3

Field Effect Transistors

Veena Misra and
Mehmet C. Öztürk
Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, North Carolina, USA

3.1 Introduction 109
3.2 Metal-Oxide-Silicon Capacitor 109
3.2.1 The Ideal MOS Capacitor • 3.2.2 Deviations from the Ideal Capacitor • 3.2.3 Small-Signal Capacitance • 3.2.4 Threshold Voltage
3.3 Metal-Oxide-Silicon Field Effect Transistor 113
3.3.1 Device Operation • 3.3.2 Effective Channel Mobility • 3.3.3 Nonuniform Channels • 3.3.4 Short Channel Effects • 3.3.5 MOSFET Scaling • 3.3.6 Modern MOSFETs
3.4 Junction Field Effect Transistor 122
3.5 Metal-Semiconductor Field Effect Transistor 123
3.6 Modulation-Doped Field Effect Transistor 124
References 126

3.1 Introduction

The concept of modulating the electrical current in a semiconductor by an external electrical field was first proposed by Julius Lilienfeld in the 1930s. Years later, William Shockley, a scientist at Western Electric, led a research program to create a semiconductor device based on this "field effect" concept. To replace bulky vacuum tubes then used in telephone switching. Two scientists in Shockley's team, W. Brattain and J. Bardeen, invented the point contact transistor in 1947. Subsequently, Shockley invented the bipolar-junction transistor (BJT). Shockley later developed the junction field effect transistor (JFET); however, the JFET was not able to challenge the dominance of BJT in many applications. The development of field effect devices continued with moderate progress until the early 1960s, when the metal-oxide-silicon field effect transistor (MOSFET) emerged as a prominent device. This device quickly became popular in semiconductor memories and digital integrated circuits. Today, the MOSFET dominates the integrated circuit technology, and it is responsible for the computer revolution of the 1990s.

The first half of this chapter is dedicated to the basic theory of the MOSFET, beginning with its fundamental building block, the MOS capacitor. The second half of the chapter presents an overview of less common field effect devices used only in specific applications. Their coverage is limited to a qualitative explanation of the operating principles and basic equations. The reader is referred to other books available in literature to obtain detailed information.

3.2 Metal-Oxide-Silicon Capacitor

The metal-oxide-silicon (MOS) capacitor is at the core of the complementary metal-oxide-silicon (CMOS) technology. MOSFETs rely on the extremely high quality of the interface between SiO_2, the standard gate dielectric, and silicon (Si). Before presenting the MOSFET in detail, it is essential to achieve a satisfactory understanding of the MOS capacitor fundamentals.

The simplified schematic of the MOS capacitor is shown in Figure 3.1. The structure is similar to a parallel plate capacitor in which the bottom electrode is replaced by a semiconductor. When Si is used as the substrate material, the top electrode, known as the gate, is typically made of polycrystalline silicon (polysilicon), and the dielectric is thermally grown silicon dioxide. In MOS terminology, the substrate is commonly referred to as the *body.*

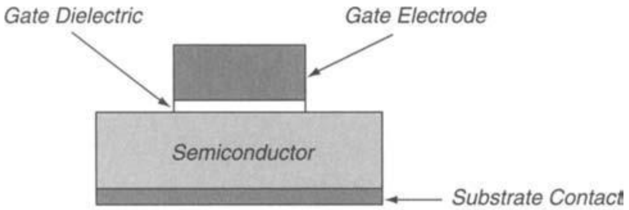

FIGURE 3.1

3.2.1 The Ideal MOS Capacitor

The energy band diagram of an ideal p-type substrate MOS capacitor at zero bias is shown in Figure 3.2. In an ideal MOS capacitor, the metal work function, ϕ_m, is equal to the semiconductor work function, ϕ_s. Therefore, when the Fermi level of the semiconductor, E_{FS}, is aligned with the Fermi level of the gate, E_{Fm}, there is no band bending in any region of the MOS capacitor. Furthermore, the gate dielectric is assumed to be free of any charges, and the semiconductor is uniformly doped.

Figure 3.3 shows the energy band diagram and the charge distribution in an ideal MOS structure for different gate-to-body voltages (V_{GB}).

With a negative gate bias (Figure 3.3(A)), the gate charge, Q_G, is negative. The source of this negative charge is electrons supplied by the voltage source. In an MOS capacitor, charge neutrality is always preserved. This requires:

$$Q_G + Q_C = 0, \tag{3.1}$$

where Q$_C$ is the charge induced in the semiconductor. Therefore, a net positive charge, Q_C, must be generated in the silicon substrate to counterbalance the negative charge on the gate. This is achieved by accumulation of the positively charged holes under the gate. This condition, where the majority carrier concentration is greater near the Si–SiO$_2$ interface compared to the bulk, is called **accumulation**.

Under an applied negative gate bias, the Fermi level of the gate is raised with respect to the Fermi level of the substrate by an amount equal to qV_{GB}. The energy bands in the semiconductor bend upward, bringing the valence band closer to the Fermi level that is indicative of a higher hole concentration under the dielectric. It is important to note that the Fermi level in the substrate remains invariant even under an applied bias since no current can flow through the device due to the presence of an insulator.

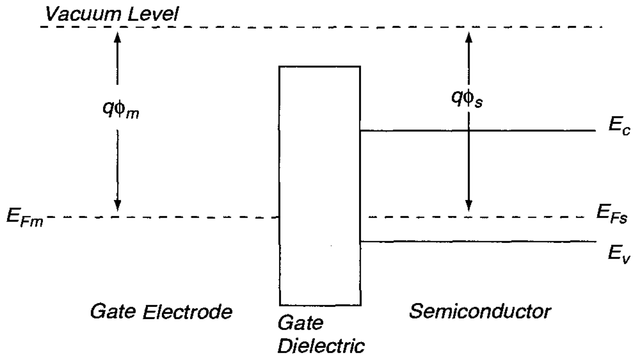

FIGURE 3.2

The applied gate voltage is shared between the gate dielectric and the semiconductor such that:

$$V_G = V_{ox} + V_c, \tag{3.2}$$

where V_{ox} and V_c are the voltages that drop across the oxide and the semiconductor, respectively. The band bending in the oxide is equal to qV_{ox}. The electric field in the oxide can be expressed as:

$$E_{ox} = \frac{V_{ox}}{t_{ox}}, \tag{3.3}$$

where t_{ox} is the oxide thickness. The amount of band bending in the semiconductor is equal to $q\psi_s$, where ψ_s is the surface potential and is negative when the band bending is upward.

Figure 3.3(B) shows the energy band diagram and the charge distribution for a positive gate bias. To counterbalance the positive gate charge, the holes under the gate are pushed away, leaving behind ionized, negatively charged acceptor atoms, which creates a **depletion region**. The charge in the depletion region is exactly equal to the charge on the gate to preserve charge neutrality. With a positive gate bias, the Fermi level of the gate is lowered with respect to the Fermi level of the substrate. The bands bend downward, resulting in a positive surface potential. Under the gate, the valence band moves away from the Fermi level indicative of hole depletion. When the band bending at the surface is such that the intrinsic level coincides with the Fermi level, the surface resembles an intrinsic material. The surface potential required to have this condition is given by:

$$\psi_s = \phi_F = \frac{1}{q}(E_i - E_F), \tag{3.4}$$

where

$$\phi_F = \frac{kT}{q}\ln\frac{N_A}{n_i} \tag{3.5}$$

Under a larger positive gate bias, the positive charge on the gate increases further, and the oxide field begins to collect thermally generated electrons under the gate. With electrons, the intrinsic surface begins to change into an n-type inversion layer. The negative charge in the semiconductor is comprised of ionized acceptor atoms in the depletion region and free electrons in the inversion layer. As noted above, at this point, the electron concentration at the surface is still less than the

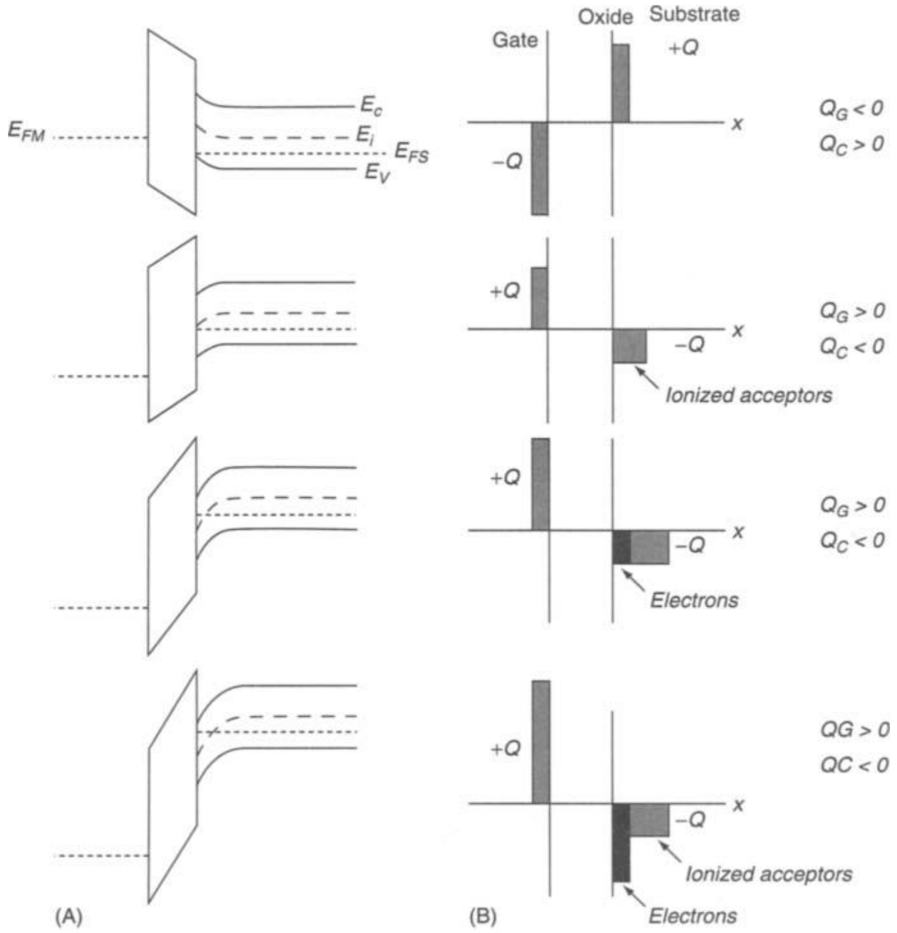

FIGURE 3.3

hole concentration in the neutral bulk. Thus, this condition is referred to as **weak inversion** and $\psi_s = \phi_F$ is defined as the onset of weak inversion and is shown in Figure 3.3(B).

