instead, the following, approximate, analytical expression is widely used:

$$I_{\text{substr}} = I_{\text{ds}} \frac{A_i}{B_i} (V_{\text{DS}} - V_{\text{SAT}}) \exp\left(-\frac{l_{\text{d}} B_i}{V_{\text{DS}} - V_{\text{SAT}}}\right), \tag{1.71}$$

where $I_{\text{ds}}$ is the channel current, $A_i$ and $B_i$ are the ionization constants, $V_{\text{SAT}}$ is the saturation voltage, and $l_{\text{d}}$ is the effective ionization length. This expression is also applicable in the subthreshold regime by using $V_{\text{SAT}} = 0$ (Iñiguez and Fjeldly 1997).

In FETs fabricated on an insulating layer, such as silicon-on-insulator (SOI) MOSFETs, impact ionization may give rise to a charging of the transistor body, causing a shift in the threshold voltage. This effect results in an increased drain current in saturation (floating body effect). Related mechanisms are also observed in amorphous TFTs (Wang *et al.* 2000) and polysilicon TFTs (Iñiguez *et al.* 1999).

At sufficiently high drain bias, we have impact ionization and avalanche breakdown in all types of FETs. In MOSFETs, a substantial amount of the majority carriers created by impact ionization near drain will flow toward source and forward-bias the source–substrate junction, causing injection of minority carriers into the substrate. This effect can be modeled in terms of conduction in a parasitic bipolar transistor, as described by Sze (1981). In MESFETs, the breakdown usually takes place in the high-field depletion extension toward the drain.

Electron trapping in the oxide and generation of interface traps caused by hot-electron emission induce degradation of the MOSFET channel near drain in conventional MOSFETs or cause changes in the parasitic drain resistances in low-doped drain (LDD) MOSFET (Ytterdal *et al.* 1995). Reduced current drive capability and transconductance degradation are manifestations of interface traps in $n$-MOSFET characteristics. Reduced current also leads to circuit speed degradation, such that the circuits may fail to meet speed specifications after aging.

Photon emission and subsequent absorption in a different location of the device may cause unwanted photocurrent, which, for example, may degrade the performance of memory circuits.

*Temperature dependence and self-heating* Since electronic devices and circuits have to operate in different environments, including a wide range of temperatures, it is imperative to establish reliable models for such eventualities. Heat generated from power dissipation in an integrated circuit chip can be considerable, and the associated temperature rise must be accounted for both in device and circuit design. In conventional silicon substrates, the thermal conductivity is relatively high such that a well-designed chip placed on a good heat sink may achieve a reasonably uniform and tolerable operating temperature. However, such design becomes increasingly difficult as the device dimensions are scaled down and power dissipation increases. The thermal behavior of MOSFETs has been extensively studied in the past, and the temperature dependencies of major model parameters have been incorporated in SPICE models.

Circuits fabricated on substrates that are poor heat conductors, such as GaAs and silicon dioxide, are more susceptible to a significant self-heating effect (SHE). In thin film SOI CMOS, the buried $SiO_2$ layer inhibits an effective heat dissipation, and the self-heating manifests itself as a reduced drain current and even as a negative differential conductance at high power inputs. Hence, for a reliable design of SOI circuits, accurate and self-consistent device models that account for SHE are needed for use in circuit simulation.

The influence of SHE on the electrical characteristics of SOI MOSFETs can be evaluated using a two-dimensional device simulator incorporating heat flow or by combining a temperature rise model with an $I$–$V$ expression through an iteration procedure. But the effect can also be described in terms of a temperature-dependent model for the device's $I$–$V$ characteristics combined with the following simplified relationship between temperature rise and power dissipation:

$$T - T_o = R_{th} I_d V_{ds}. \tag{1.72}$$

Here $T$ is the actual temperature, $T_o$ is the ambient (substrate) temperature, and $R_{th}$ is a thermal resistance that contains information on thermal conductivity and geometry. The equations can be solved self-consistently, either numerically or analytically (see Cheng and Fjeldly 1996). Once the temperature dependence of the device parameters are established, the same procedure can also be used for describing self-heating in other types of devices, such as amorphous TFTs (Wang *et al.* 2000), GaAs MESFETs and HFETs.

*Gate bias–dependent mobility* In submicron MOSFETs, scaling dictates that the gate dielectric must be made very thin. In a sub-0.1-μm device, the gate dielectric may be as thin as a few nanometers. With a gate-source voltage of 1 V, this corresponds to a transverse electric field of nearly 500 kV/cm. In this case, electrons are confined to a very narrow region at the silicon–silicon dioxide interface, and their motion in the direction perpendicular to the gate oxide is quantized. This close proximity of the carriers to the interface enhances the scattering rate by surface nonuniformities, drastically reducing the field effect mobility in comparison to that of bulk silicon.

