MOS Transistor Theory

2

2.1 Introduction

In Chapter 1, the Metal-Oxide-Semiconductor (MOS) transistor was introduced in terms of its operation as an ideal switch. As we saw in Section 1.9, the performance and power of a chip depend on the current and capacitance of the transistors and wires. In this chapter, we will examine the characteristics of MOS transistors in more detail; Chapter 6 addresses wires.

Figure 2.1 shows some of the symbols that are commonly used for MOS transistors. The three-terminal symbols in Figure 2.1(a) are used in the great majority of schematics. If the body (substrate or well) connection needs to be shown, the four-terminal symbols in Figure 2.1(b) will be used. Figure 2.1(c) shows an example of other symbols that may be encountered in the literature.

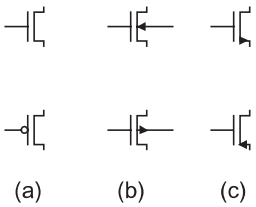


FIGURE 2.1
MOS transistor symbols

The MOS transistor is a *majority-carrier* device in which the current in a conducting channel between the source and drain is controlled by a voltage applied to the gate. In an nMOS transistor, the majority carriers are electrons; in a pMOS transistor, the majority carriers are holes. The behavior of MOS transistors can be understood by first examining an isolated MOS structure with a gate and body but no source or drain. Figure 2.2 shows a simple MOS structure. The top layer of the structure is a good conductor called the *gate*. Early transistors used metal gates. Transistor gates soon changed to use polysilicon, i.e., silicon formed from many small crystals, although metal gates are making a resurgence at 65 nm and beyond, as will be seen in Section 3.4.1.3. The middle layer is a very thin insulating film of SiO_2 called the *gate oxide*. The bottom layer is the doped silicon body. The figure shows a p-type body in which the carriers are holes. The body is grounded and a voltage is applied to the gate. The gate oxide is a good insulator so almost zero current flows from the gate to the body.[1]

In Figure 2.2(a) , a negative voltage is applied to the gate, so there is negative charge on the gate. The mobile positively charged holes are attracted to the region beneath the gate. This is called the *accumulation* mode. In Figure 2.2(b), a small positive voltage is applied to the gate, resulting in some positive charge on the gate. The holes in the body are repelled from the region directly beneath the gate, resulting in a *depletion* region forming below the gate. In Figure 2.2(c), a higher positive potential exceeding a critical threshold voltage V_t is applied, attracting more positive charge to the gate. The holes are repelled further and some free electrons in the body are attracted to the region beneath the gate. This conductive layer of electrons in the p-type body is called the *inversion* layer. The threshold

[1]Gate oxides are now only a handful of atomic layers thick and carriers sometimes tunnel through the oxide, creating a current through the gate. This effect is explored in Section 2.4.4.2.

