# 6

# MOS Transistor

## CHAPTER OBJECTIVES

This chapter provides a comprehensive introduction to the modern MOSFETs in their on state. (The off state theory is the subject of the next chapter.) It covers the topics of surface mobility, body effect, a simple IV theory, and a more complete theory applicable to both long- and short-channel MOSFETs. It introduces the general concept of CMOS circuit speed and power consumption, voltage gain, high-frequency operation, and topics important to analog circuit designs such as voltage gain and noise. The chapter ends with discussions of DRAM, SRAM, and flash nonvolatile memory cells.

The **MOSFET** is by far the most prevalent semiconductor device in ICs. It is the basic building block of digital, analog, and memory circuits. Its small size allows the making of inexpensive and dense circuits such as giga-bit (Gb) memory chips. Its low power and high speed make possible chips for gigahertz (GHz) computer processors and radio-frequency (RF) cellular phones.

## 6.1 ● INTRODUCTION TO THE MOSFET ●

Figure 6–1 shows the basic structure of a MOSFET. The two PN junctions are the **source** and the **drain** that supplies the electrons or holes to the transistor and drains them away respectively. The name **field-effect transistor** or **FET** refers to the fact that the gate turns the transistor (inversion layer) on and off with an electric *field* through the oxide. A **transistor** is a device that presents a high input resistance to the signal source, drawing little input power, and a low resistance to the output circuit, capable of supplying a large current to drive the circuit load. The hatched regions in Fig. 6–1a are the **shallow-trench-isolation** oxide region. The silicon surfaces under the thick isolation oxide have very high threshold voltages and prevent current flows between the $N^+$ (and $P^+$) diffusion regions along inadvertent surface inversion paths in an IC chip.

Figure 6–1 also shows the MOSFET IV characteristics. Depending on the gate voltage, the MOSFET can be off (conducting only a very small **off-state leakage current, $I_{off}$**) or on (conducting a large **on-state current, $I_{on}$**).

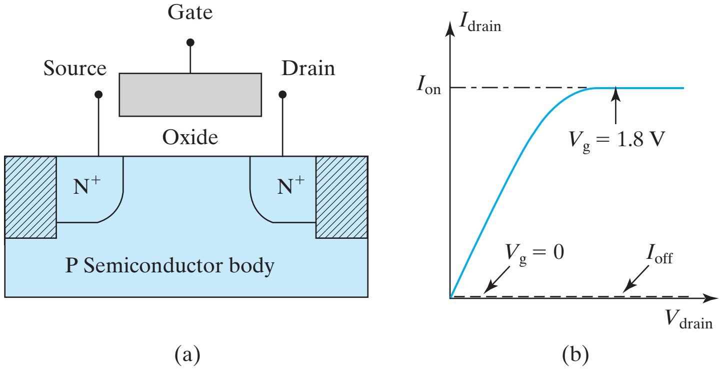


FIGURE 6–1 (a) Basic MOSFET structure and (b) IV characteristics.

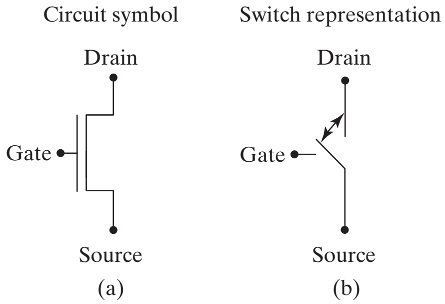


FIGURE 6–2 Two ways of representing a MOSFET: (a) a circuit symbol and (b) as an on/off switch.

At the most basic level, a MOSFET may be thought of as an on–off switch as shown in Fig. 6–2(b). The gate voltage determines whether a current flows between the drain and source or not. The circuit symbol shown in Fig. 6–2a connotes the much more complex characteristics of the MOSFET.

## ● Early Patents on the FET ●

The transistor and IC technologies owe their success mainly to the effort and ingenuity of a large number of technologists since the mid-1900s. Two early FET patents are excerpted here. These earliest patents are presented for historical interest only. Many more conceptual and engineering innovations and efforts were required to make MOSFETs what they are today.

J. E. Lilienfeld's 1930 U.S. patent is considered the first teaching of the FET. In Fig. 6–3, 10 is a glass substrate while 13 is the gate electrode (in today's terminology) and "consists of an … aluminum foil… ." 11 and 12 are metal contacts to the source and drain. 15 is a thin film of semiconductor (copper sulfide). Lilienfeld taught the following novel method of making a small (short) gate, the modern photolithography technique being yet unavailable to him. The glass substrate is broken into two pieces and then reassembled (glued back) with a thin aluminum foil inserted between the two pieces. The edge of the Al foil is used as the gate. The semiconductor film is deposited over the glass substrate and the gate, and source and drain contacts are provided. There is no oxide between the gate electrode and the semiconductor. The insulator in this FET would be the depletion layer at the metal–semiconductor junction (see Section 6.3.2).

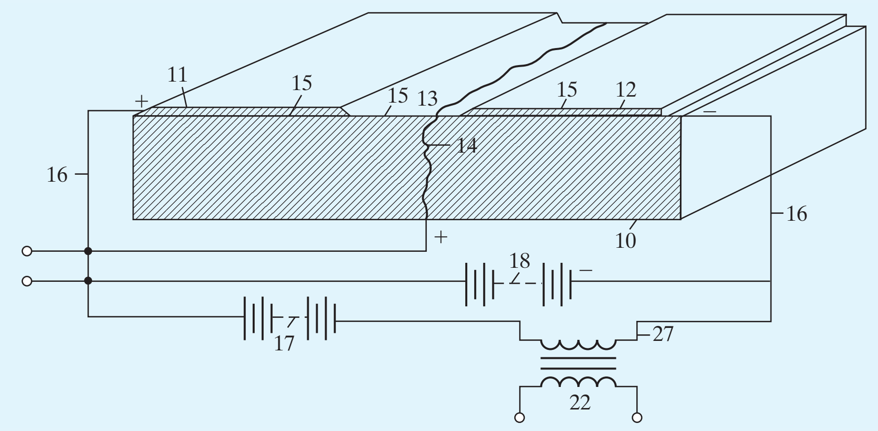


FIGURE 6–3 "A perspective view, on a greatly enlarged scale and partly in section, of the novel apparatus as embodied by way of example in an amplifier." (From [1].)

In a 1935 British patent, Oskar Heil gave a lucid description of a MOSFET. Referring to Fig. 6–4, "1 and 2 are metal electrodes between which is a thin layer 3 of semiconductor. A battery 4 sends a current through the thin layer of semiconductor and this current is measured by the ammeter 5. If, now, an electrode 6 in electro-static association with the layer 3 is charged positively or negatively in relation to the said layer 3, the electrical resistance of this layer is found to vary and the current strength as measured by the ammeter 5 also to vary."

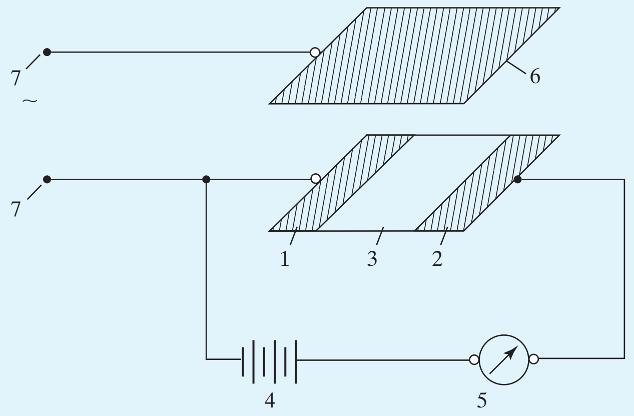


FIGURE 6–4 This 1935 drawing is a good illustration of a MOSFET even by today's standards. (From [2].)

## 6.2 • COMPLEMENTARY MOS (CMOS) TECHNOLOGY •

Modern MOSFET technology has advanced continually since its beginning in the 1950s. Figure 6–5 is a transmission electron microscope view of a part of a MOSFET. It shows the poly-Si gate and the single-crystalline Si body with visible individual Si atoms and a 1.2 nm amorphous $SiO_2$ film between them. 1.2 nm is the size of four $SiO_2$ molecules.

The basic steps of fabricating the MOSFET shown in Fig. 6–1 is to first make **shallow-trench-isolation** by etching a trench that defines the boundary of the transistor and filling the trench with chemical vapor deposition (CVD) oxide (see Section 3.7.2). Next, planarize the wafer with CMP (see Section 3.8), grow a thin layer of oxide (gate oxide) over the exposed silicon surface, deposit a layer of polycrystalline silicon as the gate material (Section 3.7.2), use optical lithography to pattern a piece of photoresist, and use the photoresist as a mask to etch the poly-Si to define the gate in Fig. 6–1 (Section 3.4). Finally, implant As into the source and drain (Section 3.5.1). The implantation is masked by the gate on one side and the trench isolation on the other. Rapid thermal annealing (see text box in Section 3.6) is applied to activate the dopant and repair the implantation damage to the crystal. Contacts can then be made to the source, drain, and the gate.

Figure 6–6a is an **N-channel MOSFET**, or **N-MOSFET** or simply **NFET**. It is called N-channel because the conduction channel (i.e., the inversion layer) is electron rich or N-type as shown in Fig. 6–6b. Figure 6–6c and d illustrate a **P-channel MOSFET**, or **P-MOSFET**, or **PFET**. In both cases, $V_g$ and $V_d$ swing between 0 V and $V_{dd}$, the power-supply voltage. The body of an NFET is connected to the lowest voltage in the circuit, 0 V, as shown in (b). Consequently, the PN junctions are always reverse-biased or unbiased and do not conduct forward diode current. When $V_g$ is equal to $V_{dd}$ as shown in (b), an inversion layer is present and the

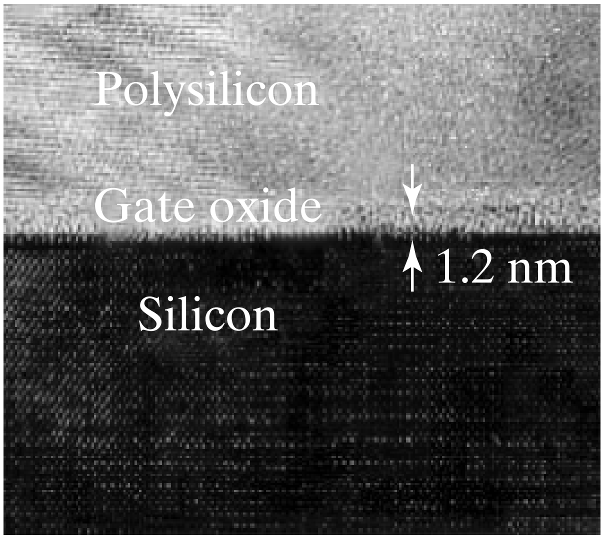


FIGURE 6–5 Gate oxides as thin as 1.2 nm can be manufactured reproducibly. Individual Si atoms are visible in the substrate and in the polycrystalline gate. (From [3]. © 1999 IEEE.)

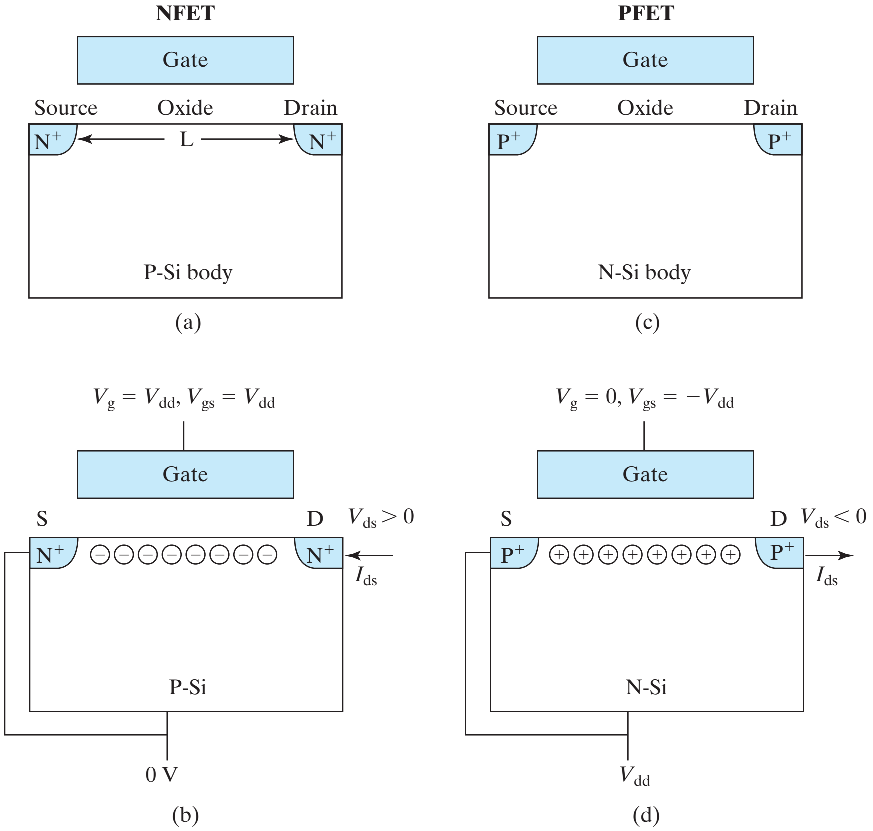


FIGURE 6–6 Schematic drawing of an N-channel MOSFET in the off state (a) and the on state (b). (c) and (d) show a P-channel MOSFET in the off and the on states.

NFET is turned on. With its body and source connected to $V_{dd}$, the PFET shown in (d) responds to $V_g$ in exactly the opposite manner. *When $V_g = V_{dd}$, the NFET is on and the PFET is off. When $V_g = 0$, the PFET is on and the NFET is off.*

The complementary nature of NFETs and PFETs makes it possible to design low-power circuits called **CMOS** or **complementary MOS** circuits as illustrated in Fig. 6–7a. The circuit symbol of PFET has a circle attached to the gate. The example is an inverter. It charges and discharges the output node with its load capacitance, $C$, to either $V_{dd}$ or 0 under the command of $V_g$. When $V_g = V_{dd}$, the NFET is on and the PFET is off (think of them as simple on–off switches), and the output node is pulled down to the ground ($V_{out} = 0$). When $V_g = 0$, the NFET is off and the PFET is on; the output node is pulled up to $V_{dd}$. In either static case, one of the two transistors is off and there is no current flow from $V_{dd}$ through the two transistors directly to the ground. Therefore, CMOS circuits consume much less power than other types of circuits. Figure 6–7b illustrates how NFET and PFET can be fabricated on the same chip. Portions of the P-type substrate are converted into N-type wells by donor implantation and diffusion. Contacts to the P substrate and N well are included in the figure. Figure 6–7c illustrates the basic **layout** of a CMOS

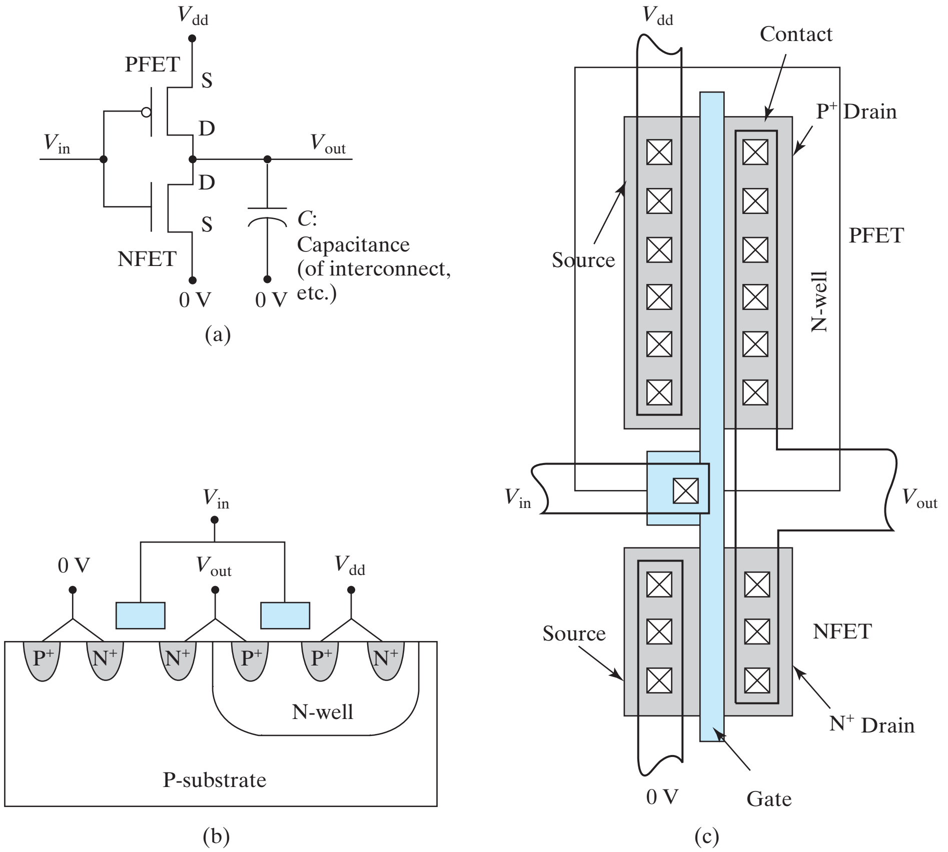


FIGURE 6–7 Three views of a CMOS inverter. (a) A CMOS inverter consists of a PFET **pull-up device** and an NFET **pull-down device**. (b) Integration of NFET and PFET on the same chip. For simplicity, trench isolation (see Fig. 6–1), which fills all the surface area except for the diffusion regions and the channel regions, is not shown. (c) Layout of a CMOS inverter.

inverter. It is a view of the circuit from above the Si wafer and may be thought of as a composite drawing of several photomasks used to fabricate the inverter. $V_{in}$, $V_{out}$, $V_{dd}$, and ground voltage are carried by metal lines. The poly-Si gate is the vertical bar connected to $V_{in}$. The metal to semiconductor contacts are usually made in multiple identical holes because it is more difficult to fabricate contact holes of varying sizes and shapes.

## 6.3 • SURFACE MOBILITIES AND HIGH-MOBILITY FETs •

It is highly desirable to have a large transistor current so that the MOSFET can charge and discharge the circuit capacitances ($C$ in Fig. 6–7a) quickly and achieve a high circuit speed. An important factor that determines the MOSFET current is the electron or hole mobility in the surface inversion layer.

### 6.3.1 Surface Mobilities

When a small $V_{ds}$ is applied, the drain to source current, $I_{ds}$,[1] in Fig. 6–6b is

$$\begin{aligned} I_{ds} &= W \cdot Q_{inv} \cdot v = WQ_{inv}\mu_{ns}\mathscr{E} = WQ_{inv}\mu_{ns}V_{ds}/L \\ &= WC_{oxe}(V_{gs}-V_t)\mu_{ns}V_{ds}/L \end{aligned} \tag{6.3.1}$$

$W$ is the **channel width,** i.e., the channel dimension perpendicular to the page in Fig. 6–6 and the vertical dimension of the channel in Fig. 6–7c. $Q_{inv}$ (C/cm$^2$) is the inversion charge density [Eq. (5.5.3)]. $\mathscr{E}$ is the channel electric field, and $L$ is the **channel length.** $\mu_{ns}$ is the electron **surface mobility**, or the **effective mobility**. In MOSFETs, $\mu_{ns}$ and $\mu_{ps}$ (hole surface mobility) are several times smaller than the bulk mobilities presented in Section 2.2. In Eq. (6.3.1), all quantities besides $\mu_{ns}$ are known in Eq. (6.3.1) or can be measured, and therefore $\mu_{ns}$ can be determined.

$\mu_{ns}$ is a function of the average of the electric fields at the bottom and the top of the inversion charge layer, $\mathscr{E}_b$ and $\mathscr{E}_t$ in Fig. 6–8 [4]. From Gauss's Law, using the depletion layer as the Gaussian box

$$\mathscr{E}_b = -Q_{dep}/\varepsilon_s \tag{6.3.2}$$

From Eq. (5.4.4)

$$V_t = V_{fb} + \phi_{st} - Q_{dep}/C_{oxe} \tag{6.3.3}$$

Therefore,

$$\mathscr{E}_b = \frac{C_{oxe}}{\varepsilon_s}(V_t - V_{fb} - \phi_{st}) \tag{6.3.4}$$

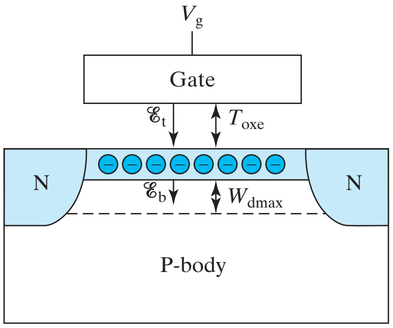


FIGURE 6–8 Surface mobility is a function of the average of the electric fields at the bottom and the top of the inversion charge layer, $\mathscr{E}_b$ and $\mathscr{E}_t$.

[1] We will follow the convention that positive $I_{ds}$ refers to the normal direction of channel current from $V_{dd}$ to ground, i.e., drain to source in NFET and source to drain in PFET. Therefore, $I_{ds}$ is always positive.

Apply Gauss's Law to a box that encloses the depletion layer and the inversion layer.

$$
\begin{aligned}
\mathscr{E}_{\mathrm{t}} &= -(Q_{\mathrm{dep}} + Q_{\mathrm{inv}})/\varepsilon_{\mathrm{s}} \\
&= \mathscr{E}_{\mathrm{b}} - Q_{\mathrm{inv}}/\varepsilon_{\mathrm{s}} = \mathscr{E}_{\mathrm{b}} + \frac{C_{\mathrm{oxe}}}{\varepsilon_{\mathrm{s}}}(V_{\mathrm{gs}} - V_{\mathrm{t}}) \\
&= \frac{C_{\mathrm{oxe}}}{\varepsilon_{\mathrm{s}}}(V_{\mathrm{gs}} - V_{\mathrm{fb}} - \phi_{\mathrm{st}}) \qquad (6.3.5)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{2}(\mathscr{E}_{\mathrm{b}} + \mathscr{E}_{\mathrm{t}}) &= \frac{C_{\mathrm{oxe}}}{2\varepsilon_{\mathrm{s}}}(V_{\mathrm{gs}} + V_{\mathrm{t}} - 2V_{\mathrm{fb}} - 2\phi_{\mathrm{st}}) \\
&\approx \frac{C_{\mathrm{oxe}}}{2\varepsilon_{\mathrm{s}}}(V_{\mathrm{gs}} + V_{\mathrm{t}} + 0.2\ \mathrm{V}) \\
&= \frac{\varepsilon_{\mathrm{ox}}}{2\varepsilon_{\mathrm{s}} T_{\mathrm{oxe}}}(V_{\mathrm{gs}} + V_{\mathrm{t}} + 0.2\ \mathrm{V}) \\
&= \frac{V_{\mathrm{gs}} + V_{\mathrm{t}} + 0.2\ \mathrm{V}}{6T_{\mathrm{oxe}}} \quad \text{for N}^{+}\text{ poly-gate NFET} \qquad (6.3.6)
\end{aligned}
$$

$\mu_{ns}$ has been found to be a function of the average of $\mathscr{E}_{\mathrm{b}}$ and $\mathscr{E}_{\mathrm{t}}$. (This conclusion is sometimes presented with the equivalent statement that $\mu_{ns}$ is a function of $Q_{\mathrm{dep}} + Q_{\mathrm{inv}}/2$.) The measured $\mu_{ns}$ is plotted in Fig. 6–9 and can be fitted with [4]:[2]

$$
\mu_{ns} = \frac{540\ \mathrm{cm}^2/\mathrm{Vs}}{1 + \left(\dfrac{V_{\mathrm{gs}} + V_{\mathrm{t}} + 0.2\ \mathrm{V}}{5.4 T_{\mathrm{oxe}}}\right)^{1.85}} \qquad (6.3.7)
$$

Empirically, the hole surface mobility is a function of $(\mathscr{E}_{\mathrm{t}} + 1.5\mathscr{E}_{\mathrm{b}})/2$ [5].

$$
\mu_{ps} = \frac{185\ \mathrm{cm}^2/\mathrm{Vs}}{1 - \left(\dfrac{V_{\mathrm{gs}} + 1.5\mathrm{V}_{\mathrm{t}} - 0.25\ \mathrm{V}}{3.38 T_{\mathrm{oxe}}}\right)} \qquad (6.3.8)
$$

$T_{\mathrm{oxe}}$ is defined in Eq. (5.9.2). Normally, $V_{\mathrm{gs}}$ and $V_{\mathrm{t}}$ are negative for a PFET, i.e., in Eq. (6.3.8). This mobility model accounts for the effects of the major variables on the surface mobility. When device variables $V_{\mathrm{gs}}$, $V_{\mathrm{t}}$, and $T_{\mathrm{oxe}}$ are properly considered, all silicon MOSFETs exhibit essentially the same surface mobility as illustrated in Fig 6–9. This is said to be Si's **universal effective mobility**. The surface mobility is lower than the bulk mobility because of **surface roughness scattering** [5, 6]. It makes the mobilities

[2] Equation (6.3.7) is for the common case of NMOSFET with N$^{+}$ poly-Si gate. In general, the 0.2 V term should be replaced with $-2(V_{\mathrm{fb}} + \phi_{\mathrm{st}})$. See Eq. (5.4.2) for $\phi_{\mathrm{st}}$. Eq. (6.3.8) is for the common case of PMOSFET with P$^{+}$ poly-Si gate. In general, the –0.25 V term should be replaced with $2.5(V_{\mathrm{fb}} + \phi_{\mathrm{st}})$.

