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.

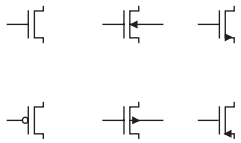


FIGURE 2.1
MOS transistor symbols

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.

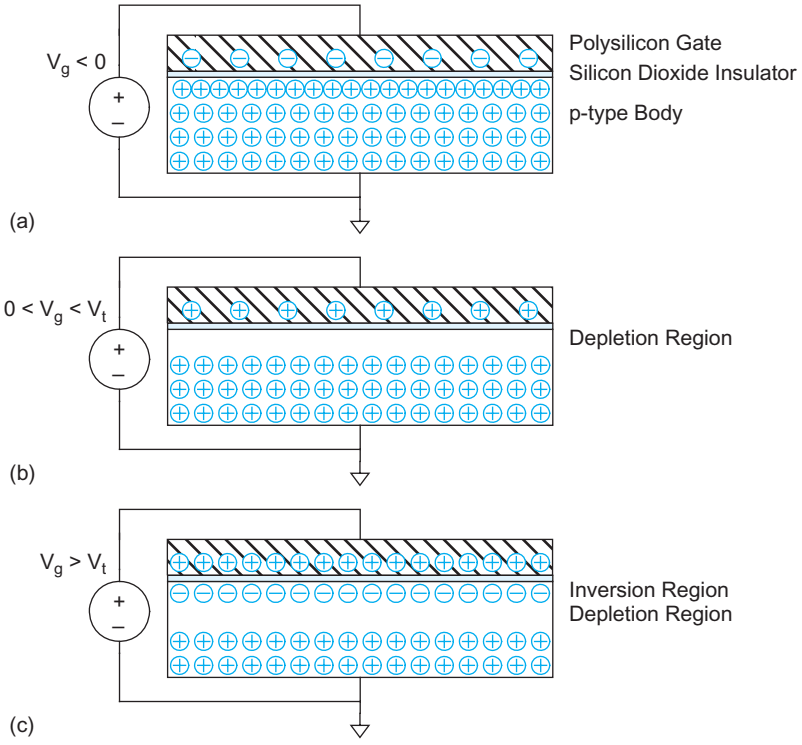


FIGURE 2.2 MOS structure demonstrating (a) accumulation, (b) depletion, and (c) inversion

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.