To a first-order approximation, the following simple expression accurately describes the dependence of the field effect mobility on the gate bias (Park *et al.* 1991)

$$\mu_n = \mu_{on} - \kappa_n (V_{GS} + V_T). \tag{1.73}$$

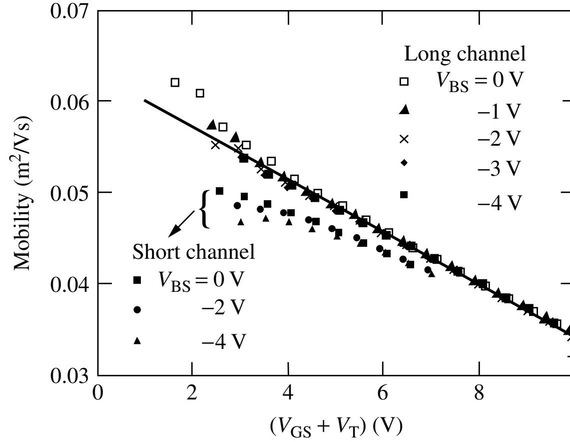


**Figure 1.24** Electron mobility versus $V_{GS} + V_T$ for a long-channel ($L = 20\,\mu$m) and a short-channel ($L = 1\,\mu$m) NMOS for different values of substrate bias. The solid line corresponds to the linear approximation used in (1.73). Reproduced from Park C. K. *et al.* (1991) A unified charge control model for long channel n-MOSFETs, *IEEE Trans. Electron Devices*, **ED-38**, 399–406

The experimental MOSFET mobility data in Figure 1.24 shows that this expression can be applied with the same set of parameters for different values of substrate bias. The parameter values are fairly close even for devices with quite different gate lengths. All in all, this leads to a reduction in the number of parameters needed for accurate modeling of the MOSFET characteristics.

More complete expression for the MOSFET field effect mobility, which takes both temperature variations and scaling into account, are used in the BSIM models (see Chapter 6).

### *1.5.2.2 Short-channel effects*

*Aspect ratio* To first order, FET dimensions are scaled by preserving the device aspect ratio, that is, the ratio between the gate length and the active vertical dimension of the device. In MOSFETs, the vertical dimension accounts for the oxide thickness $d_i$, the source and drain junction depths $r_j$, and the source and drain junction depletion depths $W_s$ and $W_d$. A low aspect ratio is synonymous with short-channel behavior. The following empirical relationship indicates the transition from long-channel to short-channel behavior (Brews *et al.* 1980):

$$L < L_{min}(\mu m) = 0.4[r_j(\mu m)d_i(\text{Å})(W_d + W_s)^2(\mu m^2)]^{1/3} \tag{1.74}$$

When $L < L_{min}$, the MOSFET threshold voltage $V_T$ will be affected in several ways as a result of reduced gate control. First, the *depletion* charges near source and drain are under the shared control of these contacts and the gate. In a short-channel device, the shared charge will constitute a relatively large fraction of the total gate depletion charge, giving rise to an increasingly large shift in $V_T$ with decreasing $L$. Also, the shared depletion charge near drain expands with increasing drain-source bias, resulting in an additional $V_{DS}$-dependent shift in $V_T$ (DIBL effect, see following text).

*Drain-induced barrier lowering* The threshold voltage is a measure of the strength of the barrier against carrier injection from source to channel. In the short-channel regime ($L < L_{\min}$), this barrier may be significantly modified by the application of a drain bias, as was schematically depicted in Figure 1.23. In $n$-channel FETs, this drain-induced barrier lowering (DIBL) translates into a lowering of the threshold voltage ($n$-channel MOSFET) and a concomitant rise in the subthreshold current with increasing $V_{DS}$. The combined scaling and DIBL effect on the threshold voltage may be expressed as follows:

$$V_T(L) = V_{To}(L) - \sigma(L)V_{DS} \tag{1.75}$$

where $V_{To}(L)$ describes the scaling of $V_T$ at zero drain bias resulting from charge sharing and $\sigma(L)$ is the channel-length-dependent DIBL parameter. In the long-channel case, where $L > L_{\min}$, $V_T$ should become independent of $L$ and $V_{DS}$. This behavior can be modeled by letting both $V_{To}(L)$ and $\sigma(L)$ scale approximately as $\exp(-L/L_{\min})$. In BSIM, somewhat more detailed scaling functions and also a dependence on substrate bias are used (see Chapter 6).

In Figure 1.25(a), we show experimental data of $V_T$ versus $V_D$ for two $n$-channel MOSFETs with short gate lengths. A good agreement with the linear relationship of (1.75) is indicated. Also, the exponential scaling for $V_T$ versus $L$ is confirmed by experiments, except for a deviation at the shortest gate lengths, as shown in Figure 1.25(b) (Fjeldly and Shur 1993).

As stated above, DIBL vanishes well above threshold. For modeling purposes, we therefore adopt the following empirical expression for $\sigma$ (Lee *et al.* 1993):

$$\sigma = \frac{\sigma_0}{1 + \exp\left(\dfrac{V_{gto} - V_{\sigma t}}{V_\sigma}\right)} \tag{1.76}$$

where $V_{gto}$ is the gate voltage overdrive at zero drain-source bias and the parameters $V_{\sigma t}$ and $V_\sigma$ determine the voltage and the width of the DIBL fade-out, respectively. We note that $\sigma \to \sigma_0$ for $V_{gto} < V_{\sigma t}$ and $\sigma \to 0$ for $V_{gto} > V_{\sigma t}$.