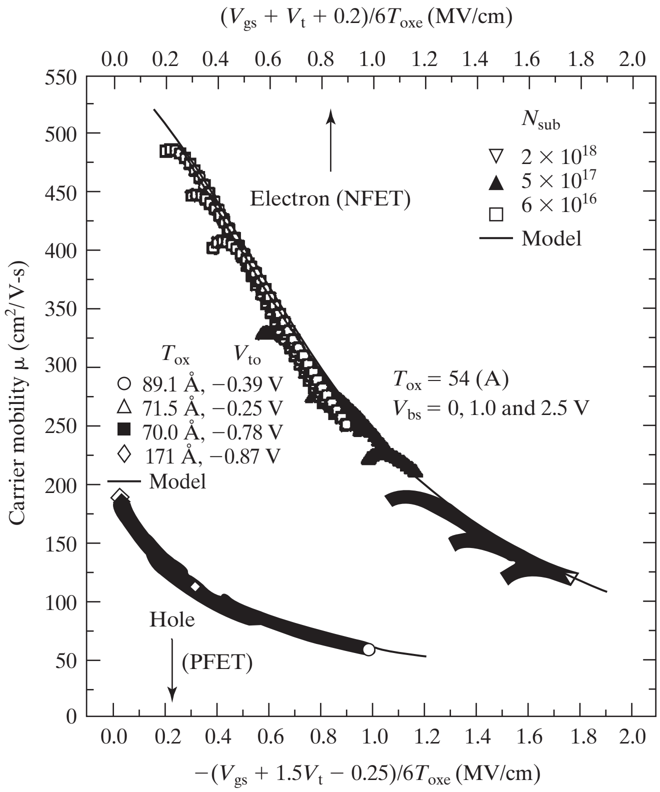


FIGURE 6–9 Electron and hole surface mobilities are determined by $V_{gs}$, $V_t$, and $T_{oxe}$. $T_{oxe}$ is the $SiO_2$ equivalent electrical oxide thickness. (From [4]. © 1996 IEEE.)

### Effect of Wafer Surface Orientation and Drift Direction

The surface mobility is a function of the surface orientation and the drift direction. The standard CMOS technology employs the [100] surface silicon wafers, and the transistors are laid out so that the electrons and holes flow along the identical (0 ±1 ±1) directions on the wafer surface. (See Section 1.1 for explanation of the notation). One of the reasons for the choice is that this combination provides the highest $\mu_{ns}$, though not the highest $\mu_{ps}$. The mobility data in Fig. 6–9 are for this standard choice. The wafer orientation and current direction also determine how $\mu_{ns}$ and $\mu_{ps}$ respond to mechanical stress (see Section 7.1.2). These orientation effects can be explained by the solution of the Schrödinger's wave equation.

decrease as the field in the inversion layer ($\mathscr{E}_b$, $\mathscr{E}_t$) becomes stronger and the charge carriers are confined closer to the Si–$SiO_2$ interface.

$\mu_{ns}$ and $\mu_{ps}$ still roughly follow the $T^{-3/2}$ temperature dependence that is characteristic of phonon scattering (see Eq. 2.2.5). In Fig. 6–9, the surface mobility around $V_g \approx V_t$, especially in the heavily doped semiconductor ($2 \times 10^{18}$ cm$^{-3}$), is lower than the universal mobility. Dopant ion scattering is the culprit. At higher $V_g$, dopant ion scattering effect is screened out by the inversion layer carriers (see Section 2.2.2).

**EXAMPLE 6–1**

What is the surface mobility at $V_{gs}$ = 1V in an N-channel MOSFET with $V_t$ = 0.3 V and $T_{oxe}$ = 2 nm?

**SOLUTION:**

$$(V_{gs} + V_t + 0.2)/6\mathrm{T}_{oxe} = (1.5\mathrm{V}/12 \times 10^{-7}\mathrm{cm} = 1.25\,\mathrm{MV/cm})$$

A megavolt ($10^6$ V) is 1 MV. From Fig. 6–9, $\mu_{ns} \approx 190$ cm$^2$/V·s. To the dismay of MOSFET engineers, this is several times smaller than $\mu_n$, the bulk mobility. $\mu_{ps}$ for a PMOSFET of similar design is only 60 cm$^2$/V·s.

### 6.3.2 GaAs MESFET

Higher carrier mobility allows the carriers to travel faster and the transistors to operate at higher speeds. High-speed devices not only improve the throughput of electronic equipment but also open up new applications such as inexpensive microwave communication. The most obvious way to improve speed is to use a semiconductor having higher mobility than silicon such as germanium, Ge (see Table 2–1) or strained Si (see Section 7.1.2). Single-crystalline Ge and SiGe alloy films can be grown epitaxially over Si substrates. The extension of Si technology to include Ge or SiGe transistor is a promising way to improve the device speed.

Table 2–1 indicates that GaAs and some other compound semiconductors have much higher electron mobilities than Si. For some applications, only N-channel FETs are needed and the hole mobility is of no importance. Unfortunately, it is very difficult to produce high-quality MOS transistors in these materials. There are too many charge traps at the semiconductor/dielectric interface for MOSFET application. Fortunately, a Schottky junction can serve as the control gate of a GaAs FET in place of an MOS gate. The device, called **MESFET** for **metal–semiconductor field-effect transistor**, is shown in Fig. 6–10. Because GaAs has a large $E_g$ and small $n_i$, undoped GaAs has a very high resistivity and can be considered an insulator. The metal gate may be made of Au, for example. A large Schottky barrier height is desirable for minimizing the input gate current, i.e., the Schottky diode current.

When a reverse-bias voltage or a small forward voltage (small enough to keep the gate diode current acceptable) is applied to the gate, the depletion region under the gate expands or contracts. This modulates the thickness of the conductive channel, the part that is not depleted. This change, in turn, modulates the channel

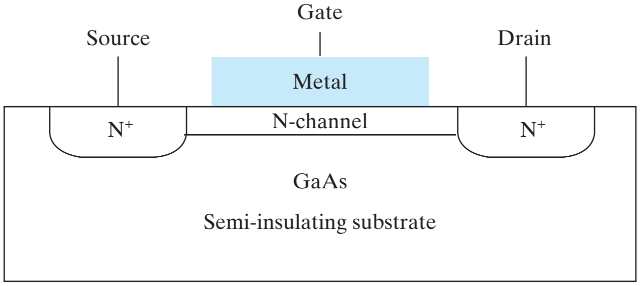


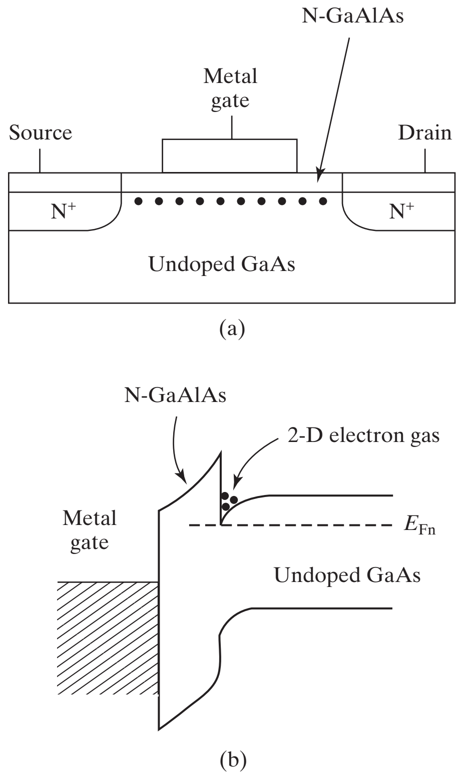
FIGURE 6–10 Schematic of a Schottky gate FET called **MESFET**.

current $I_{ds}$. Because $I_{ds}$ does not flow in a surface inversion layer, the electron mobility is not degraded by surface scattering. This fact further enhances GaAs MESFET's speed advantage.

If the N-channel thickness is larger than the depletion-layer width at $V_g = 0$, the MESFET is conductive at $V_g = 0$ and requires a (reverse bias) gate voltage to turn it off. It is called a **depletion-mode transistor**. If the N-channel is thinner than the depletion-layer width at $V_g = 0$, a (forward) gate voltage is needed to turn the transistor on. This is known as an **enhancement-mode transistor**. Modern Si MOSFETs are all enhancement-mode transistors, which make circuit design much easier. GaAs FETs of both depletion-mode and enhancement-mode types are used. The depletion-type device is easier to make.

### 6.3.3 HEMT

The dopants in the channel in Fig. 6–10 significantly reduce the electron mobility through impurity scattering (see Section 2.2.2). If the channel is undoped, the mobility can be much higher. A MOSFET does not rely on doping to provide the conduction channel. Can GaAs FET do the same? The answer is yes. A MOS-like structure can be made by growing a thin epitaxial layer of GaAlAs over the undoped GaAs substrate as shown in Fig. 6–11a. Under the gate the GaAlAs film is



(a)

(b)

FIGURE 6–11 (a) The basic HEMT structure. The large band gap GaAlAs functions like the $SiO_2$ in a MOSFET. The conduction channel is in the undoped GaAs. (b) The energy diagram confirms the similarity to a MOSFET.

depleted. GaAlAs has a larger band gap than GaAs and Fig. 6–11b shows that it functions like the oxide in a MOSFET (see Fig. 5–9) in that it creates an energy well and a thin layer of electrons at the GaAs–GaAlAs interface. The curvature in the GaAlAs band diagram is due to the presence of the dopant ions as in the depletion layer of a PN junction. $E_F$ is the Fermi level of the N+ source and it (with $E_c$) determines the electron concentration in the conduction channel. The channel electrons come from the $N^+$ source. Because the epitaxial interface of the two semiconductors is smoother than the Si–$SiO_2$ interface, this device does not suffer from mobility degradation by surface scattering as MOSFET does. This device is called **HEMT** or **high electron-mobility transistor**, or **MODFET** for **modulation-doped FET**. It is used in microwave communication, satellite TV receivers, etc.

### 6.3.4 JFET

If the Schottky junction in Fig. 6–10 is replaced with a $P^+N$ junction, the new structure is called a **JFET** or **junction field-effect transistor**. The $P^+$ gate is of course connected to a metal for circuit connections. As in a MESFET, a reverse bias would expand the depletion layer and constrict the conduction channel. In this manner, the JFET current can be controlled with the gate voltage. Before the advent of MOSFET, ICs were built

## How to Measure the $V_t$ of a MOSFET

$V_t$ is rarely determined from the *CV* data. Instead it can be more easily measured from the $I_{ds} - V_{gs}$ plot shown in Fig. 6–12.

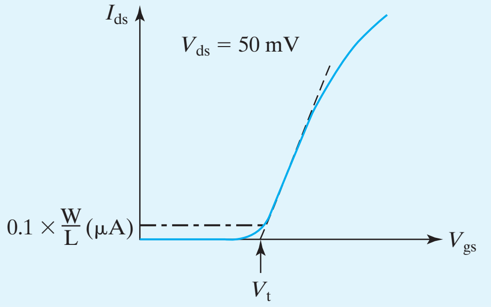


FIGURE 6–12 $V_t$ can be measured by extrapolating the $I_{ds}$ vs. $V_{gs}$ curve to $I_{ds} = 0$. Alternatively, it can be defined as the $V_{gs}$, at which $I_{ds}$ is a small fixed amount.

$I_{ds}$ measured at a small $V_{ds}$ such as 50 mV is plotted against $V_{gs}$. At $V_{gs} > V_t$, $I_{ds}$ increases linearly with $(V_{gs} - V_t)$ according to Eq. (6.3.1), if $\mu_{ns}$ were a constant. Because $\mu_{ns}$ decreases with increasing $V_{gs}$ (see Section 6.3), the curve is sublinear. It is a common practice to extrapolate the curve at the point of maximum slope and take the intercept with the *x*-axis as $V_t$.

An increasingly popular alternative is to define $V_t$ as the $V_{gs}$ at which $I_{ds}$ is equal to a small value such as

$$I_{ds} = 0.1\ \mu A \times \frac{W}{L}$$

Also see Fig. 7–2 d.

with bipolar transistors, which have forward-biased diodes at the input and draw significant input current (see Chapter 8). The high input currents and capacitances were quite undesirable for some circuits. JFET provided a low input current and capacitance device because its input is a reverse-biased diode. JFET can be fabricated with bipolar transistors and coexist in the same IC chip.

## 6.4 ● MOSFET $V_t$, BODY EFFECT, AND STEEP RETROGRADE DOPING ●

The inversion layer of a MOSFET can be thought of as a resistive N-type film (1–2 nm thin) that connects the source and the drain as shown in Fig. 6–13. This film, at potential $V_s$, forms a capacitor with the gate, the oxide being the capacitor

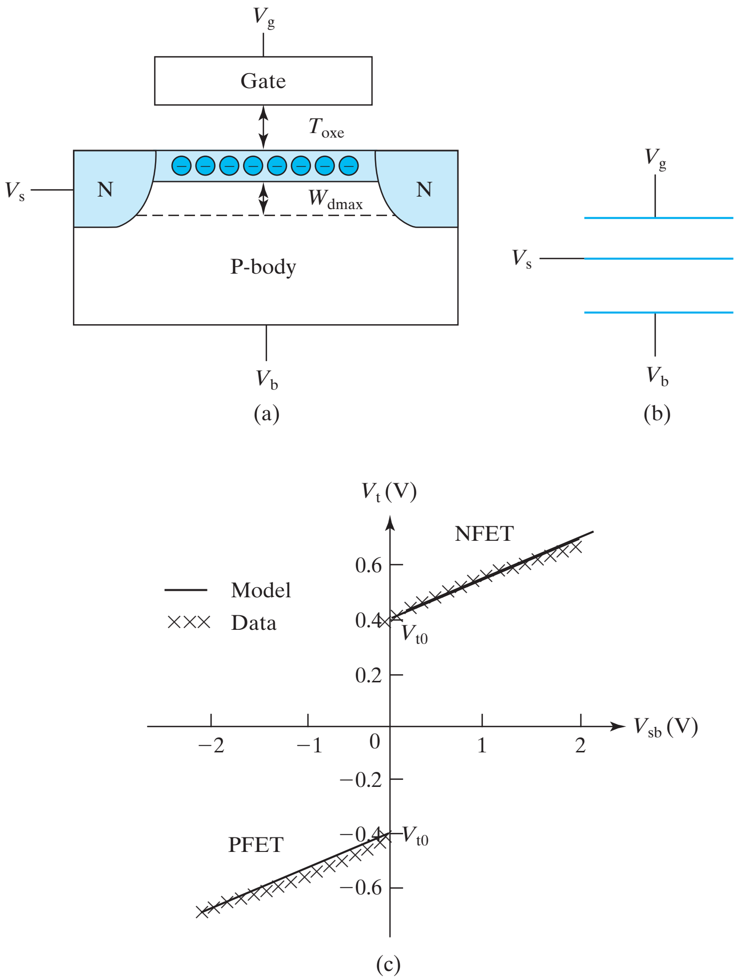


FIGURE 6–13 (a) and (b) The inversion layer can be viewed as a conductive film that is coupled to $V_g$ through the oxide capacitance and coupled to $V_b$ through the depletion-layer capacitance. The drain is open-circuited. (c) $V_t$ is an approximately linear function of the body to source bias voltage. The polarity of the body bias is normally that which would reverse bias the body-source junction.

dielectric. It also forms a second capacitor with the body and the capacitor dielectric is the depletion layer. The depletion-layer capacitance is

$$C_{\mathrm{dep}} = \frac{\varepsilon_s}{W_{\mathrm{dmax}}} \tag{6.4.1}$$

In Chapter 5, with $V_b = V_s$, we concluded that the gate voltage induces a charge in the invesion layer,

$$Q_{\mathrm{inv}} = -C_{\mathrm{oxe}}(V_{\mathrm{gs}} - V_t) \tag{6.4.2}$$

Let us now assume that there is also a voltage between the source and the body, $V_{\mathrm{sb}}$. Since the body and the channel are coupled by $C_{\mathrm{dep}}$, $V_{\mathrm{sb}}$ induces a charge in the inversion layer, $C_{\mathrm{dep}}V_{\mathrm{sb}}$. Therefore

$$Q_{\mathrm{inv}} = -C_{\mathrm{oxe}}(V_{\mathrm{gs}} - V_t) + C_{\mathrm{dep}}V_{\mathrm{sb}} \tag{6.4.3}$$

$$= -C_{\mathrm{oxe}}\left(V_{\mathrm{gs}} - \left(V_t + \frac{C_{\mathrm{dep}}}{C_{\mathrm{oxe}}}V_{\mathrm{sb}}\right)\right) \tag{6.4.4}$$

Equation (6.4.4) can be rewritten in the simple form of Eq. (6.4.2) if we adopt a modification to $V_t$. (What we have called $V_t$ up to this point will henceforth be called $V_{t0}$.)

$$Q_{\mathrm{inv}} = -C_{\mathrm{oxe}}(V_{\mathrm{gs}} - V_t(V_{\mathrm{sb}})) \tag{6.4.5}$$

$$\boxed{V_t(V_{\mathrm{sb}}) = V_{t0} + \frac{C_{\mathrm{dep}}}{C_{\mathrm{oxe}}}V_{\mathrm{sb}} = V_{t0} + \alpha V_{\mathrm{sb}}} \tag{6.4.6}$$

$$\alpha = C_{\mathrm{dep}}/C_{\mathrm{oxe}} = 3\mathrm{T}_{\mathrm{oxe}}/W_{\mathrm{dmax}} \tag{6.4.7}$$

The factor 3 is the ratio of the relative dielectric constants of silicon (11.9) and $SiO_2$ (3.9). Figure 6–13c illustrates the conclusion that $V_t$ is a function of $V_{\mathrm{sb}}$. *When the source-body junction is reverse-biased, the NFET $V_t$ becomes more positive and the PFET $V_t$ becomes more negative.* Normally, the source-body junctions are never forward biased so that there is no forward diode current.

The fact that $V_t$ is a function of the body bias is called the **body effect**. When multiple NFETs (or PFETs) are connected in series in a circuit, they share a common body (the silicon substrate) but their sources do not have the same voltage. Clearly some transistors' source–body junctions are reversed biased. This raises their $V_t$ and reduces $I_{\mathrm{ds}}$ and the circuit speed. Circuits therefore perform best when $V_t$ is as insensitive to $V_{\mathrm{sb}}$ as possible, i.e., the body effect should be minimized. This can be accomplished by minimizing the $T_{\mathrm{ox}}/W_{\mathrm{dmax}}$ ratio. (We will see again and again that a thin oxide is desirable.) $\alpha$ in Eq. (6.4.6) can be extracted from the slope of the curve in Fig. 6–13c and is called the **body-effect coefficient**.

Modern transistors employ steep **retrograde body doping profiles** (light doping in a thin surface layer and very heavy doping underneath) illustrated in Fig. 6–14. Steep retrograde doping allows transistor shrinking to smaller sizes for cost reduction and reduces impurity scattering. Section 7.5 explains why. The depletion-layer thickness is basically the thickness of the lightly doped region. As $V_{\mathrm{sb}}$ increases, the depletion layer does not change significantly. Therefore $C_{\mathrm{dep}}$ and

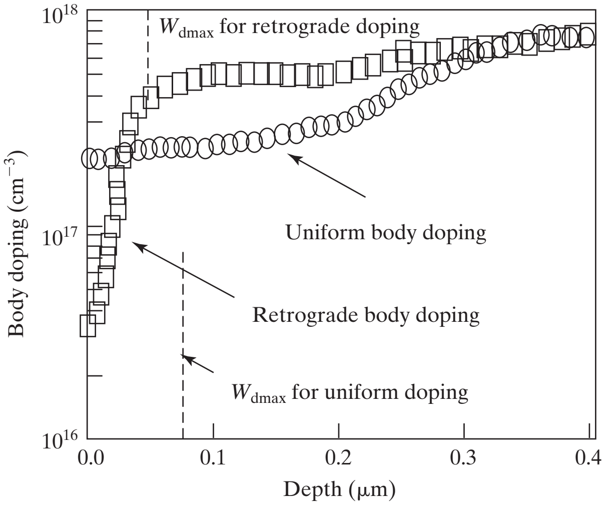


FIGURE 6–14 Comparison of a steep retrograde doping profile and a uniform doping profile.

$\alpha$ are basically constants. As a result, modern transistors exhibit a more or less linear relationship between $V_t$ and $V_{sb}$. A linear relationship means that $W_{\text{dmax}}$ and therefore the $C_{\text{dep}}/C_{\text{oxe}}$ ratio are independent of the body bias.

In earlier generations of MOSFETs, the body doping density is more or less uniform (see the lower curve in Fig. 6–14) and $W_{\text{dmax}}$ varies with $V_{sb}$. In that case, the theory for the body effect is more complicated. $V_t$ can be obtained by replacing the $2\phi_B$ term (band bending in the body) in Eq. (5.4.3) with $2\phi_B + V_{sb.}$[3]

$$V_t = V_{t0} + \frac{\sqrt{qN_a 2\varepsilon_s}}{C_{oxe}}(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B})$$

$$\equiv V_{t0} + \gamma(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}) \qquad (6.4.8)$$

$\gamma$ is called the **body-effect parameter**. Equation (6.4.8) predicts that $V_t$ is a sublinear function of $V_{sb}$. A hint of the sublinearity is observable in the data in Fig. 6–13c. Equation (6.4.8) is sometimes linearized by Taylor expansion so that $V_t$ is expressed as a linear function of $V_{sb}$ in the form of Eq. (6.4.6).

## 6.5 ● $Q_{\text{INV}}$ IN MOSFET ●

Let us consider Fig. 6–15 with $V_d > V_s$. The channel voltage, $V_c$, is now a function of $x$. $V_c = V_s$ at $x = 0$ and $V_c = V_d$ at $x = L$. Compare a point in the middle of the channel where $V_c > V_s$ with a point at the source-end of the channel, where

[3] When the source–body junction is reverse biased, there are two quasi-Fermi levels, $E_{Fn}$ and $E_{Fp}$ (similar to Fig. 4–7c with the P-region being the MOSFET body and the N-region being the source), which are separated by $qV_{sb}$. The inversion layer does not appear when $E_c$ at the interface is close to $E_{Fp}$ ($E_F$ in Fig. 5–7). It appears when $E_F$ is close to $E_{Fn}$ ($qV_{sb}$ below $E_F$ in Fig. 5–7). This requires the band bending to be $2\phi_B + V_{sb}$, not $2\phi_B$.

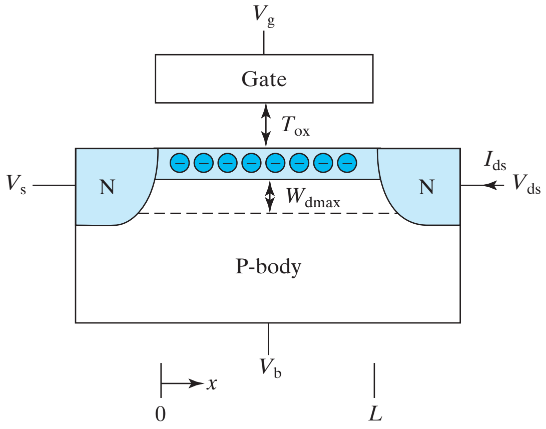


FIGURE 6–15 When $V_{ds} \neq 0$, the channel voltage $V_c$ is a function of $x$.