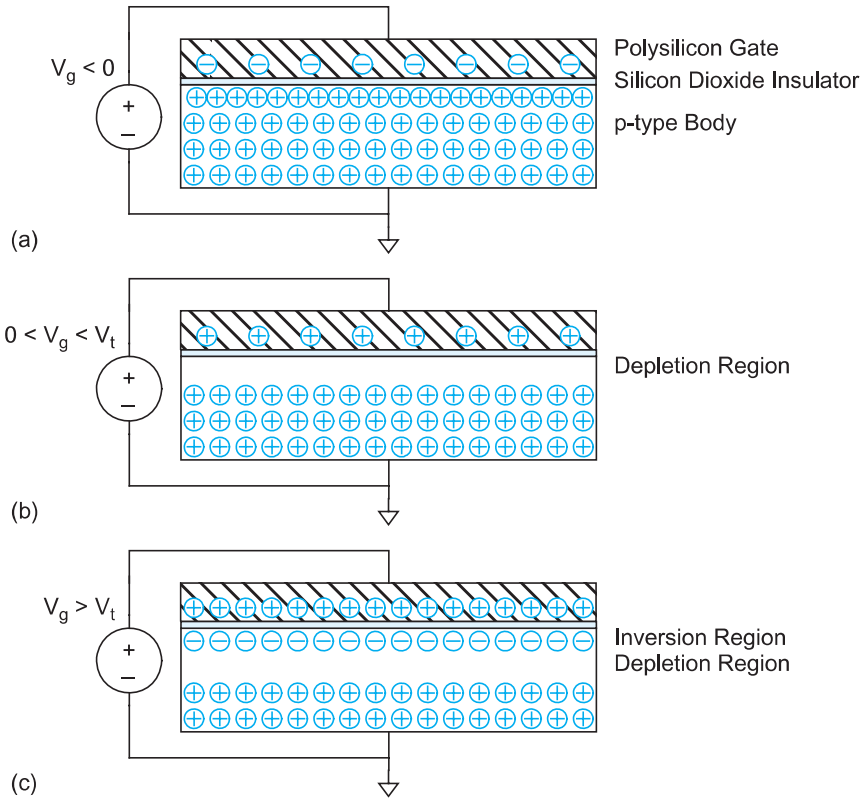


FIGURE 2.2 MOS structure demonstrating (a) accumulation, (b) depletion, and (c) inversion

voltage depends on the number of dopants in the body and the thickness t_{ox} of the oxide. It is usually positive, as shown in this example, but can be engineered to be negative.

Figure 2.3 shows an nMOS transistor. The transistor consists of the MOS stack between two n-type regions called the *source* and *drain*. In Figure 2.3(a), the gate-to-source voltage V_{gs} is less than the threshold voltage. The source and drain have free electrons. The body has free holes but no free electrons. Suppose the source is grounded. The junctions between the body and the source or drain are zero-biased or reverse-biased, so little or no current flows. We say the transistor is OFF, and this mode of operation is called *cutoff*. It is often convenient to approximate the current through an OFF transistor as zero, especially in comparison to the current through an ON transistor. Remember, however, that small amounts of current leaking through OFF transistors can become significant, especially when multiplied by millions or billions of transistors on a chip. In Figure 2.3(b), the gate voltage is greater than the threshold voltage. Now an inversion region of electrons (majority carriers) called the *channel* connects the source and drain, creating a conductive path and turning the transistor ON. The number of carriers and the conductivity increases with the gate voltage. The potential difference between drain and source is $V_{ds} = V_{gs} - V_{gd}$. If $V_{ds} = 0$ (i.e., $V_{gs} = V_{gd}$), there is no electric field tending to push current from drain to source.

When a small positive potential V_{ds} is applied to the drain (Figure 2.3(c)), current I_{ds} flows through the channel from drain to source.[2] This mode of operation is termed *linear*,

[2]The terminology of source and drain might initially seem backward. Recall that the current in an nMOS transistor is carried by moving electrons with a negative charge. Therefore, positive current from drain to source corresponds to electrons flowing from their source to their drain.