The DIBL effect can be accounted for in our $I$–$V$ models by adjusting the threshold voltage according to (1.75) in the expressions for the saturation current and the linear channel conductance. Likewise, the UCCM expression in (1.68) is modified as follows:

$$V_{GTo} + \sigma V_{DS} - \alpha V_F \approx \eta V_{th} \ln\left(\frac{n_s}{n_o}\right) + a(n_s - n_o) \tag{1.77}$$

where $V_{GTo}$ is the intrinsic threshold voltage overdrive at zero drain-source bias.

A related effect of device miniaturization is observed in narrow-channel FETs, where charges associated with the extension of the gate depletion regions beyond the nominal width of the gate may become a significant fraction of the total gate depletion charge. In this case, a one-dimensional analysis will underestimate the total depletion charge and give a wrong prediction of the threshold voltage. In practice, the threshold voltage increases ($n$-MOSFET) as the channel width is reduced. A common method of modeling this effect is to add an additional term in the threshold voltage expression containing a $1/W$ term, where $W$ is the effective width of the gate.

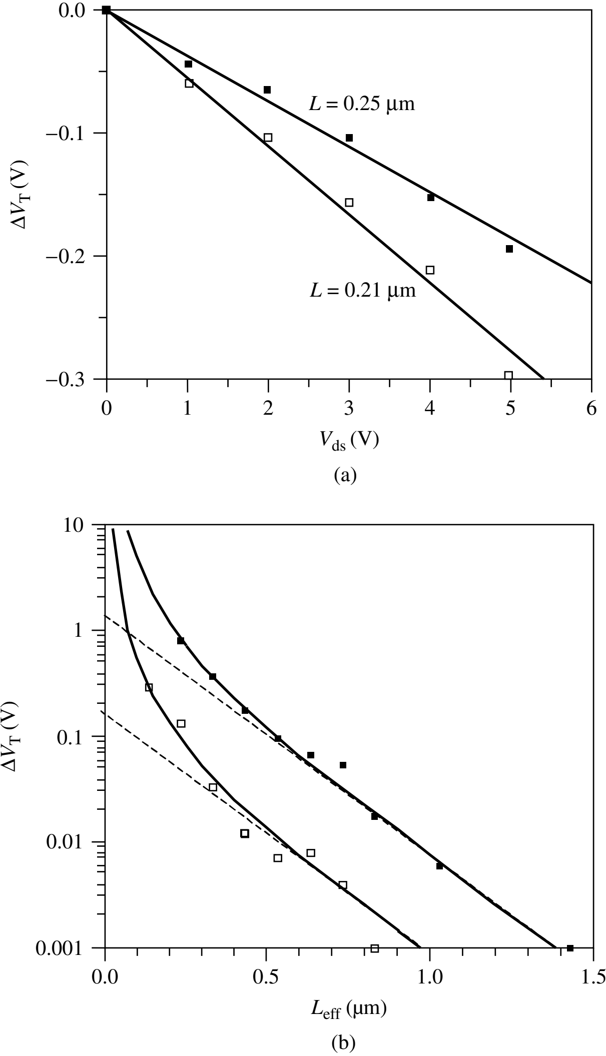


**Figure 1.25** DIBL effect: (a) experimental threshold voltage shift versus drain-source voltage for two $n$-MOSFETs with different gate lengths and (b) experimental threshold voltage shifts versus gate length compared with exponential scaling. Reproduced from Fjeldly T. A. and Shur M. (1993) Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs, *IEEE Trans. Electron Devices*, **TED-40**, 137–145

### *1.5.2.3 Gate leakage and effective oxide thickness*

The basic properties and the integrity of the silicon dioxide gate dielectric has been essential for the success of the silicon MOSFETs. However, as the CMOS technology node (half-pitch size in DRAMs) of MOSFETs in large-scale integration moves into the sub-100-nm range, this very success factor harbors one of the most difficult issues facing a continued evolution along the trend described by Moore's law. The reason for this

lies in the nonyielding rules of device scaling combined with the well-known quantum-mechanical phenomenon of tunneling. In fact, to derive sufficient advantage of sub-100 nm technology, the gate oxide thickness has to be scaled down to just a couple of nanometers or less, corresponding to only a handful of atomic layers (see ITRS 2001). At such small dimensions, the tunneling leakage current through the oxide from the gate to the semiconductor becomes significant enough to add noticeably to the power consumption and to interfere with the device operation.

An additional problem arises from the long-term reliability of such ultrathin dielectric films (Stathis 2002). These problems grow rapidly with further scaling, ultimately with completely debilitating consequences. The limits for viable scaling have recently been predicted to be at a technology node of 50 nm (gate length of 25 nm) and a silicon dioxide thickness of 1 nm (Wu *et al.* 2002).