As the gate bias is increased further, the band bending increases. The depletion region becomes wider, and the electron concentration in the inversion layer increases. When the electron concentration is equal to the hole concentration in the bulk, a **strong inversion** layer is said to form. The surface potential required to achieve strong inversion is equal to:[1]

$$\psi_s = 2\phi_F = \frac{2}{q}(E_i - E_F). \tag{3.6}$$

The inversion and depletion charge variation with ψ_s is shown in Figure 3.4. The electron concentration in the inversion layer is an exponential function of the surface potential and is given by:

$$n_{inv} \approx N_A e^{q(\psi_s - 2\phi_F)/kT} \tag{3.7}$$

On the other hand, the charge density in the depletion region is written as:

$$Q_D = -qN_A W_D, \tag{3.8}$$

where W_D is the depletion region width is given by:

$$W_D = \sqrt{\frac{2\varepsilon_s}{qN_A}}\sqrt{\psi_s}. \tag{3.9}$$

Therefore, the charge density in the depletion region is a weak function of the surface potential. Consequently, when the gate bias is further increased beyond the value required to reach strong inversion, the extra positive charge on the gate can be easily compensated by new electrons in the inversion layer. This eliminates the need to uncover additional acceptor atoms in the depletion region. After this point, the depletion region width and, hence, the band bending can increase only

[1] In some texts, $\psi_s = 2\phi_F$ is taken as the onset of moderate inversion. For strong inversion, the surface potential is required to be $\sim$ 6 kT/q above this level.

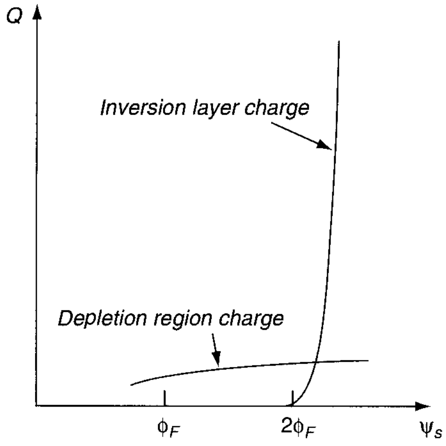

FIGURE 3.4

negligibly. The surface potential is pinned to its maximum value, which is a few kT/q over $2\phi_F$. For simplicity, $2\phi_F$ is generally taken as the maximum value for the surface potential, ψ_s.[2] The maximum depletion layer width, $W_{D,max}$, reached at the onset of strong inversion is given by:

$$W_{D,max} = \sqrt{\frac{2\varepsilon_s}{qN_A}}\sqrt{2\phi_F}. \tag{3.10}$$

3.2.2 Deviations from the Ideal Capacitor

In practical devices, the work function of the metal, ϕ_m, is not equal to that of the semiconductor, ϕ_s. Consider a MOS capacitor with $\phi_m < \phi_s$ at zero gate bias (gate connected to the substrate). Electrons in the metal reside at energy levels above the conduction band of the semiconductor. As a result, electrons flow from the metal into the semiconductor until a potential that counterbalances the difference in the work functions is built up between the two plates. This induces a negative charge under the gate dielectric accompanied by a downward band bending and, hence, a positive surface potential.

If an external voltage equal to this difference is applied to the gate, the net charge in the semiconductor disappears and the bands return to their flat position. This voltage is defined as the flatband voltage, V_{FB}, and can be expressed as:

$$V_{FB} = q\phi_{ms} = q(\phi_m - \phi_s). \tag{3.11}$$

Another modification to the above picture comes from charges that reside in the dielectric. These charges may originate from processing, from defects in the bulk, or from charges that exist at the interface between Si and SiO$_2$. The charges in the oxide or at the interface induce opposite charges in the semiconductor and the gate. This can cause the bands to bend up or down. Therefore, the flat-band voltage will have to be adjusted to take this into account. Assuming that all oxide fixed charges reside at the Si–SiO$_2$ interface, the flat-band voltage can be expressed as:

$$V_{FB} = \phi_{ms} - \frac{Q_o}{C_{ox}}, \tag{3.12}$$

where Q_o is the oxide charge and C_{ox} is the oxide capacitance.

In addition, defects located at the Si–SiO$_2$ interface may not be fixed in their charge state and may vary with the surface potential of the substrate. These defects are referred to as **fast interface traps** and have an impact on switching characteristics of MOSFETs as will be discussed later.

Based on the values of ψ_s, different regions of operation can be defined and are shown in Table 3.1.

3.2.3 Small-Signal Capacitance

If the gate voltage is changed by a small positive amount, a small positive charge will flow into the gate. To preserve charge neutrality, an equal amount of charge must flow out of the semiconductor. The relation of the small change in charge due to a small change in voltage is defined as the small-signal capacitance and can be written as:

$$C_{gb} \equiv \frac{dQ_G}{dV_{GB}}. \tag{3.13}$$

The equivalent circuit for the total gate capacitance is shown in Figure 3.5. The total gate capacitance consists of the oxide and semiconductor capacitances, where the semiconductor capacitance is the sum of the depletion, the inversion capacitance, and the interface states capacitance.

3.2.4 Threshold Voltage

The threshold voltage of an MOS capacitor is the gate voltage, V_{GB}, required to create strong inversion (i.e., $\psi_s = 2\phi_F$) under the gate. Figure 3.6 shows the inversion charge as a function of V_{GB}. The straight-line extrapolation of this charge to the x-axis

TABLE 3.1 Regions of Operation of MOS Capacitor for p-Substrate

Accumulation	$\psi_s < 0$	$V_{GB} < V_{FB}$	$Q_c > 0$
Depletion	$\psi_s > 0$	$V_{GB} > V_{FB}$	$Q_c < 0$
Inversion	$\psi_s > 0$	$V_{GB} >> V_{FB}$	$Q_c < 0$

[2] A more accurate definition for the onset of strong inversion is the surface potential at which the inversion layer charge is equal to the depletion region charge, which is ~ 6 kT/q above the onset of moderate inversion.

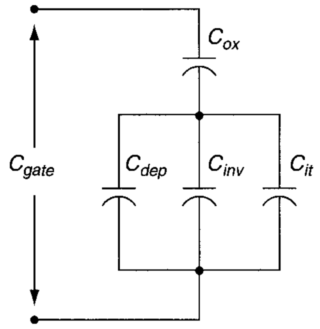

FIGURE 3.5

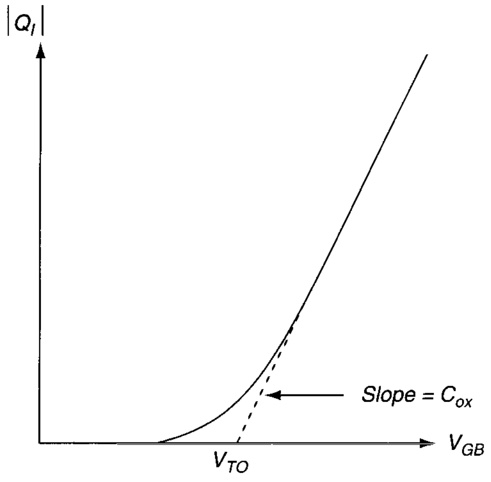

FIGURE 3.6

is called the extrapolated threshold voltage, V_{TO}. When the semiconductor (body) is at ground potential ($V_B = 0$), V_{TO} is given by:

$$V_{TO} = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F}, \qquad (3.14)$$

where γ is referred to as the body-effect coefficient and can be expressed as:

$$\gamma = \frac{\sqrt{2q\varepsilon_S N_A}}{C_{OX}}. \qquad (3.15)$$

With a body bias, the threshold voltage is given by:

$$V_T = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_B}. \qquad (3.16)$$

As shown in the above equation, the threshold voltage increases when a back bias is applied. A positive bias on the substrate results in a wider depletion region and assists in balancing the gate charge. This causes the electron concentration in the inversion layer to decrease. Thus, a higher gate voltage is needed to achieve the onset of inversion resulting in an increase of the threshold voltage. In addition, the doping concentration and oxide thickness can also have an impact on the threshold voltage dependence on back bias. Lower doping concentrations and thinner oxides result in a weaker dependence of back bias on threshold voltage.

3.3 Metal-Oxide-Silicon Field Effect Transistor

The MOSFET is the building block of the current ultra large-scale integrated circuits (ICs). The growth and complexity of MOS ICs have been continually increasing over the years to enable faster and denser circuits. Shown in Figure 3.7 is the simplified schematic of a MOSFET. The device consists of a MOS gate stack formed between two *pn* junctions called the **source** and **drain** junctions. The region under the gate is often referred to as the channel region. The length and the width of this region are called the **channel length** and **channel width**, respectively. An inversion layer under the gate creates a conductive path (i.e., channel) from source to drain and turns the transistor on. When the channel forms right under the gate dielectric, the MOSFET is referred to as a **surface channel MOSFET**. Buried channel MOSFETs in which the channel forms slightly beneath the surface are also used, but they are becoming less popular with the continuous downscaling of MOSFETs. A fourth contact to the substrate is also formed and referred to as the **body contact**.