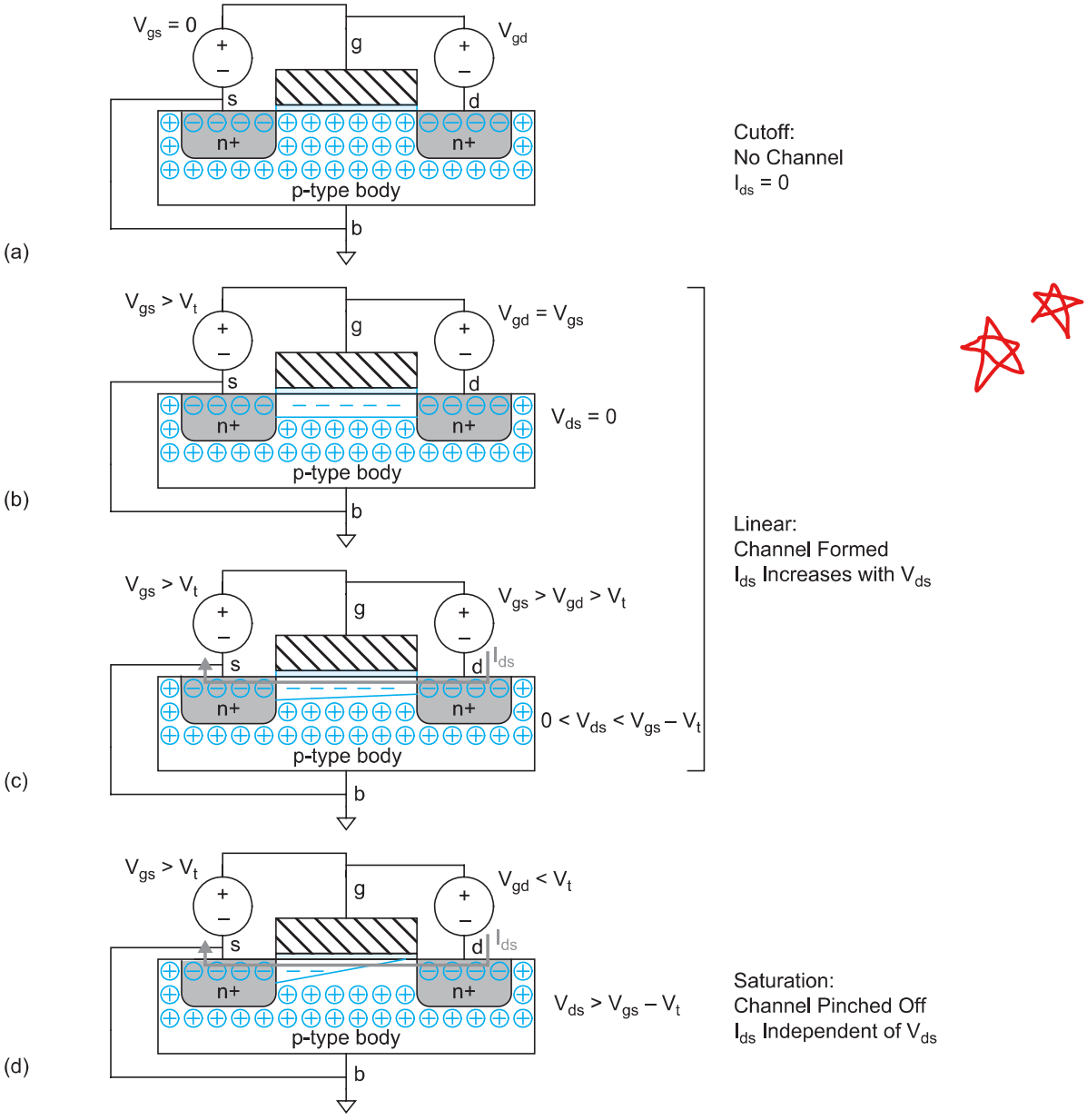


FIGURE 2.3 nMOS transistor demonstrating cutoff, linear, and saturation regions of operation

resistive, *triode*, *nonsaturated*, or *unsaturated*; the current increases with both the drain voltage and gate voltage. If V_{ds} becomes sufficiently large that $V_{gd} < V_t$, the channel is no longer inverted near the drain and becomes *pinched off* (Figure 2.3(d)). However, conduction is still brought about by the drift of electrons under the influence of the positive drain voltage. As electrons reach the end of the channel, they are injected into the depletion region near the drain and accelerated toward the drain. Above this drain voltage the current I_{ds} is controlled only by the gate voltage and ceases to be influenced by the drain. This mode is called *saturation*.

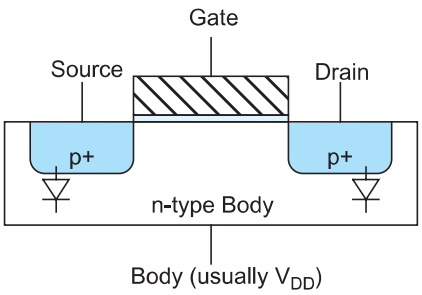


FIGURE 2.4 pMOS transistor

In summary, the nMOS transistor has three modes of operation. If $V_{gs} < V_t$, the transistor is cutoff (OFF). If $V_{gs} > V_t$, the transistor turns ON. If V_{ds} is small, the transistor acts as a linear resistor in which the current flow is proportional to V_{ds}. If $V_{gs} > V_t$ and V_{ds} is large, the transistor acts as a current source in which the current flow becomes independent of V_{ds}.