$V_c = V_s$. Because the voltage in the middle of the channel is higher at $V_c(x)$, there is less voltage across the oxide capacitor (and across the depletion layer capacitor). Therefore, there will be fewer electrons on the capacitor electrode (the inversion layer). Specifically, the $V_{gs}$ term in Eq. (6.4.5) should be replaced by $V_{gc}(x)$ or $V_{gs} - V_{cs}(x)$ and $V_{sb}$ by $V_{sb} + V_{cs}(x)$.

$$\begin{aligned} Q_{inv}(x) &= -C_{oxe}(V_{gs} - V_{cs} - V_{t0} - \alpha(V_{sb} + V_{cs})) \\ &= -C_{oxe}(V_{gs} - V_{cs} - (V_{t0} + \alpha V_{sb}) - \alpha V_{cs}) \\ &= -C_{oxe}(V_{gs} - mV_{cs} - V_t) \qquad (6.5.1) \end{aligned}$$

$$m \equiv 1 + \alpha = 1 + C_{dep}/C_{oxe} = 1 + 3\mathrm{T}_{oxe}/W_{dmax} \qquad (6.5.2)$$

$m$ is typically around 1.2. It is acceptable and easier at the beginning to simply assume $m = 1$. However, including $m$ in the equations significantly improves their accuracies for later reference. The body is sometimes called the **back gate** since it clearly has a similar though weaker effect on the channel charge. The back-gate effect on $Q_{inv}$ is often called the **bulk-charge effect**. $m$ is called the **bulk-charge factor**. Clearly the bulk-charge effect is closely linked to the body-effect of Section 6.4.

## 6.6 • BASIC MOSFET IV MODEL •

Using Eq. (6.5.1) and dropping the negative sign for simplicity ($I_{ds}$ in Fig. 6–15 is understood to flow from the high-voltage terminal to the low-voltage terminal).

$$\begin{aligned} I_{ds} &= W \cdot Q_{inv}(x) \cdot v = W \cdot Q_{inv}\mu_{ns}\mathscr{E} \\ &= WC_{oxe}(V_{gs} - mV_{cs} - V_t)\mu_{ns}\mathrm{d}V_{cs}/\mathrm{d}x \qquad (6.6.1) \end{aligned}$$

$$\int_0^L I_{ds}dx = WC_{oxe}\mu_{ns}\int_0^{V_{ds}}(V_{gs}-mV_{cs}-V_t)dV_{cs} \tag{6.6.2}$$

$$I_{ds}L = WC_{oxe}\mu_{ns}\left(V_{gs}-V_t-\frac{m}{2}V_{ds}\right)V_{ds} \tag{6.6.3}$$

$$\boxed{I_{ds} = \frac{W}{L}C_{oxe}\mu_{ns}\left(V_{gs}-V_t-\frac{m}{2}V_{ds}\right)V_{ds}} \tag{6.6.4}$$

Equation (6.6.4) shows that $I_{ds}$ is proportional to $W$ (channel width), $\mu_{ns}$, $V_{ds}/L$ (the average field in the channel), and $C_{ox}(V_g - V_t - mV_{ds}/2)$, which may be interpreted as the average $Q_{inv}$ in the channel. When $V_{ds}$ is very small, the $mV_{ds}/2$ term is negligible and $I_{ds} \propto V_{ds}$, i.e., the transistor behaves as a resistor. As $V_{ds}$ increases, the average $Q_{inv}$ decreases and $dI_{ds}/dV_{ds}$ decreases. By differentiating Eq. (6.6.4) with respect to $V_{ds}$, it can be shown that $dI_{ds}/dV_{ds}$ becomes zero at a certain $V_{ds}$.

$$\frac{dI_{ds}}{dV_{ds}} = 0 = \frac{W}{L}C_{ox}\mu_{ns}(V_{gs}-V_t-mV_{ds}) \quad \text{at} \quad V_{ds} = V_{dsat}$$

$$\boxed{V_{dsat} = \frac{V_{gs}-V_t}{m}} \tag{6.6.5}$$

$V_{dsat}$ is called the **drain saturation voltage**, beyond which the drain current is saturated as shown in Fig. 6–16. For each $V_g$, there is a different $V_{dsat}$. The part of the *IV* curves with $V_{ds} << V_{dsat}$ is the **linear region**, and the part with $V_{ds} > V_{dsat}$ is the **saturation region**. Analog designers often refer to the regions as the **Ohmic region** and the **active region**.

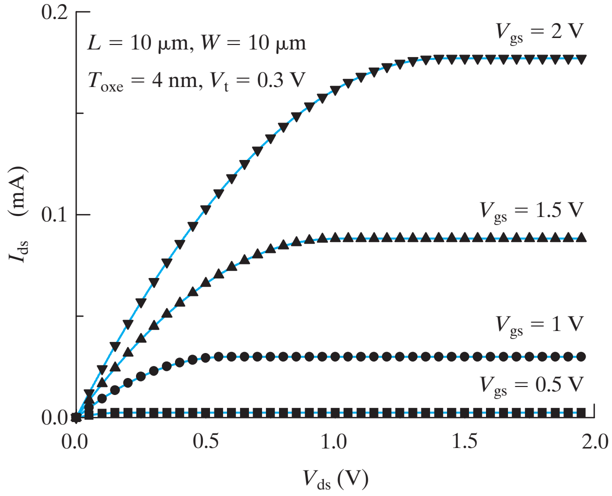


FIGURE 6–16 MOSFET IV characteristics.

The saturation current can be obtained by substituting $V_{\text{dsat}}$ [Eq. (6.6.5)] for $V_{\text{ds}}$ in Eq. (6.6.4).

$$I_{\text{dsat}} = \frac{W}{2\ \text{mL}} C_{\text{oxe}} \mu_{ns} (V_{\text{gs}} - V_{\text{t}})^2 \qquad (6.6.6)$$

What happens at $V_{\text{d}} = V_{\text{dsat}}$ and why does $I_{\text{ds}}$ stay constant beyond $V_{\text{dsat}}$? The first question can be answered by substituting $V_{\text{dsat}}$ [Eq. (6.6.5)] for $V_{\text{cs}}$ in Eq. (6.5.1). $Q_{\text{inv}}$ at the drain end of the channel, when $V_{\text{ds}} = V_{\text{dsat}}$, is zero! This disappearance of the inversion layer is called channel **pinch-off**. Figure 6–17 plots $V_{\text{cs}}$, $Q_{\text{inv}}$, and $I_{\text{ds}}$ at $V_{\text{ds}} = V_{\text{dsat}}$ and $V_{\text{ds}} > V_{\text{dsat}}$. In these two cases, $V_{\text{cs}}(x)$, $Q_{\text{inv}}(x)$ and therefore $I_{\text{ds}}$ are the same. This explains why $I_{\text{ds}}$ does not change with $V_{\text{ds}}$ beyond $V_{\text{dsat}}$. The only difference is that, at $V_{\text{ds}} > V_{\text{dsat}}$, there exists a short, high-field **pinch-off region** where $Q_{\text{inv}} = 0$ and across which the voltage $V_{\text{ds}} - V_{\text{dsat}}$ is dropped. Section 6.9.1 will present an improvement to the concept of pinch-off such that $Q_{\text{inv}}$ does not drop to zero. For now, the concept of pinch-off is useful for introducing the phenomenon of current saturation.

How can a current flow through the pinch-off region, which is similar to a depletion region? The fact is that a depletion region does not stop current flow as long as there is a supply of the right carriers. For example, in solar cells and photodiodes, current can flow through the depletion region of PN junctions. Similarly, when the electrons reach the pinch-off region of a MOSFET, they are swept down the steep potential drop in Fig. 6–17h. Therefore, the pinch-off region does not present a barrier to current flow. Furthermore, Fig. 6–17d and h show that the electron flow rates (current) are equal in the two cases because they have the same drift field and $Q_{\text{inv}}$ in the channel. In other words, the current is independent of $V_{\text{ds}}$ beyond $V_{\text{dsat}}$. The situation is like a mountain stream feeding into a waterfall. The slope of the river bed ($\text{d}E_{\text{c}}/\text{d}x$) and the amount of water in the stream determine the water flow rate in the stream, which in turn determines the flow rate down the waterfall. The height of the waterfall ($V_{\text{ds}} - V_{\text{dsat}}$), whether 1 or 100 m, has no influence over the flow rate.

## ● Channel Voltage Profile ●

First consider the case of $V_{\text{ds}} = V_{\text{dsat}}$. Substituting the upper limits of integration in Eq. (6.6.2), $L$ and $V_{\text{ds}}$, with $x$ and $V_{\text{cs}}$ and using $I_{\text{ds}} = I_{\text{dsat}}$ = Eq. (6.6.6), you can show that (see Problem 6.9 at the end of the chapter).

$$V_{\text{cs}} = \frac{V_{\text{gs}} - V_{\text{t}}}{m} \left( 1 - \sqrt{1 - \frac{x}{L}} \right) \qquad (6.6.7)$$

As expected, $V_{\text{cs}} = 0$ at $x = 0$ and $V_{\text{cs}} = V_{\text{dsat}} = (V_{\text{g}} - V_{\text{t}})/\text{m}$ at $x = L$. From this, you can show that $WQ_{\text{inv}}\mu_{\text{s}}\mathcal{E}$ or $WC_{\text{ox}}(V_{\text{gs}} - \text{m}V_{\text{cs}} - V_{\text{t}})\mu_{\text{s}}\text{d}V_{\text{cs}}/\text{d}x$ is independent of $x$ and yields the $I_{\text{dsat}}$ expressed in Eq. (6.6.6). Equation (6.6.7) is plotted in Fig. 6–17a.

See Fig. 6–17e for the $V_{\text{ds}} > V_{\text{dsat}}$ case. $V_{\text{cs}}$ still follows Eq. (6.6.7) from the source to the beginning of the pinch-off region. $V_{\text{ds}} - V_{\text{dsat}}$ is dropped in a narrow pinch-off region next to the drain.

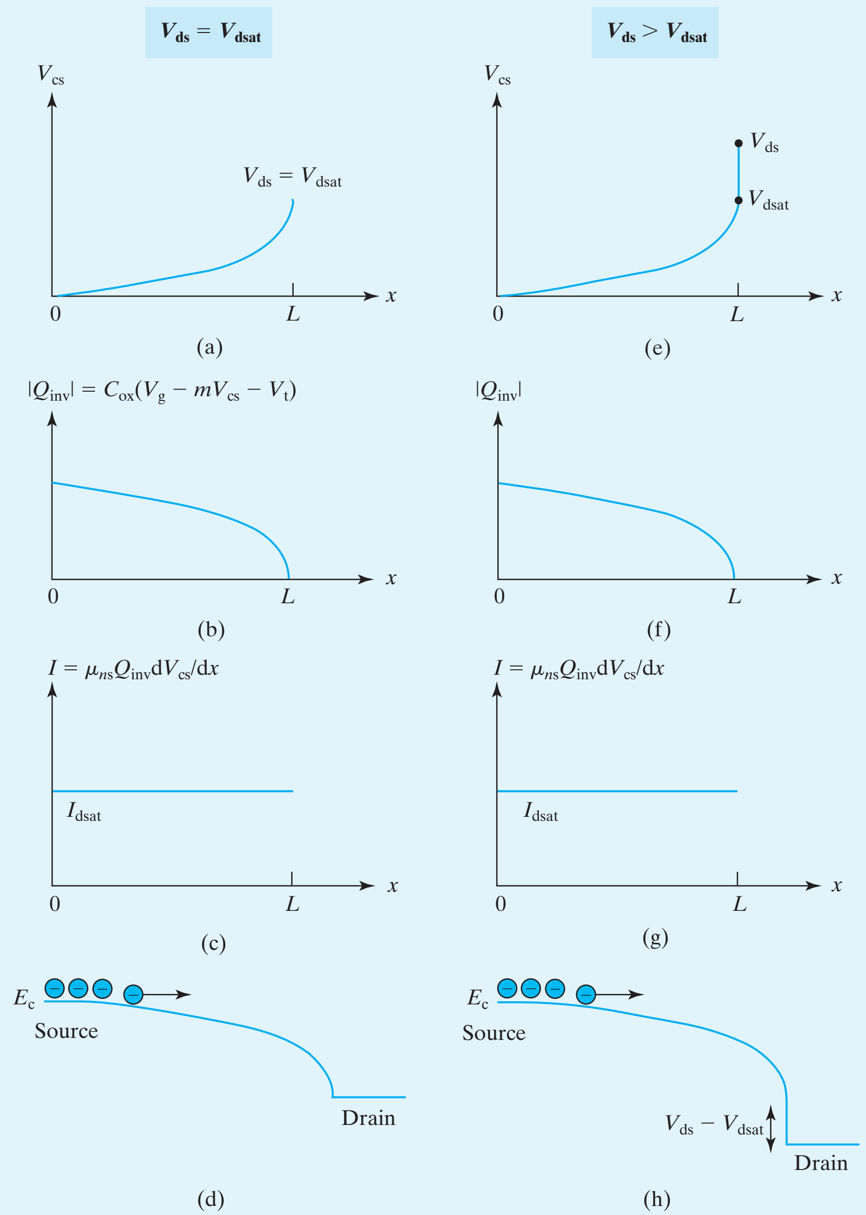


FIGURE 6–17 (a)–(d) $V_{ds} = V_{dsat}$ and (e)–(h) $V_{ds} > V_{dsat}$. Current does not change when $V_{ds}$ increases beyond $V_{dsat}$. (d) and (h) are $E_c(x)$ from the energy band diagrams.

**Transconductance**, defined as

$$g_m \equiv dI_{ds}/dV_{gs}\big|_{V_{ds}} \qquad (6.6.8)$$

is a measure of a transistor's sensitivity to the input voltage. In general, a large $g_m$ is desirable. Substituting Eq. (6.6.6) into Eq. (6.6.8), we find

$$g_{\mathrm{msat}} = \frac{W}{mL} C_{\mathrm{oxe}} \mu_{ns}(V_{\mathrm{gs}} - V_{\mathrm{t}}) \tag{6.6.9}$$

## 6.7 ● CMOS INVERTER—A CIRCUIT EXAMPLE ●

Transistors' influences on circuits will be illustrated using CMOS inverters, which were introduced in Section 6.2. They consume little power and have the important property of regenerating or cleaning up the digital signal. The latter property will be discussed in detail in Section 6.7.1. The speed of the inverters is analyzed in Section 6.7.2.

### 6.7.1 Voltage Transfer Curve (VTC)

Consider the CMOS inverter shown in Fig. 6–18a. The NFET IV characteristics are similar to those shown in Fig. 6–16 and are plotted on the right half of Fig. 6–18b. Assume that the PFET has identical (symmetric) IV as plotted on the left half of the figure. From (a), the $V_{\mathrm{ds}}$ of the PFET and NFET are related to $V_{\mathrm{out}}$ by $V_{\mathrm{dsN}} = V_{\mathrm{out}}$ and $V_{\mathrm{dsP}} = V_{\mathrm{out}} - 2$ V. Therefore, the two halves of (b) can be replotted in (c) using $V_{\mathrm{out}}$ as the common variable. For example, at $V_{\mathrm{out}} =$ 2V in (c), $V_{\mathrm{dsN}} = 2$V and $V_{\mathrm{dsP}} = 0$ V.

The two $V_{\mathrm{in}} = 0$ curves in (c) intersect at $V_{\mathrm{out}} = 2$ V. This means $V_{\mathrm{out}} = 2$ V when $V_{\mathrm{in}} = 0$ V. This point is recorded in Fig. 6–19. The two $V_{\mathrm{in}} = 0.5$ V curves intersect at around $V_{\mathrm{out}} = 1.9$ V. The two $V_{\mathrm{in}} = 1$ V curves intersect at $V_{\mathrm{out}} = 1$ V. All the $V_{\mathrm{in}}/V_{\mathrm{out}}$ pairs are represented by the curve in Fig. 6–19, which is the **voltage transfer characteristic** or **voltage transfer curve** or **VTC** of the inverter. The VTC provides the important noise margin of the digital circuits. $V_{\mathrm{in}}$ may be anywhere between 0 V and the NFET $V_{\mathrm{t}}$ and still produce a perfect $V_{\mathrm{out}} = V_{\mathrm{dd}}$. Similarly, $V_{\mathrm{in}}$ may be anywhere between 2V and 2 V plus the PFET $V_{\mathrm{t}}$ and produce a perfect $V_{\mathrm{out}} = 0$ V. Therefore, perfect "0" and "1" outputs can be produced by somewhat corrupted inputs. This regenerative property allows complex logic circuits to function properly in the face of inductive and capacitive noises and IR drops in the signal lines. A VTC with a narrow and steep middle region would maximize the noise tolerance. Device characteristics that contribute to a desirable VTC include a large $g_m$, low leakage in the off state, and a small $\partial I_{\mathrm{ds}}/\partial V_{\mathrm{ds}}$ in the saturation region. The latter two device properties will be discussed further in the next chapter.

For optimal circuit operation, the sharp transition region of the VTC should be located at or near $V_{\mathrm{in}} = V_{\mathrm{dd}}/2$. To achieve this symmetry, the *IV* curves of NFET and PFET Fig. 6–18b need to be closely matched (symmetric). This is accomplished by choosing a larger *W* for the PFET than the NFET. The $W_{\mathrm{P}}/W_{\mathrm{N}}$ ratio is usually around two to compensate for the fact that $\mu_{ps}$ is smaller than $\mu_{ns}$.

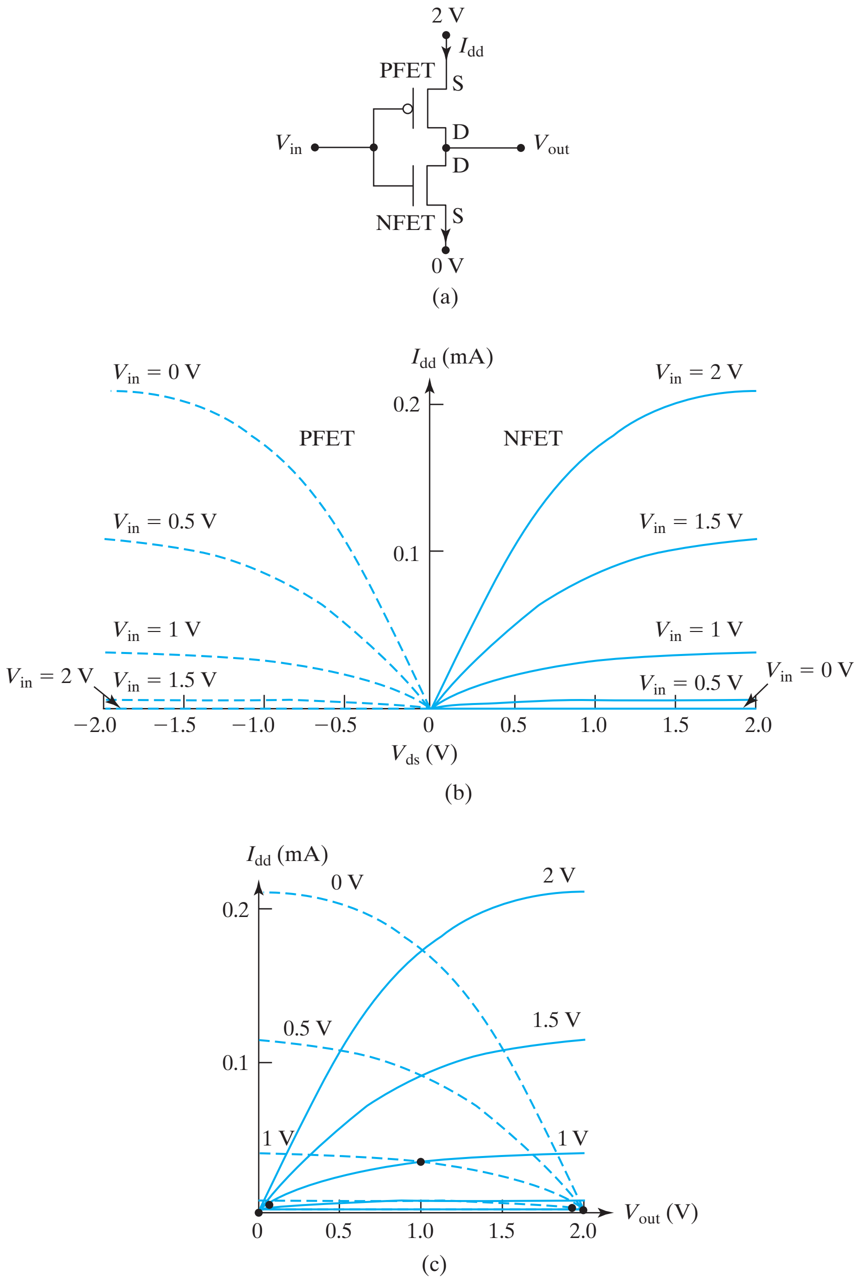


**FIGURE 6–18** (a) CMOS inverter; (b) IV characteristics of NFET and PFET; and (c) $V_{out} = V_{dsN} = 2\text{ V} + V_{dsP}$ according to (a).

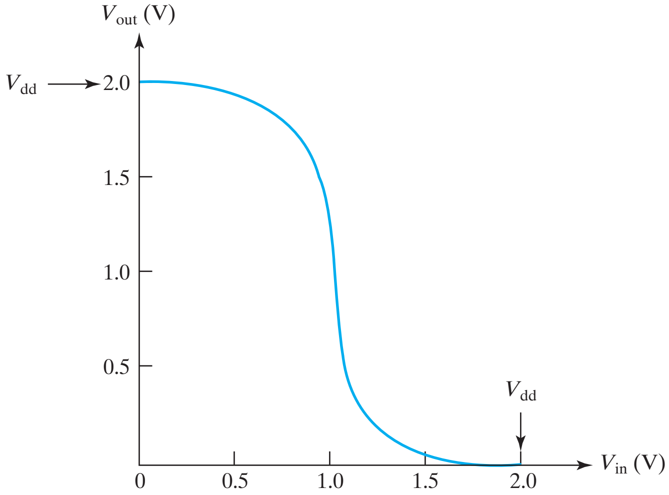


FIGURE 6–19 The VTC of a CMOS inverter.

### 6.7.2 Inverter Speed—The Importance of $I_{on}$

Propagation delay is the time delay for a signal to propagate from one gate to the next in a chain of identical gates as shown in Fig. 6–20.

$\tau_d$ is the average of the delays of pull-down (rising $V_1$ pulling down the output, $V_2$) and pull-up (falling $V_2$ pulling up the output, $V_3$). The propagation delay of an inverter may be expressed as [7]

$$\tau_d \approx \frac{CV_{dd}}{4}\left(\frac{1}{I_{onN}} + \frac{1}{I_{onP}}\right) \tag{6.7.1}$$

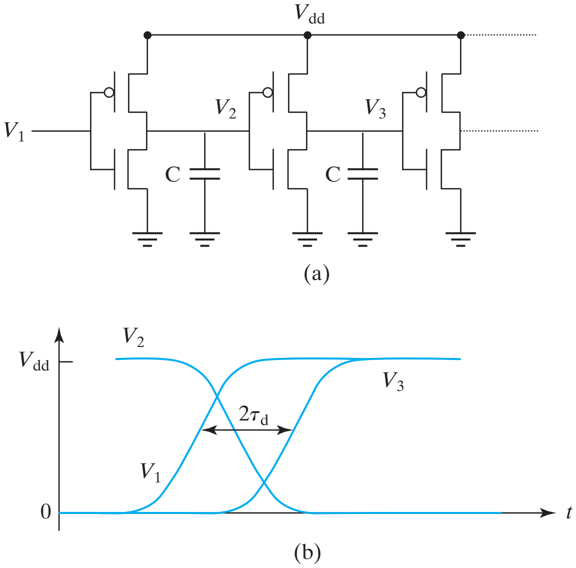


FIGURE 6–20 (a) A CMOS inverter chain. A circle on the gate indicates a PFET. (b) Propagation delay, $\tau_d$, defined.

where $I_{\text{onN}}$ is taken at $V_{\text{gs}} = V_{\text{dd}}$ and $I_{\text{onP}}$ taken at $V_{\text{gs}} = -V_{\text{dd}}$. They are called the **on-state current**, of the NFET and the PFET

$$I_{\text{on}} \equiv I_{\text{dsat}}\big|_{\text{maximum}\,|V_{\text{gs}}|} \tag{6.7.2}$$

Equation (6.7.1) has a simple explanation

$$\tau_{\text{d}} = \frac{1}{2}(\text{pull-down delay} + \text{pull-up delay}) \tag{6.7.3}$$

$$\text{pull-down delay} \approx \frac{CV_{\text{dd}}}{2I_{\text{onN}}}$$

$$\text{pull-up delay} \approx \frac{CV_{\text{dd}}}{2I_{\text{onP}}} \tag{6.7.4}$$

The delay is the time for the on-state transistor supplying a current, $I_{\text{on}}$, to change the output by $V_{\text{dd}}/2$ (not $V_{\text{dd}}$). $V_{\text{dd}}/2$ is plausible in view of Fig. 6–17. The charge drained from (or supplied to) $C$ by the FET during the delay is $CV_{\text{dd}}/2$. Therefore, the delay is $Q/I = CV_{\text{dd}}/2I_{\text{on}}$. One may interpret the delay as RC with $V_{\text{dd}}/2I_{\text{on}}$ as the switching resistance of the transistor. In order to maximize circuit speed it is clearly important to maximize $I_{\text{on}}$. We will further improve the $I_{\text{on}}$ model in the next two sections.