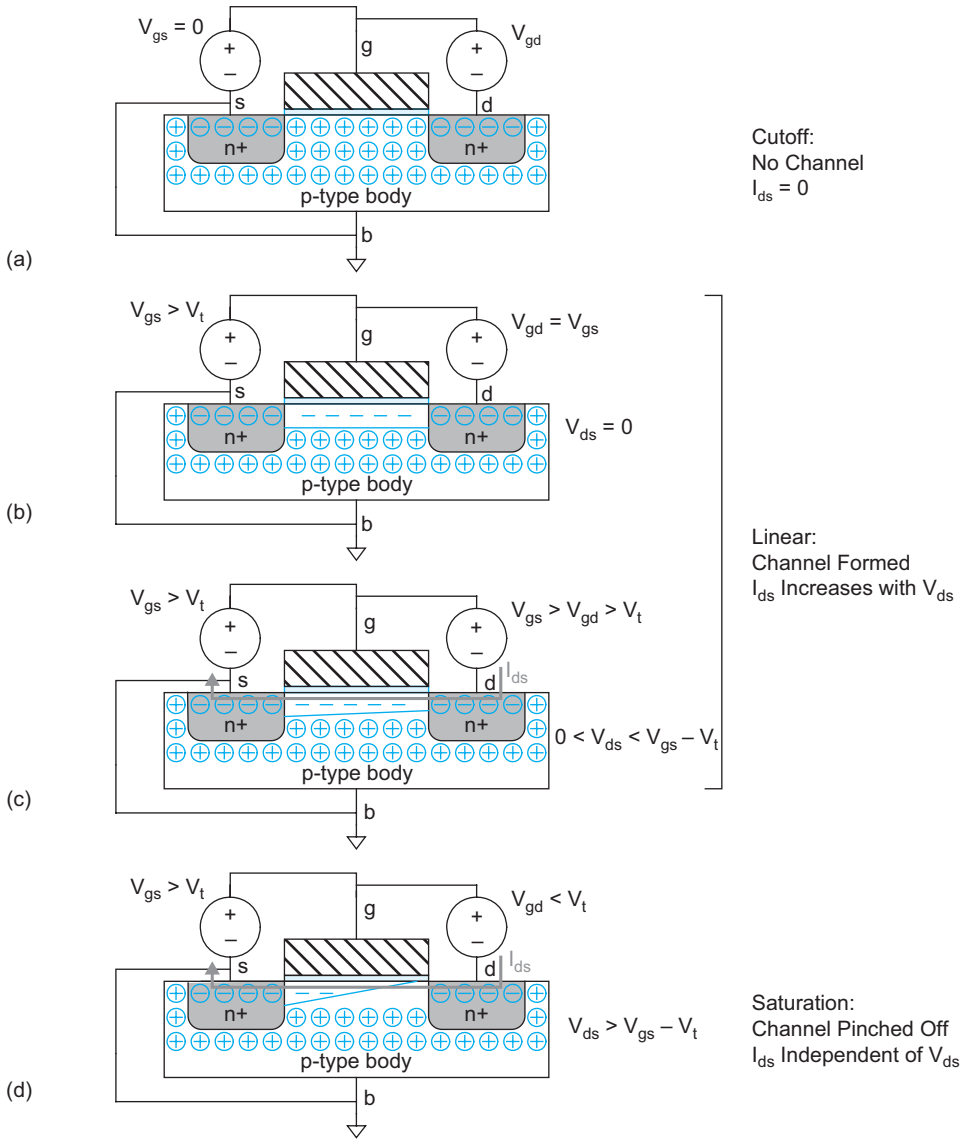


FIGURE 2.3 nMOS transistor demonstrating cutoff, linear, and saturation regions of operation

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*.

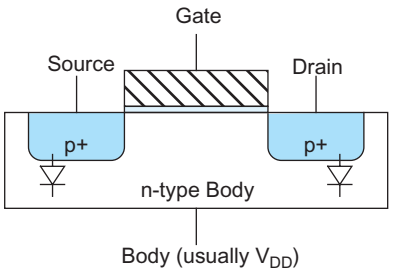


FIGURE 2.4 pMOS transistor

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:

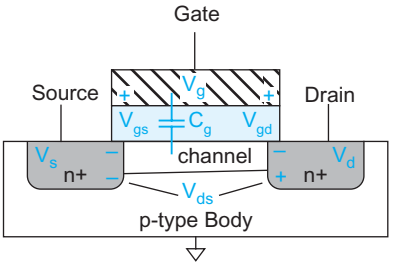
- Cutoff or subthreshold region
- Linear region
- Saturation region

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.

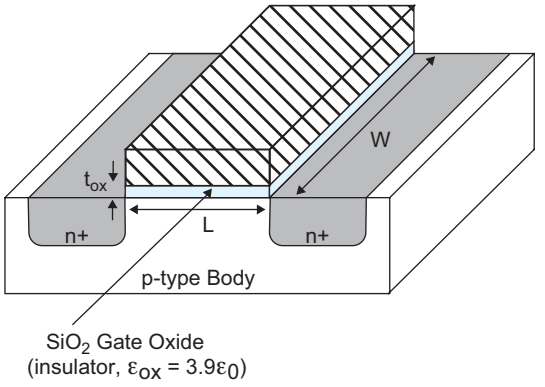


FIGURE 2.6 Transistor dimensions

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 cm^2/V·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 $\mu_p = 40$ cm^2/V·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.