The pMOS transistor in Figure 2.4 operates in just the opposite fashion. The n-type body is tied to a high potential so the junctions with the p-type source and drain are normally reverse-biased. When the gate is also at a high potential, no current flows between drain and source. When the gate voltage is lowered by a threshold V_t, holes are attracted to form a p-type channel immediately beneath the gate, allowing current to flow between drain and source. The threshold voltages of the two types of transistors are not necessarily equal, so we use the terms V_{tn} and V_{tp} to distinguish the nMOS and pMOS thresholds.

Although MOS transistors are symmetrical, by convention we say that majority carriers flow from their source to their drain. Because electrons are negatively charged, the source of an nMOS transistor is the more negative of the two terminals. Holes are positively charged so the source of a pMOS transistor is the more positive of the two terminals. In static CMOS gates, the source is the terminal closer to the supply rail and the drain is the terminal closer to the output.

We begin in Section 2.2 by deriving an ideal model relating current and voltage (I-V) for a transistor. The delay of MOS circuits is determined by the time required for this current to charge or discharge the capacitance of the circuits. Section 2.3 investigates transistor capacitances. The gate of an MOS transistor is inherently a good capacitor with a thin dielectric; indeed, its capacitance is responsible for attracting carriers to the channel and thus for the operation of the device. The p–n junctions from source or drain to the body contribute additional *parasitic* capacitance. The capacitance of wires interconnecting the transistors is also important and will be explored in Section 6.2.2.

This idealized I-V model provides a general qualitative understanding of transistor behavior but is of limited quantitative value. On the one hand, it neglects too many effects that are important in transistors with short channel lengths L. Therefore, the model is not sufficient to calculate current accurately. Circuit simulators based on SPICE [Nagel75] use models such as BSIM that capture transistor behavior quite thoroughly but require entire books to fully describe [Cheng99]. Chapter 8 discusses simulation with SPICE. The most important effects seen in these simulations that impact digital circuit designers are examined in Section 2.4. On the other hand, the idealized I-V model is still too complicated to use in back-of-the-envelope calculations tuning the performance of large circuits. Therefore, we will develop even simpler models for performance estimation in Chapter 4.

Section 2.5 wraps up this chapter by applying the I-V models to understand the DC transfer characteristics of CMOS gates and pass transistors.

2.2 Long-Channel I-V Characteristics

As stated previously, MOS transistors have three regions of operation:

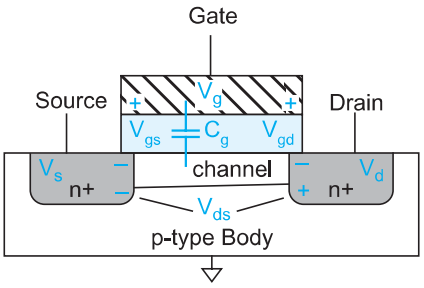
- Cutoff or subthreshold region
- Linear region
- Saturation region

Let us derive a model [Shockley52, Cobbold70, Sah64] relating the current and voltage (I-V) for an nMOS transistor in each of these regions. The model assumes that the channel length is long enough that the lateral electric field (the field between source and drain) is relatively low, which is no longer the case in nanometer devices. This model is variously known as the *long-channel*, *ideal*, *first-order*, or *Shockley* model. Subsequent sections will refine the model to reflect high fields, leakage, and other nonidealities.

The long-channel model assumes that the current through an OFF transistor is 0. When a transistor turns ON ($V_{gs} > V_t$), the gate attracts carriers (electrons) to form a channel. The electrons drift from source to drain at a rate proportional to the electric field between these regions. Thus, we can compute currents if we know the amount of charge in the channel and the rate at which it moves. We know that the charge on each plate of a capacitor is $Q = CV$. Thus, the charge in the channel Q_{channel} is

$$Q_{\text{channel}} = C_g\left(V_{gc} - V_t\right) \tag{2.1}$$

where C_g is the capacitance of the gate to the channel and $V_{gc} - V_t$ is the amount of voltage attracting charge to the channel beyond the minimum required to invert from p to n. The gate voltage is referenced to the channel, which is not grounded. If the source is at V_s and the drain is at V_d, the average is $V_c = (V_s + V_d)/2 = V_s + V_{ds}/2$. Therefore, the mean difference between the gate and channel potentials V_{gc} is $V_g - V_c = V_{gs} - V_{ds}/2$, as shown in Figure 2.5.