A temporary solution to this impasse is to replace the silicon dioxide with materials that have much larger dielectric constants, so-called high-k insulators. This way, the same scaling advantage can be derived using a correspondingly thicker insulator with reduced tunneling current. Many such materials are presently being investigated, but it is hard to find candidates that can match the chemical and electrical properties of silicon dioxide and its excellent interface with silicon. If this development meets with success, the end of the present evolutionary trend in MOSFET/CMOS technology may be extended for yet another decade, bringing the technology node to about 20 nm (see ITRS 2001). Within this time frame, alternative MOSFET architectures with improved short-channel properties will also have to be developed for large-scale integration, including Vertical, FinFET, and planar double gate structures.

Another problem related to the thin dielectrics in MOSFETs is the relative importance of the inversion layer thickness in the semiconductor and the depletion layer thickness when using polysilicon gate electrodes. The former is a result of the lack of vertical confinement of the carriers, especially electrons, owing to the quantum-mechanical uncertainty and exclusion principles in combination with the finite steepness of the semiconductor band bending at the interface. In terms of device performance, these layers add to the effective oxide thickness $d_{\mathrm{eff}}$, thereby reducing the gate's field effect coupling to the channel. The two layers may contribute a few tenths of a nanometer each to $d_{\mathrm{eff}}$, which is quite significant for radically scaled MOSFETs. Development of suitable metal gate electrodes will alleviate some of the problem.

## 1.5.3 Unified MOSFET $C-V$ Model

### *1.5.3.1 Unified Meyer C–V model*

In order to develop unified expressions for the intrinsic MOSFET capacitances, we return to the Meyer capacitances discussed in Section 1.4. The Meyer large-signal equivalent is shown in Figure 1.16 and the gate-source and gate-drain capacitances are given by (1.59) and (1.60), respectively. The only device-specific part of these equations is the gate-channel capacitance $C_{\mathrm{ch}}$ at zero drain bias ($V_{\mathrm{F}} = 0$), which can be derived from the UCCM expression of (1.69), that is,

$$C_{\mathrm{ch}} = WLq\frac{\mathrm{d}n_{\mathrm{s}}}{\mathrm{d}V_{\mathrm{GT}}} \approx C_{\mathrm{i}}\left[1 + 2\exp\left(-\frac{V_{\mathrm{GT}}}{\eta V_{\mathrm{th}}}\right)\right]^{-1}. \tag{1.78}$$

Well above or below threshold, this expression has the familiar asymptotic forms

$$C_{ch} \approx C_i, \tag{1.79}$$

$$C_{ch} \approx \frac{C_i}{2} \exp\left(\frac{V_{GT}}{\eta V_{th}}\right), \tag{1.80}$$

respectively. Figure 1.26 shows $C_{ch}$ versus $V_{GT}$ in a linear and a semilogarithmic plot. From UCCM and from (1.78), we find that $C_{ch} = C_i/3$ at threshold, which may serve as a convenient and straightforward way of determining the threshold voltage from experimental $C_{ch}$ versus $V_{GS}$ curves.

In the subthreshold regime, the gate-substrate capacitance $C_{GB}$ of (1.65) is the dominant Meyer capacitance in MOSFETs. Above threshold, $C_{GB}$ vanishes in the ideal long-channel case. A unified version of $C_{GB}$ that includes a gradual phase-out above threshold can be modeled as follows:

$$C_{GB} = \frac{C_i/(1 + n_s/n_o)}{\sqrt{1 + 4(V_{GS} - V_{BS} - V_{FB})/\gamma^2}}. \tag{1.81}$$

Here $n_s$ is the unified electron density given by UCCM in (1.68) or its approximate solution (1.69). Equation (1.81) utilizes the fact that the increasing density of inversion charge above threshold gradually shields the substrate from the influence of the gate electrode. A typical plot of $C_{GB}$ versus $V_{GT}$ is shown in Figure 1.27.

Using this unified gate-channel capacitance in conjunction with Meyer's capacitance model, we obtain the following continuous expressions for the intrinsic gate-source capacitance $C_{GS}$ and the gate-drain capacitance $C_{GD}$, valid for all regimes of operation:

$$C_{GS} = \frac{2}{3} C_{ch} \left[1 - \left(\frac{V_{GTe} - V_{DSe}}{2V_{GTe} - V_{DSe}}\right)^2\right], \tag{1.82}$$

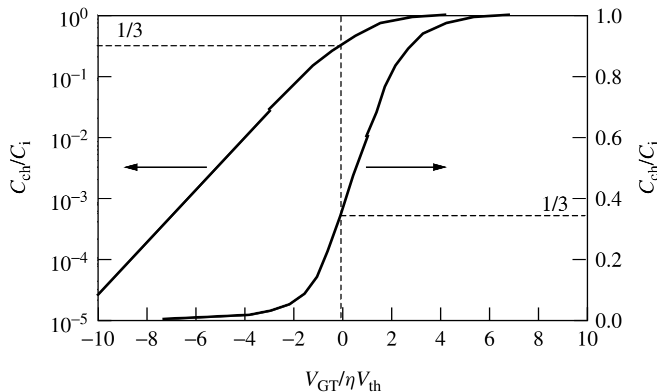


**Figure 1.26** Normalized channel capacitance versus $V_{GT}/\eta V_{th}$ according to (1.78) in a linear plot (right) and a semilog plot (left). The condition $C_{ch}/C_i = 1/3$ at threshold is indicated. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