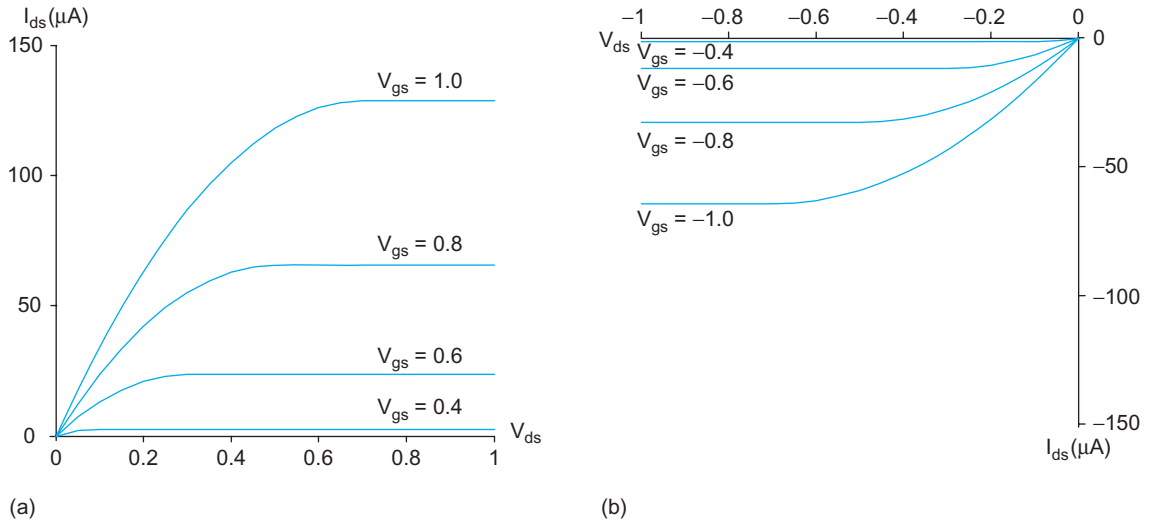


FIGURE 2.7 I-V characteristics of ideal 4/2 λ (a) nMOS and (b) pMOS transistors

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.

mum L as a constant for a particular process, we can define

$$C_g = C_{\text{permicron}} \times W \tag{2.13}$$

where

$$C_{\text{permicron}} = C_{\text{ox}} L = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} L \tag{2.14}$$

Notice that if we develop a more advanced manufacturing process in which both the channel length and oxide thickness are reduced by the same factor, $C_{\text{permicron}}$ remains unchanged. This relationship is handy for quick calculations but not exact; $C_{\text{permicron}}$ has fallen from about 2 fF/μm in old processes to about 1 fF/μm at the 90 and 65 nm nodes. Table 8.5 lists gate capacitance for a variety of processes.

In addition to the gate, the source and drain also have capacitances. These capacitances are not fundamental to operation of the devices, but do impact circuit performance and hence are called *parasitic* capacitors. The source and drain capacitances arise from the p–n junctions between the source or drain diffusion and the body and hence are also called *diffusion*[5] capacitance C_{sb} and C_{db}. A *depletion region* with no free carriers forms along the junction. The depletion region acts as an insulator between the conducting p- and n-type regions, creating capacitance across the junction. The capacitance of these junctions depends on the area and perimeter of the source and drain diffusion, the depth of the diffusion, the doping levels, and the voltage. As diffusion has both high capacitance and high resistance, it is generally made as small as possible in the layout. Three types of diffusion regions are frequently seen, illustrated by the two series transistors in Figure 2.8. In Figure

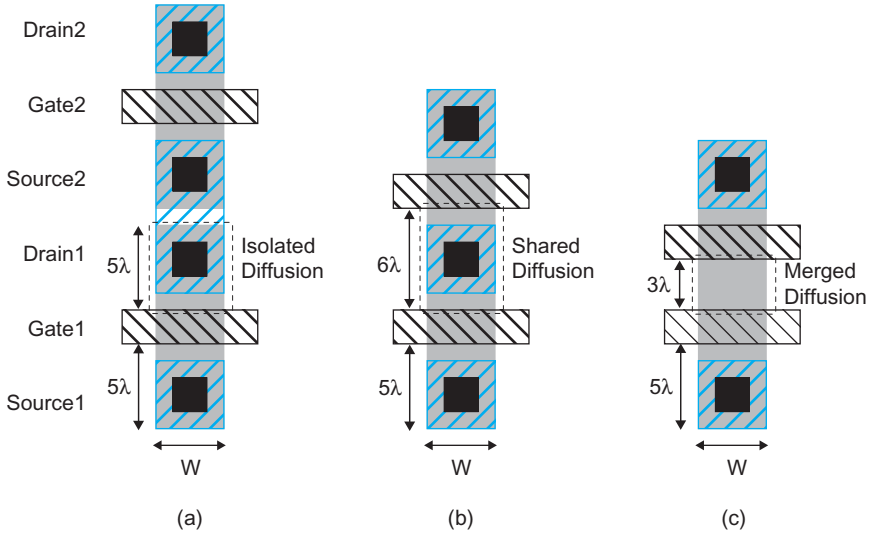


FIGURE 2.8 Diffusion region geometries

[5]Device engineers more properly call this *depletion* capacitance, but the term *diffusion* capacitance is widely used by circuit designers.

2.8(a), each source and drain has its own *isolated* region of contacted diffusion. In Figure 2.8(b), the drain of the bottom transistor and source of the top transistor form a *shared* contacted diffusion region. In Figure 2.8(c), the source and drain are *merged* into an uncontacted region. The average capacitance of each of these types of regions can be calculated or measured from simulation as a transistor switches between V_{DD} and GND. Table 8.5 also lists the capacitance for each scenario for a variety of processes.