Average gate to channel potential:

$V_{gc} = (V_{gs} + V_{gd})/2 = V_{gs} - V_{ds}/2$

FIGURE 2.5 Average gate to channel voltage

We can model the gate as a parallel plate capacitor with capacitance proportional to area over thickness. If the gate has length L and width W and the oxide thickness is t_{ox}, as shown in Figure 2.6, the capacitance is

$$C_g = k_{ox}\varepsilon_0 \frac{WL}{t_{ox}} = \varepsilon_{ox}\frac{WL}{t_{ox}} = C_{ox}WL \tag{2.2}$$

where ε_0 is the permittivity of free space, 8.85×10^{-14} F/cm, and the permittivity of SiO_2 is $k_{ox} = 3.9$ times as great. Often, the ε_{ox}/t_{ox} term is called C_{ox}, the capacitance per unit area of the gate oxide.

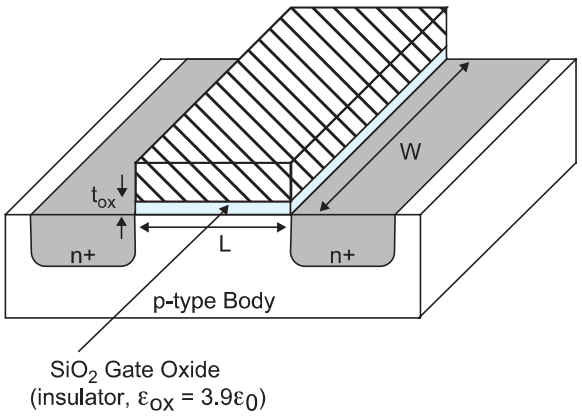


FIGURE 2.6 Transistor dimensions

Some nanometer processes use a different gate dielectric with a higher dielectric constant. In these processes, we call t_{ox} the *equivalent oxide thickness* (EOT), the thickness of a layer of SiO_2 that has the same C_{ox}. In this case, t_{ox} is thinner than the actual dielectric.

Each carrier in the channel is accelerated to an average velocity, v, proportional to the lateral electric field, i.e., the field between source and drain. The constant of proportionality μ is called the *mobility*.

$$v = \mu E \tag{2.3}$$

A typical value of μ for electrons in an nMOS transistor with low electric fields is 500–700 $cm^2/V \cdot s$. However, most transistors today operate at far higher fields where the mobility is severely curtailed (see Section 2.4.1).

The electric field E is the voltage difference between drain and source V_{ds} divided by the channel length

$$E = \frac{V_{ds}}{L} \tag{2.4}$$

The time required for carriers to cross the channel is the channel length divided by the carrier velocity: L/v. Therefore, the current between source and drain is the total amount of charge in the channel divided by the time required to cross

$$\begin{aligned} I_{ds} &= \frac{Q_{\text{channel}}}{L/v} \\ &= \mu C_{ox} \frac{W}{L}\left(V_{gs} - V_t - V_{ds}/2\right)V_{ds} \\ &= \beta\left(V_{GT} - V_{ds}/2\right)V_{ds} \end{aligned} \tag{2.5}$$

where

$$\beta = \mu C_{ox} \frac{W}{L};\ V_{GT} = V_{gs} - V_t \tag{2.6}$$

The term $V_{gs} - V_t$ arises so often that it is convenient to abbreviate it as V_{GT}. EQ (2.5) describes the linear region of operation, for $V_{gs} > V_t$, but V_{ds} relatively small. It is called *linear* or *resistive* because when $V_{ds} << V_{GT}$, I_{ds} increases almost linearly with V_{ds}, just like an ideal resistor. The geometry and technology-dependent parameters are sometimes merged into a single factor β. Do not confuse this use of β with the same symbol used for the ratio of collector-to-base current in a bipolar transistor. Some texts [Gray01] lump the technology-dependent parameters alone into a constant called "k prime."[3]

$$k' = \mu C_{ox} \tag{2.7}$$

If $V_{ds} > V_{dsat} \equiv V_{GT}$, the channel is no longer inverted in the vicinity of the drain; we say it is pinched off. Beyond this point, called the *drain saturation voltage*, increasing the drain voltage has no further effect on current. Substituting $V_{ds} = V_{dsat}$ at this point of maximum current into EQ (2.5), we find an expression for the saturation current that is independent of V_{ds}.

$$I_{ds} = \frac{\beta}{2} V_{GT}^2 \tag{2.8}$$

[3]Other sources (e.g., MOSIS) define $k' = \frac{\mu C_{ox}}{2}$; check the definition before using quoted data.