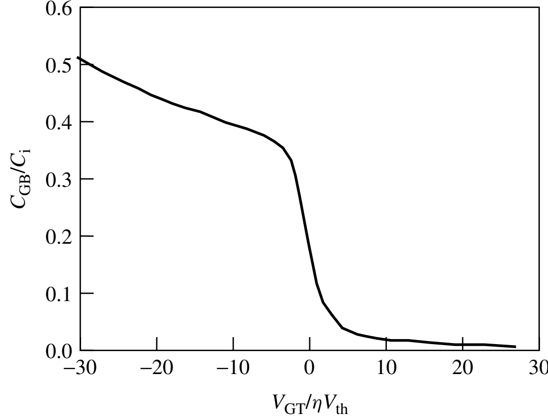


**Figure 1.27** Normalized and unified Meyer-type gate-substrate capacitance versus $V_{GT}/\eta V_{th}$ according to (1.81) for $V_{BS} = 0$. Typical values for an $n$-channel MOSFET with a polysilicon gate were used: $V_T = 0.7\,\mathrm{V}$, $V_{FB} = -1\,\mathrm{V}$, $\gamma = 1\,\mathrm{V}^{1/2}$, and $\eta = 1.33$. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

$$C_{GD} = \frac{2}{3} C_{ch} \left[ 1 - \left( \frac{V_{GTe}}{2V_{GTe} - V_{DSe}} \right)^2 \right]. \tag{1.83}$$

Here, $V_{DSe}$ is an effective intrinsic drain-source voltage that is equal to $V_{DS}$ for $V_{DS} < V_{GTe}$ and is equal to $V_{GTe}$ for $V_{DS} > V_{GTe}$. $V_{GTe}$ is the effective gate voltage overdrive, which equals $V_{GT}$ above threshold and is of the order of the thermal voltage in the subthreshold regime. The following expression is used to model this behavior:

$$V_{GTe} = V_{th} \left[ 1 + \frac{V_{GT}}{2V_{th}} + \sqrt{\delta^2 + \left( \frac{V_{GT}}{2V_{th}} - 1 \right)^2} \right], \tag{1.84}$$

where $\delta$ determines the width of the transition region at threshold ($V_{GT} = 0$). Typically, $\delta = 3$ is a good choice.

A smooth transition between the nonsaturated and the saturated regimes is assured by using the following type of interpolation expression for effective intrinsic drain-source voltage:

$$V_{DSe} = \frac{1}{2} \left[ V_{DS} + V_{GTe} - \sqrt{V_\delta^2 + (V_{DS} - V_{GTe})^2} \right], \tag{1.85}$$

where $V_\delta$ is a constant voltage that determines the width of the transition region. This parameter may be treated as an adjustable parameter to be extracted from experiments. $V_{GTe}$ is needed to assure a smooth transition between the correct limiting $I$–$V$ and $C$–$V$ expressions above and below threshold.

A comparison of the normalized dependencies of $C_{GS}$ and $C_{GD}$ on $V_{DS}$ is shown in Figure 1.28 for $V_\delta / V_{GTe} = 0$, corresponding to the nonunified Meyer capacitances, and for

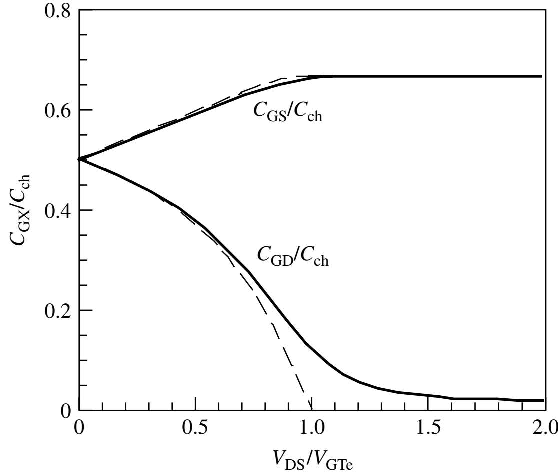


**Figure 1.28** Normalized and nonunified Meyer capacitances according to (1.59) and (1.60) (dashed lines) and unified Meyer capacitances according to (1.82) and (1.83) (solid lines), using a transition width parameter $V_\delta = 0.2\,V_{GTe}$. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

a more realistic value of $V_\delta/V_{GTe} = 0.2$. On the basis of the discussion in Section 1.4.4, we can conclude that the present unified version of the Meyer capacitances is applicable also for short-channel devices. Still more flexible expressions for the capacitances are obtained by substituting $V_{GT}$ by $\chi V_{GT}$ in (1.82) and (1.83), where $\chi$ is an adjustable parameter close to unity.