For the purposes of hand estimation, you can observe that the diffusion capacitance C_{sb} and C_{db} of contacted source and drain regions is comparable to the gate capacitance (e.g., 1–2 fF/μm of gate width). The diffusion capacitance of the uncontacted source or drain is somewhat less because the area is smaller but the difference is usually unimportant for hand calculations. These values of $C_g = C_{sb} = C_{db} \approx 1$fF/$\mu$m will be used in examples throughout the text, but you should obtain the appropriate data for your process using methods to be discussed in Section 8.4.

2.3.2 Detailed MOS Gate Capacitance Model

The MOS gate sits above the channel and may partially overlap the source and drain diffusion areas. Therefore, the gate capacitance has two components: the intrinsic capacitance C_{gc} (over the channel) and the overlap capacitances C_{gol} (to the source and drain).

The intrinsic capacitance was approximated as a simple parallel plate in EQ (2.12) with capacitance $C_0 = WLC_{ox}$. However, the bottom plate of the capacitor depends on the mode of operation of the transistor. The intrinsic capacitance has three components representing the different terminals connected to the bottom plate: C_{gb} (gate-to-body), C_{gs} (gate-to-source), and C_{gd} (gate-to-drain). Figure 2.9(a) plots capacitance vs. V_{gs} in the cutoff region and for small V_{ds}, while 2.9(b) plots capacitance vs. V_{ds} in the linear and saturation regions [Dally98].

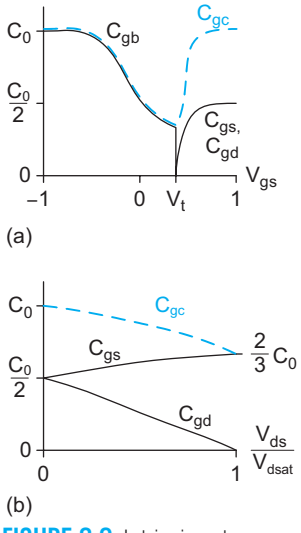


FIGURE 2.9 Intrinsic gate capacitance $C_{gc} = C_{gs} + C_{gd} + C_{gb}$ as a function of (a) V_{gs} and (b) V_{ds}

1. *Cutoff.* When the transistor is OFF ($V_{gs} < V_t$), the channel is not inverted and charge on the gate is matched with opposite charge from the body. This is called C_{gb}, the gate-to-body capacitance. For negative V_{gs}, the transistor is in accumulation and $C_{gb} = C_0$. As V_{gs} increases but remains below a threshold, a depletion region forms at the surface. This effectively moves the bottom plate downward from the oxide, reducing the capacitance, as shown in Figure 2.9(a).

2. *Linear.* When $V_{gs} > V_t$, the channel inverts and again serves as a good conductive bottom plate. However, the channel is connected to the source and drain, rather than the body, so C_{gb} drops to 0. At low values of V_{ds}, the channel charge is roughly shared between source and drain, so $C_{gs} = C_{gd} = C_0/2$. As V_{ds} increases, the region near the drain becomes less inverted, so a greater fraction of the capacitance is attributed to the source and a smaller fraction to the drain, as shown in Figure 2.9(b).

3. *Saturation.* At $V_{ds} > V_{dsat}$, the transistor saturates and the channel pinches off. At this point, all the intrinsic capacitance is to the source, as shown in Figure 2.9(b). Because of pinchoff, the capacitance in saturation reduces to $C_{gs} = 2/3\ C_0$ for an ideal transistor [Gray01].

The behavior in these three regions can be approximated as shown in Table 2.1.

TABLE 2.1 Approximation for intrinsic MOS gate capacitance

Parameter	Cutoff	Linear	Saturation
C_{gb}	$\leq C_0$	0	0
C_{gs}	0	$C_0/2$	2/3 C_0
C_{gd}	0	$C_0/2$	0
$C_g = C_{gs} + C_{gd} + C_{gb}$	C_0	C_0	2/3 C_0

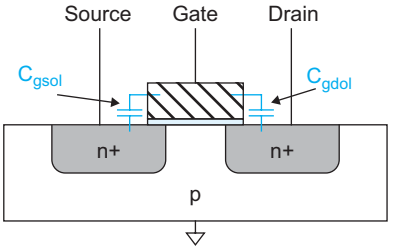


FIGURE 2.10 Overlap capacitance

The gate overlaps the source and drain in a real device and also has fringing fields terminating on the source and drain. This leads to additional overlap capacitances, as shown in Figure 2.10. These capacitances are proportional to the width of the transistor. Typical values are $C_{gsol} = C_{gdol} = 0.2 - 0.4$ fF/μm. They should be added to the intrinsic gate capacitance to find the total.

$$C_{gsol(\text{overlap})} = C_{gsol}W$$
$$C_{gdol(\text{overlap})} = C_{gdol}W \qquad \textbf{(2.15)}$$

It is convenient to view the gate capacitance as a single-terminal capacitor attached to the gate (with the other side not switching). Because the source and drain actually form second terminals, the effective gate capacitance varies with the switching activity of the source and drain. Figure 2.11 shows the effective gate capacitance in a 0.35 μm process for seven different combinations of source and drain behavior [Bailey98].