This expression is valid for $V_{gs} > V_t$ and $V_{ds} > V_{\text{dsat}}$. Thus, long-channel MOS transistors are said to exhibit *square-law behavior* in saturation.

Two key figures of merit for a transistor are I_{on} and I_{off}. I_{on} (also called I_{dsat}) is the ON current, I_{ds}, when $V_{gs} = V_{ds} = V_{DD}$. I_{off} is the OFF current when $V_{gs} = 0$ and $V_{ds} = V_{DD}$. According to the long-channel model, $I_{\text{off}} = 0$ and

$$I_{\text{on}} = \frac{\beta}{2}\left(V_{DD} - V_t\right) \tag{2.9}$$

EQ (2.10) summarizes the current in the three regions:

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{Cutoff} \\ \beta\left(V_{GT} - V_{ds}/2\right)V_{ds} & V_{ds} < V_{\text{dsat}} & \text{Linear} \\ \dfrac{\beta}{2}V_{GT}^2 & V_{ds} > V_{\text{dsat}} & \text{Saturation} \end{cases} \tag{2.10}$$

Example 2.1

Consider an nMOS transistor in a 65 nm process with a minimum drawn channel length of 50 nm ($\lambda = 25$ nm). Let $W/L = 4/2\ \lambda$ (i.e., 0.1/0.05 μm). In this process, the gate oxide thickness is 10.5 Å. Estimate the high-field mobility of electrons to be 80 $\text{cm}^2/\text{V}\cdot\text{s}$ at 70 °C. The threshold voltage is 0.3 V. Plot I_{ds} vs. V_{ds} for $V_{gs} = 0$, 0.2, 0.4, 0.6, 0.8, and 1.0 V using the long-channel model.

SOLUTION: We first calculate β.

$$\beta = \mu C_{\text{ox}}\frac{W}{L} = \left(80\frac{\text{cm}^2}{\text{V}\cdot\text{s}}\right)\left(\frac{3.9\times 8.85\times 10^{-14}\ \frac{\text{F}}{\text{cm}}}{10.5\times 10^{-8}\text{cm}}\right)\left(\frac{W}{L}\right) = 262\frac{W}{L}\frac{\text{A}}{\text{V}^2} \tag{2.11}$$

Figure 2.7(a) shows the I-V characteristics for the transistor. According to the first-order model, the current is zero for gate voltages below V_t. For higher gate voltages, current increases linearly with V_{ds} for small V_{ds}. As V_{ds} reaches the saturation point $V_{\text{dsat}} = V_{GT}$, current rolls off and eventually becomes independent of V_{ds} when the transistor is saturated. We will later see that the Shockley model overestimates current at high voltage because it does not account for mobility degradation and velocity saturation caused by the high electric fields.

pMOS transistors behave in the same way, but with the signs of all voltages and currents reversed. The I-V characteristics are in the third quadrant, as shown in Figure 2.7(b). To keep notation simple in this text, we will disregard the signs and just remember that the current flows from source to drain in a pMOS transistor. The mobility of holes in silicon is typically lower than that of electrons. This means that pMOS transistors provide less current than nMOS transistors of comparable size and hence are slower. The symbols μ_n and μ_p are used to distinguish mobility of electrons and of holes in nMOS and pMOS transistors, respectively. The *mobility ratio* μ_n/μ_p is typically 2–3; we will generally use 2 for examples in this book. The pMOS transistor has the same geometry as the nMOS in Figure 2.7(a), but with $\mu_p = 40$ $\text{cm}^2/\text{V}\cdot\text{s}$ and $V_{tp} = -0.3$ V. Similarly, β_n, β_p, k'_n, and k'_p are sometimes used to distinguish nMOS and pMOS I-V characteristics.