The capacitance $C$ represents the sum of all the capacitances that are connected to the output node of the inverter. They are the input capacitance of the next inverter in the chain, all the parasitic capacitances of the drain, and the capacitance of the metal interconnect that feeds the output voltage to the next inverter. In a large circuit, some interconnect metal lines can be quite long and their capacitances slow down the circuit significantly. This is ameliorated with the low-$k$ dielectric technology described in Section 3.8 and circuit design techniques such as using a transistor with large $W$ (a large $I_{\text{on}}$) to drive a longer interconnect and using repeaters.

Although the inverter is a very simple circuit, it is the basis of other more complex logic gates and memory cells. For example, Fig. 6–21 shows a NAND gate with two inputs. It is an inverter circuit with two series transistors in the pull-down path and two parallel transistors in the pull-up path.

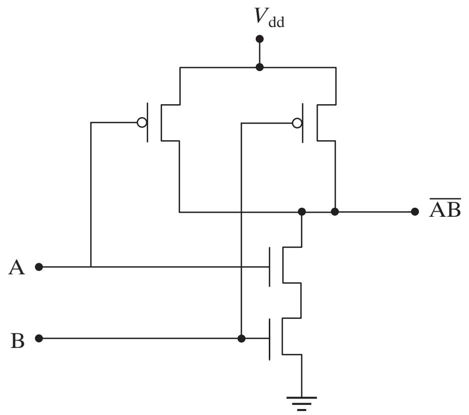


FIGURE 6–21 Inverters are the foundation of more complex circuits such as this two-input NAND gate.

### ● Ring Oscillators ●

$\tau_d$ of a logic gate can be conveniently measured by connecting the end of a chain of identical logic gates (see Fig. 6–20a, for example) to the beginning of the chain to form a **ring oscillator**. The signal of any of the drain nodes in the ring oscillates with a period equal to $\tau_d$ times the number of gates in the ring. By using a large number of gates in the ring, the oscillation frequency can be conveniently low for easy measurement. Dividing the measured period of oscillation by the number of gates yields $\tau_d$.

The number of gates in a ring oscillator must be an odd number such as 91. If the number is an even number such as 92, the circuit will not oscillate. Instead, it will be static at one of two stable states.

## 6.7.3 Power Consumption

An important goal of device design is to minimize circuit power consumption. In each switching cycle, a charge $CV_{dd}$ is transferred from the power supply to the load, $C$. The charge taken from the power supply in each second, $kCV_{dd}f$, is the average current provided by the power supply. Here, $f$ is the clock frequency and $k(<1)$ is an **activity factor** that represents the fact that a particular gate in a given circuit is not switched every clock cycle all the time. Therefore

$$P_{dynamic} = V_{dd} \times \text{average current} = kCV_{dd}^2 f \qquad (6.7.6)$$

This **dynamic power** dominates the power consumption when the inverter is switched frequently. *Power consumption can be reduced by lowering $V_{dd}$ and by minimizing all capacitances in the circuit as well as by reducing $k$*. It is interesting to note that *making $I_{on}$ large by using a small L or improving the carrier mobility does not increase $P_{dynamic}$*.

It is desirable for a transistor to provide a large $I_{on}$ (to reduce circuit switching delay) at a low $V_{dd}$ (to reduce circuit power consumption). Reducing the transistor L and W, other parameters being equal, would lower $C$ through reduction in the gate capacitance and the source–drain junction capacitance. Furthermore, smaller transistors make the chip smaller and therefore reduce the interconnect capacitance, too. Both device size reduction and $V_{dd}$ reduction have been powerful means of lowering the power consumption per circuit function.

Another component of power consumption is the **static power**, or **leakage power** or **stand-by power** that is consumed when the inverter is static.

$$P_{static} = V_{dd}I_{off} \qquad (6.7.7)$$

$I_{off}$ is the off-state leakage current when the transistor is supposed to be off. In an ideal transistor, $I_{off}$ would be zero. It is difficult to keep $I_{off}$ low in very high speed IC technologies as explained in detail in Chapter 7. The total power consumption is

$$P_{static} = P_{static} + P_{dynamic} \qquad (6.7.8)$$

## 6.8 ● VELOCITY SATURATION ●

A major weakness of the basic MOSFET IV model is that a finite current flows through the pinch-off region, where $Q_{\text{inv}} = 0$. This requires the carrier velocity to be infinite, a physical impossibility. We will now remove this shortcoming.

When the electric field is low, the carrier drift velocity, $v$, is $\mu\mathscr{E}$. As $\mathscr{E}$ increases, the kinetic energy of the carriers rises. When the energy of a carrier exceeds the optical phonon energy,[4] it generates an optical phonon and loses much of its velocity. Consequently, the kinetic energy and therefore the drift velocity cannot exceed a certain value. The limiting velocity is called the **saturation velocity**. The $v$–$\mathscr{E}$ relationship is shown in Fig. 6–22.

The flattening of the $v$–$\mathscr{E}$ curve is called **velocity saturation** and can be approximated with

$$\text{v} = \frac{\mu_{ns}\mathscr{E}}{1 + \mathscr{E}/\mathscr{E}_{\text{sat}}} \tag{6.8.1}$$

where $\mu_{ns}$ is the electron surface mobility and $\mathscr{E}_{\text{sat}}$ is the field at which velocity saturation becomes significant or dominant. When $\mathscr{E} << \mathscr{E}_{\text{sat}}$, Eq. (6.8.1) reduces to $v = \mu\mathscr{E}$. When $\mathscr{E} >> \mathscr{E}_{\text{sat}}$, $v$ is a constant regardless of how large $\mathscr{E}$ is. *Velocity saturation has a large and deleterious effect on the $I_{\text{on}}$ of MOSFETs.*

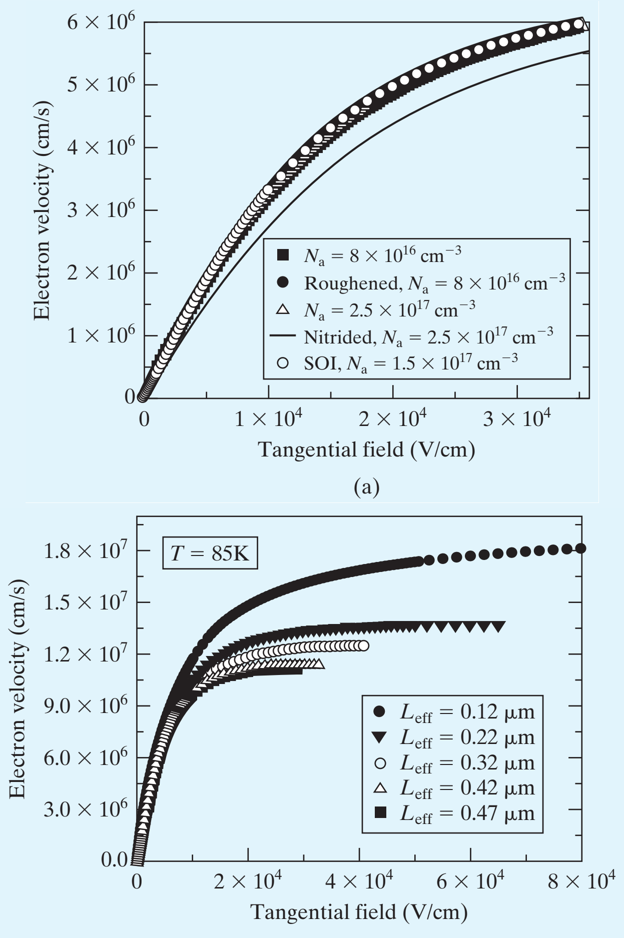
### ● Velocity Overshoot ●

Figure 6–22b shows the $v$–$\mathscr{E}$ characteristics of inversion-layer electrons at 85 K [8]. This is offered as clearer evidence that velocity saturates at high field than the room-temperature data (Fig. 6–22a). Because the velocity saturation phenomenon is clearer, we can see an important detail—$v_{\text{sat}}$ is larger in transistors with very small channel lengths.

In the basic velocity-saturation model, $v_{\text{sat}}$ is independent of the channel length. However, this figure shows that $v_{\text{sat}}$ becomes larger when $L$ is very small. When the channel length is sufficiently small, electrons may pass through the channel in too short a time for all the energetic carriers to lose energy by emitting optical phonons. As a result, the carriers can attain somewhat higher velocities in very small devices. This phenomenon is called **velocity overshoot**.

Velocity overshoot frees the extremely short transistors from the limit of velocity saturation. Unfortunately, another velocity limit (see Section 6.12) sets in before velocity overshoot offers a lot of relief.

[4] Optical phonon is a type of phonons (atom vibration) that has much higher energy than the acoustic phonons that are partially responsible for the low-field mobility (see Section 2.2.2). The optical phonons involve large displacements of neighboring atoms. These displacements create electrical dipole field that interact very strongly with electrons and holes. An electron or a hole that has enough energy to generate an optical phonon will do so readily and lose its kinetic energy in the process.



(a)

(b)

FIGURE 6–22 (a) The inversion-layer electron velocity saturates at high field regardless of the body doping concentration and surface treatment. (b) Velocity saturation is more prominent at low temperature. Velocity overshoot is also evident. (From [8]). © 1997 IEEE.)

## 6.9 • MOSFET IV MODEL WITH VELOCITY SATURATION •

The basic MOSFET IV theory presented in Section 6.6 assumes a constant mobility. It provides an excellent introduction to the theory of MOSFET. The present section refines the theory by including the important velocity saturation effect. If we apply Eq. (6.8.1) to Eq. (6.6.1), using an NMOSFET for example

$$I_{ds} = WC_{oxe}(V_{gs} - mV_{cs} - V_t)\frac{\mu_{ns}\mathrm{d}V_{cs}/\mathrm{d}x}{1 + \frac{\mathrm{d}V_{cs}}{\mathrm{d}x}/\mathscr{E}_{sat}} \qquad (6.9.1)$$

$$\int_0^L I_{\mathrm{ds}}\mathrm{d}x = \int_0^{V_{\mathrm{ds}}}[WC_{\mathrm{oxe}}\mu_{ns}(V_{\mathrm{gs}}-mV_{\mathrm{cs}}-V_{\mathrm{t}})-I_{\mathrm{ds}}/\mathscr{E}_{\mathrm{sat}}]\mathrm{d}V_{\mathrm{cs}} \tag{6.9.2}$$

$$I_{\mathrm{ds}} = \frac{\frac{W}{L}C_{\mathrm{oxe}}\mu_{ns}\left(V_{\mathrm{gs}}-V_{\mathrm{t}}-\frac{m}{2}V_{\mathrm{ds}}\right)V_{\mathrm{ds}}}{1+\frac{V_{\mathrm{ds}}}{\mathscr{E}_{\mathrm{sat}}L}} \tag{6.9.3}$$

When $L$ is large, Eq. (6.9.3) reduces to Eq. (6.6.4). Therefore the latter is known as the **long-channel IV model**.

$$\boxed{I_{\mathrm{ds}} = \frac{\text{long-channel } I_{\mathrm{ds}} \text{ (Eq. (6.6.4))}}{1+V_{\mathrm{ds}}/\mathscr{E}_{\mathrm{sat}}L}} \tag{6.9.4}$$

The effect of velocity saturation is to reduce $I_{\mathrm{ds}}$ by a factor of $1 + V_{\mathrm{ds}}/\mathscr{E}_{\mathrm{sat}}L$. This factor reduces to one (i.e., velocity saturation becomes negligible) when $V_{\mathrm{ds}}$ is small or $L$ is large. This factor may be interpreted as $1 + \mathscr{E}_{\mathrm{ave}}/\mathscr{E}_{\mathrm{sat}}$, where $\mathscr{E}_{\mathrm{ave}} \equiv V_{\mathrm{ds}}/L$ is the average channel field. The saturation voltage, $V_{\mathrm{dsat}}$, can be found by solving $\mathrm{d}I_{\mathrm{ds}}/\mathrm{d}V_{\mathrm{ds}} = 0$:

$$V_{\mathrm{dsat}} = \frac{2(V_{\mathrm{gs}}-V_{\mathrm{t}})/m}{1+\sqrt{1+2(V_{\mathrm{gs}}-V_{\mathrm{t}})/m\mathscr{E}_{\mathrm{sat}}L}} \tag{6.9.5}$$

Equation (6.9.5) is rather inconvenient to use. A simpler and even more accurate $V_{\mathrm{dsat}}$ model may be derived from a piece-wise model that actually fits the $v$–$\mathscr{E}$ data better than Eq. (6.8.1)[9]. It assumes that

$$v = \frac{\mu_{ns}\mathscr{E}}{1+\mathscr{E}/\mathscr{E}_{\mathrm{sat}}} \qquad \text{for} \quad \mathscr{E} \le \mathscr{E}_{\mathrm{sat}} \tag{6.9.6}$$

$$v = v_{\mathrm{sat}} \qquad \text{for} \quad \mathscr{E} \ge \mathscr{E}_{\mathrm{sat}} \tag{6.9.7}$$

Equating Eqs. (6.9.6) and (6.9.7) at $\mathscr{E} = \mathscr{E}_{\mathrm{sat}}$ yields

$$\mathscr{E}_{\mathrm{sat}} = 2v_{\mathrm{sat}}/\mu_{ns} \tag{6.9.8}$$

Equation (6.9.6) leads to Eq. (6.9.3), which is valid when the carrier speed is less than $v_{\mathrm{sat}}$, i.e., $V_{\mathrm{ds}} \le V_{\mathrm{dsat}}$. Equation (6.9.7) leads to the following equation describing the current at the drain end of the channel at the onset of velocity saturation (i.e., at $V_{\mathrm{d}} = V_{\mathrm{dsat}}$):

$$\begin{aligned} I_{\mathrm{ds}} &= WQ_{\mathrm{inv}}v \\ &= WC_{\mathrm{oxe}}(V_{\mathrm{g}}-V_{\mathrm{t}}-mV_{\mathrm{dsat}})v_{\mathrm{sat}} \end{aligned} \tag{6.9.9}$$

Equating Eqs. (6.9.3) and (6.9.9) leads to

$$\boxed{\frac{1}{V_{\mathrm{dsat}}} = \frac{m}{V_{\mathrm{gs}}-V_{\mathrm{t}}}+\frac{1}{\mathscr{E}_{\mathrm{sat}}L}} \tag{6.9.10}$$

$V_{dsat}$ in Eq. (6.9.6) is an average of $\mathcal{E}_{sat}L$ and the long-channel $V_{dsat}$, $(V_{gs} - V_t)/m$ [Eq. (6.6.5)]. It is smaller than the latter. Note that $\mathcal{E}_{sat}$ is defined with Eq. (6.9.8).[5] It is known that $v_{sat}$ is $8 \times 10^6$ cm/s for electrons and $6 \times 10^6$ cm/s for holes.

### EXAMPLE 6–2 Drain Saturation Voltage

At $V_{gs} = 1.8$ V, what is the $V_{dsat}$ of an NMOSFET with $T_{oxe} = 3$ nm, $V_t = 0.25$ V, and $W_{dmax} = 45$ nm for (a) $L = 10$ µm, (b) $L = 1$ µm, (c) $L = 0.1$ µm, and (d) $L = 0.05$ µm?

**SOLUTION:**

From Fig. 6–9 or Eq. (6.3.7), $\mu_n$ is 200 cm$^2$/V/s. Using Eq. (6.9.8)

$$\mathcal{E}_{sat} = 2\text{v}_{sat}/\mu_{ns} = 2 \times 8 \times 10^6\text{cm/s} \div 200\text{ cm}^2/\text{Vs} = 8 \times 10^4\text{V/cm}$$

Using Eq. (6.5.2)

$$m = 1 + 3T_{oxe}/W_{dmax} = 1 + 9\text{ nm}/45\text{ nm} = 1.2$$

Using Eq. (6.9.10)

$$V_{dsat} = \left(\frac{m}{V_{gs} - V_t} + \frac{1}{\mathcal{E}_{sat}L}\right)^{-1}$$

**a.** $L = 10$ µm,

$$V_{dsat} = \left(\frac{1.2}{1.55\text{ V}} + \frac{1}{8 \times 10^4\text{ V/cm} \cdot L}\right)^{-1} = \left(\frac{1}{1.3\text{ V}} + \frac{1}{80\text{ V}}\right)^{-1} = 1.3\text{ V}$$

**b.** $L = 1$ µm,

$$V_{dsat} = \left(\frac{1}{1.3\text{ V}} + \frac{1}{8\text{ V}}\right)^{-1} = 1.1\text{ V}$$

**c.** $L = 0.1$ µm

$$V_{dsat} = \left(\frac{1}{1.3\text{ V}} + \frac{1}{0.8\text{ V}}\right)^{-1} = 0.5\text{ V}$$

**d.** $L = 0.05$ µm

$$V_{dsat} = \left(\frac{1}{1.3\text{ V}} + \frac{1}{0.4\text{ V}}\right)^{-1} = 0.3\text{ V}$$

Clearly, short-channel $V_{dsat}$ is much smaller than long-channel $V_{dsat}$, $V_g - V_t$.

Substituting Eq. (6.9.10) for $V_{ds}$ in Eq. (6.9.3)

$$\boxed{I_{dsat} = \frac{W}{2\text{mL}}C_{oxe}\mu_{ns}\frac{(V_{gs} - V_t)^2}{1 + \dfrac{V_{gs} - V_t}{\text{m}\mathcal{E}_{sat}L}} = \frac{\text{long channel } I_{dsat}\text{ (Eq. (6.6.6))}}{1 + \dfrac{V_{gs} - V_t}{\text{m}\mathcal{E}_{sat}L}}} \quad (6.9.11)$$

[5] You may find this $\mathcal{E}_{sat}$ definition to be inconsistent with Eq. (6.8.1). Equations (6.9.6)–(6.9.8) match the sharp curvature and the asymptotic values of the velocity-field data better than Eq. (6.8.6) [9].

Two special cases of Eqs. (6.9.10) and (6.9.11) are discussed below.

1. Long-channel or low $V_{gs}$ case, $\mathcal{E}_{sat}L >> V_{gs} - V_t$

$$V_{dsat} = (V_{gs} - V_t)/m \tag{6.9.12a}$$

$$I_{dsat} = \frac{W}{2\mathrm{mL}} C_{oxe} \mu_{ns} (V_{gs} - V_t)^2 \tag{6.9.12b}$$

These are identical to Eqs. (6.6.5) and (6.6.6). The long-channel model is valid when $L$ is large.

### ● How Large Must *L* Be to Be "Long Channel"? ●

The condition $\mathcal{E}_{sat}L >> V_{gs} - V_t$ can be satisfied when $L$ is large or when $V_{gs}$ is close to $V_t$. The latter case is frequently encountered in analog circuits where the gate is biased close to $V_t$ to reduce power consumption. Assuming $\mathcal{E}_{sat} = 6 \times 10^4$ V/cm and $V_{gs}\ \ V_t = 2$ V (for digital circuits), a 0.2 µm channel length would not satisfy the condition of $\mathcal{E}_{sat}L >> V_{gs} - V_t$. Therefore, it exhibits significant short-channel behaviors. But, read on. If $V_{gs} - V_t = 0.1$ V (for low-power analog circuits), even a 0.1 µm channel length would satisfy the inequality and the transistor would exhibit some *long-channel* characteristics, i.e., $I_{dsat} \propto (V_{gs} - V_t)^2/L$ and $V_{dsat} = (V_{gs} - V_t)/m$. For applications to this low-power analog circuit, the "long-channel" equations such as Eq. (6.6.6) may be used even if $L$ is 0.05 µm.

There are other short-channel behaviors that are observable even at small $V_{gs} - V_t$, e.g., a larger leakage current and a larger slope in the $I_d - V_d$ plot at $V_{ds} > V_{dsat}$. These other behaviors are sensitive to transistor design parameters such as $\mathrm{T}_{oxe}$ as explained in the next chapter.

2. Very short-channel case, $\mathcal{E}_{sat}L << V_{gs} - V_t$

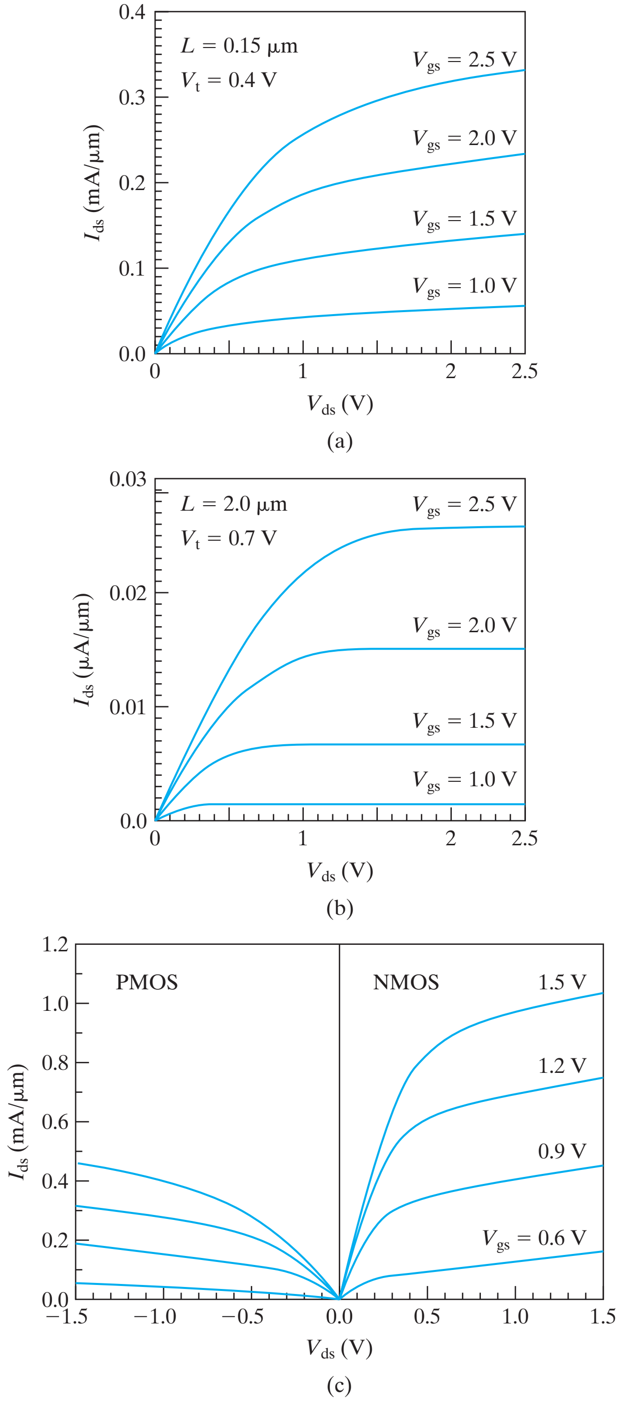
$$V_{dsat} \approx \mathcal{E}_{sat}L < \frac{(V_g - V_t)}{\mathrm{m}} \tag{6.9.13}$$

$$I_{dsat} \approx W v_{sat} C_{oxe} (V_{gs} - V_t - m\mathcal{E}_{sat}L) \tag{6.9.14}$$

$I_{dsat}$ is proportional to $V_{gs} - V_t$ rather than $(V_{gs} - V_t)^2$ and is less sensitive to $L$ than the long-channel $I_{dsat}$ ($\propto 1/L$). Equation (6.9.14), derived from Eq. (6.9.11) by Taylor expansion, is quite easy to understand. $I_{dsat}$ is proportional to $W$. Carriers travel at the saturation velocity at the drain end of the channel where $Q_{inv} = C_{oxe}(V_{gs} - V_t - mV_{dsat})$ and $V_{dsat}$ is $\mathcal{E}_{sat}L$.

Figure 6–23a and b compare the measured IV characteristics of two NFETs with $L = 0.15$ µm and $L = 2$ µm. The shorter channel device shows an approximately linear relationship between $I_{dsat}$ and $V_{gs}$ in agreement with Eq. (6.9.14). $V_{dsat}$ is significantly less than $(V_{gs} - V_t)/m$. (The behavior at $V_{ds} > V_{dsat}$ is explained in Sec. 7.9.) The 2 µm channel device shows a superlinear increase of $I_{dsat}$ with increasing $V_g$ in rough agreement with Eq. (6.9.12).

To raise $I_{dsat}$, we must increase $C_{oxe}(V_{gs} - V_t)$, i.e., reduce $T_{oxe}$, minimize $V_t$, and use high $V_{gs}$. The limit of $T_{oxe}$ is set by oxide tunneling leakage and reliability. The lower limit of $V_t$ is set by MOSFET leakage in the off state. These will be discussed in the next chapter. The maximum $V_{gs}$ is the power supply voltage, $V_{dd}$, which is limited by concerns over circuit power consumption and device reliability.