More accurate modeling of the gate capacitance may be achieved by using a charge-based model [Cheng99]. For the purpose of delay calculation of digital circuits, we usually approximate $C_g = C_{gs} + C_{gd} + C_{gb} \approx C_0 + 2C_{gol}W$ or use an effective capacitance extracted

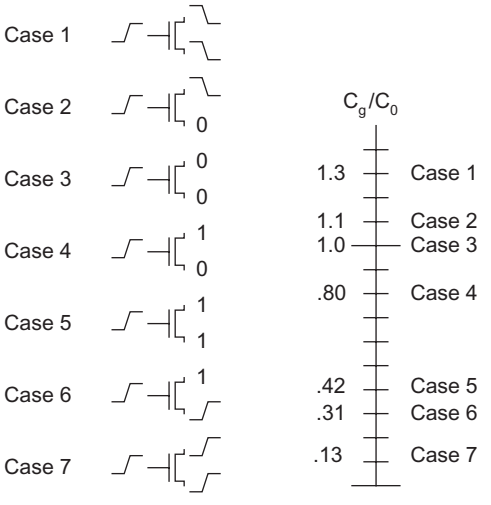


FIGURE 2.11 Data-dependent gate capacitance

from simulation [Nose00b]. It is important to remember that this model significantly overestimates the capacitance of transistors operating just below threshold.

2.3.3 Detailed MOS Diffusion Capacitance Model

As mentioned in Section 2.3.1, the p–n junction between the source diffusion and the body contributes parasitic capacitance across the depletion region. The capacitance depends on both the *area AS* and *sidewall perimeter PS* of the source diffusion region. The geometry is illustrated in Figure 2.12. The area is $AS = WD$. The perimeter is $PS = 2W + 2D$. Of this perimeter, W abuts the channel and the remaining $W + 2D$ does not.

The total source parasitic capacitance is

$$C_{sb} = AS \times C_{jbs} + PS \times C_{jbssw} \tag{2.16}$$

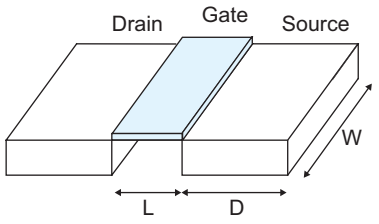


FIGURE 2.12 Diffusion region geometry

where C_{jbs} (the capacitance of the junction between the body and the bottom of the source) has units of capacitance/area and C_{jbssw} (the capacitance of the junction between the body and the side walls of the source) has units of capacitance/length.

Because the depletion region thickness depends on the bias conditions, these parasitics are nonlinear. The area junction capacitance term is [Gray01]

$$C_{jbs} = C_J \left(1 + \frac{V_{sb}}{\psi_0}\right)^{-M_J} \tag{2.17}$$

C_J is the junction capacitance at zero bias and is highly process-dependent. M_J is the *junction grading coefficient*, typically in the range of 0.5 to 0.33 depending on the abruptness of the diffusion junction. ψ_0 is the *built-in potential* that depends on doping levels.

$$\psi_0 = v_T \ln \frac{N_A N_D}{n_i^2} \tag{2.18}$$

v_T is the *thermal voltage* from thermodynamics, not to be confused with the threshold voltage V_t. It has a value equal to kT/q (26 mV at room temperature), where $k = 1.380 \times 10^{-23}$ J/K is Boltzmann's constant, T is absolute temperature (300 K at room temperature), and $q = 1.602 \times 10^{-19}$ C is the charge of an electron. N_A and N_D are the doping levels of the body and source diffusion region. n_i is the intrinsic carrier concentration in undoped silicon and has a value of 1.45×10^{10} cm^{-3} at 300 K.