The standard gate dielectric used in Si-integrated circuit industry is SiO_2. The gate electrode is heavily doped *n*-type or *p*-type polysilicon. The source and drain regions are formed by ion-implantation.

We shall first consider MOSFETs with uniform doping in the channel region. Later in the chapter, however, we shall learn how nonuniform doping profiles are used to enhance the performance of MOSFETs.

The MOSFET shown in Figure 3.7 is an *n*-channel MOSFET, in which electrons flow from source to drain in the channel

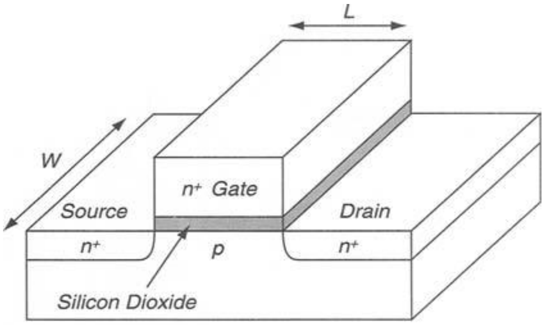

FIGURE 3.7

TABLE 3.2 Dopant Types for Different Regions of *n*-Channel and *p*-Channel MOSFETs

	n-channel MOSFET	*p*-channel MOSFET
Substrate (channel)	*p*	*N*
Gate electrode	*n*+	*p*+
Source and drain	*n*+	*p*+

induced under the gate oxide. Both *n*-channel and *p*-channel MOSFETs are extensively used. In fact, CMOS IC technology relies on the ability to use both devices on the same chip. Table 3.2 shows the dopant types used in each region of the two structures.

3.3.1 Device Operation

Figure 3.8 shows an *n*-channel MOSFET with voltages applied to its four terminals. Typically, $V_S = V_B$ and $V_D > V_S$. Shown in Figure 3.9 is the typical $I_{DS} - V_{DS}$ characteristics of such a device. For simplicity, we assume that the body and the source terminals are tied to the ground (i.e. $V_{SB} = 0$. This yields.

$$\begin{aligned} V_{GS} &= V_{GB} - V_{SB} = V_{GB} = V_G. \\ V_{DS} &= V_{DB} - V_{SB} = V_{DB} = V_D. \end{aligned} \tag{3.17}$$

As shown, at low V_{DS}, the drain current increases almost linearly with V_{DS}, resulting in a series of straight lines with slopes increasing with V_{GS}. At high V_{DS}, the drain current saturates and becomes independent of V_{DS}.

In this section, we present a description of the device operation complete with important equations valid in different regions of operation. We shall refer to Figure 3.10, which shows the inversion layer and the depletion regions that form under the inversion layer and around the source and drain junctions. Now we can begin to discuss MOSFET regions of operation in detail.

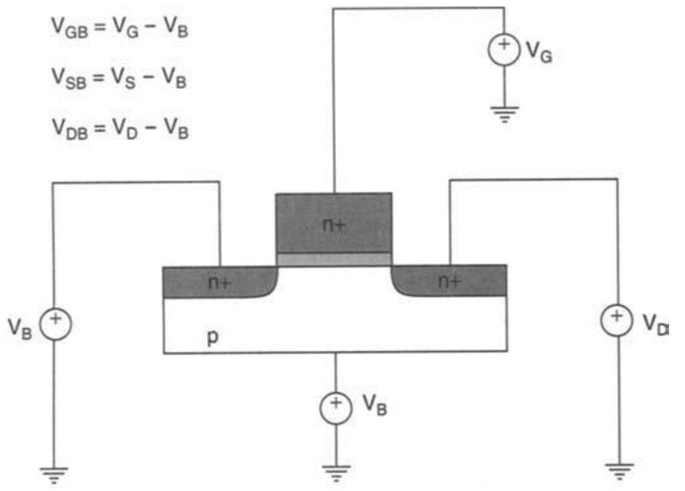

FIGURE 3.8

Linear Region

With a small positive voltage on the drain and no bias on the gate (i.e., $V_{DS} > 0$ and $V_{GS} = 0$), the drain is a reverse-biased *pn* junction. Conduction band electrons in the source region encounter a potential barrier determined by the built-in potential of the source junction. As a result, electrons cannot enter the channel region and, hence, no current flows from the source to the drain. This is referred to as the *off* state. With a small positive bias on the gate, band bending in the channel region ($\psi_s > 0$) brings the conduction band in the channel region closer to the conduction band in the source region, thus reducing the height of the potential barrier to electrons. Electrons can now enter the channel and a current flow from source to drain is established.

In the low-drain-bias regime, the drain current increases almost linearly with drain bias. Indeed, here the channel resembles an ideal resistor obeying Ohm's law. The channel resistance is determined by the electron concentration in the channel, which is a function of the gate bias. Therefore, the channel acts like a voltage-controlled resistor whose resistance is determined by the applied gate bias. As shown in Figure 3.9, as the gate bias is increased, the slope of the *I–V* characteristic gradually increases due to the increasing conductivity of the channel. We obtain different slopes for different gate biases. This region where the channel behaves like a resistor is referred to as the **linear region** of operation. The drain current in the linear regime is given by:

$$I_{D,lin} = \frac{W}{L}\mu C'_{ox}\left[(V_{GS} - V_T)V_{DS} - \frac{\alpha}{2}V_{DS}^2\right], \tag{3.18}$$

where V_T is the threshold voltage, C'_{ox} is the gate capacitance per unit area, α is a constant, and μ is the effective channel mobility (which differs from bulk mobility). We shall deal with the concept of effective channel mobility later in this chapter.

Threshold voltage in the above equation is defined as:

$$V_T = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{SB}}. \tag{3.19}$$

For small V_{DS}, the second term in the parenthesis can be ignored, and the expression for drain current reduces to:

$$I_{D,lin} = \frac{W}{L}\mu C'_{ox}(V_{GS} - V_T)V_{DS}, \tag{3.20}$$

which is a straight line with a slope equal to the channel conductance:

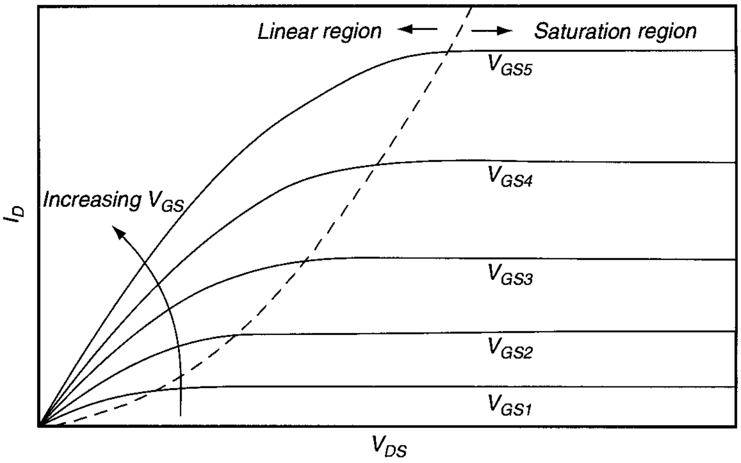

FIGURE 3.9

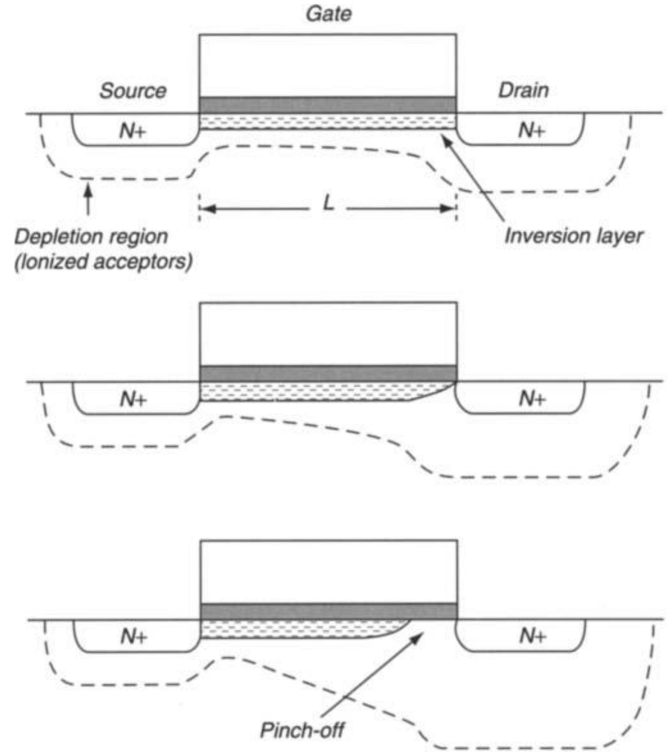

FIGURE 3.10

$$\sigma_c = \frac{W}{L} \mu C'_{ox} (V_{GS} - V_T). \tag{3.21}$$

Saturation Region

For larger drain biases, the drain current saturates and becomes independent of the drain bias. Naturally, this region is referred to as the **saturation region**. The drain current in saturation is derived from the linear region current shown in equation 3.18, which is a parabola with a maximum occurring at $V_{D,\,sat}$ given by:[3]

$$V_{D,\,sat} = \frac{V_{GS} - V_T}{\alpha}. \tag{3.22}$$

To obtain the drain current in saturation, this $V_{D,\,sat}$ value can be substituted in the linear region expression, which gives:

$$I_{D,\ sat} = \frac{W}{L} \mu C'_{ox} \frac{(V_{GS} - V_T)^2}{2\alpha}. \tag{3.23}$$

As V_{DS} increases, the number of electrons in the inversion layer decreases near the drain. This occurs due to two reasons. First, because both the gate and the drain are positively biased, the potential difference across the oxide is smaller near the drain end. Because the positive charge on the gate is determined by the potential drop across the gate oxide, the gate charge is smaller near the drain end. This implies that the amount of negative charge in the semiconductor needed to preserve charge neutrality will also be smaller near the drain. Consequently, the electron concentration in the inversion layer drops. Second, increasing the voltage on the drain increases the depletion width around the reverse-biased drain junction. Since more negative acceptor ions are uncovered, a fewer number of inversion layer electrons are needed to balance the gate charge. This implies that the electron density in the inversion layer near the drain would decrease even if the charge density on the gate was constant. The reduced number of carriers causes a reduction in the channel conductance, which is reflected in the smaller slope of $I_{DS} - V_{DS}$ characteristics as V_{DS} approaches $V_{D,\,sat}$, and the MOSFET enters the saturation region. Eventually, the inversion layer completely

[3] The V_{DS} at which the linear drain current parabola reaches its maximum can be found by setting $\partial I_{DS}/\partial V_{DS}$ equal to zero.

disappears near the drain. This condition is called **pinch-off**, and the channel conductance becomes zero. As shown in Figure 3.9, $V_{D,sat}$ increases with gate bias. This results because a larger gate bias requires a larger drain bias to reduce the voltage drop across the oxide near the drain end. As given in equation 3.22, $V_{D,sat}$ increases linearly with V_{GS}.