(a)

(b)

(c)

FIGURE 6–23 **Measured IV characteristics**. (a) A 0.15 μm channel device ($V_t = 0.4$ V) shows a linear relationship between $I_{dsat}$ and $V_{gs}$. $V_{dsat}$ is significantly less than $V_{gs} - V_t$. (b) A 2 μm device ($V_t = 0.7$ V) exhibits the $I_{dsat} \propto (V_{gs} - V_t)^2$ relationship. (c) IV characteristics of PFET and NFET with $T_{oxe} = 3$ nm and $L \approx 100$ nm.

Figure 6–23c shows that PFET and NFET have similar IV characteristics, e.g., both exhibit a linear $I_{\mathrm{dsat}}$–$V_{\mathrm{g}}$ relationship. $I_{\mathrm{P}}$ is about half of $I_{\mathrm{N}}$. The holes' mobility is three times smaller and their saturation velocity is 30% smaller than that of the electrons.

### 6.9.1 Velocity Saturation vs. Pinch-Off

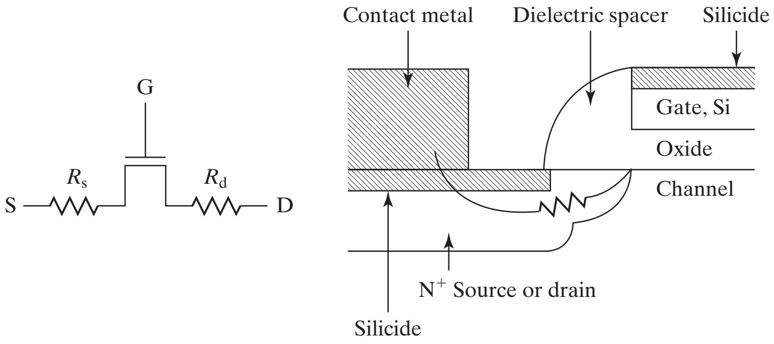
The concept of pinch-off in Section 6.6 suggests that $I_{\mathrm{ds}}$ saturates when $Q_{\mathrm{inv}}$ becomes zero at the drain end of the channel. A more accurate description of the cause of current saturation is that the carrier velocity has reached $v_{\mathrm{sat}}$ at the drain. Instead of the pinch-off region, there is a **velocity saturation region** next to the drain where $Q_{\mathrm{inv}}$ is a constant ($I_{\mathrm{dsat}}/Wv_{\mathrm{sat}}$). The series of plots in Fig. 6–17 are still valid with one modification. In (b) and (f), $Q_{\mathrm{inv}} = I_{\mathrm{dsat}}/Wv_{\mathrm{sat}}$ at $L$. In (f), of course, there is a very short region next to $L$, the velocity saturation region, where $Q_{\mathrm{inv}}$ remains constant. This region is not shown in Fig. 6–17 for simplicity.

## 6.10 ● PARASITIC SOURCE-DRAIN RESISTANCE ●

The main effect of the parasitic resistance shown in Fig. 6–24a is that $V_{\mathrm{gs}}$ in the $I_{\mathrm{ds}}$ equations is reduced by $R_{\mathrm{s}} \cdot I_{\mathrm{ds}}$. For example, Eq. (6.9.14) becomes

$$I_{\mathrm{dsat}} = \frac{I_{\mathrm{dsat0}}}{1 + R_{\mathrm{s}} I_{\mathrm{dsat0}}/(V_{\mathrm{gs}} - V_{\mathrm{t}})} \tag{6.10.1}$$

$I_{\mathrm{dsat0}}$ is the current in the absence of $R_{\mathrm{s}}$. $I_{\mathrm{dsat}}$ may be significantly reduced by the parasitic resistance, and the impact is expected to rise in the future. The shallow diffusion region under the dielectric spacer is a contributor to the parasitic resistance. The shallow junction is needed to prevent excessive off-state leakage $I_{\mathrm{ds}}$ in short-channel transistors (see Section 7.6). The silicide (e.g., $\mathrm{TiSi_2}$ or $\mathrm{NiSi_2}$) reduces the **sheet resistivity**[6] of the $\mathrm{N^+}$ (or $\mathrm{P^+}$) source–drain regions by a factor of ten. It also reduces the



(a) (b)

FIGURE 6–24 Source–drain series resistance.

[6] If the sheet resistivity of a film is 1 Ω per square, the resistance between two opposite edges of a square-shaped piece of this film (regardless of the size of the square) will be 1 Ω.

**contact resistance** between the silicide and the N$^+$ or P$^+$ Si. The contact resistance is another main source of resistance and more on this subject may be found in Section 4.21. The dielectric spacer is produced by coating the structure in Fig. 5–1 with a conformal film of dielectric followed by anisotropic dry etching to remove the dielectric from the horizontal surfaces. The silicides over the source/drain diffusion regions and over the gate are formed simultaneously by reaction between metal and silicon at a high temperature. The unreacted metal over the surface of the dielectric spacer is removed with acid. A second effect of the series resistance is an increase in $V_{\text{dsat}}$:

$$V_{\text{dsat}} = V_{\text{sat0}} + I_{\text{dsat}}(R_{\text{s}} + R_{\text{d}}) \qquad (6.10.2)$$

where $V_{\text{dsat0}}$ is the $V_{\text{dsat}}$ in the absence of $R_{\text{s}}$ and $R_{\text{d}}$.

## 6.11 • EXTRACTION OF THE SERIES RESISTANCE AND THE EFFECTIVE CHANNEL LENGTH[7] •

Figure 6–25 illustrates the channel length and two other related quantities. A circuit designer specifies a channel length in the circuit layout, called the drawn gate length, $L_{\text{drawn}}$. This layout is transferred to a photomask, then to a photoresist pattern, and finally to the physical gate. The final physical **gate length**, $L_{\text{g}}$, may not be equal to $L_{\text{drawn}}$ because each pattern transfer can introduce some dimensional change. However, engineers devote extraordinary efforts, e.g., by OPC (optical proximity correction) (see Section 3.3) to minimize the difference between $L_{\text{drawn}}$ and $L_{\text{g}}$. As a result, one may assume $L_{\text{drawn}}$ and $L_{\text{g}}$ to be equal. $L_{\text{g}}$ can be measured using scanning electron microscopy (SEM).

For device analysis and modeling, it is necessary to know the channel length, $L$, also called the effective channel length ($L_{\text{eff}}$) or the electrical channel length ($L_{\text{e}}$) to differentiate it from $L_{\text{drawn}}$ and $L_{\text{g}}$. It is particularly useful to know the

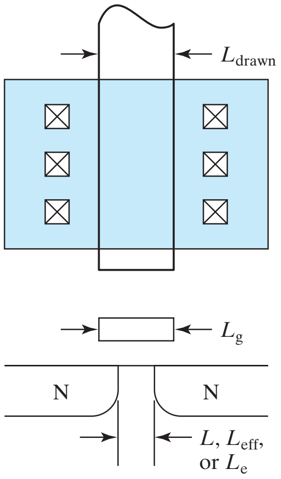


FIGURE 6–25 $L_{\text{drawn}}$, $L_{\text{g}}$, and $L$ (also known as $L_{\text{eff}}$ or $L_{\text{e}}$) are different in general.

[7] This section may be omitted in an accelerated course.

difference between $L_{\text{drawn}}$ and $L$. This difference is called $\Delta L$, which is assumed to be a constant, independent of $L_{\text{drawn}}$

$$L = L_{\text{drawn}} - \Delta L \tag{6.11.1}$$

Measuring $\Delta L$ in short transistors is quite difficult. There are several imperfect options. The following method is the oldest and still commonlly used. From Eq. (6.3.1),

$$V_{\text{ds}} = \frac{I_{\text{ds}}(L_{\text{drawn}} - \Delta L)}{WC_{\text{oxe}}(V_{\text{gs}} - V_{\text{t}})\mu_{ns}} \tag{6.11.2}$$

When the series resistance, $R_{\text{ds}} \equiv R_{\text{d}} + R_{\text{s}}$, shown is Fig. 6–24a is included, Eq. (6.11.2) becomes

$$V_{\text{ds}} = I_{\text{ds}}R_{\text{ds}} + \frac{I_{\text{ds}}(L_{\text{drawn}} - \Delta L)}{WC_{\text{oxe}}(V_{\text{gs}} - V_{\text{t}})\mu_{ns}} \tag{6.11.3}$$

$$\frac{V_{\text{ds}}}{I_{\text{ds}}}(= R_{\text{ds}} + \text{channel resistance }) = R_{\text{ds}} + \frac{L_{\text{drawn}} - \Delta L}{WC_{\text{oxe}}(V_{\text{gs}} - V_{\text{t}})\mu_{ns}} \tag{6.11.4}$$

Figure 6–26 plots the measured $V_{\text{ds}}/I_{\text{ds}}$ against $L_{\text{drawn}}$ using three MOSFETs that are identical (fabricated on the same test chip) except for their $L_{\text{drawn}}$s. $I_{\text{ds}}$ is measured at a small $V_{\text{ds}}$ (≤ 50 mV) and at least two values of $V_{\text{gs}} - V_{\text{t}}$. $V_{\text{ds}}/I_{\text{ds}}$ is a linear function of $L_{\text{drawn}}$. The two straight lines intersect at a point where $V_{\text{ds}}/I_{\text{ds}}$ is independent of $V_{\text{gs}} - V_{\text{t}}$ according to Eq. (6.11.4), i.e., where $L_{\text{drawn}} = \Delta L$ and $V_{\text{ds}}/I_{\text{ds}} = R_{\text{ds}}$. Once $\Delta L$ is known, $L$ can be calculated using Eq. (6.11.1).

Detailed measurements indicate that $R_{\text{ds}}$ tends to decrease with increasing $V_{\text{g}}$. One reason is that the gate voltage induces more (accumulation) electrons in the source–drain diffusion region and therefore reduces $R_{\text{ds}}$. More puzzling is the observation that $\Delta L$ decreases (or $L$ increases) with increasing $V_{\text{g}}$. The dependence of both $\Delta L$ and $R_{\text{ds}}$ on $V_{\text{g}}$ suggests the interpretation of channel length illustrated in Fig. 6–27 [10]. The sheet conductivities (inverse of sheet resistivity, introduced in Section 6.10) of the source–drain diffusion regions and the channel inversion layer (the horizontal lines) are plotted. The inversion-layer sheet conductivity increases with increasing $V_{\text{g}}$, of course. The channel length may be interpreted as the length of the part of the channel where the inversion-layer sheet conductivity is larger than the source/drain sheet conductivity. In other words, the channel is where the

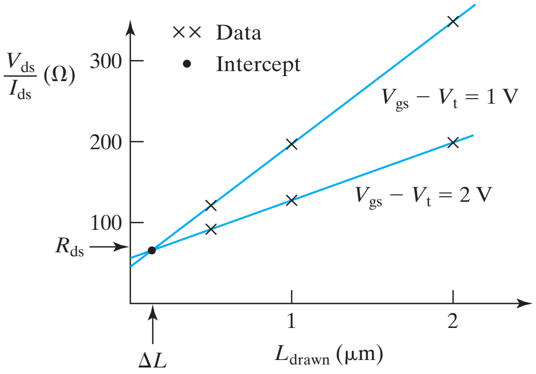


FIGURE 6–26 Method of extracting $R_{\text{ds}}$ and $\Delta L$.

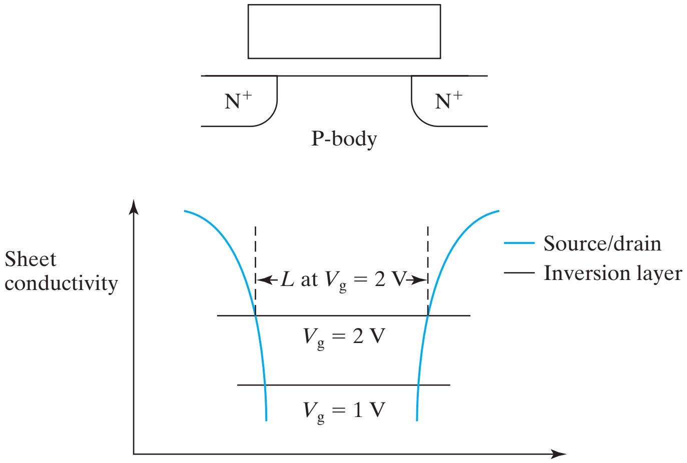


FIGURE 6–27 Interpretation of channel length and its dependence on $V_g$.

conductivity is determined by $V_g$, not by the source–drain doping profiles. Any resistance from outside the "channel" is attributed to $R_{ds}$. It is clear from Fig. 6–27 that the channel expands (i.e., $L$ increases and $R_{ds}$ decreases) with increasing $V_g$.

## 6.12 • VELOCITY OVERSHOOT AND SOURCE VELOCITY LIMIT[8] •

The concept of mobility is dubious when the channel length is comparable to or smaller than the mean free path (see Section 2.2.2). For this reason, Eq. (6.9.14) is particularly interesting because it does not contain mobility. The carrier velocity at the drain end of the channel is limited by the saturation velocity, which determines $I_{dsat}$. However, when the channel length is reduced much below 100 nm, the saturation velocity may be greatly raised by velocity overshoot as explained in Section 6.8. In that case, some other limit on $I_{dsat}$ may set in.

The carrier velocity at the source becomes the limiting factor. There, the velocity is limited by the thermal velocity, with which the carriers enter the channel from the source. This is known as the **source injection velocity** limit.

The source is a reservoir of carriers moving at the thermal velocity. As the channel length approaches zero, all the carriers moving from the source into the channel are captured by the drain. No carriers flow from the drain to the source due to the voltage difference (or energy barrier) shown in Fig. 6–28.

$$I_{dsat} = WBv_{thx}Q_{inv} = WBv_{thx}C_{oxe}(V_{gs} - V_t) \qquad (6.12.1)$$

Equation (6.12.1) is similar to Eq. (6.9.14) except that $v_{sat}$ is replaced by $v_{thx}$, the $x$-direction component of the thermal velocity. Thorough analysis of $v_{thx}$ shows that $v_{thx}$ is about $1.6 \times 10^7$ cm/s for electrons and $1 \times 10^7$ cm/s for holes in silicon MOSFETs [11]. $B$ is the fraction of carriers captured by the drain in a real transistor. The rest of the injected carriers are scattered back toward the source.

[8] This section may be omitted in an accelerated course.

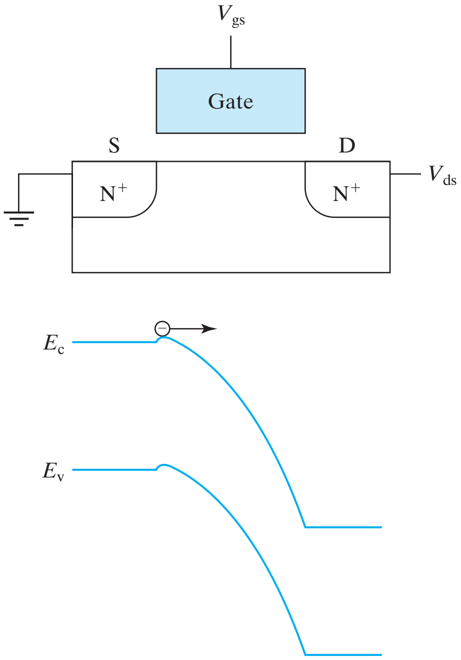


FIGURE 6–28 In the limit of no scattering in a very short channel, carriers are injected from the source into the channel at the thermal velocity and travel ballistically to the drain.

A particle simulation technique called the Monte Carlo simulation arrived at 0.5 as a typical value of $B$ [11]. This makes Eq. (6.12.1) practically identical to Eq. (6.9.14) because $v_{sat}$ is about $8 \times 10^6$ cm/s for electrons and $6 \times 10^6$ cm/s for holes. Both the drain-end velocity saturation limit and the source-end injection velocity limit predict similar $I_{dsat}$. $B$ in Eq. (6.12.1) is expected to increase somewhat with decreasing $L$ as $v_{sat}$ in Eq. (6.9.14) is expected to do, too.

## 6.13 • OUTPUT CONDUCTANCE •

The saturation of $I_{ds}$ (at $V_{ds} > V_{dsat}$) is rather clear in Fig. 6–23b. The saturation of $I_{ds}$ in Fig. 6–23a is gradual and incomplete. The cause for the difference is that the channel length is long in the former case and short in the latter. The slope of the $I$–$V$ curve is called the **output conductance**

$$g_{ds} = \frac{dI_{dsat}}{dV_{ds}} \tag{6.13.1}$$

A clear saturation of $I_{ds}$, i.e., a small $g_{ds}$ is desirable. The reason can be explained with the simple amplifier circuit in Fig. 6–29. The bias voltages are chosen such that the transistor operates in the saturation region. A small-signal input, $v_{in}$, is applied.

$$i_{ds} = g_{msat} \cdot v_{gs} + g_{ds} \cdot v_{ds} \tag{6.13.2}$$

$$= g_{msat} \cdot v_{in} + g_{ds} \cdot v_{out}$$

$$v_{out} = -R \times i_{ds} \tag{6.13.3}$$

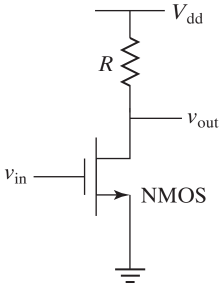


FIGURE 6–29 A simple MOSFET amplifier.

Eliminate $i_{ds}$ from the last two equations and we obtain

$$v_{out} = \frac{-g_{msat}}{g_{ds} + 1/R} \times v_{in} \tag{6.13.4}$$

The magnitude of the output voltage, according to Eq. (6.13.4) is amplified from the input voltage by a **gain** factor of $\frac{g_{msat}}{g_{ds} + 1/R}$. The gain can be increased by using a large $R$. Even with $R$ approaching infinity, the voltage gain cannot exceed

$$\text{Maximum Voltage Gain} = \frac{g_{msat}}{g_{ds}} \tag{6.13.5}$$

This is the **intrinsic voltage gain** of the transistor. If $g_{ds}$ is large, the voltage gain will be small. As an extreme example, the maximum gain will be only 1 if $g_{ds}$ is equal to $g_{msat}$. A large gain is obviously beneficial to analog circuit applications. A reasonable large gain is also needed to obtain a steep transition in the VTC, i.e., needed for digital circuit applications to enhance noise immunity. Therefore, $g_{ds}$ must be kept much lower than $g_{msat}$.

The physical causes of the output conductance are the influence of $V_{ds}$ on $V_t$ and a phenomenon called channel length modulation. They are discussed in Section 7.9. The conclusions may be summed up this way. In order to achieve a small $g_{ds}$ and a large voltage gain, $L$ should be large and/or $T_{ox}$, $W_{dep}$, and $X_j$ should be small.

## 6.14 • HIGH-FREQUENCY PERFORMANCE •

The high-frequency performance of the MOSFET shown in Fig. 6–30a is limited by the input RC time constant. $C$ is the gate capacitance, $C_{ox}WL_g$. At high frequencies, the gate capacitive impedance, 1/2πfC, decreases and the gate AC current increases. More of the gate signal voltage is dropped across $R_{in}$, and the output current is reduced. At some high frequency, the output current becomes equal to the input current. This unit current-gain frequency is called the ***cutoff frequency***, $f_T$. In narrow-band analog circuits operating at a particular high frequency, the gate capacitance may be compensated with an on-chip inductor at that frequency to

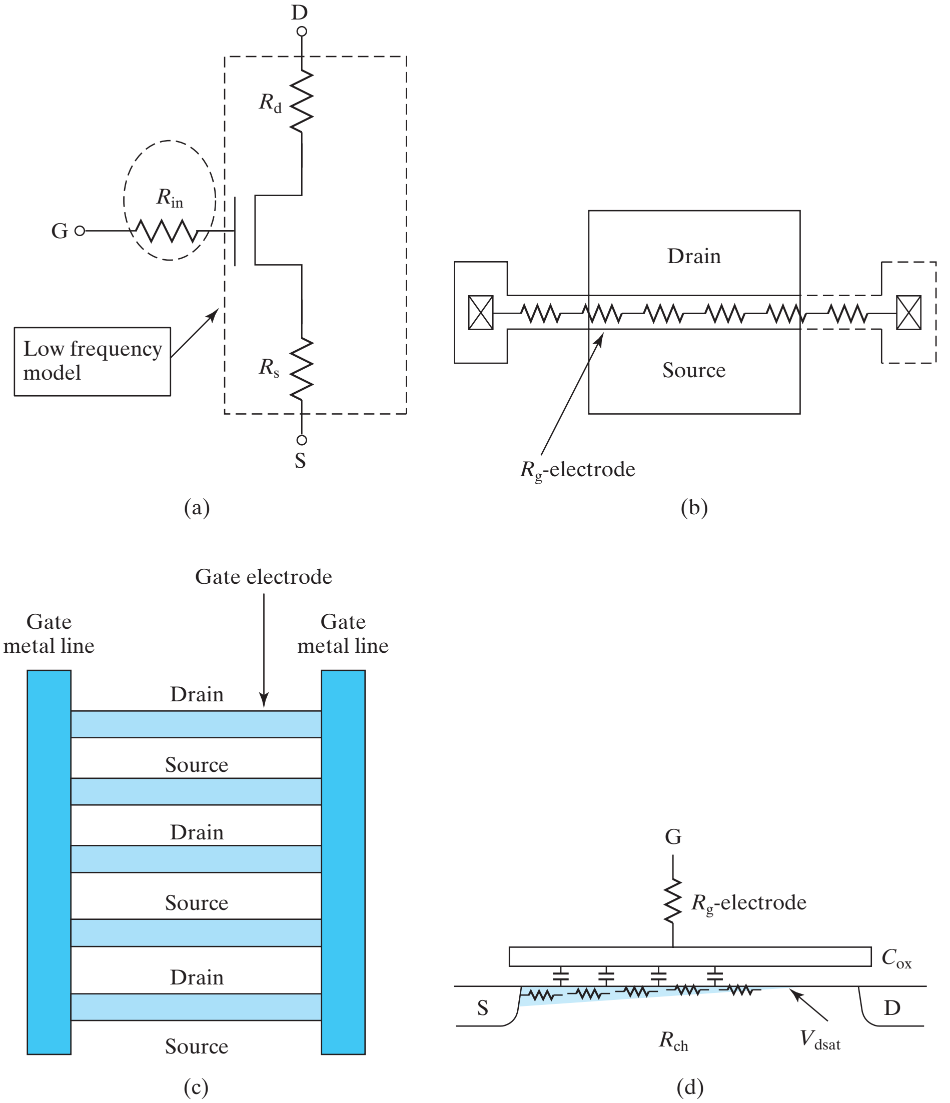


FIGURE 6–30 (a) The input resistance together with the input capacitance sets the high-frequency limit. (b) One component of $R_{in}$ is the gate-electrode resistance. (c) The multi-finger layout dramatically reduces the gate-electrode resistance. (d) The more fundamental and important component of $R_{in}$ is the channel resistance, which is also in series with the gate capacitor.

overcome the $f_T$ limit. In that case, $R_{in}$ still consumes power and at some frequency, typically somewhat higher than $f_T$, the power gain drops to unity. This frequency is called the **maximum oscillation frequency**, $f_{max}$. In either case, it is important to minimize $R_{in}$.

$R_{in}$ consists of two components, the **gate-electrode resistance**, $R_{g\text{-electrode}}$, and the **intrinsic input resistance**, $R_{ii}$.

$$R_{in} = R_{g-\text{electrode}} + R_{ii} \qquad (6.14.1)$$

The gate-electrode resistance is straightforward as shown in Fig. 6–30b. A powerful way to reduce the gate-electrode resistance is **multi-finger layout** shown in Fig. 6–30c, which means designing a MOSFET with a large channel width, say 10 μm, as 10 MOSFETs connected in parallel each having a width of 1 μm. This reduces the gate-electrode resistance by a factor of 100 because each finger's resistance is ten times smaller and there are now ten finger resistors in parallel.

$$R_{g-\text{electrode}} = \rho W/12 T_g L_g N_{f2} \tag{6.14.2}$$

$\rho$ is the gate resistivity of the gate material, $W$, is the total channel width, $T_g$ is the gate thickness, $L_g$ is the gate length, and $N_f$ is the number of fingers. The factor 12 comes from two sources. A factor of three comes from the fact that the gate current is distributed over the finger width and all the gate capacitor current does not flow through the entire finger resistor. The remaining factor of four arises from contacting the gate fingers at both the left and the right ends of the fingers as shown in Fig. 6–30c. Doing so effectively doubles the number of fingers and halves the finger width as if each finger is further divided into two at the middle of the finger. Using multifinger layout, the gate-electrode resistance can be quite low if the gate material is silicided poly-silicon. If the gate material is metal, this component of $R_{in}$ becomes negligible.