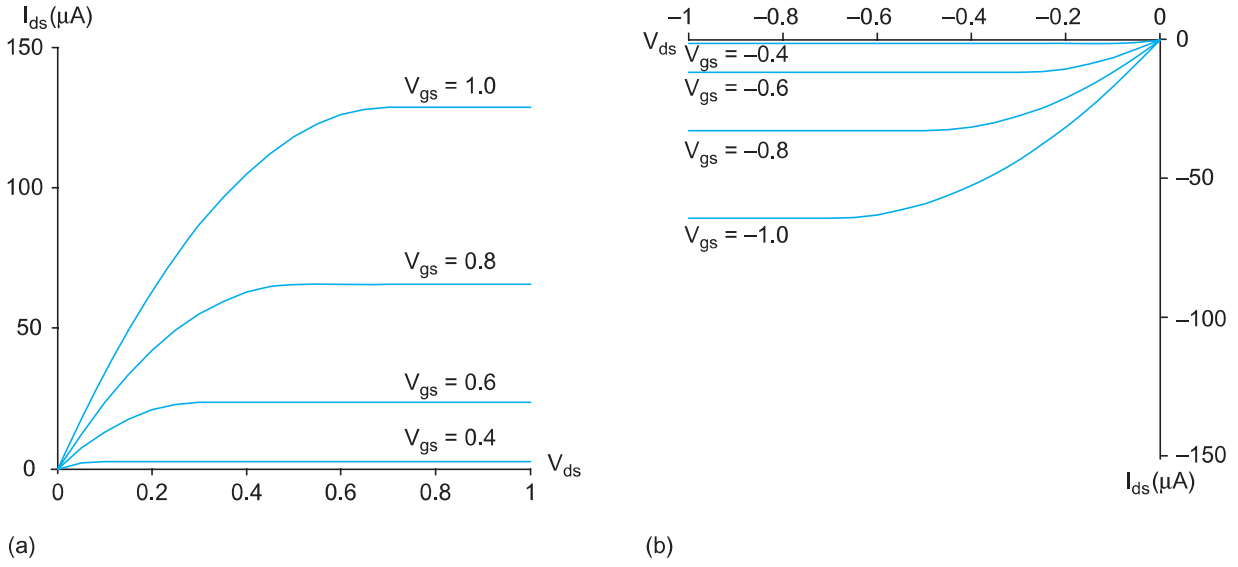


FIGURE 2.7 I-V characteristics of ideal 4/2 λ (a) nMOS and (b) pMOS transistors

2.3 C-V Characteristics

Each terminal of an MOS transistor has capacitance to the other terminals. In general, these capacitances are nonlinear and voltage dependent (C-V); however, they can be approximated as simple capacitors when their behavior is averaged across the switching voltages of a logic gate. This section first presents simple models of each capacitance suitable for estimating delay and power consumption of transistors. It then explores more detailed models used for circuit simulation. The more detailed models may be skipped on a first reading.

2.3.1 Simple MOS Capacitance Models

The gate of an MOS transistor is a good capacitor. Indeed, its capacitance is necessary to attract charge to invert the channel, so high gate capacitance is required to obtain high I_{ds}. As seen in Section 2.2, the gate capacitor can be viewed as a parallel plate capacitor with the gate on top and channel on bottom with the thin oxide dielectric between. Therefore, the capacitance is

$$C_g = C_{ox}WL \tag{2.12}$$

The bottom plate of the capacitor is the channel, which is not one of the transistor's terminals. When the transistor is on, the channel extends from the source (and reaches the drain if the transistor is unsaturated, or stops short in saturation). Thus, we often approximate the gate capacitance as terminating at the source and call the capacitance C_{gs}.

Most transistors used in logic are of minimum manufacturable length because this results in greatest speed and lowest dynamic power consumption.[4] Thus, taking this mini-

[4]Some designs use slightly longer than minimum transistors that have higher thresholds because of the short-channel effect (see Sections 2.4.3.3 and 5.3.3). This avoids the cost of an extra mask step for high-V_t transistors. The change in channel length is small (~5–10%), so the change in gate capacitance is minor.