As V_{DS} is increased beyond $V_{D,sat}$, the width of the pinch-off region increases. However, the voltage that drops across the inversion layer remains constant and equal to $V_{D,sat}$. The portion of the drain bias in excess of $V_{D,sat}$ appears across the pinch-off region. In a long channel MOSFET, the width of the pinch-off region is assumed small relative to the length of the channel. Thus, neither the length nor the voltage across the inversion layer change beyond the pinch-off, resulting in a drain current independent of drain bias. Consequently, the drain current saturates. In smaller devices, this assumption falls apart and leads to **channel length modulation**, which is discussed later in the chapter.

From the above discussion, it is also evident that the electron distribution is highest near the source and lowest near the drain. To keep a constant current throughout the channel, the electrons travel slower near the source and speed up near the drain. In fact, in the pinch-off region, the electron density is negligibly small. Therefore, in this region, to maintain the same current level, the electrons have to travel at much higher speeds to transport the same magnitude of charge.

An important figure of merit for MOSFETs is transconductance, g_m, in the saturation regime, which is defined as:

$$g_m = \frac{\partial I_{D,sat}}{\partial V_{GS}} = \frac{W}{L}\frac{\mu C'_{ox}}{\alpha}(V_{GS} - V_T). \qquad (3.24)$$

Transconductance is a measure of the responsiveness of the drain current to variations in gate bias.

Subthreshold Region: MOSFET in Weak Inversion

When the surface potential at the source end is sufficient to form an inversion layer but the band bending is less than what is needed to reach strong inversion (i.e., $\phi_F < \Psi_s < 2\phi_F$), the MOSFET is said to operate in weak inversion. This region of operation is commonly called the **subthreshold region** and plays an important role in determining switching characteristics of logic circuits. When an n-channel MOSFET is in weak inversion, the drain current is determined by diffusion of electrons from the source to the drain. This is because the drift current is negligibly small due to the low lateral electric field and small electron concentration in weak inversion.

Even though the electron concentration in weak inversion is small, it increases exponentially with gate bias. Consequently, the drain current in weak inversion also rises exponentially, and it can be expressed as:

$$I_{DS} = \frac{W}{L} I' e^{q(V_{GS}-V_T)/nkT}\left(1 - e^{-q^{V_{DS}/kT}}\right), \qquad (3.25)$$

where I' and n are constants defined as:

$$I' = \mu \frac{\sqrt{2q\varepsilon_S N_A}}{2\sqrt{2\phi_F + V_{SB}}}\phi^2 \qquad (3.26)$$

$$n = 1 + \frac{\gamma}{2\sqrt{2\phi_F + V_{SB}}}. \qquad (3.27)$$

When the drain bias is larger than a few kT/q, the dependence on the drain bias can be neglected, and the above equation reduces to:

$$I_{DS} = \frac{W}{L} I' e^{q^{(V_{GS}-V_T)/nkT}}, \qquad (3.28)$$

which yields an exponential dependence on gate bias. When $\log I_{DS}$ is plotted against gate bias, we obtain:

$$\log I_{DS} = \log\left(\frac{W}{L}I'\right) + \frac{q}{kT}\frac{V_{GS} - V_T}{n}, \qquad (3.29)$$

which is a straight line with a slope:

$$\frac{1}{S} = \frac{1}{n}\frac{q}{kT}. \qquad (3.30)$$

The parameter S in the above equation is the MOSFET **subthreshold swing**, which is one of the most critical performance figures of MOSFETs in logic applications. It is highly desirable to have a subthreshold swing as small as possible since this is the parameter that determines the amount of voltage swing necessary to switch a MOSFET from its *off* state to its *on* state. This is especially important for modern MOSFETs with supply voltages approaching 1.0 V.

Figure 3.11 shows typical subthreshold characteristics of a MOSFET. As predicted by the above model, $\log I_{DS}$ increases linearly with gate bias up to V_T. In strong inversion, the subthreshold model is no longer valid, and either equation 20 or 23 must be used depending on the drain bias. When $V_{GS} = V_T$, $\log I_{DS}$ deviates from linearity. In practice, this is a commonly used method to measure the threshold voltage. In terms of device parameters, subthreshold swing can be expressed as:

$$S \cong \frac{kT}{q}\left(1 + \frac{C'_{dep} + C'_{it}}{C'_{ox}}\right)\ln(10), \qquad (3.31)$$

where C'_{dep} is depletion region capacitance per unit area of the MOS gate determined by the doping density in the channel region, and C'_{it} is the interface trap capacitance. Lower

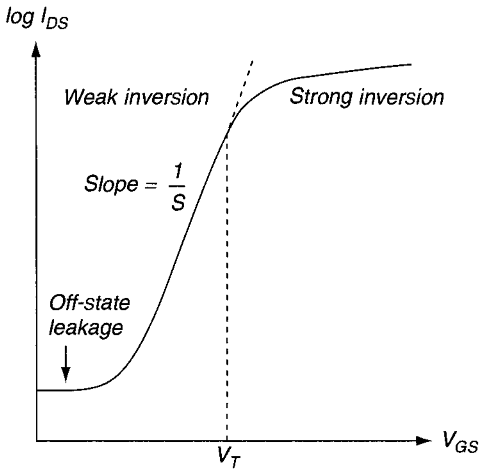

FIGURE 3.11

channel doping densities yield wider depletion region widths and hence smaller values for C'_{dep}. Another critical parameter is the gate oxide thickness, which determines C'_{ox}. To minimize S, the thinnest possible oxide must be used. The most widely used unit for S is mV/decade. Typical values range from 60 to 100 mV/decade.

The lower limit of the subthreshold current is determined by the leakage current that flows from source to drain when the transistor is off. This current, which is referred to as the **off-state leakage**, is determined by the source-to-channel potential barrier as well as the leakage currents of the source and drain junctions. For this reason, it is critical to be able to form low leakage junctions.

3.3.2 Effective Channel Mobility

The mobility term used in the previous equations differs from the bulk mobility due to additional scattering mechanisms associated with the surface, including interface charge scattering, coulombic scattering, and surface roughness scattering. These mechanisms tend to lower the carrier mobility in the channel. These mechanisms also tend to exhibit a dependence on the vertical electric field, E_y. At low fields, interface and coulombic scattering are the dominant mechanisms. At high fields, the surface roughness scattering dominates. Since the vertical field varies along the channel, mobility also varies. The dependence on the vertical field is generally expressed as:

$$\mu = \frac{\mu_o}{1 + \alpha E_y}, \qquad (3.32)$$

where μ_o is roughly half of the bulk mobility and α is roughly 0.025 μm/V at room temperature. This mobility equation is not useful in device equations because E_y varies along the channel. To be able to use the standard equations, the field dependence is lumped into a constant termed the **effective channel mobility**, μ_{eff}. A simple yet intuitive model for μ_{eff} *is* as follows:

$$\mu_{eff} \approx \frac{\mu_o}{1 + \theta(V_{GS} - V_T)}, \qquad (3.33)$$

where θ is a constant inversely proportional to the gate oxide thickness, t_{ox}. A typical plot of effective mobility as a function of gate bias is shown in Figure 3.12 (A). The effect of mobility degradation on $I_{DS} - V_{GS}$ characteristics is shown in Figure 3.12(B).

3.3.3 Nonuniform Channels

Typical channels in today's MOSFETs are nonuniform, which consist of a higher doped region placed underneath the Si–SiO$_2$ interface. This is done primarily to optimize the threshold voltage value while keeping the substrate concentration low. The new threshold voltage can be expressed as:

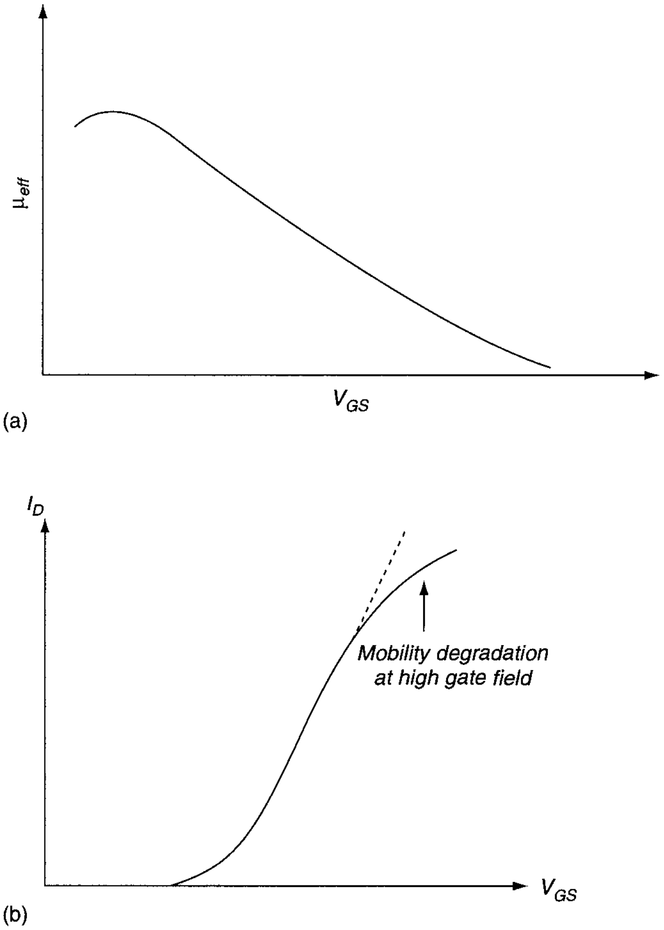

FIGURE 3.12

$$V_T = V'_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_B}, \tag{3.34}$$

where, V'_{FB} is given by:

$$V'_{FB} = V_{FB} + \frac{qM}{C_{ox}}. \tag{3.35}$$

The variable M is the implant dose. This equation assumes that the threshold adjust implant is very shallow and behaves as a sheet of fixed charge located under the gate oxide. In some cases, additional doped layers are placed in the channel to suppress punchthrough, a term that will be discussed later.