The more important, fundamental, and interesting component is the intrinsic input resistance. The concept is illustrated in Fig. 6–30d. Even if $R_{\text{g-electrode}}$ is zero, there is still a resistor in series with the gate capacitor. The gate capacitor current flows through the channel resistance, $R_{ch}$, to the source, then through the input signal source (not shown) back to the gate to complete the current loop. $R_{ii}$ is a resistance in the path of the gate current[12].

$$R_{ii} = \kappa\int dR_{ch} = \kappa\frac{V_{ds}}{I_{ds}} \tag{6.14.3}$$

k is a number smaller than one [12] because due to the distributed nature of the RC network in Fig. 6–30d, the capacitance current does not flow through the entire channel resistance. $V_{ds}$ Eq. (6.14.3) saturates at $V_{dsat}$ when $V_{ds} > V_{dsat}$.

With each new generation of MOSFET technology, the gate length is reduced making $R_{ii}$ smaller for a fixed $W$ due to larger $I_{ds}$ and smaller $V_{dsat}$. Furthermore, the input capacitance $C_{ox}WL_g$ is reduced somewhat when $L_g$ is made smaller although $C_{ox}$ is made larger ($T_{oxe}$ thinner) at the same time. As a result, $f_T$ and $f_{max}$ have been improving linearly with the gate length. They are about 200 GHz in the 45 nm technology node, sufficient for a wide range of new applications.

## 6.15 ● MOSFET NOISES ●

**Noise** is whatever that corrupts the desired signal. One type of noise, the inductive and capacitive interferences or **cross talk** created by the interconnect network, may be called external noise. This kind of noise is important but can be reduced in principle by careful shielding and isolation by the circuit designers. The other noise category is called **device noise** that is inherent to the electronic devices. This kind of noise is due to the random behaviors of the electric carriers inside the device that create voltage and current fluctuations measurable at the terminals of the device.

This section is concerned with the device noise. *Noise, power consumption, speed, and circuit size (cost) are the major circuit-design constraints.*

### 6.15.1 Thermal Noise of a Resistor

If a resistor is connected to the input of an oscilloscope, the noise voltage across the resistor can be observed as shown in Fig. 6–31a. The origin of the noise is the random thermal motion of the charge carriers shown in Fig. 2–1, and the noise is called the **thermal noise**. The noise contains many frequency components. If one inserts a frequency filter with bandwidth $\Delta f$ and measures the root-mean-square value of the noise in this frequency band, the results are

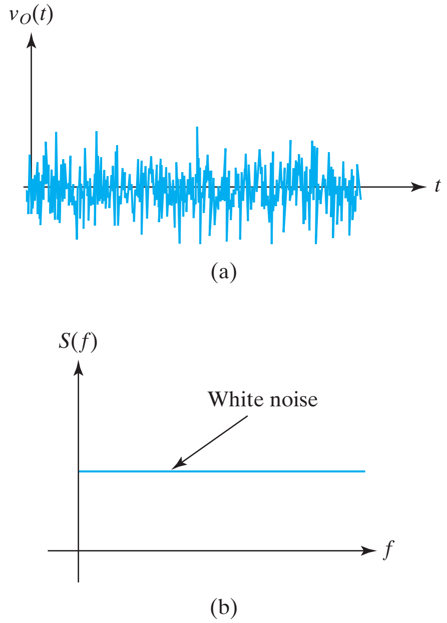
$$\overline{v_n^2} = S_{\mathrm{v}_n}\Delta f = 4\mathrm{kT}\Delta f R \qquad (6.15.1)$$

$$\overline{i_n^2} = \mathrm{S}_{\mathrm{i}_n}\Delta f = 4\mathrm{kT}\Delta f / R \qquad (6.15.2)$$

where $R$ is the resistance and Eq. (6.15.2) presents the noise current that would flow if the resistor's terminals were short-circuited. Clearly, the noise is proportional to $\Delta f$ but is independent of $f$. This characteristic is called **white noise** and its **noise spectral density** is shown in Fig. 6–31b. S is called the noise power density.

### 6.15.2 MOSFET Thermal Noise

The intrinsic thermal noise of MOSFETs originates from the channel resistance. The channel may be divided into many segments as shown in Fig. 6–32 and each contributes some noise. The channel noise voltage can be expressed by Eq. (6.15.1).



(a)

(b)

FIGURE 6–31 (a) The thermal noise voltage across a resistor and (b) the spectral density of white noise.

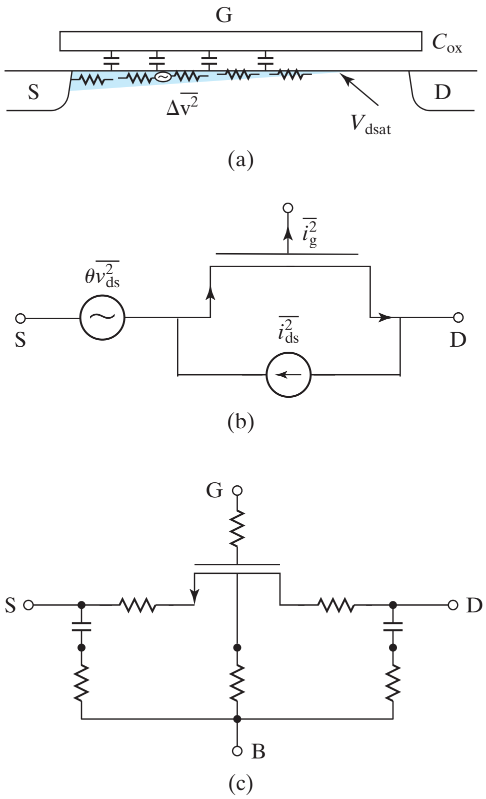


**FIGURE 6–32** (a) Each segment of the channel may be considered a resistor that contributes thermal noise. (b) The noise current is added to the normal MOSFET current as a parallel current source. The noise voltage is multiplied by the transconductance into another component of noise current. (c) Parasitic resistances also contribute to the thermal noises.

However, there are several theories of what value should be assigned to $R$. A classical and popular theory interprets it as $dV_{ds}/dI_{ds}$, or $1/g_{ds}$ in the linear (small $V_{ds}$) region, as shown in Eqs. (6.15.3) and (6.15.4). $\gamma$ is a function of $V_{ds}$ and $V_{gs}$. At $V_{ds} > V_{dsat}$, $\gamma$ saturates at 2/3. While this model works well at long-channel length, it underestimates the noise in short-channel MOSFETs. In circuit design practice, $\gamma$ is chosen to fit noise measurements to improve the accuracy of the noise model.

$$\overline{v_{ds}^2} = 4\gamma kT\Delta f/g_{ds} \tag{6.15.3}$$

$$\overline{i_{ds}^2} = 4\gamma kT\Delta f g_{ds} \tag{6.15.4}$$

As in a resistor, this white noise of Eqs. (6.15.4) presents itself as a parallel current source added to the regular MOSFET current in Fig. 6–32b.

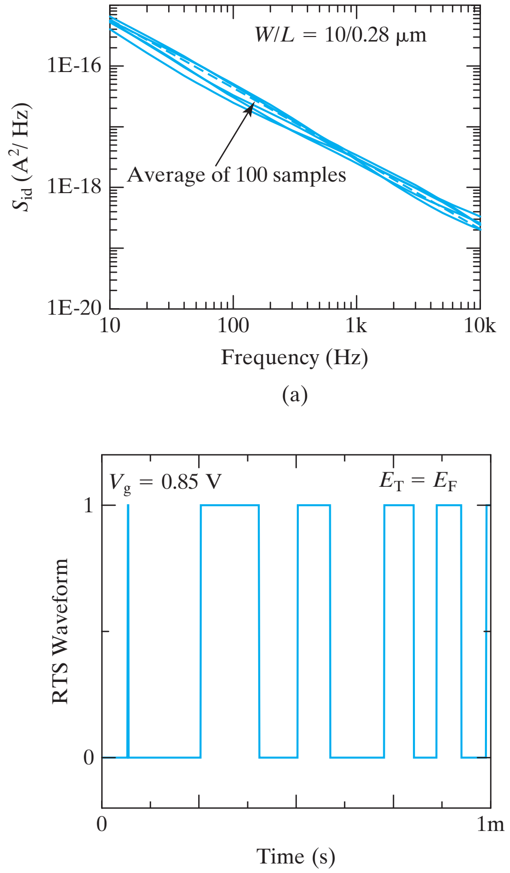
The channel noise voltage also induces a gate current through the gate capacitance. As a result, a portion of the channel noise current flows into the gate network. The gate noise current multiplied by the impedance of the gate input network and the transconductance produces a second noise current at the output. The complete model of the MOSFET noise therefore includes a partially correlated noise source appearing at the gate terminal. This effect can be approximately modeled by lumping the channel noise voltage at the source. $\theta$ in Fig. 6–32b is a function of $L$ and $V_{gs}$ and accounts for the fact that the noise voltage is actually distributed throughout the channel rather than lumped at the source [12,13]. Due to the partial correlation between the gate noise and the channel noise, the channel and gate noises can partially cancel each other at the output of the device. By optimizing the gate network impedance, design engineers can minimize the output noise.

The gate electrode, source, drain, and substrate parasitic resistances shown in Fig. 6–31c also contribute thermal noises. These resistances are usually minimized through careful MOSFET layout. It is important to reduce the gate electrode resistance as its noise is amplified by $g_{msat}$ into the $I_{ds}$ noise. The gate resistance can be minimized with the same **multifinger layout** discussed in Section 6.14.

### 6.15.3 MOSFET Flicker Noise

**Flicker noise**, also known as **1/*f* noise**, refers to a noise spectral density that is inversely proportional to the frequency as shown in Fig. 6–33a. The mechanism for flicker noise is the random capture and release of electrons by traps located in the gate dielectric. When a trap captures an electron from the inversion layer, there is one less electron to conduct current. Also the trap becomes charged and reduces the channel carrier mobility due to **Coulombic scattering** similar to the effect of an impurity ion (see Section 2.2.2). In other words, both the carrier number and the mobility fluctuate due to charge trapping and detrapping. In a MOFSET with very small $W$ and $L$, there is often only a single operative trap at a given bias condition and $I_{ds}$ fluctuates between a high and a low current level with certain average cycle period as shown in Fig. 6–33b. This noise is called the **random telegraph noise**. The two current states reflect the empty and filled states of the trap. In a larger area ($W \times L$) MOSFET, there are many traps. The traps located at or near the oxide–semiconductor interface can capture and release electrons with short time constants and they contribute mostly high-frequency noises. Traps located far from the interface have long time constants and contribute mostly low-frequency noises. It can be shown that adding these contributions up with the assumption of a uniform distribution of traps in the oxide leads to the $1/f$ noise spectrum [14].

$$\overline{i_{ds}^2} = \frac{KF \cdot W}{fL^2 C_{ox}}\left(\frac{I_{ds}}{W}\right)^{AF} \cdot kT\Delta f \qquad (6.15.7)$$



(a)

(b)

**FIGURE 6–33** (a) Flicker noise is also known as 1/*f* noise because the noise power density is proportional to 1/frequency. (b) In a MOSFET with very small *W* and *L*, there may be only one operative trap and $I_{ds}$ fluctuates between two levels. This is the random telegraph noise.

The constant *KF* is proportional to the oxide trap density, which is technology specific. *AF* is between 1 and 2 depending on the importance of Coulombic scattering to carrier mobility and *W*, *L*, and $C_{ox}$ are the width, length, and per-area oxide capacitance of the MOSFET. The flicker noise is the dominant noise at low frequency. At frequencies above 100 MHz, one can safely ignore the flicker noise as it is much smaller than the thermal noise. In non-linear or time-varying circuits such as oscillators and mixers, which operate periodically with a large-amplitude high-frequency signal, the flicker noise is shifted up or down in frequency to the beat (sum and difference) frequencies of the signal and the noise. This creates a noise in the oscillator output, for example. HEMT (see Section 6.3.3) and bipolar transistors (see Chapter 8) have significantly lower flicker noise than MOSFET because they do not employ the MOS structure.

### 6.15.4 Signal to Noise Ratio, Noise Factor, Noise Figure

The input to a device or a circuit is in general a combination of the desired signal and some noise. The ratio of the signal power to the noise power is called the **signal**

**to noise ratio** or **SNR**. SNR is a measure of the detectability of the signal in the presence of noise. The device or circuit also has some internal noise that is added to the amplified input noise and forms the total noise at the output. As a result, the SNR at the output of a linear device or circuit is smaller than the SNR at the input. The ratio of the input SNR and output SNR is called the **noise factor**.

$$F = \frac{S_i/N_i}{S_0/N_0} \tag{6.15.8}$$

The **noise figure** is defined as ten times the base-10 logarithm of the noise factor.

$$N_F = 10 \times \log F \tag{6.15.9}$$

The unit of noise figure is decibel or dB. As discussed earlier (see Sec. 6.15.2), the noise can be minimized with an optimum gate network impedance. Achieving this $N_{F\text{-min}}$ is an important goal of low-noise circuit design.

### ● Noise and Digital Circuits ●

The above discussion of MOSFET noise is more relevant to analog circuits than digital circuits. For a linear circuit such as a linear amplifier that must faithfully preserve the input waveform while amplifying its magnitude, the SNR at the output is at best the same as at the input. A digital circuit such as an inverter can generate an output that is 0 or $V_{dd}$ even when the input is somewhat lower than $V_{dd}$ or higher than 0. It eliminates the small noise at the input with its nonlinear voltage-transfer characteristic (see Fig. 6–19). In other words, a digital circuit has no gain for the small-amplitude noise at the input and has gain only for the larger real digital signal.

You may have had the pleasant experience of getting a photocopy of a black and white document that is cleaner looking than the original. The light smudges or erased pencil writings on the original are absent in the copy. That photocopier is a nonlinear system as is the digital circuit. If a photocopier is called on to reproduce a gray tone photograph as a linear system, it cannot reduce the noise in the original photograph because the copier cannot tell whether a smudge in the original is noise or part of the photograph.

This signal sharpening property of the digital circuits makes it possible to pack the digital circuits densely with long signal wires running close to each other. The dense wiring creates large cross-talk noise that is typically much larger than the thermal noise and flicker noise. Engineers reduce the cross talk by electrically shielding the sensitive lines, using low-*k* dielectrics between the lines (to reduce capacitive coupling), and limiting the line lengths.

When the MOSFET becomes very small as in advanced flash memory cells (see Section 6.16.3), a single trap can produce enough random telegraph noise (see Fig. 6–33b) to cause difficulty reading the 1 and 0 stored in a cell. Although this happens to only a very small portion of the memory cells, it is a concern for high-density memory design [15].

## 6.16 ● SRAM, DRAM, NONVOLATILE (FLASH) MEMORY DEVICES ●

Most of the transistors produced every year are used in semiconductor memories. Memory devices are commonly embedded in digital integrated circuits (ICs). For example, memory can occupy most of the area of a computer processor chip. Memory devices are also available in stand-alone memory chips that only perform the memory function. There are three types of semiconductor memories—**static RAM** or **SRAM**, **dynamic RAM** or **DRAM**, and **nonvolatile memory** with **flash memory** being the most prevalent nonvolatile memory. RAM stands for **random-access memory** meaning every data byte is accessible any time unlike hard disk memory, which has to move the read head and the disk to fetch new data with a significant delay. "Nonvolatile" means that data will not be lost when the memory is disconnected from electrical power source. The three types coexist because each has its own advantages and limitations. Table 6–1 summarizes their main differences.

SRAM only requires the same transistors and fabrication processes of the basic CMOS technology. It is therefore the easiest to integrate or embed into CMOS circuits. A DRAM cell is many times smaller than an SRAM cell but requires some special fabrication steps. High-density stand-alone DRAM chips are produced at large specialized DRAM fabrication plants. Low cost DRAMs has helped to proliferate PCs. A flash memory cell employs one of a variety of physical mechanisms to perform nonvolatile storage and has even smaller size than DRAM. Flash memory has not replaced DRAM or SRAM because of its slower writing speed and limited write cycles. Flash memory is economical and compact and has enabled advanced portable applications such as cell phones, media players, and digital cameras. Less aggressive (larger cell size) versions of DRAM and flash memory can be embedded in CMOS logic chips with some modification of the CMOS process technology. **Embedded DRAM** can be more economical than embedded SRAM when the required number of memory bits is very large.

### 6.16.1 SRAM

A basic SRAM cell uses six transistors to store one bit of data. As shown in Fig. 6–34a, its core consists of two cross-coupled inverters. $M_1$ and $M_3$ make up the left inverter. $M_2$ and $M_4$ make up the right inverter. The output of the left inverter is connected to the input of the right inverter and vice versa. If the left-inverter output, which is the input of the right inverter is high (hi), the right-inverter output would be low. This low output in turn makes the left-inverter out high. The positive feedback ensures that this state is stored and stable. If we change the left-inverter output to low and the right-inverter output to high, that would be a second stable state. Therefore this cell has two stable states, which represent the "0" and "1" and can store one bit of data. Many identical SRAM cells are arranged in an XY array. Each row of cells is connected to one word line (WL) and each column of cells is connected to a pair of bit lines (BL and BLC).

Two pass transistors $M_5$ and $M_6$ connect the outputs of the inverters to the bit lines. In order to read the stored data (determine the inverter state), the selected cell's WL is raised to turn on the pass transistors. A sensitive **sense amplifier** circuit compares the voltages on BL and BLC to determine the stored state.

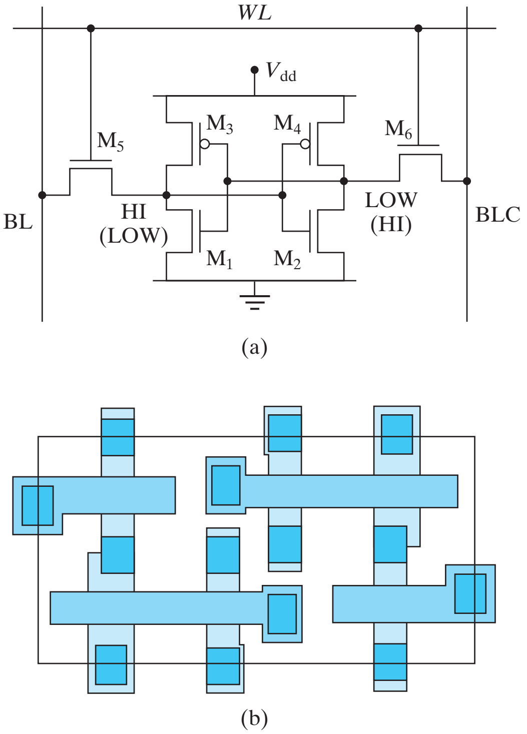


FIGURE 6–34 (a) Schematic of an SRAM cell. (b) Layout of a 32 nm technology SRAM, from [16]. The dark rectangles are the contacts. The four horizontal pieces are the gate electrodes and the two PFETs have larger Ws than the six NFETs. Metal interconnects (not shown) cross couple the two inverters.

TABLE 6–1 • The differences among three types of memories.

| | Keep Data Without Power? | Cell Size and Cost/bit | Rewrite Cycles | Write-One-byte Speed | Compatible with Basic CMOS Manufacturing | Main Applications |
|---|---|---|---|---|---|---|
| SRAM | No | Large | Unlimited | Fast | Totally | Embedded in logic chips |
| DRAM | No | Small | Unlimited | Fast | Need modifications | Stand-alone chips and embedded |
| Flash memory | Yes | Smallest | Limited | Slow | Need extensive modifications | Nonvolatile storage stand-alone |

In order to write the left-low state into the cell, for example, BL is set to low and BLC is set to high. Next, the word-line voltage is raised and the inverters will be forced into this (new) state.

SRAM cells provide the fastest operation among all memories. But since it requires six transistors to store one bit of data, the cost per bit is the largest. SRAM cells are often used as cache memory embedded in a processing unit where speed is critical. The steady increase in the clock speed of the processors requires the cache size to increase as well. Much effort is spent on size reduction, called ***scaling***, for SRAM and for other types of memories. Figure 6–34b shows the layout of the six transistors of a 32 nm technology node SRAM cell [16].

### 6.16.2 DRAM

A DRAM cell contains only one transistor and one capacitor as shown in Fig. 6–35. Therefore it can provide a large number of bits per area and therefore lower cost per bit. Figure 6–35 is a portion of a schematic DRAM cell array. One end of the cell capacitor is grounded. The states "1" and "0" are represented by charging the cell capacitor to $V_{dd}$ or zero. To write data into the upper-left cell, WL 1 is raised high to turn on the transistor (connecting the capacitor to bit line 1) and bit line 1 is set to $V_{dd}$ to write "1" or 0 V to write "0." The cell to the right can be written at the same time by setting bit line 2 to the appropriate value ($V_{dd}$ or 0 V).

Each bit line has its own (unavoidable) capacitance, $C_{\text{bit line}}$. In order to read the stored data from the upper-left cell, bit line 1 is precharged to $V_{dd}/2$ and then left floating. WL 1 voltage is raised to connect the cell capacitor in parallel with the larger $C_{\text{bit line}}$. Depending on the cell capacitor voltage ($V_{dd}$ or 0), the cell capacitor either raises or lowers the bit line voltage by $C \cdot V_{dd}/2(C + C_{\text{bit line}})$, usually tens of milivolts. A sense amplifier circuit connected to the bit line monitors this voltage change to determine (read) the stored data. All cells connected to one WL are read at the same time. After each read operation the same data are automatically written back to the cell because the capacitor charge has been corrupted by the read.

The DRAM capacitor can only hold the data for a limited time because its charge gradually leaks through the capacitor dielectric, the PN junction (transistor S/D), and the transistor subthreshold leakage (see Section 7.2). To prevent data loss, the change must be refreshed (read and rewritten) many times each second.

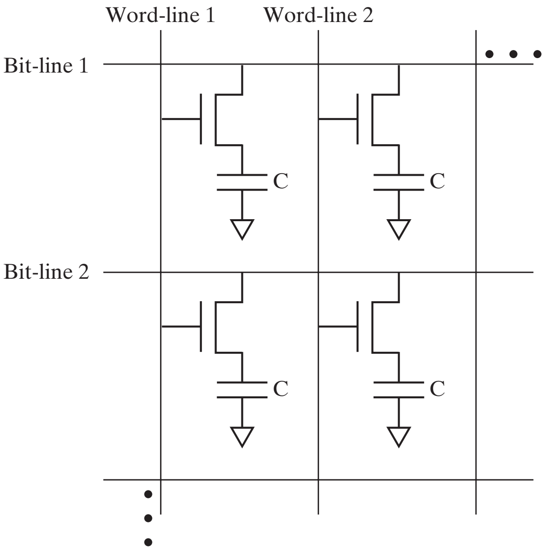


FIGURE 6–35 A schematic DRAM cell array. Each cell consists of a transistor and a capacitor.

The D in DRAM refers to this dynamic **refresh** action. Refresh consumes stand-by power. To increase the refresh interval, the cell capacitance should be large so that more charge is stored.

A large cell capacitance (not too much smaller than $C_{\text{bit line}}$) is also important for generating a large read signal for fast and reliable reading. However, it has become increasingly difficult to provide a large $C$ while the cell area has been reduced to a few percent of 1 μm$^2$. Besides deploying very thin capacitor dielectrics, engineers have resorted to complex three-dimensional capacitor structures that provide capacitor areas larger than the cell area. Figure 6–36a shows a cup-shaped capacitor and Fig. 6–36b shows a scanning electron microscope view of the cross section of

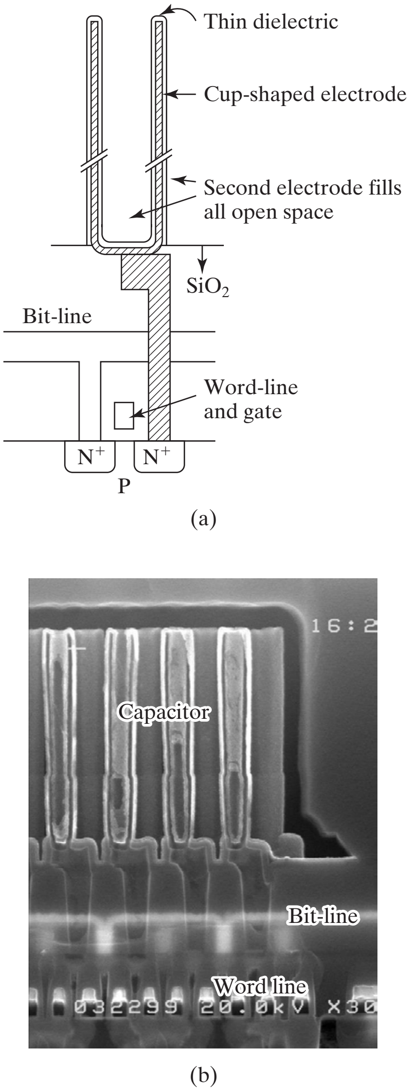


FIGURE 6–36 (a) Schematic drawing of a DRAM cell with a cup-shaped capacitor. (b) Cross-sectional image of DRAM cells. The capacitors are on top and the transistors are near the bottom. (From [17].)

several DRAM cells. The four deep-cup shaped elements are four capacitors. Each capacitor has two electrodes. One electrode is cup-shaped and made of polysilicon or metal. It is connected at the bottom by a poly-Si post to the transistor below. Both the inside and the outside of the cup electrode are coated with a thin dielectric film. The other electrode is also made of poly-Si and it fills the inside of the cup as well as all the spaces between the cups. This second electrode is grounded (see Fig. 6–35). This complex structure provides the necessary large capacitor area.