### *1.5.3.2 Ward–Dutton model*

As we discussed in Section 1.4, an accurate modeling of the intrinsic capacitances associated with the gate region of FETs requires an analysis of the charge distribution in the channel versus the terminal bias voltages. Normally, the problem is simplified by assigning the distributed charges to the various "intrinsic" terminals. Hence, the mobile charge $Q_I$ of a MOSFET is divided into a source charge $Q_S = F_p Q_I$ and a drain charge $Q_D = (1 - F_p)Q_I$, where $F_p$ is a partitioning factor. The depletion charge $Q_B$ under the gate is assigned to the MOSFET substrate terminal. The total gate charge $Q_G$ is the negative sum of these charges, that is, $Q_G = -Q_I - Q_B = -Q_S - Q_D - Q_B$. Note that by assigning the charges this way, charge conservation is always assured.

The net current flowing into terminal $X$ can now be written as

$$I_X = \frac{dQ_X}{dt} = \sum_Y \frac{\partial Q_X}{\partial V_Y}\frac{\partial V_Y}{\partial t} = \sum_Y \chi_{XY} C_{XY}\frac{\partial V_Y}{\partial t}, \tag{1.86}$$

where the indices $X$ and $Y$ run over the terminals $G$, $S$, $D$, and $B$. In this expression, we have introduced a set of intrinsic capacitance elements $C_{XY}$, the so-called transcapacitances, defined by

$$C_{XY} = \chi_{XY}\frac{\partial Q_X}{\partial V_Y} \quad \text{where} \quad \chi_{XY} = \begin{cases} -1 & \text{for } X \neq Y \\ 1 & \text{for } X = Y \end{cases}. \tag{1.87}$$

These are equivalent to the charge-based nonreciprocal capacitances introduced by Ward and Dutton (1978) and by Ward (1981). The term *nonreciprocal* means that we have $C_{XY} \neq C_{YX}$ when $X \neq Y$. The elements $C_{XX}$ are called self-capacitances. $C_{XY}$ contain information on how much the charge $Q_X$ assigned to terminal $X$ changes by a small variation in the voltage $V_Y$ at terminal $Y$. To illustrate why $C_{XY}$ may be different from $C_{YX}$, assume a MOSFET in saturation. Then the gate charge changes very little when the drain voltage is perturbed since the inversion charge is very little affected, making $C_{\mathrm{GD}}$ small. However, if $V_{\mathrm{G}}$ is perturbed, the inversion charge changes significantly and so does $Q_{\mathrm{D}}$, making $C_{\mathrm{DG}}$ large.

For the four-terminal MOSFET, the Ward–Dutton description leads to a total of 16 transcapacitances. This set of 16 elements can be organized as follows in a 4 × 4 matrix, a so-called indefinite admittance matrix:

$$\mathbf{C} = \begin{bmatrix} C_{\mathrm{GG}} & C_{\mathrm{GS}} & C_{\mathrm{GD}} & C_{\mathrm{GB}} \\ C_{\mathrm{SG}} & C_{\mathrm{SS}} & C_{\mathrm{SD}} & C_{\mathrm{SB}} \\ C_{\mathrm{DG}} & C_{\mathrm{DS}} & C_{\mathrm{DD}} & C_{\mathrm{DB}} \\ C_{\mathrm{BG}} & C_{\mathrm{BS}} & C_{\mathrm{BD}} & C_{\mathrm{BB}} \end{bmatrix}. \tag{1.88}$$

Here, the elements in each column and each row must sum to zero owing to the constraints imposed by charge conservation (which is equivalent to obeying Kirchhoff's current law) and for the matrix to be reference independent, respectively (see Arora 1993). This means that some of the transconductances will be negative, and of the 16 MOSFET elements, only 9 are independent. The complete MOSFET large-signal equivalent circuit, including the 16 transcapacitances, is shown in Figure 1.29. This compares with the simple Meyer model in Figure 1.16, which comprises 3 capacitances.

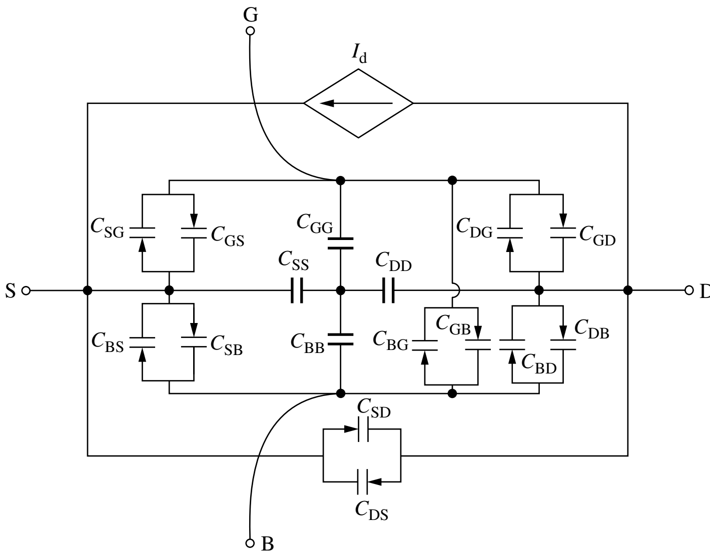


**Figure 1.29** Intrinsic large-signal MOSFET equivalent circuit including a complete set of nonreciprocal and charge-conserving transcapacitances. The transcapacitances $C_{XY}$ are defined in the text. Reproduced from Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York

We note that in three-terminal FETs, such as HFETs, MESFETs, and TFTs, we have a total of 9 transcapacitances, of which 4 are independent (Nawaz and Fjeldly 1997). The equivalent circuit for this case is obtained from Figure 1.29 by removing the substrate terminal $B$ and all elements connected to it. Also, the $4 \times 4$ matrix in (1.88) reduces to a $3 \times 3$ matrix.