3.3.4 Short Channel Effects

MOSFETs are continually downscaled for higher packing density, higher device speed, and lower power consumption. The scaling methods are covered later in this chapter in a dedicated subsection. When physical dimensions of MOSFETs are reduced, the equations for drain current have to be modified to account for the so-called **short channel effects**. The three primary short channel effects included in this chapter are the following:

1. *Velocity saturation*: Limits the maximum carrier velocity in the channel to the saturation velocity in Si
2. *Channel length modulation*: Causes the drain current to increase with drain bias in the saturation region
3. *Drain-induced barrier lowering (DIBL)*: Causes the threshold voltage to change from its long channel value with dependence on device geometry as well as drain bias

Velocity Saturation

As discussed earlier, in long channel MOSFETs, the drain current saturates for V_{DS} larger than $V_{D,sat}$. The potential drop across the inversion layer remains at $V_{D,sat}$, and the horizontal electric field, E_x, along the channel is fixed at:

$$E_x = \frac{V_{Dsat}}{L}. \tag{3.36}$$

The electron drift velocity as a function of applied field is shown in Figure 3.13. At low fields, the drift velocity is given by:

$$v_d = \mu E_x, \tag{3.37}$$

where μ is generally referred to as the low-field mobility. At high fields, the velocity saturates due to phonon scattering. In a short channel device, the electric field across the channel can become sufficiently high such that carriers can suffer from velocity saturation. The effect of velocity saturation on MOSFET drain current can be severe. In short channel MOSFETs, it is impossible to overcome this effect. Thus, it is safe to say that all devices used in modern logic circuits suffer from velocity saturation to some extent.

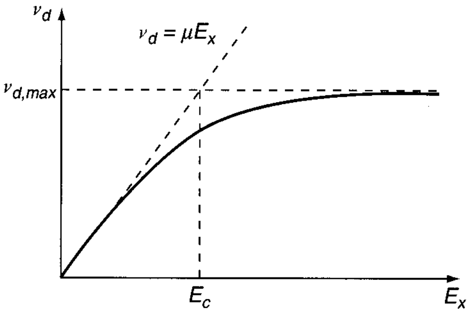

FIGURE 3.13

To account for this phenomenon, I_{DS} has to be obtained using velocity dependence on the electric field. A simple yet intuitive model for I_{DS} contains the original long channel drain current expression in the linear regime (equation 18) divided by a factor that accounts for velocity saturation. This formula is written as:

$$I_{DS}(\text{with velocity saturation}) = \frac{I_{DS}(\text{without velocity saturation})}{1 + V_{DS}/LE_c} \tag{3.38}$$

where E_c is the critical field above which velocity saturation occurs (Figure 3.13) and is given by:

$$E_c = \frac{v_{sat}}{\mu}. \tag{3.39}$$

The $V_{D,sat}$ value obtained using the modified current is smaller than the long channel $V_{D,sat}$ value. The drain current in saturation can be obtained from the linear region expression by assuming the drain current saturates for $V_{DS} > V_{D,sat}$:

$$I_{D,sat} \approx W\mu C_{ox}(V_{GS} - V_T)E_C. \tag{3.40}$$

It is important to note that with velocity saturation, the dependence of drain current on V_{GS} is linear instead of square, as it is the case for the long channel MOSFET.

Channel Length Modulation

After pinch-off occurs at the drain end, the length of the inversion layer and, hence, the channel resistance continually decrease as the drain bias is raised above $V_{D,sat}$. In long channel devices, the length of the pinch-off region is negligibly small. Since the voltage drop across the channel is pinned at $V_{D,sat}$, drain current increases only negligibly. However, in short channel devices, this pinch-off region can become a significant fraction of the total channel length. This reduction of channel length manifests itself as a finite slope in the

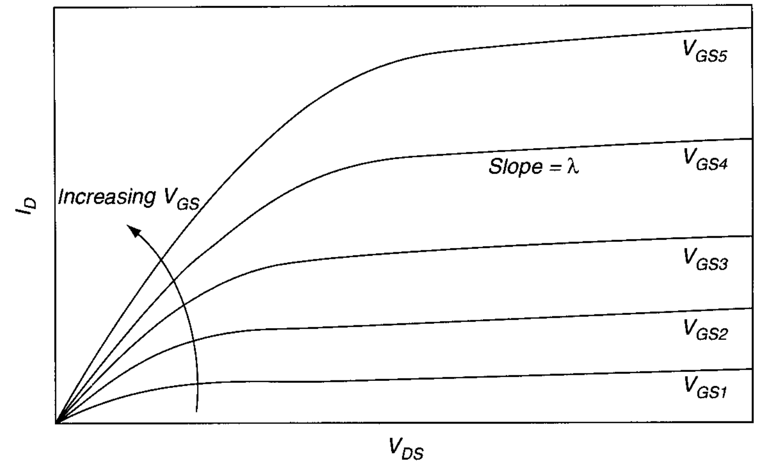

FIGURE 3.14

$I_{DS} - V_{DS}$ characteristics beyond $V_{D,sat}$ as shown in Figure 3.14. A commonly used expression for MOSFET drain current in saturation with channel length modulation is the following:

$$I_{D,sat} = \frac{W}{L}\mu C'_{ox}\frac{(V_{GS} - V_T)^2}{2\alpha}(1 + \lambda V_{DS}), \tag{3.41}$$

where λ is known as the channel length modulation factor expressed as:

$$\lambda = \frac{1}{E_o L}, \tag{3.42}$$

and E_o is the magnitude of the electric field at the pinch-off point. If both velocity saturation and channel length modulation are taken into account, the saturation drain current can be written as:

$$I_{D,sat} \approx W\mu C_{ox}(V_{GS} - V_T)E_C(1 + \lambda V_{DS}) \tag{3.43}$$

Drain-Induced Barrier Lowering

Another major short channel effect deals with the reduction of the threshold voltage as the channel length is reduced. In long channel devices, the influence of source and drain on the channel depletion layer is negligible. However, as channel lengths are reduced, overlapping source and drain depletion regions start having a large effect on the channel depletion region. This causes the depletion region under the inversion layer to increase. The wider depletion region is accompanied by a larger surface potential, which makes the channel more attractive to electrons. Therefore, a smaller amount of charge on the gate is needed to reach the onset of strong inversion, and the threshold voltage decreases. This effect is worsened when there is a larger bias on the drain since the depletion region becomes even wider. This phenomenon is called drain-induced barrier lowering (DIBL). Figure 3.15 shows the variation of surface potential from source to drain for a short channel and a long channel MOSFET. As shown, for the long channel device, the potential is nearly constant throughout the channel and is determined only by the gate bias.[4] The height of the potential barrier for electrons at the source end is given by:

$$\phi_{BL} \approx 2\phi_F. \tag{3.44}$$

For the smaller device, the surface potential is larger due to additional band bending. In addition, the surface potential is

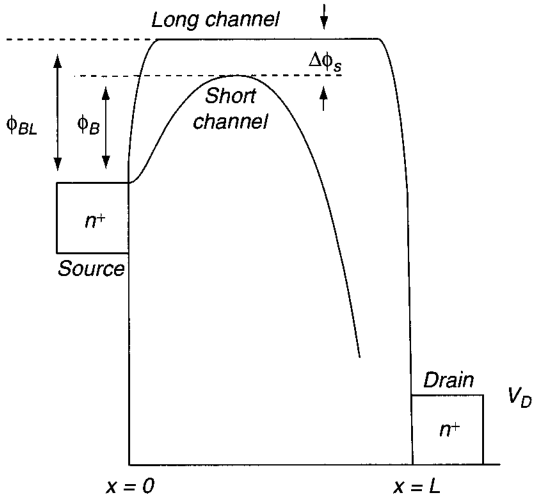

FIGURE 3.15

[4] The channel potential actually increases gradually toward the drain due to small yet finite resistance of the inversion layer. When pinch-off occurs at the drain end, majority of the drain-to-source bias appears across the pinch-off region due to its much larger resistivity.

no longer constant throughout the channel. With a larger surface potential, the barrier for electrons between the source and the channel is reduced and is given by:

$$\phi_B = \phi_{BL} - \Delta\phi_B. \tag{3.45}$$

The change in threshold voltage due to DIBL has been modeled Liu *et al.* (1993) as:

$$\Delta V_T \approx -[3(\phi_{bi} - 2\phi_F) + V_{DS}]e^{-L/\lambda_{\mathrm{DIBL}}}, \tag{3.46}$$

where ϕ_{bi} is the built-in potential of the drain junction and:

$$\lambda_{\mathrm{DIBL}} = \sqrt{\frac{\varepsilon Si^{t}_{ox} W_D}{\varepsilon_{ox}\beta}}. \tag{3.47}$$

In the above equation, W_D is the depletion region width under the inversion layer, and β is a fitting parameter. The model indicates that ΔV_T is a strong function of the λ_{DIBL} term in the exponential and must be minimized. This requires higher doping density in the channel to reduce W_D and downscaling of the gate oxide thickness, t_{ox}. The model does not include the well-known dependence of ΔV_T on source/drain junction depths although some modifications have been suggested.