The sidewall capacitance term is of a similar form but uses different coefficients.

$$C_{jbssw} = C_{JSW} \left(1 + \frac{V_{sb}}{\psi_{SW}}\right)^{-M_{JSW}} \tag{2.19}$$

In processes below about 0.35 μm that employ shallow trench isolation surrounding transistors with an SiO_2 insulator (see Section 3.2.6), the sidewall capacitance along the nonconductive trench tends to be minimal, while the sidewall facing the channel is more significant. In some SPICE models, the capacitance of this sidewall abutting the gate and channel is specified with another set of parameters:

$$C_{jbsswg} = C_{JSW} \left(1 + \frac{V_{sb}}{\psi_{SWG}}\right)^{-M_{JSWG}} \tag{2.20}$$

Section 8.3.4 discusses SPICE perimeter capacitance models further.

The drain diffusion has a similar parasitic capacitance dependent on *AD*, *PD*, and V_{db}. Equivalent relationships hold for pMOS transistors, but doping levels differ. As the capacitances are voltage-dependent, the most useful information to digital designers is the value averaged across a switching transition. This is the C_{sb} or C_{db} value that was presented in Section 2.3.1.

Example 2.2

Calculate the diffusion parasitic C_{db} of the drain of a unit-sized contacted nMOS transistor in a 65 nm process when the drain is at 0 V and again at $V_{DD} = 1.0$ V. Assume the substrate is grounded. The diffusion region conforms to the design rules from Figure 2.8 with $\lambda = 25$ nm. The transistor characteristics are $CJ = 1.2$ fF/μm^2, $MJ = 0.33$, $CJSW = 0.1$ fF/μm, $CJSWG = 0.36$ fF/μm, $MJSW = MJSWG = 0.10$, and $\Psi_0 = 0.7$ V at room temperature.

SOLUTION: From Figure 2.8, we find a unit-size diffusion contact is $4 \times 5\ \lambda$, or $0.1 \times 0.125\ \mu$m. The area is 0.0125 μm^2 and perimeter is 0.35 μm plus 0.1 μm along the channel. At zero bias, $C_{jbd} = 1.2$ fF/μm^2, $C_{jbdsw} = 0.1$ fF/μm, and $C_{jbdswg} = 0.36$ fF/μm. Hence, the total capacitance is

$$C_{db}(0\text{ V}) = (0.0125\mu\text{m}^2)\left(1.2\frac{\text{fF}}{\mu\text{m}^2}\right) + (0.35\mu\text{m})\left(0.1\frac{\text{fF}}{\mu\text{m}}\right) + (0.1\mu\text{m})\left(0.36\frac{\text{fF}}{\mu\text{m}}\right) = 0.086\text{ fF} \tag{2.21}$$

At a drain voltage of V_{DD}, the capacitance reduces to

$$C_{db}(1\text{ V}) = (0.0125\mu\text{m}^2)\left(1.2\frac{\text{fF}}{\mu\text{m}^2}\right)\left(1+\frac{1.0}{0.7}\right)^{-0.33} + \left[(0.35\mu\text{m})\left(0.1\frac{\text{fF}}{\mu\text{m}}\right) + (0.1\mu\text{m})\left(0.36\frac{\text{fF}}{\mu\text{m}}\right)\right]\left(1+\frac{1.0}{0.7}\right)^{-0.10} = 0.076\text{ fF} \tag{2.22}$$

For the purpose of manual performance estimation, this nonlinear capacitance is too much effort. An effective capacitance averaged over the switching range is quite satisfactory for digital applications. In this example, the effective drain capacitance would be approximated as the average of the two extremes, 0.081 fF.

Diffusion regions were historically used for short wires called *runners* in processes with only one or two metal levels. Diffusion capacitance and resistance are large enough that such practice is now discouraged; diffusion regions should be kept as small as possible on nodes that switch.

In summary, an MOS transistor can be viewed as a four-terminal device with capacitances between each terminal pair, as shown in Figure 2.13. The gate capacitance includes an intrinsic component (to the body, source and drain, or source alone, depending on operating regime) and overlap terms with the source and drain. The source and drain have parasitic diffusion capacitance to the body.

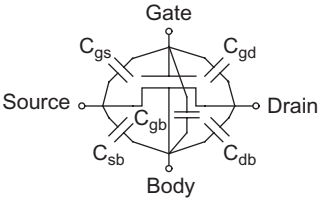


FIGURE 2.13 Capacitance of an MOS transistor

2.4 Nonideal I-V Effects

The long-channel I-V model of EQ (2.10) neglects many effects that are important to devices with channel lengths below 1 micron. This section summarizes the effects of greatest significance to designers, then models each one in more depth.