As explained in Section 1.4.4, the simplified $C$–$V$ model by Meyer is obtained by taking derivatives of the total gate charge with respect to the various terminal voltages. The Meyer capacitances can be viewed as a subset of the Ward–Dutton capacitances. Although charge conservation is not assured in the Meyer model, the resulting errors in circuit simulations are usually small, but can in some cases lead to serious errors. The unified transcapacitances needed for the complete Ward–Dutton model can be obtained along the same lines as described for $C_{GS}$ and $C_{DS}$. The accuracy of the model depends on the quality of the charge and current models used and on the partitioning of the inversion charge between the source and the drain terminal.

#### *1.5.3.3 Non-quasi-static modeling*

For very high-frequency operation of the MOSFET, comparable to the inverse carrier transport time of the channel (non-quasi-static (NQS) regime), we have to consider the temporal relaxation of the inversion and depletion charges. Most of the MOSFET models used in SPICE are based on the quasi-static assumption (QSA), in which an instantaneous charging of the inversion layer is assumed. Hence, circuit simulations will fail to accurately predict the performance of high-speed circuits.

The channel of a MOSFET is analogous to a bias-dependent distributed $RC$ network as indicated schematically in Figure 1.30. In QSA, the distributed gate-channel capacitance is instead lumped into discrete capacitances between the gate and source and drain nodes, ignoring the finite charging time arising from the $RC$ product associated with the channel resistance and the gate-channel capacitance.

The inclusion of the so-called Elmore equivalent circuit shown in Figure 1.31 can be viewed as a first step toward an NQS model. Using this equivalent circuit, the channel charge buildup is modeled with reasonable accuracy because the lowest frequency pole of the original $RC$ network is retained. The Elmore resistance $R_{\text{Elmore}}$ is calculated from the channel resistance in strong inversion as

$$R_{\text{Elmore}} \approx \frac{L_{\text{eff}}^2}{e\mu_{\text{eff}} Q_{\text{ch}}}. \tag{1.89}$$

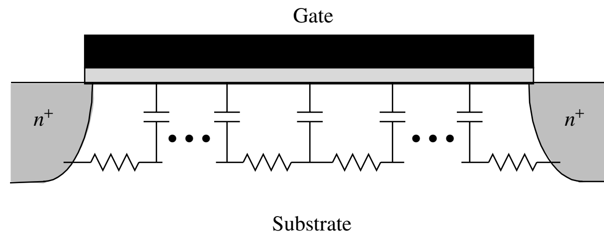


**Figure 1.30** Equivalent $RC$ network representing the MOSFET channel

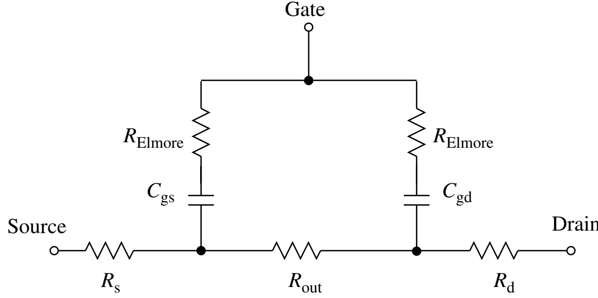


**Figure 1.31** Elmore non-quasi-static equivalent circuit

where $e$ is the Elmore constant with a theoretical value close to 5 and $Q_{\text{ch}}$ is the total charge in the channel. This formulation is only valid above threshold where the drift current dominates.

To obtain a unified expression, including the subthreshold diffusion current, a relaxation time–based approach is adapted. The overall relaxation time for channel charging and discharging can be written as a combination of the contributions due to drift and diffusion as follows:

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{drift}}} + \frac{1}{\tau_{\text{diff}}}, \tag{1.90}$$

where

$$\tau_{\text{drift}} = R_{\text{Elmore}} C_{\text{i}}, \tag{1.91}$$

$$\tau_{\text{diff}} = \frac{q(L_{\text{eff}}/4)^2}{\mu_{\text{eff}} k_{\text{B}} T}. \tag{1.92}$$

On the basis of this relaxation time concept, the NQS effect can be implemented in the SPICE MOSFET model using the subcircuit shown in Figure 1.32. The variable $Q_{\text{def}}$ is an additional node created to keep track of the amount of deficit or surplus channel charge needed to achieve equilibrium. $Q_{\text{def}}$ will decay exponentially into the channel with a bias-dependent NQS relaxation time $\tau$, and the terminal currents can be written as

$$I_{\text{d}} = I_{\text{d}}(dc) + X_{\text{d}} \frac{Q_{\text{def}}}{\tau}, \tag{1.93}$$