The impact of DIBL on V_T can be measured by plotting the subthreshold characteristics for increasing values of V_{DS} as shown in Figure 3.16. The subthreshold slope degrades and the impact of drain bias on ΔV_T increases as the channel length is reduced. This methodology is valid only until the subthreshold slope remains intact. A very large subthreshold swing implies that the device cannot be turned off. This phenomenon is called **punchthrough** and occurs roughly when the source and drain depletion regions meet. Since the depletion region becomes wider at larger drain biases, the onset of punchthrough is reached sooner. Punchthrough can occur either at or below the surface depending on the doping profile in the channel region. Surface punchthrough occurs for uniformly doped substrates and results in the loss of gate control of the channel region, causing the device to fail. Bulk punchthrough occurs for ion-implanted channels that have a higher doping concentration at the surface. In this condition, there still exists a channel that is gate controlled, but the background leakage becomes very high and is strongly dependent on V_{DS}. The common solution is to add a shallow punchthrough implant typically placed immediately below the threshold adjust implant. This raises the doping density under the channel while keeping the bulk doping density as low as possible. The goal is to minimize junction depletion region capacitance, which has an impact on device switching speed.

Similarly, for a fixed drain bias, a larger body bias increases the reverse bias on the drain junction and, thus, has the same effect as increasing the drain bias. Lowering of the threshold voltage due to DIBL is demonstrated in Figure 3.17. Note that the effect of body bias on threshold voltage for large channel MOSFETs is determined by equation 3.15. Here we are interested in lowering of the threshold voltage due to body bias only for short channel lengths.

In addition to the short channel effects described above, the scaling of the channel width can also have a large effect on the threshold voltage, and its nature depends on the isolation technology (Tsividis, 1999). In the case of LOCOS isolation, the threshold voltage increases as the channel width decreases. On the other hand, for shallow trench isolation, the threshold voltage decreases as the channel width decreases. In today's technologies, shallow trench isolation is the predominant isolation technique.

3.3.5 MOSFET Scaling

Scaling of MOSFETs is necessary to achieve (a) low power, (b) high speed, and (c) high packing density. Several scaling methodologies have been proposed and used, which are

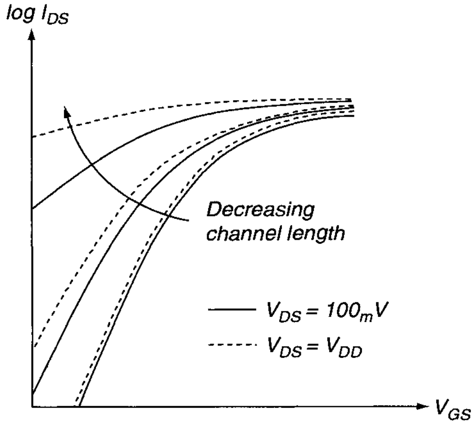

FIGURE 3.16

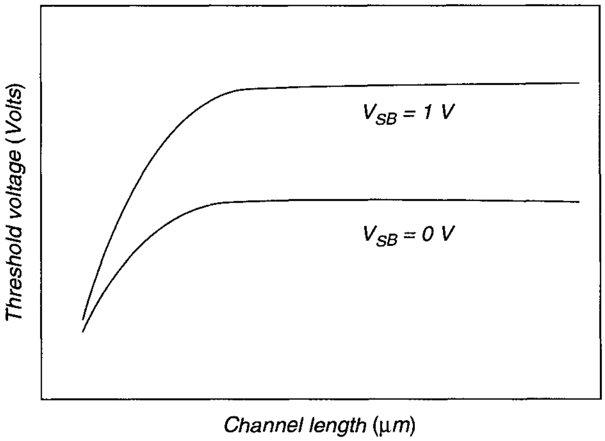

FIGURE 3.17

derivatives of constant field scaling. In constant-field scaling, the supply voltage and the device dimensions (both lateral and vertical) are scaled by the same factor, κ such that the electric field remains unchanged. This results in scaling of the drain current or current drive. At the same time, gate capacitance is also scaled due to reduced device size. This provides a reduction in the gate charging time, which directly translates into higher speed. Furthermore, power dissipation per transistor is also reduced.

Constant field scaling provides a good framework for CMOS scaling without degrading reliability. However, there are several parameters, such as the kT/q and an energy gap, that do not scale with reduced voltages or dimensions and present challenges in device design. Since the junction built-in potential, ϕ_{bi}, and the surface potential, ψ_s, are both determined by the bandgap, they do not scale either. Consequently, depletion region widths do not scale as much as other parameters, which results in worsened short channel effects.

Another parameter that does not scale easily is the threshold voltage. This sets the lower limit for the power-supply voltage since a safe margin between the two parameters is required for reliable device operation. Other parameters that do not scale well include the off-state current and subthreshold slope.

Although **constant field scaling** provides a reasonable guideline, scaling of the voltages by the same factor as the physical dimensions is not always practical. This is due to the inability to scale the subthreshold slope properly as well as standardize voltage levels of prior generations. This is the basis for **constant voltage scaling** and other modified scaling methodologies. One problem with constant voltage scaling is that the oxide field increases since the t_{ox} is also scaled by κ. To reduce this problem, the oxide thickness is reduced by κ' where $\kappa' < \kappa$. Modifications of the constant voltage and constant field scaling have also been tried to avoid high-field problems. For example, in **quasi-constant voltage scaling**, physical dimensions are scaled by κ, whereas the voltages are scaled by a different factor, κ'. Furthermore, since depletion layers do not scale proportionately, the doping concentration, N_A, must be increased by more than what is suggested by constant-field scaling. This is called **generalized scaling** and the doping scaling factor is greater than κ. The scaling rules discussed above are shown in Table 3.3.

It should be noted that there are other effects that can limit the performance of scaled devices. These include (a) reduction of the effective oxide capacitance due to finite thickness of the inversion layer, (b) depletion in the polysilicon gate, (c) quantum effects in the inversion layer that increase V_T, and (d) quantum mechanical tunneling of carriers through the thin oxide, which is a critical issue for $t_{ox} < 50$ A.

3.3.6 Modern MOSFETs

Today's MOSFET has evolved through many modifications to provide solutions for parasitic problems that emerge during scaling. A modern *n*-channel MOSFET structure is shown in Figure 3.18. The gate dielectric is SiO_2 and is grown by thermal oxidation. As a result of continued scaling, the gate oxide has thinned down to the extent that direct tunneling has become a serious concern. Currently, alternate high-κ dielectrics are being considered as replacements for SiO_2. The goal is to achieve higher capacitance without compromising gate leakage. The standard gate electrode is heavily doped polysilicon defined by anisotropic reactive ion etching. Extension junctions are formed by ion-implantation using the polysilicon gate as an implant mask. Because a separate masking step is not needed to define the junction regions, the approach is named **self-aligned polysilicon gate technology**. At the same time, the junction sheet resistance must be sufficiently low not to increase the device series resistance. Extension junctions must be as shallow as possible to reduce DIBL. When formed by ion-implantation, shallow junctions require low doses and low energies. Unfortunately, low dose results in higher series resistance. Sidewall spacers are formed by deposition of a dielectric (SiO_2 or Si_3N_4) followed by anisotropic reactive ion etching. The spacers must be as thin as possible to minimize the series resistance contribution of the extension junctions. Deep source and drain junctions are formed by ion-implantation following spacer formation. Because they are separated from the channel by the spacers, their contribution to DIBL is negligible. These junctions must be sufficiently deep to allow formation of good quality silicide contacts. Both Ti and Co silicides are used as source/drain contact materials. Silicides are formed by an approach referred to as **self-aligned-silicide** (SALICIDE), which selectively forms the silicide on the junctions as well as polysilicon. Silicide formation consumes the Si substrate in the deep source/drain regions and can lead to excessive junction leakage. To avoid this, the deep source/ drain regions are required to be about 50 nm deeper than the

TABLE 3.3 Scaling Rules for Four Different Methodologies

Parameters	Constant field scaling $1 < \kappa' < \kappa$	Constant voltage scaling $1 < \kappa' < \kappa$	Quasi-constant voltage scaling $1 < \kappa' < \kappa$	Generalized scaling $1 < \kappa' < \kappa$
W, L	$1/\kappa$	$1/\kappa$	$1/\kappa$	$1/\kappa$
t_{ox}	$1/\kappa$	$1/\kappa'$	$1/\kappa$	$1/\kappa$
N_A	κ	κ	κ	κ^2/κ'
$Vdd,\ V_T$	$1/\kappa$	1	$1/\kappa'$	$1/\kappa'$

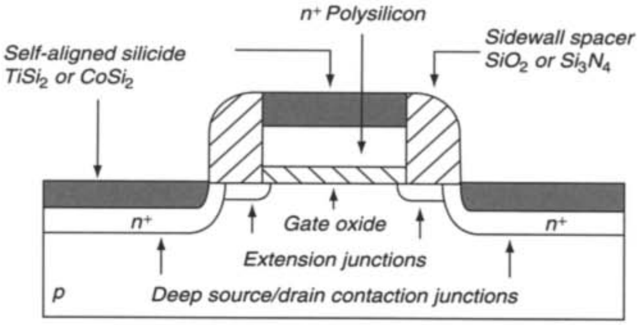

FIGURE 3.18

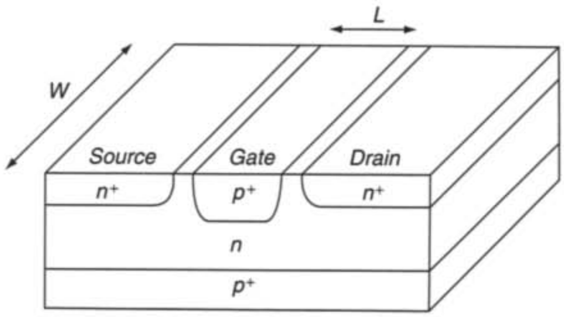

FIGURE 3.19

amount consumed. An important advantage of the SALICIDE process is that the entire junction area up to the sidewall spacer is used for contact formation, which translates into low contact resistance. The channel doping profile typically consists of the threshold adjust and punchthrough stop implants. The gate electrode is very heavily doped to avoid polysilicon depletion.