Figure 2.14 compares the simulated I-V characteristics of a 1-micron wide nMOS transistor in a 65 nm process to the ideal characteristics computed in Section 2.2. The saturation current increases less than quadratically with increasing V_{gs}. This is caused by two effects: velocity saturation and mobility degradation. At high lateral field strengths (V_{ds}/L), carrier velocity ceases to increase linearly with field strength. This is called *velocity saturation* and results in lower I_{ds} than expected at high V_{ds}. At high vertical field strengths (V_{gs}/t_{ox}), the carriers scatter off the oxide interface more often, slowing their progess. This *mobility degradation* effect also leads to less current than expected at high V_{gs}. The saturation current of the nonideal transistor increases somewhat with V_{ds}. This is caused by *channel length modulation*, in which higher V_{ds} increases the size of the depletion region around the drain and thus effectively shortens the channel.

The threshold voltage indicates the gate voltage necessary to invert the channel and is primarily determined by the oxide thickness and channel doping levels. However, other fields in the transistor have some effect on the channel, effectively modifying the threshold voltage. Increasing the potential between the source and body raises the threshold through the *body effect*. Increasing the drain voltage lowers the threshold through *drain-induced barrier lowering*. Increasing the channel length raises the threshold through the *short channel effect*.

Several sources of leakage result in current flow in nominally OFF transistors. When $V_{gs} < V_t$, the current drops off exponentially rather than abruptly becoming zero. This is called *subthreshold conduction*. The current into the gate I_g is ideally 0. However, as the

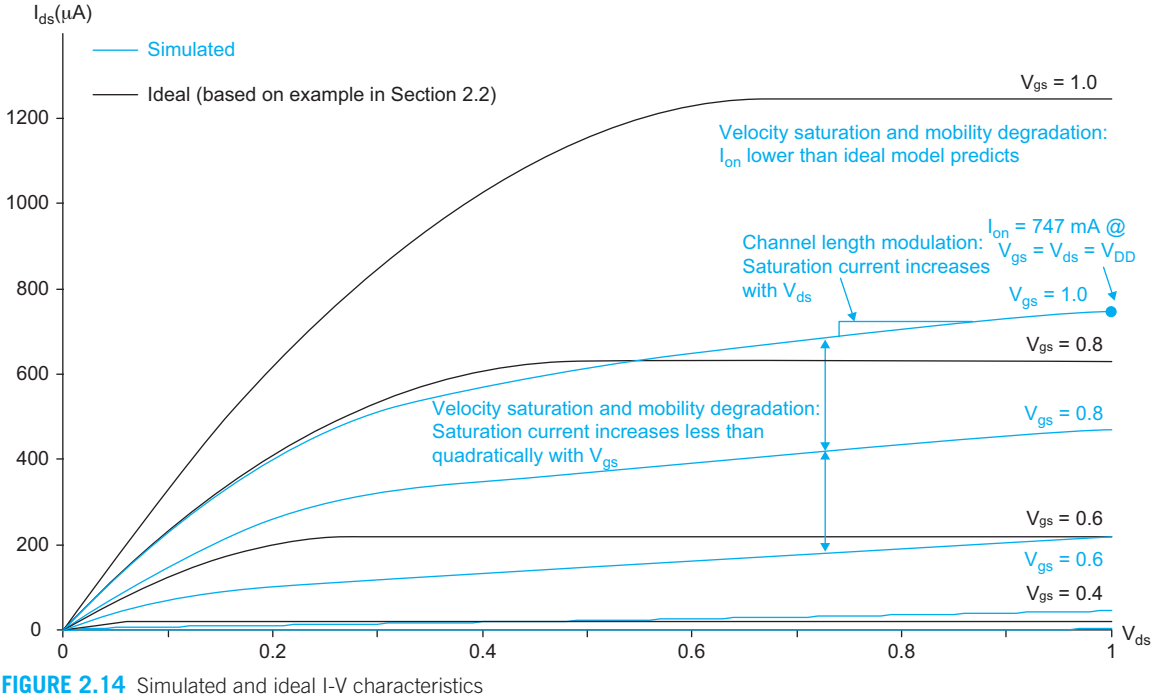


FIGURE 2.14 Simulated and ideal I-V characteristics

thickness of gate oxides reduces to only a small number of atomic layers, electrons *tunnel* through the gate, causing some *gate leakage* current. The source and drain diffusions are typically reverse-biased diodes and also experience *junction leakage* into the substrate or well.

Both mobility and threshold voltage decrease with rising temperature. The mobility effect tends to dominate for strongly ON transistors, resulting in lower I_{ds} at high temperature. The threshold effect is most important for OFF transistors, resulting in higher leakage current at high temperature. In summary, MOS characteristics degrade with temperature.