A much simplified DRAM process technology can be integrated into logic CMOS technology at significant sacrifice of the cell area. Such an embedded DRAM technology is an attractive alternative to embedded SRAM when the number of bits required is large.

### 6.16.3 Nonvolatile (Flash) Memory

SRAMs and DRAMs lose their stored content if they are not connected to an electric power source. **Nonvolatile memory** or **NVM** is a memory device that keeps its content without power for many years. NVMs are used for program **code storage** in cell phones and most microprocessor based systems. They are also the preferred **data storage** medium (over hard disks and CDs) in portable applications for storing documents, photos, music, and movies because of their small size, low power consumption, and absence of moving parts. There are many variations of NVM devices [18], but the prevalent type is illustrated in Fig. 6–37a.

The structure may be understood as a MOSFET with one modification. The gate insulator is replaced with two insulators sandwiching a charge-storage layer. For example, the charge-storage layer can be silicon nitride or another insulator with a high density of electron traps. When the traps are empty or neutral, the transistor has a low $V_t$. When electrons are trapped in the insulator, the transistor has a high $V_t$ as discussed in Section 5.7 and illustrated in Fig. 6–37b. The low and high $V_t$ states represent the “0” and “1,” respectively, and can be easily read with a sense circuit that checks the $V_t$. The charge storage layer may be a conductor, and in fact the most important and prevalent charge storage layer material is the familiar polycrystalline Si. NVM employing a poly-Si charge storage layer is called the **floating-gate memory** because the poly-Si layer is a transistor gate that is electrically floating.

Figure 6–37c shows how to put electrons into the charge-storage layer, i.e., how to write “1” into the NVM cell. About 20 V is applied to the gate and the high field causes electrons to tunnel (see Section 4.20) from the inversion layer into the charge storage layer. In Fig. 6–37d the cell is erased into “0” when the stored electrons tunnel into the substrate (the P-type accumulation layer).

Because the erase operation by tunneling is slow (taking milliseconds compared to nano-seconds for SRAM and DRAM), these NVMs are erased in blocks of kilobytes rather than byte by byte. Electrical erase by large memory blocks is called **flash erase** and this type of memory is called **flash memory**. Flash memory is the dominant type of NVM so that the two terms are often used interchangeably. Writing by tunneling is also slow so that it is also performed on hundreds of bytes at the same time.

There is another way of writing the cell in Fig. 6–37 (a and e). When the source is grounded and higher-than-normal voltages are applied to the gate and the drain, a high electric field exists in the pinch-off (or velocity-saturation) region near

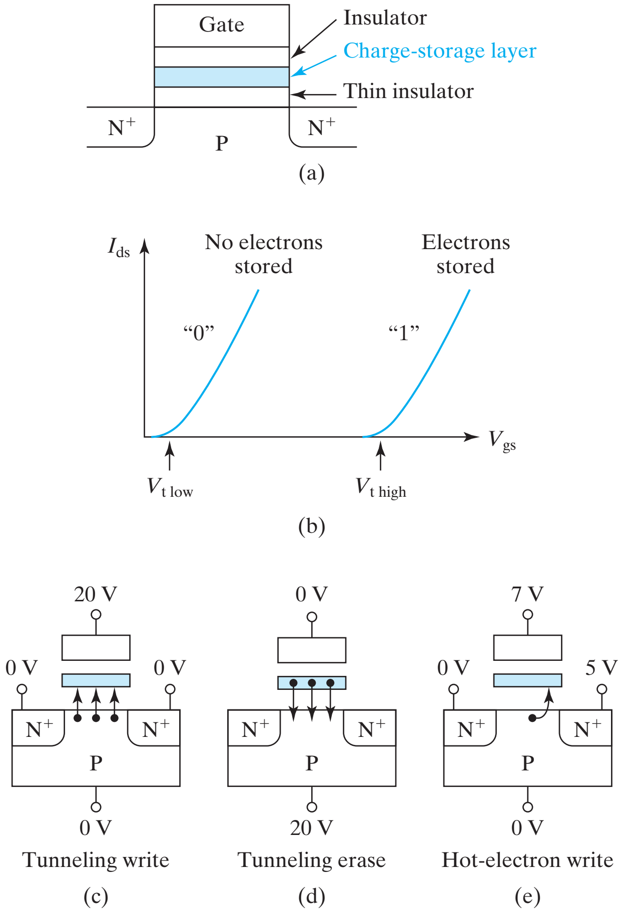


FIGURE 6–37 (a) A charge-storage NVM cell has a charge-storage layer in the gate dielectric stack; (b) $V_t$ is modified by trapping electrons; (c) electron injection by tunneling; (d) electron removal by tunneling; and (e) electron injection by hot-electron injection.

the drain. A small fraction of electrons traveling through this region can gain enough energy to jump over the insulator energy barrier into the charge-storage layer. This method of writing is faster than tunneling but takes more current and power. The energetic electrons are called the **hot electrons** and this writing mechanism is called **hot carrier injection** or **HCI**.

## • Hot-Carrier-Injection Reliability of MOSFETs •

The high-quality gate oxide of the best CMOS transistors still contains charge traps. Even under normal CMOS circuit operation, a small number of hot carriers may be injected and trapped in the oxide. Over many years the trapped charge may change $V_t$ and the *I–V* characteristics. Before releasing a CMOS technology for production, engineers must carry out accelerated tests of **hot-carrier reliability** and conduct careful analysis of the data to ensure that circuit performance will not change appreciably [19] over the product lifetime.

A limitation of the flash memory is that repeated write and erase cycling under high-electric field can break chemical bonds in the insulator and create leakage paths with diameters of a few atoms and at random locations. A single leakage path can discharge a floating gate and cause data loss. This sets an NVM endurance limit of less than $10^6$ write/erase cycles. If the floating gate is replaced with a dielectric film containing many isolated electron traps or isolated nanocrystals of metal or semiconductor, one leakage path can only discharge a fraction of the stored electrons in the cell. Endurance may be improved. They are called **charge-trap NVM** and the **nano-crystal NVM**.

For several reasons, NVMs can store larger numbers of bits per centimeter square than DRAMs and SRAMs. First, the NVM cell (see Fig. 6–37a) is simple and small even in comparison with a DRAM cell. Second, it is possible to write and store more than two $V_t$ values (see Fig. 6–37b) in a flash memory cell by controlling the number of stored electrons. Two $V_t$s can code one bit of data. Four $V_t$s can code two bits of data (00, 01, 10, and 11). This technique is called the **multilevel cell** technology. **NAND flash** memory gets even higher integration density (measured in bits/cm$^2$) by stringing dozens of flash memory cells in series. Imagine a long and narrow silicon strip area covered with the gate dielectric stack and flanked by shallow-trench-isolation oxide on its left and its right. Thirty-two parallel poly-Si gate lines, separated by minimum spacing, cross over the silicon strip. The spaces between the poly-Si gates are doped into N$^+$ regions by ion implantation. This creates 32 NFETs (NVM cells) connected in series. Doing so eliminates the need to make metal contacts to every cell because the N$^+$ source of one cell doubles as the drain of the next cell and so on. To illustrate the operation, let us consider only two cells in series. To read the data (the $V_t$) stored in the top cell, the gate voltage of the bottom cell is raised to higher than $V_{t\text{-high}}$. Similarly, reading the other cell as well as writing and erasing the cells can be performed by cleverly choosing the control voltages. It is call NAND flash because the string of transistors resembles a part of the NAND logic gate.

Charge storage is the most common but not the only mechanism for data storage. Figure 6–38a shows a **resistance-change NVM** or **RRAM** cell employing a programmable resistor. The resistor can be made of metal oxide or other inorganic or organic materials and programmed by electric field or current and sits over the transistor to save area. In one version, it is programmed by a heat pulse and made

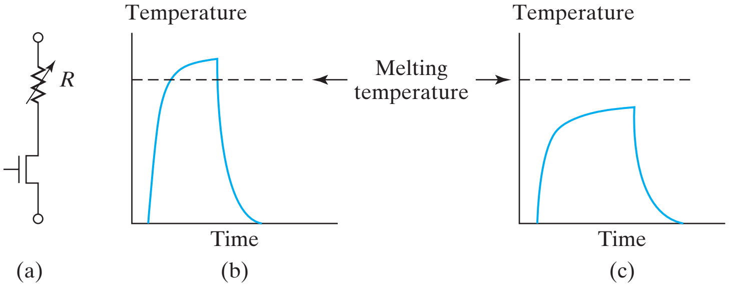


FIGURE 6–38 (a) Concept of a resistance-change memory such as a PCM. (b) Program the PCM into high-resistance state by rapid solidification, producing a highly resistive amorphous phase. (c) Program the PCM into low-resistance state by annealing, turning the amorphous material into a conductive crystalline phase.

of an alloy of Ge, Sb, and Te. If a current pulse is applied to heat the material above its melting temperature as shown in Fig. 6–38b, the subsequent rapid solidification creates an amorphous phase (see Fig. 3–15) of the material that is highly resistive. In Fig. 6–38c, another current pulse heats the resistor to a below-melting temperature, at which the amorphous material is annealed into a (poly)crystalline phase that has order-of-magnitude lower resistivity. The $R_{\text{low}}$ and $R_{\text{high}}$ states represent the "0" and "1." Reading is performed at a much lower current level with less heating. This memory is known as the **phase change memory** or **PCM**. PCM can be written and erased at SRAM speed and has much better endurance than the charge-storage memory.

In another technology, the resistor in Fig. 6–38a is an extremely thin filament of metal ions. The filament can be broken to create $R_{\text{high}}$ by moving just a few metal ions with an electrical pulse. It can be restored with an electrical pulse of the opposite polarity. This memory concept is called **metal migration memory**.

## 6.17 • CHAPTER SUMMARY •

The basic CMOS technology is presented in Fig. 6–7. The CMOS inverter, as a representative digital gates, is analyzed in Section 6.7. The PFET pull-up device and the NFET pull-down device create a highly nonlinear **VTC**. This nonlinearity gives the inverter its ability to refresh digital signals and provides the much-needed noise margin in a noisy digital circuit. The inverter *propagation delay* is

$$\tau_{\text{d}} \approx \frac{CV_{\text{dd}}}{4}\left(\frac{1}{I_{\text{onN}}} + \frac{1}{I_{\text{onP}}}\right) \tag{6.7.1}$$

CMOS circuits' power consumption is

$$P = kCV_{\text{dd}}^2 f + V_{\text{dd}} I_{\text{off}} \tag{6.7.9}$$

where $k < 1$ accounts for the *activity of the circuit*. The first term is the *dynamic power* and the second, the *static power*.

It is highly desirable to have large $I_{\text{on}}s$ without using a large power supply voltage, $V_{\text{dd}}$. It is also desirable to reduce the total load capacitance, $C$ (including the junction capacitance of the driver devices, the gate capacitance of the driven devices, and the interconnect capacitance). Both capacitance and cost reductions provide strong motivations for reducing the size of the transistors and therefore the size of the chip. In addition, speed has benefited from the relentless push for smaller $L$, thinner $T_{\text{ox}}$, and lower $V_{\text{t}}$; and power consumption has benefited greatly from the lowering of $V_{\text{dd}}$.

Electron and hole *surface mobilities*, $\mu_{ns}$ and $\mu_{ps}$, are well-known functions of the average electric field in the inversion layer, which can be roughly expressed as $(V_{\text{gs}} + V_{\text{t}})/6T_{\text{ox}}$. As this *effective vertical field* increases, the surface mobility decreases. At typical operating fields, surface mobilities are only fractions of the bulk mobilities. All of these are captured in Fig. 6–9.

GaAs has a high electron mobility but poor quality of dielectric–semiconductor interface. GaAs MESFET is an FET structure that does not require an MOS structure. Instead, the channel conductance is controlled by a Schottky contact gate.

HEMT uses an epitaxial high-band-gap semiconductor as an insulator in a MOSFET-like structure. The epitaxial interface is smooth. The electron mobility is very high and the device speed is very fast.

The $V_t$ of a MOSFET can be easily measured from the $I_{ds}$ vs. $V_g$ plot. $V_t$ increases with increasing body-to-source reverse bias, $V_{sb}$. This *body effect* is deleterious to circuit speed.

$$\boxed{V_t(V_{sb}) = V_{t0} + \alpha V_{sb}} \quad \text{for steep retograde body doping} \tag{6.4.6}$$

$$\alpha = 3T_{oxe}/W_{dmax} \tag{6.4.7}$$

$$V_t = V_{t0} + \frac{\sqrt{qN_a 2\varepsilon_s}}{C_{oxe}}(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}) \quad \text{for uniform body doping} \tag{6.4.8}$$

where $V_{t0}$ is the threshold voltage in the absence of body bias.

The basic $I_{ds}$ model is

$$\boxed{I_{ds} = \frac{W}{L}C_{oxe}\mu_{ns}\left(V_{gs} - V_t - \frac{m}{2}V_{ds}\right)V_{ds}} \tag{6.6.4}$$

$$m = 1 + 3T_{oxe}/W_{dmax} \approx 1.2 \tag{6.5.2}$$

The IV characteristics may be divided into the *linear region* and the *saturation region*. $I_{ds}$ saturates at

$$\boxed{\begin{aligned} V_{dsat} &= \frac{V_{gs} - V_t}{m} && (6.6.5) \\ I_{dsat} &= \frac{W}{2mL}C_{oxe}\mu_{ns}(V_{gs} - V_t)^2 && (6.6.6) \end{aligned}}$$

The *transconductance* of a MOSFET in the saturation region is

$$g_{msat} \equiv dI_{dsat}/dV_{gs} = \frac{W}{mL}C_{oxe}\mu_{ns}(V_{gs} - V_t) \qquad (6.6.8), (6.6.9)$$

*The above basic* $I_{ds}$ *model can be significantly improved by considering velocity saturation*. The result is

$$\boxed{V_{dsat} = \left(\frac{m}{V_{gs} - V_t} + \frac{1}{\mathscr{E}_{sat}L}\right)^{-1}} \tag{6.9.10}$$

$$\begin{aligned} \mathscr{E}_{sat} &= 1.6 \times 10^7 \text{ cm/s} \div \mu_{ns} && \text{for electrons, and} \\ & \phantom{=}\ 1.2 \times 10^7 \text{ cm/s} \div \mu_{ns} && \text{for holes.} \end{aligned}$$

$$\boxed{I_{dsat} = \frac{\text{long channel } I_{dsat} \text{ (Eq. (6.6.6))}}{1 + \dfrac{V_{gs} - V_t}{m\mathscr{E}_{sat}L}}} \tag{6.9.11}$$

If $\mathscr{E}_{sat}L >> V_{gs} - V_t$, Eqs. (6.9.10) and (6.9.11) reduce to the long-channel model, Eqs. (6.6.5) and (6.6.6). If $\mathscr{E}_{sat}L << V_{gs} - V_t$

$$V_{dsat} \approx \mathscr{E}_{sat}L < \text{long-channel } V_{dsat} \qquad (6.9.13)$$

$$I_{dsat} = Wv_{sat}C_{oxe}(V_{gs} - V_t - \mathscr{E}_{sat}L) \qquad (6.9.14)$$

If $L$ is reduced to tens of nanometers, *velocity overshoot* will raise $\mathscr{E}_{sat}$ and $v_{sat}$ in the above equations somewhat. Eventually, the *carrier injection velocity* at the source will limit $I_{dsat}$. Interestingly, The present estimate of this limit is not significantly different from what Eq. (6.9.14) would predict.

The *intrinsic voltage gain* of a MOSFET is $g_{msat}/g_{ds}$ . $g_{ds} = dI_{dsat}/dV_d$ is the output conductance. To achieve a small $g_{ds}$ requires a large $L$ and/or small $T_{ox}$, $W_{dep}$, and $X_j$ (see Section 7.9).

For high-frequency applications, it is important to reduce the (poly-Si) gate electrode resistance by breaking a wide-$W$ transistor into a large number of smaller-$W$ transistors connected in parallel. Reducing the channel length can reduce the intrinsic input resistance as shown in Eq. (6.14.3).

MOSFET noise arises from the channel, gate, substrate thermal noise, and the flicker noise. While the thermal noise is a white noise, the flicker noise per bandwidth is proportional to $1/f$. The flicker ($1/f$) noise is reduced if the trap densities in the gate dielectric or the oxide–semiconductor interface are reduced.

A basic SRAM cell employs six MOSFETs. SRAM is commonly embedded in logic chips. DRAM cell consists of one transistor and one capacitor. Its size is very small. DRAM requires refreshing and a specialized technology, partly because of the complex capacitor structure that has a large surface area. The prevalent NVM is the flash memory. It uses even smaller Si area per bit than DRAM and can store data without power for many years. While floating-gate NAND is the dominant NVM, several new NVM concepts are under active investigation.

## ● PROBLEMS ●

### ● MOSFET AND MESFET $V_t$ ●

**6.1** An N-channel MOSFET with N$^+$-poly gate is fabricated on a 15 Ω cm P-type Si wafer with oxide fixed charge density = $q \times 8 \times 10^{10}\text{cm}^{-2}$, $W = 50\ \mu\text{m}$, $L = 2\ \mu\text{m}$, $T_{ox} = 5$ nm.

**(a)** Determine the flat-band voltage, $V_{fb}$.

**(b)** What is the threshold voltage, $V_t$?

**(c)** A circuit designer requested N-MOSFET with $V_t = 0.5$ V from a device engineer. It was not allowed to change the gate oxide thickness. If you are the device engineer, what can you do? Give specific answers including what type of equipment to use.

**6.2** A GaAs MESFET has a 0.2 μm thick N-channel doped to $N_d = 10^{17}\ \text{cm}^{-3}$. Assume that $\phi_{Bn}$ of the Au–GaAs Schottky gate is 1 V. $\varepsilon_s$ of GaAs is 13 times the vacuum dielectric constant. $V_d = V_s = 0$.

**(a)** What is $W_{dep}$ at $V_g = 0$? (Hint: Please refer to Table 1–4 for the value of $N_c$ of GaAs at room temperature.)

**(b)** At what $V_g$ (including the sign) will $W_{dep}$ be equal to the channel thickness? This is the cut-off gate voltage of the MESFET. The channel is shut off at this $V_g$.

**(c)** Can any gate voltage of the opposite sign to (b) be applied to the gate without producing expression gate current? What is its effect on $W_{dep}$ and $I_{ds}$?

**(d)** What needs to be done to redesign this MESFET so that its channel is cut off at $V_g = 0$ and the channel only conducts current at $V_g$ larger than a threshold voltage?

*Discussion:* The device in (d) is called an enhancement-mode transistor. The device of (b) is a depletion mode transistor.

6.3 An N-MOSFET and a P-MOSFET are fabricated with substrate doping concentration of $6 \times 10^{17} cm^{-3}$ (P-type substrate for N-MOSFET and N-type substrate for P-MOSFET). The gate oxide thickness is 5 nm. See Fig. 6–39.

**(a)** Find $V_t$ of the N-MOSFET when $N^+$ poly-Si is used to fabricate the gate electrode.

**(b)** Find $V_t$ of the P-MOSFET when $N^+$ poly-Si is used to fabricate the gate electrode.

**(c)** Find $V_t$ of the P-MOSFET when $P^+$ poly-Si is used to fabricate the gate electrode.

**(d)** Assume that the only two voltages available on the chip are the supply voltage $V_{dd} = 2.5$ V and ground, 0 V. What voltages should be applied to each of the terminals (body, source, drain, and gate) to maximize the source-to-drain current of the N-MOSFET?

**(e)** Repeat part (d) for P-MOSFET.

**(f)** Which of the two transistors (b) or (c) is going to have a higher saturation current. Assuming that the supply voltage is 2.5 V, find the ratio of the saturation current of transistor (c) to that of transistor (b).

**(g)** What is the ratio of the saturation current of transistor (c) to that of transistor (a)? Use the mobility values from Fig. 6–9.

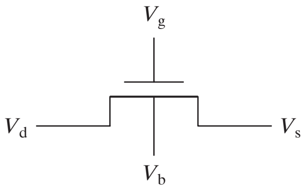


FIGURE 6–39

## Basic MOSFET IV Characteristics

6.4 CV and $I_d - V_g$ characteristics of a hypothetical MOSFET with channel length $L = 1$ µm are given in Fig. 6–40.

**(a)** Is the CV characteristic obtained at high frequency or low frequency? Or, is it impossible to determine? Explain.

**(b)** Is this a PMOSFET or an NMOSFET?

**(c)** Find the threshold voltage of this transistor.

**(d)** Determine the mobility of the carriers in the channel of the transistor.

**(e)** Plot $I_d - V_d$ curves at $V_g = 1$ V and $V_g = 2.5$ V.

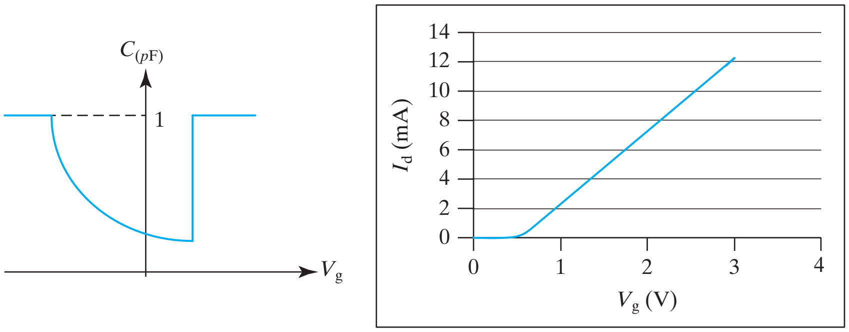


FIGURE 6–40

6.5 Figure 6–41 is the IV characteristics of an NMOSFET with $T_{ox}$ = 10 nm, $W$ = 10 µm, and $L$ = 2 µm. (Assume $m$ = 1 and do not consider velocity saturation.)

**(a)** Estimate $V_t$ from the plot.

**(b)** Estimate $\mu_{ns}$ in the inversion layer.

**(c)** Add the $I$–$V$ curve corresponding to $V_{gs}$ = 3 V to the plot.

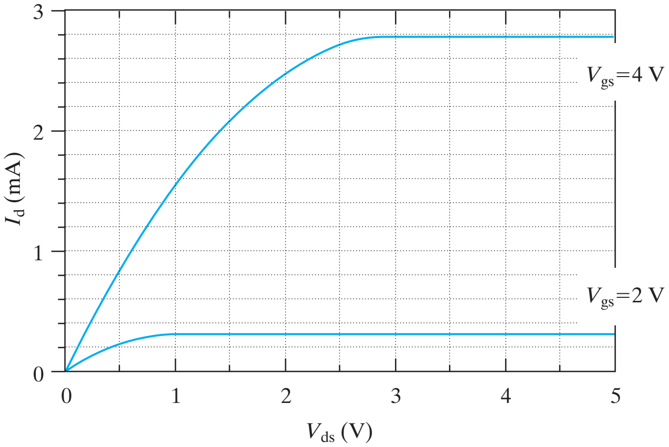
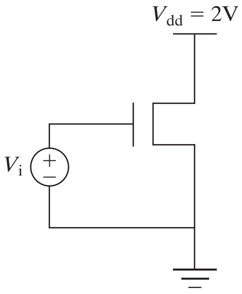


FIGURE 6–41

6.6 The MOSFET in the circuit shown in Fig. 6–42 is described by



$$I_{dsat} = \frac{k'W}{2L}(V_g - V_t)^2, \qquad \text{for } V_d > V_{dsat}$$

$$I_d = \frac{k'W}{L}\left[(V_g - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right] \qquad \text{for } V_d < V_{dsat}$$

FIGURE 6–42

where $k'$ is $\mu_{ns}C_{ox}$ and obtained in practical case by measuring $I_{dsat}$ at a given gate bias. When $k' = 25\ \mu A/V^2$, $V_t = 0.\ 5$ V, $W = 10\ \mu m$, $L = 1\ \mu m$, and $V_i$ varied from 0 to 3 V,

**(a)** Make a careful plot of $\sqrt{I_{dsat}}$ as a function of $V_i$ showing any break points on the curve.

**(b)** Make a plot of the MOSFET transconductance using a solid line.

**(c)** On the plot of part (b), use a dotted line to indicate a curve of the output conductance, $dI_{ds}/dV_{ds}$.

**6.7** One $I_{ds} - V_{ds}$ curve of an ideal MOSFET is shown in Fig. 6–43. Note that $I_{dsat} = 10^{-3}$A and $V_{dsat} = 2$ V for the given characteristic. (You may or may not need the following information: $m = 1$, $L = 0.5\ \mu m$, $W = 2.5\ \mu m$, $T_{ox} = 10$ nm. Do not consider velocity saturation.)