The MOSFET continues to maintain its dominance in today's digital ICs. To ensure continued benefits from MOSFET scaling, several structural and material changes are required. These include new gate dielectric materials to reach equivalent SiO_2 thicknesses less than 1 nm, new doping technologies that provide ultra-low resistance and ultra-shallow junctions, as well as metal gate electrodes that do not suffer from depletion effects.

3.4 Junction Field Effect Transistor

The junction field effect transistor (JFET) was first proposed by William Shockley in 1952. The first demonstration was made the following year by Dacey and Ross. Due to the large popularity of the bipolar junction transistor (BJT) at the time, major advancements in JFET fabrication did not occur until the 1970s. With the introduction of MOSFET, the use of JFET remained limited to specific applications.

Figure 3.19 shows a simplified schematic of the JFET. The device consists of an n-type channel region sandwiched between two p^+n junctions that serve as the gate. The p^+ regions can be tied together to form a dual-gate JFET. The heavily doped $n+$ source and drain regions serve as low resistivity ohmic contacts to the n-type channel. As such, a conductive path exists between the source and the drain contacts. The device is turned off by depleting the channel off the free carriers.

The $I_{DS} - V_{DS}$ characteristics of a JFET are very similar to those of a MOSFET. Figure 3.20 illustrates the operation of a JFET in different regions. The depletion region widths of the two p^+n junctions are determined by the gate bias and the potential variation along the channel. The widths of the depletion regions at the source end are determined mainly by the built-in potential of the junction. The depletion regions are wider at the drain end due to the large positive bias applied to the drain terminal. Figure 3.20(A) corresponds to the linear region of the JFET. Even though the channel depth is reduced at the drain end, the n-type channel extends from the source to the drain. The channel resembles a resistor, and the current is a linear function of the drain bias. The depth and, hence, the resistance of the channel is modulated by the gate bias. The drain current in the linear region can be expressed as:

$$I_{D,lin} = \frac{G_o}{2V_P}(V_G - V_T)V_D, \tag{3.48}$$

where V_P is the pinch-off voltage corresponding to the potential drop across the gate junction when the two depletion

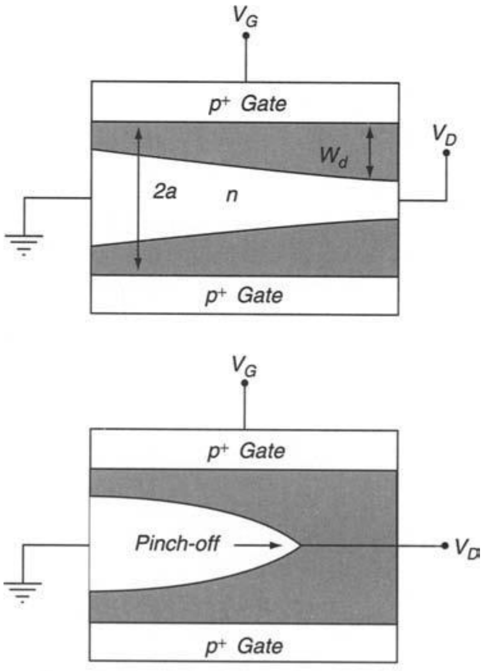

FIGURE 3.20

regions meet at the drain end. The variable G_o is the full channel conductance with a channel depth equal to $2a$. These can be expressed as:

$$G_o = 2aq\mu_n N_D \frac{W}{L} \tag{3.49}$$

and

$$V_P = \frac{qN_D a^2}{2\varepsilon_s}. \tag{3.50}$$

When the two depletion regions meet at the drain end, the JFET enters the saturation region. As shown in Figure 3.20(B), the pinch-off point moves toward the source as the drain bias is raised. The potential at the pinch-off point, however, remains pinned at:

$$V_{D,sat} = V_P - \psi_{bi} + V_G = V_G - V_T. \tag{3.51}$$

The built-in potential ψ_{bi} is a function of the doping concentrations in both sides of the junction. For large devices, the reduction in channel length due to pinch-off can be negligibly small. Hence the channel conductance remains approximately the same with a fixed voltage equal to $V_{D,sat}$ across the conductive channel. The drain current no longer increases with drain bias, and JFET is said to operate in the saturation region. The drain current after pinch-off can be expressed as:

$$I_{D,sat} = \frac{G_o}{4V_P}\left(1 + \frac{V_G}{V_P}\right)^2. \tag{3.52}$$

In JFETs with short channel lengths, the electric field along the channel can be large enough to cause carrier velocity saturation even before pinch-off. This requires a drain bias of:

$$V_D = E_c L, \tag{3.53}$$

where E_c is the critical field for velocity saturation.

3.5 Metal-Semiconductor Field Effect Transistor

A MESFET is very similar to a metal-semiconductor field effect transistor (MOSFET). The main difference is that in a MESFET, the MOS gate is replaced by a metal-semiconductor (Schottky) junction. A self-aligned GaAs MESFET structure is shown in Figure 3.21. The heavily doped source and drain junctions are formed in an n-type epitaxial layer formed on semi-insulating GaAs, which provides low parasitic capacitance. The n-type channel region has a thickness, a, which is typically less than 200 nm. The source and drain junctions are formed by ion-implantation followed by annealing. Common source and drain contact materials are AuGe alloys. Popular Schottky gate metals include Al, Ti-Pt-Au-layered structure, Pt, W, and WSi_2. However, in a self-aligned MESFET, since the source and drain implantation and implant annealing must be performed after gate formation, refractory metals that can withstand the annealing temperatures are preferred.

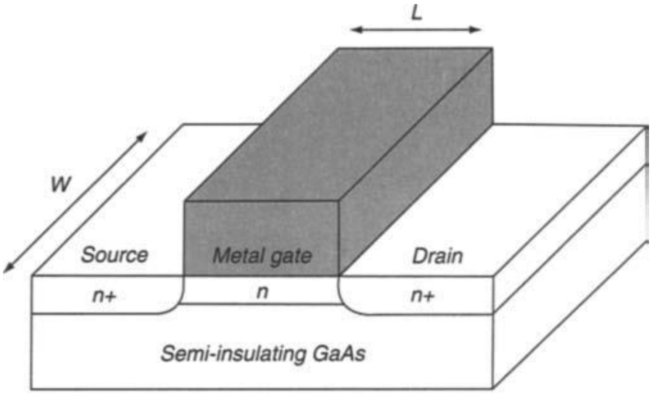

FIGURE 3.21

MESFET is better suited to materials that do not have a good dielectric to form a high-quality MOS gate. Today, MESFETs are commonly fabricated on compound semiconductors, predominantly GaAs. Because of high mobility of carriers in GaAs and low capacitance due to the semi-insulating GaAs, MESFETs have the speed advantage over MOSFETs. They are used in microwave applications, which require high-speed devices. Application areas include communication, high-speed computer, and military systems.

Although the MESFET structure resembles the MOSFET, its operation is much closer to that of the JFET. Figure 22 shows the MESFET cross-section in different regions of operation. The MESFET has a conductive path between the source and the drain since the channel is of the same conductivity type as the junctions. Therefore, to turn-off a MESFET, a sufficiently high gate bias must be applied to deplete the channel. The depletion region under the gate modulates the width and the resistance of the conductive channel from source to drain. As such, in a MESFET, the Schottky gate plays the exact same role that the pn junction gate plays in a JFET.

The width of the depletion region is wider at the drain end since the source is grounded and a finite bias is applied to the drain. At small drain biases, the channel is conductive from source to drain. However, the width of the channel is smaller at the drain end due to the wider depletion region. The depletion region width under the gate is given by:

$$W_d(x) = \sqrt{\frac{2\varepsilon_s[\psi_{bi} + \psi(x) - V_G]}{qN_D}}, \tag{3.54}$$

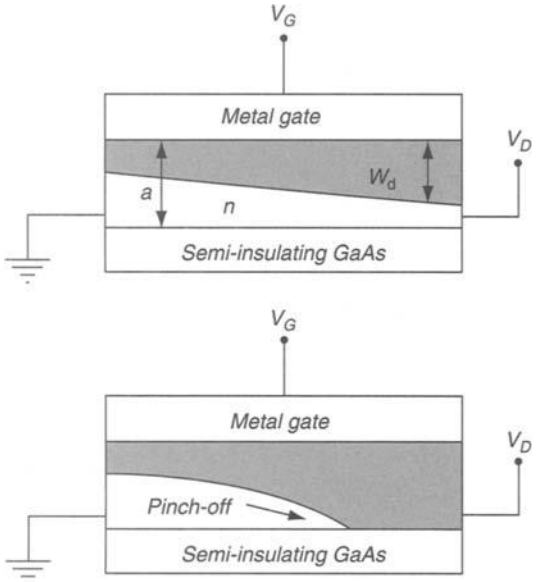

FIGURE 3.22

where ψ_{bi} is the band bending in the semiconductor due to the metal-semiconductor work function difference, and $\psi(x)$ is the channel potential with respect to the source. In this regime, the MESFET channel resembles a voltage, controlled resistor. The drain current varies linearly with the drain bias. The gate bias changes the width of the depletion region to modulate the width and the resistance of the channel. This is the linear region of operation for a MESFET.