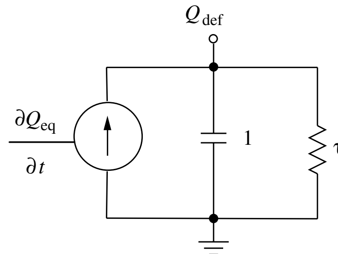


**Figure 1.32** Non-quasi-static subcircuit implementation in MOSFET SPICE models. The $RC$ time constant $\tau$ is determined by the resistance and capacitance values chosen

$$I_{\mathrm{g}} = -\frac{Q_{\mathrm{def}}}{\tau}. \tag{1.94}$$

Here $X_{\mathrm{d}} = 1 - F_p$ and $X_{\mathrm{s}} = F_p$, where $F_p$ is the charge partitioning factor introduced in Section 1.5.3.2.

## REFERENCES


Arora N. (1993) *MOSFET Models for VLSI Circuit Simulation. Theory and Practices*, Springer-Verlag, Wien.

Brews J. R., Fichtner W., Nicollian E. H., and Sze S. M. (1980) Generalized guide for MOSFET miniaturization, *IEEE Electron Device Lett.*, **EDL-1**, 2.

Byun Y., Lee K., and Shur M. (1990) Unified charge control model and subthreshold current in heterostructure field effect transistors, *IEEE Electron Device Lett.*, **EDL-11**, 50–53; (see erratum (1990) *IEEE Electron Device Lett.*, **EDL-11**, 273).

Cheng Y. and Fjeldly T. A. (1996) Unified physical I–V model including self-heating effect for fully depleted SOI/MOSFET's, *IEEE Trans. Electron Devices*, **ED-43**, 1291–1296.

Fjeldly T. A. and Shur M. (1993) Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs, *IEEE Trans. Electron Devices*, **TED-40**, 137–145.

Fjeldly T. A., Ytterdal T., and Shur M. (1998) *Introduction to Device Modeling and Circuit Simulation*, John Wiley & Sons, New York.

Fonstad C. G. (1994) *Microelectronic Devices and Circuits*, McGraw-Hill, New York.

Iñiguez B. and Fjeldly T. A. (1997) Unified substrate current model for MOSFETs, *Solid-State Electron.*, **41**, 87–94.

Iñiguez B., Xu Z., Fjeldly T. A., and Shur M. S. (1999) Unified model for short-channel poly-Si TFTs, *Solid-State Electron.*, **43**, 1821–1831.

*ITRS – International Technology Roadmap for Semiconductor*, Semiconductor Industry Assoc., Austin, TX (2001).

Lee K., Shur M., Fjeldly T. A., and Ytterdal T. (1993) *Semiconductor Device Modeling for VLSI*, Prentice Hall, Englewood Cliffs, NJ.

Meyer J. E. (1971) MOS models and circuit simulation, *RCA Rev.*, **32**, 42–63.

Nawaz M. and Fjeldly T. A. (1997) A new charge conserving capacitance model for GaAs MESFETs, *IEEE Trans. Electron Devices*, **ED-44**, 1813–1821.

Park C. K. *et al*. (1991) A unified charge control model for long channel n-MOSFETs, *IEEE Trans. Electron Devices*, **ED-38**, 399–406.

Shur M. (1990) *Physics of Semiconductor Devices*, Prentice Hall, Englewood Cliffs, NJ.

Shur M., Fjeldly T. A., Ytterdal T., and Lee K. (1992) Unified MOSFET model, *Solid-State Electron.*, **35**, 1795–1802.

Sodini C. G., Ko P. K., and Moll J. L. (1984) The effect of high fields on MOS device and circuit simulation, *IEEE Trans. Electron Devices*, **ED-31**, 1386–1393.

Stathis J. H. (2002) Reliability limits of the gate insulator in CMOS technology, *IBM J. Res. Dev.*, **46**(2/3), 265–286.

Sze S. M. (1981) *Physics of Semiconductor Devices*, Second Edition, John Wiley & Sons, New York.

Wang L. *et al*. (2000) Self-heating and kink effects in a-Si:H thin film transistors, *IEEE Trans. Electron Devices*, **ED-47**, 387–397.

Ward D. E. and Dutton R. W. (1978) A charge-oriented model for MOS transistor capacitances, *IEEE J. Solid-State Circuits*, **SC-13**, 703–708.

Ward D. E. (1981) *Charge Based Modeling of Capacitance in MOS Transistors*, Ph.D. thesis, Stanford University, Stanford, CA.

Wu E. Y. *et al*. (2002) CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics, *IBM J. Res. Dev.*, **46**(2/3), 287–298.

Ytterdal T., Kim S. H., Lee K., and Fjeldly T. A. (1995) A new approach for modeling of current degradation in hot-electron damaged LDD NMOSFETs, *IEEE Trans. Electron Devices*, **ED-42**, 362–364 (1995).