**(a)** Given a $V_t$ of 0.5 V, what is the gate voltage $V_{gs}$ one must apply to obtain the $I$–$V$ curve?

**(b)** What is the inversion-layer charge per unit area at the drain end of the channel when the MOSFET is biased at point (1) on the curve?

**(c)** Suppose the gate voltage is changed such that $V_{gs} - V_t = 3$ V. For the new condition, determine $I_{ds}$ at $V_{ds} = 4$ V.

**(d)** If $V_d = V_s = V_b = 0$ V, sketch the general shape of the gate capacitance $C$ vs. $V_g$ to be expected from the MOSFET, when measured at 1 MHz. Do not calculate any capacitance but do label the $V_g = V_t$ point in the $C$–$V$ curve.

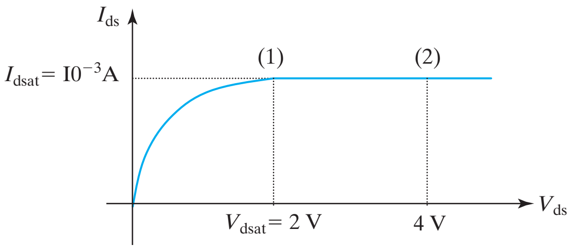


FIGURE 6–43

**6.8** An ideal N-channel MOSFET has the following parameters: $W = 50\ \mu m$, $L = 5\ \mu m$, $T_{ox} = 0.05\ \mu m$, $N_a = 10^{15}\ cm^{-3}$, N$^+$ poly-Si gate, $\mu_{ns} = 800\ cm^2/V/s$ (and independent of $V_g$). Ignore the bulk charge effect and velocity saturation.
Determine:

**(a)** $V_t$

**(b)** $I_{dsat}$ if $V_g = 2$ V

**(c)** $dI_{ds}/dV_{ds}$ if $V_g = 2$ V and $V_d = 0$

**(d)** $dI_{ds}/dV_{gs}$ if $V_g = 2$ V and $V_d = 2$ V.

## ● Potential and Carrier Velocity in MOSFET Channel ●

**6.9** Derive the equation $V_c(x) = (V_g - V_t)\ [1\ \surd\ 1 - x/L\ ]$ in Section 6.6. Assume $m = 1$. (Do not consider velocity saturation.)

**6.10** This is an expanded version of Problem 6.9.

**(a)** Provide the derivation of Eq. (6.6.7).

**(b)** Find the expression for $Q_{inv}(x)$.

**(c)** Find the expression for $v(x) = \mu_n dV_{cs}/dx$.

**(d)** Show that $WQ_{inv}(x)v(x) = I_{dsat}$ expressed in Eq. (6.6.6).

**(e)** Make a qualitative sketch of $V_{cs}(x)$.

## ● IV Characteristics of Novel MOSFET ●

**6.11** An NMOSFET has thinner $T_{ox}$ at the center of the channel and thicker $T_{ox}$ near the source and drain (Fig. 6–44). This could be approximately expressed as $T_{ox} = Ax^2 + B$. Assume that $V_t$ is independent of $x$ and $m = 1$. (Do not consider velocity saturation.)

**(a)** Derive an expression for $I_d$.

**(b)** Derive an expression for $V_{dsat}$.

**(c)** Does the assumption of nearly constant $V_t$ suggest a large or small $W_{dmax}$?

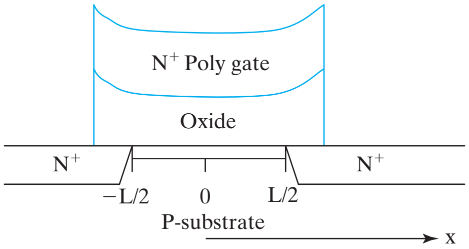


FIGURE 6–44

**6.12** Suppose you have a MOSFET whose gate width changes as a function of distance along the channel as:

$$W(x) = W_0 + x$$

where $x = 0$ at the source and $x = L$ at the drain. Except for its gate width, assume that this MOSFET is like the typical MOSFET you studied in Chapter 6. (Do not consider velocity saturation.)

**(a)** Find an expression for $I_d$ for this device. Ignore the bulk charge effect ($m = 1$).

**(b)** Derive an expression for $I_{dsat}$ for this device.

## ● CMOS ●

**6.13** MOS circuits perform best when the $V_t$ of NMOS and the $V_t$ of PMOS devices are about equal in magnitude and of opposite signs. To achieve this symmetry in $V_t$, PFET and NFET should have equal $N_{substrate}$, and symmetrical flat-band voltages, i.e., $V_{fb,\ PMOS} = -V_{fb,\ NMOS}$.

**(a)** Calculate the $V_{fb}$ of NMOS and PMOS devices if the substrate doping is $5 \times 10^{16}$ cm$^{-3}$ and the gate is N$^+$. Are the flat-band voltages symmetrical?

**(b)** Assume the NMOS and PMOS devices now have a P$^+$ gate. Redo (a).

**(c)** If you were restricted to one type of gate material, what work function value would you choose to achieve the same $|V_t|$?

**(d)** If you were allowed to use both N$^+$ and P$^+$ gates, which type of gate would you use with your NMOS and which with your PMOS devices?

(Hint: Use the results of (a) and (b). Consider the need to achieve symmetrical $V_t$ and the fact that large $|V_t|$ is bad for circuit speed.)

**6.14**

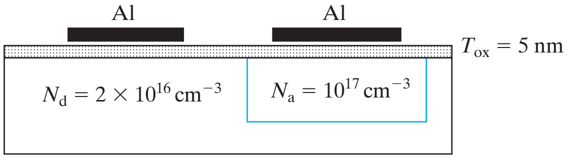

Al
Al
$T_{ox}$ = 5 nm
$N_d = 2 \times 10^{16}\,\text{cm}^{-3}$
$N_a = 10^{17}\,\text{cm}^{-3}$

FIGURE 6–45

**(a)** Determine the flat-band voltage of the NMOS and PMOS capacitors fabricated on the same chip. (The devices are shown in Fig. 6–45.)

**(b)** Find the threshold voltages of these two devices.

**(c)** It is desirable to make the NMOS and PMOS threshold voltages equal in magnitude ($V_{\text{tPMOS}} = -V_{\text{tNMOS}}$). One can in principle implant dopant with ionized dopant charge $Q_{\text{impl}}(\text{C/cm}^2)$ at the Si–SiO$_2$ interface to change the threshold voltage. Assume that such an implant is applied to PMOS only. Find the value of $Q_{\text{impl}}$ necessary to achieve $V_{\text{tPMOS}} = -V_{\text{tNMOS}}$.

**6.15** Supply the missing steps between (a) Eqs. (6.7.1) and (6.7.3) and between (b) Eqs. (6.7.3) and (6.7.4).

**6.16**

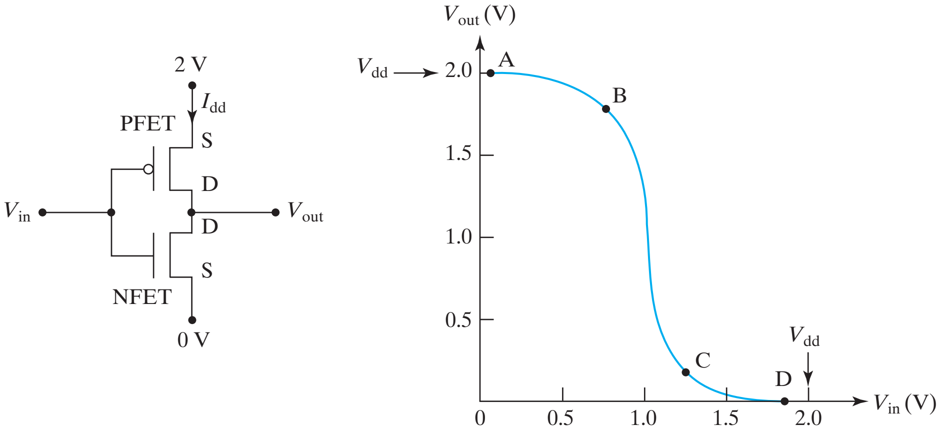


FIGURE 6–46

The voltage transfer curve of an inverter is given in Fig. 6–46. The threshold voltages of the NFET and PFET are +0.4 and –0.4 V, respectively. Determine the states of the two transistors (cut-off, linear, or saturation) at points A, B, C, and D, respectively. (Assume the output conductance of the transistor is very large.) Assume the two transistors have identical $\mu C_{ox}(W/L)$, $m = 1.333$.

| | NFET operation mode | PFET operation mode |
|---|---|---|
| A | | |
| B | | |
| C | | |
| D | | |

**6.17** Consider the CMOS inverter shown in Fig. 6–47.

**(a)** Sketch the voltage transfer characteristics (VTC), i.e., a plot of $V_0$ vs. $V_\mathrm{i}$ for this inverter, if the threshold voltages of the N-channel and P-channel MOSFETs are $V_{tn}$ and $V_{tp}$, respectively. Indicate the state (off, linear, or saturation) of each MOSFET as $V_\mathrm{i}$ is changed from 0 to $V_\mathrm{dd}$. Indicate all points on the VTC where a MOSFET changes its conduction state.

**(b)** Calculate the voltage at all points indicated in part (a) if both MOSFETs $I_\mathrm{d} - V_\mathrm{d}$ are characterized by the square-law theory with the following parameters.
For the N-channel MOSFET: $\mu_{ns} C_\mathrm{ox}(W/L) = 40$ mA /V$^{-2}$ and $V_{tn} = 1$ V.
For the P-channel MOSFET: $\mu_{ps} C_\mathrm{ox}(W/L) = 35$ mA /V$^{-2}$ and $V_{tp} = 1$ V.
The supply voltage $V_\mathrm{dd} = 5$ V.

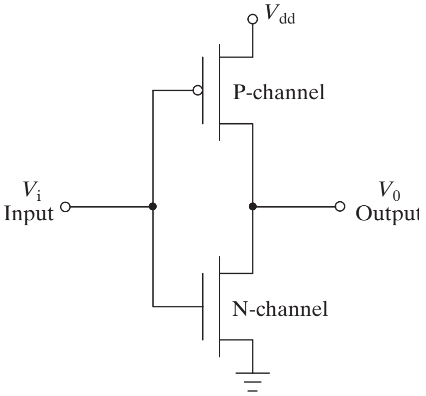


FIGURE 6–47

## ● Body Effect ●

**6.18** P-channel MOSFET with heavily doped P-type poly-Si gate has a threshold voltage of –1.5 V with $V_\mathrm{sb} = 0$ V. When a 5 V reverse bias is applied to the substrate, the threshold voltage changes to –2.3 V.

**(a)** What is the dopant concentration in the substrate if the oxide thickness is 100 nm?

**(b)** What is the threshold voltage if $V_\mathrm{sb}$ is –2.5 V?

## Velocity-Saturation Effect

**6.19** The $I_d - V_d$ characteristics of an NMOSFET are shown in Fig. 6–48.

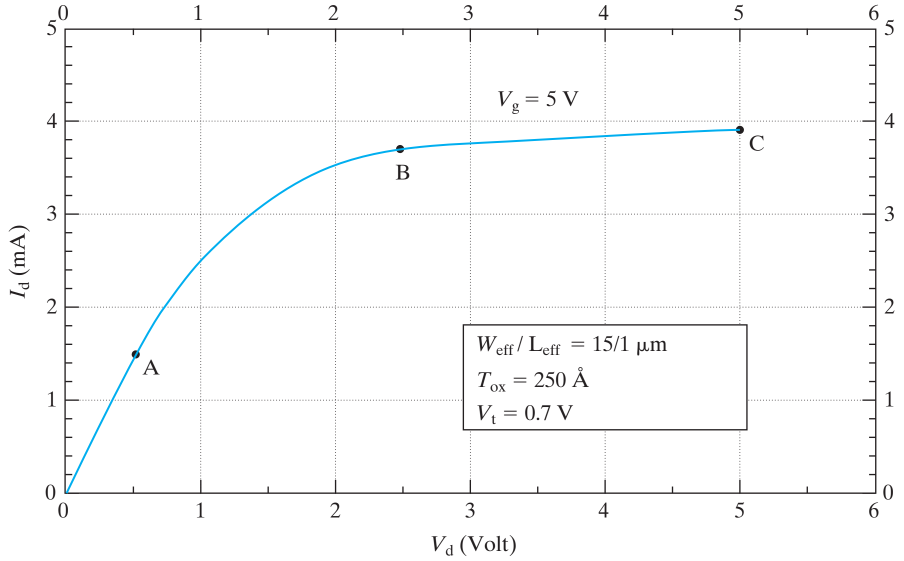


FIGURE 6–48

What are the velocities of the electrons near the drain and near the source at points A, B, and C? Use the following numbers in your calculations:

| | | |
|---|---|---|
| A: | $I_{ds} = 1.5$ mA | $V_{ds} = 0.5$ V |
| B: | $I_{ds} = 3.75$ mA | $V_{ds} = 2.5$ V |
| C: | $I_{ds} = 4.0$ mA | $V_{ds} = 5.0$ V |

(Hint: $I_d = W \times Q_{inv} \times$ v.)

**6.20** For an NMOS device with velocity saturation, indicate whether $V_{dsat}$ and $I_{dsat}$ increase, decrease, or remain unchanged when the following device parameters are reduced.

| | $T_{ox}$ | $W$ | $L$ | $V_t$ | $V_g$ |
|---|---|---|---|---|---|
| $V_{dsat}$ | | | | | |
| $I_{dsat}$ | | | | | |

**6.21** Verify Eq. (6.9.10) by equating Eqs. (6.9.3) and (6.9.9).

**6.22** Verify Eq. (6.9.11) by substituting Eq. (6.9.10) into Eq. (6.9.3).

**6.23** Consider a MOSFET with $\varepsilon_{sat} = 10^4$ V cm$^{-1}$. For $V_g - V_t = 2$ V, find $V_{dsat}$ when:

**(a)** $L = 0.1$ µm.

**(b)** $L = 10$ µm.

**(c)** For the device in part (a) with $I_{dsat} = 7$ mA, calculate the low field electron mobility if the gate capacitance is 10 fF.

**6.24** An NMOSFET with a threshold voltage of 0.5 V and oxide thickness of 6 nm has a $V_{\text{dsat}}$ of 0.75 V when biased at $V_{\text{gs}} = 2.5$ V. What is the channel length and saturation current per unit width of his device? (Hint: Use the universal mobility curve to find $\mu_s$. From $\mu_s$, you can determine

$$\varepsilon_{\text{sat}} = v_{\text{sat}}/2\mu_{ns} = \frac{8\times 10^6 \text{cm/s}}{2\mu_{ns}}$$

**6.25** The MOSFET drain current with velocity saturation is given as follows:

$$\text{In linear region,} \quad I_{\text{dlin}}(\text{velocity saturation}) = \frac{I_{\text{dlin}}(\text{no velocity saturation})}{1+\dfrac{V_{\text{ds}}}{E_{\text{sat}}L}}$$

$$\text{In saturation region,} \quad I_{\text{dsat}}(\text{velocity saturation}) = \frac{I_{\text{dsat}}(\text{no velocity saturation})}{1+\dfrac{V_{\text{gs}}-V_{\text{t}}}{mE_{\text{sat}}L}}$$

Consider a MOSFET with bulk charge factor $m = 1.2$, saturation velocity $v\text{sat} = 8\times 10^6 \text{cm/s}$ and surface mobility $\mu_{ns} = 300 \text{ cm}^2/V-s$. Under what condition will velocity saturation cause the drain current to degrade by a factor of two? Assume $mV_{\text{ds}} > V_{\text{gs}} - V_{\text{t}}$

**(a)** If $L = 100$ nm, $V_{\text{gs}} - V_{\text{th}} = ?$

**(b)** If $V_{\text{gs}} - V_{\text{t}} = 0.2$ V, $L = ?$

## ● Effective Channel Length ●

**6.26** The total resistance across the source and drain contacts of a MOSFET is ($R_{\text{s}} + R_{\text{d}} + R_{\text{Channel}}$), where $R_{\text{s}}$ and $R_{\text{d}}$ are source and drain series resistances, respectively, and $R_{\text{Channel}}$ is the channel resistance. Assume that $V_{\text{ds}}$ is very small in this problem.

**(a)** Write down an expression for $R_{\text{Channel}}$, which depends on $V_{\text{gs}}$ (Hint: $R_{\text{Channel}} = V_{\text{ds}}/I_{\text{ds}}$).

**(b)** Consider that $L_{\text{effective}} = L_{\text{gate}} - \Delta L$, where $L_{\text{gate}}$ is the known gate length and $\Delta L$ accounts for source and drain diffusion, which extend beneath the gate. Define $R_{\text{sd}}$ to be equal to ($R_{\text{s}} + R_{\text{d}}$). Explain how you can find what $R_{\text{sd}}$ and $\Delta L$ are. (Hint: Study the expression from part (a). Note that $\Delta L$ is the same for devices of all gate lengths. You may want to take measurements using a range of gate voltages and lengths.)

**(c)** Prove that

$$I_{\text{dsat}} = \frac{I_{\text{dsat0}}}{1+\dfrac{I_{\text{dsat0}}R_{\text{s}}}{(V_{\text{gs}}-V_{\text{t}})}}$$

where $I_{\text{dsat0}}$ is the saturation current in the absence of $R_{\text{s}}$.

**(d)** Given $T_{\text{ox}} = 3$ nm, $W/L = 1/0.1$ µm, $V_{\text{gs}} = 1.5$ V and $V_{\text{t}} = 0.4$ V, what is $I_{\text{dsat}}$ for $R_{\text{s}} = 0$, 100, and 1,000 Ω?

6.27 The drawn channel length of a transistor is in general different from the electrical channel length. We call the electrical channel length $L_{\text{eff}}$, while the drawn channel length is called $L_{\text{drawn}}$. Therefore the transistor $I_{\text{d}}$–$V_{\text{d}}$ curves should be represented by

$$I_{\text{dsat}} = \frac{\mu_{\text{n}} C_{\text{ox}} W}{2L_{\text{eff}}}(V_{\text{g}} - V_{\text{t}})^2 \qquad \text{for } V_{\text{d}} > V_{\text{dsat}}$$

$$I_{\text{dsat}} = \frac{\mu_{\text{n}} C_{\text{ox}} W}{L_{\text{eff}}}\left[(V_{\text{g}} - V_{\text{t}})\text{V}_{\text{ds}} - \frac{\text{V}_{\text{ds}}^2}{2}\right] \qquad \text{for } V_{\text{d}} < V_{\text{dsat}}$$

**(a)** How can you find the $L_{\text{eff}}$? (Hints: You may assume that several MOSFETs of different $L_{\text{drawn}}$, such as 1, 3, and 5 μm, are available. $W$ and $V_{\text{t}}$ are known.) Describe the procedure.

**(b)** Find the $\Delta L = L_{\text{drawn}} - L_{\text{eff}}$ and gate oxide thickness when you have three sets of $I_{\text{dsat}}$ data measured at the same $V_{\text{g}}$ as follows.

| ***L* (Drawn channel length)** | **1 (μm)** | **3 (μm)** | **5 (μm)** |
|---|---|---|---|
| $I_{\text{dsat}}$ (mA) | 2.59 | 0.8 | 0.476 |

The channel width, $W$, is 10 μm, and the mobility, μ, is 300 cm$^2$/V/s.

**(c)** If the $I_{\text{dsat}}$ of the transistor is measured at $V_{\text{gs}} = 2$ V, what is the threshold voltage of the transistor with $L_{\text{drawn}} = 1$ μm?

### ● Memory Devices ●

6.28 **(a)** Qualitatively describe the differences among SRAM, DRAM, and flash memory in terms of closeness to the basic CMOS manufacturing technology, write speed, volatility, and cell size.

**(b)** What are the main applications of SRAM, DRAM, and flash memory? Why are each suitable for the applications. Hint: Consider your answers to (a).

6.29 **(a)** Match the six transistors in Fig. 6–34b to the transistors in Fig. 6–34a. (Hint: $\text{M}_5$ and $\text{M}_6$ usually have larger $W$ than the transistors in the inverters.)

**(b)** Add the possible layout of the bit line and word line into Fig. 6–34b.

**(c)** Starting from the answer of (b), add another cell to the right and a third cell to the top of the original cell.

**(d)** Try to think of another way to arrange the six transistors (a new layout) that will pack them and the word line/bit lines into an even smaller cell area. (Hint: It is unlikely that you can pack them into a smaller area, although it should be fun spending 10 minutes trying. Furthermore, one cannot do this exercise fairly unless you know the detailed "design rules," which are the rules governing the size and spacing of all the features in a layout.)

## ● REFERENCES ●

1. Lilienfeld, J. E. "Method and Apparatus for Controlling Electronic Current." U.S. Patent 1,745,175 (1930).

2. Heil, O. "Improvements in or Relating to Electrical Amplifiers and Other Control Arrangements and Devices." British Patent 439,457 (1935).

**3.** Timp, G., et al. “The Ballistic Nano-transistor.” *International Electron Devices Meeting Technical Digest* 1999, 55–58.

**4.** Chen, K., H. C. Wann, et al. “The Impact of Device Scaling and Power Supply Change on CMOS Gate Performance,” *IEEE Electron Device Letters* 17 (5) (1996) 202–204.

**5.** Takagi, S., M. Iwase, and A. Toriumi. “On Universality of Inversion-Layer Mobility in N-and-P-channel MOSFETs.” *International Electron Devices Meeting Technical Digest* (1988), 398–401.

**6.** Komohara, S., et al. MOSFET Carrier Mobility Model Based on the Density of States at the DC Centroid in the Quantized Inversion Layer. 5th International Conference on VLSI and CAD (1997), 398–401.

**7.** Chen, K., C. Hu, et al. “Optimizing Sub-Quarter Micron CMOS Circuit Speed Considering Interconnect Loading Effects.” *IEEE Transactions on Electron Devices* 44 (9) (1997), 1556.

**8.** Assaderaghi, F., et al. “High-Field Transport of Inversion-Layer Electrons and Holes Including Velocity Overshoot.” *IEEE Transactions on Electron Devices* 44 (4) (1997), 664–671.

**9.** Toh, K. Y., P. K. Ko, and R. G. Meyer. “An Engineering Model for Short-Channel MOS Devices.” *IEEE Journal of Solid State Circuits* (1988), 23 (4), 950.

**10.** Hu, G. J., C. Chang, and Y. T. Chia. “Gate-Voltage Dependent Channel Length and Series Resistance of LDD MOSFETs.” *IEEE Transactions on Electron Devices* 34 (1985), 2469.

**11.** Assad, F., et al. “Performance Limits of Silicon MOSFETs.” *International Electron Devices Meeting Technical Digest* (1999) 547–550.

**12.** Hu, C. “A Compact Model for Rapidly Shrinking MOSFETs.” *Electron Devices Meeting Technical Digest* (2001), 13.1.1–13.1.4.

**13.** Hu, C. “BSIM Model for Circuit Design Using Advanced Technologies.” *VLSI Circuits Symposium Digest of Technical Papers* (2001), 5–10.

**14.** Hung, K. K., et al. “A Physics-Based MOSFET Noise Model for Circuit Simulations.” *IEEE Transactions on Electron Devices Technical Digest* (1990), 1323–1333.

**15.** Fukuda, K., et al. “Random Telegraph Noise in Flash Memories—Model and Technology Scaling.” *Electron Devices Meeting Technical Digest* (2007), 169–172.

**16.** Wu, S-Y., et al. “A 32 nm CMOS Low Power SoC Platform Technology for Foundry Applications with Functional High Density SRAM.” *IEDM Technical Digest* (2007), 263–266.

**17.** Park, Y. K., et al. “Highly Manufacturable 90 nm DRAM Technology.” *International Electron Devices Meeting Technical Digest* (2002), 819–822.

**18.** Brewer, J. E., and M. Gill, eds. *Nonvolatile Memory Technologies with Emphasis on Flash*. Hoboken, NJ: John Wiley & Sons, Inc., 2008.

**19.** Quader, K., et al. “Hot-Carrier Reliability Design Rules for Translating Device Degradation to CMOS Digital Circuit Degradation.” *IEEE Transactions on Electron Devices* 41 (1994), 681–691.

## ● GENERAL REFERENCES ●

**1.** Taur, Y., and T. H. Ning. *Fundamentals of Modern VLSI Devices*. Cambridge, UK: Cambridge University Press, 1998.

**2.** Pierret, R. F. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996.