For a long channel MESFET ($L \gg a$), the drain current in the linear region is approximately equal to:

$$I_{D,\,lin} \approx \frac{G_i}{2V_P}(V_G - V_T)V_D, \tag{3.55}$$

where

$$G_i = \frac{W}{L} q\mu_n N_D a \tag{3.56}$$

is the channel conductance without depletion under the Schottky gate. The variable V_P in equation 3.55 is called the pinch-off voltage, which is the net potential across the gate junction when the depletion region width under the gate is exactly equal to the channel depth, a. The pinch-off voltage is given by:[5]

$$V_P = \frac{qN_D a^2}{2\varepsilon_s}. \tag{3.57}$$

Finally, V_T is the threshold voltage, and it is given by:

$$V_T = \psi_{bi} - V_P$$

In saturation, the drain current is given by:

$$I_{D,\,sat} \approx \frac{G_i}{4V_P}(V_G - V_T)^2, \tag{3.59}$$

yielding a transconductance of:

$$g_{m,\,sat} = G_i\left(1 - \sqrt{\frac{\psi_{bi} - V_G}{V_P}}\right). \tag{3.60}$$

When pinch-off occurs, the voltage at the pinch-off point is pinned at V_P even when the channel length is continually reduced at higher drain biases. In long channel MESFETs, the length of the pinch-off region is negligibly small and can be ignored. This results in saturation of the drain current for voltages beyond V_P. In smaller devices, however, the length and the resistance of the channel decreases as the pinch-off region becomes wider. Similar to channel length modulation in MOSFETs, this results in a gradual increase in drain current with applied drain bias.

In short channel devices, the horizontal field in the channel is high enough to reach the velocity saturation regime, which degrades the drain current as well as the transconductance. Velocity saturation can be reached even before pinch-off. With velocity saturation, the drain current in the saturation region can be expressed as:

$$I_{D,\,sat} = qv_s WN_D a\left(1 - \sqrt{\frac{\psi_{bi} - V_G}{V_P}}\right), \tag{3.61}$$

which gives a transconductance of:

$$g_{m,\,sat} \approx \frac{qv_s WaN_D}{2\sqrt{V_P(\psi_{bi} - V_G)}}. \tag{3.62}$$

3.6 Modulation-Doped Field Effect Transistor

The modulation-doped field effect transistor (MODFET) is also known as the high-electron mobility transistor (HEMT). The device relies on the ability to form a high-quality heterojunction between a wide bandgap material lattice matched to a narrow bandgap material. The preferred material system is AlGaAs–GaAs; however, MODFETs have also been demonstrated using other material systems including $\mathrm{Si{-}Si_xGe_{1-x}}$. The device was developed in the

[5] This comes from the depletion region of equation 3.54 by replacing the potential term with V_P and the depletion region width with the channel depth.

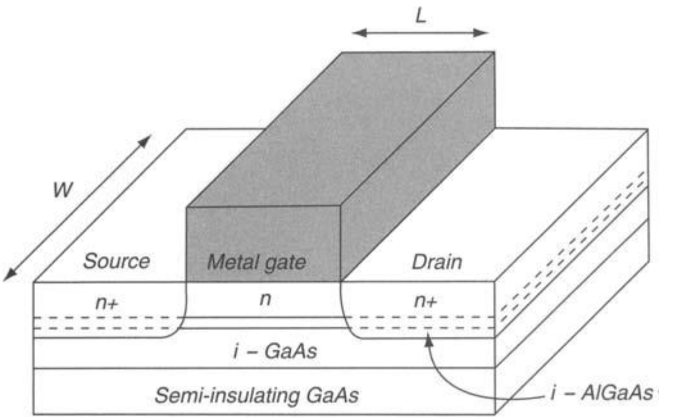

FIGURE 3.23

1970s. Enhanced mobility was first demonstrated by Dingle in 1978.

Figure 3.23 shows the simplified schematic of a self-aligned MODFET. The current conduction takes place in the undoped GaAs layer. The *n*-type AlGaAs layer located under the metal (Schottky barrier) gate is separated from the undoped GaAs by a thin undoped AlGaAs that acts as a buffer layer. A two-dimensional (2-D) electron gas is formed in GaAs immediately under the AlGaAs. High mobility results from the absence of ionized impurity scattering in the undoped layer. The thickness of the undoped AlGaAs buffer layer is critical. The buffer layer must be sufficiently thin to allow electrons to diffuse from the *n*-type AlGaAs into GaAs. At the same time, it must be sufficiently thick to place the 2-D electron gas sufficiently far away from the ionized impurities in the doped AlGaAs layer. The thickness of this layer is typically around 80 A.

The semiconductor layers are typically grown by molecular beam epitaxy (MBE); however, metal-organic chemical vapor deposition (MOCVD) has also shown to be feasible.

The energy band diagram at the onset of threshold is shown in Figure 3.24. The threshold is typically taken as the gate bias at which the conduction band of the GaAs layer coincides with the Fermi level. The 2-D electron gas forms at the heterointerface immediately under the undoped AlGaAs buffer layer and resembles the inversion layer that forms under the gate dielectric of a MOSFET.

The threshold voltage can be expressed as:

$$V_T = \phi_{bn} - V_P - \frac{\Delta E_C}{q}, \tag{3.63}$$

where V_P is the pinch-off voltage for the AlGaAs layer given by:

$$V_P = \frac{q_N D x_d^2}{2\varepsilon_s}. \tag{3.64}$$

The drain current in the linear regime is given by:

$$I_{D,\,lin} \cong \frac{\mu_n C_o W (V_G - V_T) V_D}{L}, \tag{3.65}$$

where

$$C_o = \frac{\varepsilon_S}{x_d + x_{ud} + \Delta d}. \tag{3.66}$$

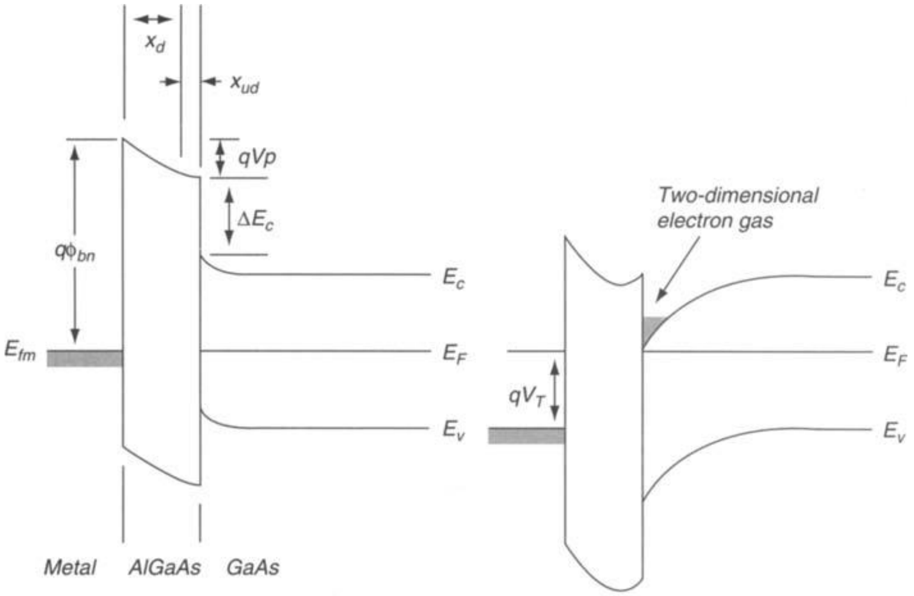

FIGURE 3.24

In the above equations, x_d and x_{ud} are the doped and undoped AlGaAs thicknesses, and Δd is the thickness of the 2-D electron gas that is typically less than 100 A. The threshold voltage can be made positive or negative by adjusting x_d. Hence, both enhancement and depletion mode devices are possible. The transconductance in the linear region is given by:

$$g_{m,\,lin} \equiv \frac{dI_{D,\,lin}}{dV_G} = \frac{\mu_n C_o W V_D}{L}. \tag{3.67}$$

When the drain bias is sufficiently large, electron concentration in the 2-D gas is reduced to zero at the drain end and the drain current saturates with V_D. This occurs at a drain bias of:

$$V_{D,\,sat} = V_G - V_T. \tag{3.68}$$

The drain current in saturation is given by:

$$I_{D,\,sat} = \frac{\mu_n C_o W}{2L}(V_G - V_T)^2, \tag{3.69}$$

yielding a transconductance of:

$$g_{m,\,sat} \equiv \frac{dI_{D,\,sat}}{dV_G} = \frac{\mu_n C_o W(V_G - V_T)}{L}. \tag{3.70}$$

In practical devices, the lateral electric field can be high enough that the carriers suffer from velocity saturation even before the drain current saturates. This is especially an issue for MODFETs due to high mobilities achieved in the 2-D electron gas since high mobility also implies low E_c. With velocity saturation, the drain current saturates at a drain bias of:

$$V_{D,\,sat} \approx E_c L. \tag{3.71}$$

The drain current is then given by:

$$I_{D,\,sat} = WC_o(V_G - V_T)v_s, \tag{3.72}$$

where v_s is the saturated drift velocity. This yields a transconductance of:

$$g_{m,\,sat} \equiv \frac{dI_{D,\,sat}}{dV_G} = WC_o v_s, \tag{3.73}$$

It is interesting to note that the saturation current is independent of channel length, and the transconductance is independent of both L and V_G. The device, however, benefits from a velocity overshoot that provides higher drive current.

References

Arora, N. (1983). *MOSFET models for VLSI circuits simulation: Theory and practice.* New York: Springer-Verlag Wien.

Grove, A.S. (1967) *Physics and technology of semiconductor devices.* New York: John Wiley & Sons.

Ng, K.K. (1995). *Complete guide to semiconductor devices.* New York: McGraw-Hill.

Tsividis, Y.P. (1999). *Operation and modeling of the MOS transistor.* New York: McGraw-Hill.

Z.H. Liu, et al., Threshold voltage model for deep submicrometer MOSFETs. *IEEE Transactions on Electron Devices 40,* 86–95.