It is useful to have a qualitative understanding of nonideal effects to predict their impact on circuit behavior and to be able to anticipate how devices will change in future process generations. However, the effects lead to complicated I-V characteristics that are hard to directly apply in hand calculations. Instead, the effects are built into good transistor models and simulated with SPICE or similar software.

2.4.1 Mobility Degradation and Velocity Saturation

Recall from EQ (2.3) that carrier drift velocity, and hence current, is proportional to the lateral electric field $E_{\text{lat}} = V_{ds}/L$ between source and drain. The constant of proportionality is called the carrier mobility, μ. The long-channel model assumed that carrier mobility is independent of the applied fields. This is a good approximation for low fields, but breaks down when strong lateral or vertical fields are applied.

As an analogy, imagine that you have been working all night in the VLSI lab and decide to run down and across the courtyard to the coffee cart.[6] The number of hours you have been up is analogous to the lateral electric field. The longer you have been up, the faster you want to reach coffee: Your speed equals your fatigue times your mobility. There is a strong wind blowing in the courtyard, analogous to the vertical electric field. This wind buffets you against the wall, slowing your progress. In the same way, a high voltage at the gate of the transistor attracts the carriers to the edge of the channel, causing collisions with the oxide interface that slow the carriers. This is called mobility degradation. Moreover, freshman physics is just letting out of the lecture hall. Occasionally, you bounce off a confused freshman, fall down, and have to get up and start running again. This is analogous to carriers scattering off the silicon lattice (technically called collisions with optical phonons). The faster you try to go, the more often you collide. Beyond a certain level of fatigue, you reach a maximum average speed. In the same way, carriers approach a maximum velocity v_{sat} when high fields are applied. This phenomenon is called velocity saturation.[7]

Mobility degradation can be modeled by replacing μ with a smaller μ_{eff} that is a function of V_{gs}. A universal model [Chen96, Chen97] that matches experimental data from multiple processes reasonably well is

$$\mu_{\text{eff}-n} = \frac{540\,\frac{\text{cm}^2}{\text{V}\cdot\text{s}}}{1+\left(\frac{V_{gs}+V_t}{0.54\,\frac{\text{V}}{\text{nm}}\,t_{\text{ox}}}\right)^{1.85}} \qquad \mu_{\text{eff}-p} = \frac{185\,\frac{\text{cm}^2}{\text{V}\cdot\text{s}}}{1+\frac{\left|V_{gs}+1.5V_t\right|}{0.338\,\frac{\text{V}}{\text{nm}}\,t_{\text{ox}}}} \tag{2.23}$$

[6] This practice has been observed empirically, but is not recommended. Productivity decreases with fatigue. Beyond a certain point of exhaustion, the net work accomplished per hour becomes negative because so many mistakes are made.

[7] Do not confuse the *saturation* region of transistor operation (where $V_{ds} > V_{gs} - V_t$) with *velocity saturation* (where $E_{\text{lat}} = V_{ds}/L$ approaches E_c). In this text, the word "saturation" alone refers to the operating region while "velocity saturation" refers to the limiting of carrier velocity at high field.

Example 2.3

Compute the effective mobilities for nMOS and pMOS transistors when they are fully ON. Use the physical parameters from Example 2.1.

SOLUTION: Use $V_{gs} = 1.0$ for ON transistors, remembering that we are treating voltages as positive in a pMOS transistor. Substituting $V_t = 0.3$ V and $t_{ox} = 1.05$ nm into EQ (2.23) gives:

$$\mu_{\text{eff-n}}(V_{gs} = 1.0) = 96 \text{ cm}^2/\text{V}, \mu_{\text{eff-p}}(V_{gs} = 1.0) = 36 \text{ cm}^2/\text{V}$$

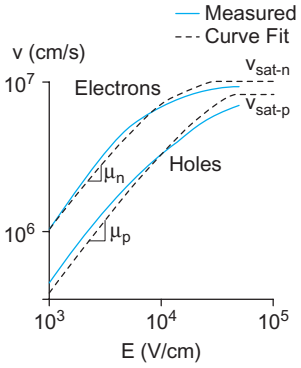


FIGURE 2.15 Carrier velocity vs. electric field at 300 K, adapted from [Jacoboni77]. Velocity saturates at high fields.

Figure 2.15 shows measured data for carrier velocity as a function of the electric field, E, between the drain and source. At low fields, the velocity increases linearly with the field. The slope is the mobility, μ_{eff}. At fields above a critical level, E_c, the velocity levels out at v_{sat}, which is approximately 10^7 cm/s for electrons and 8×10^6 cm/s for holes [Muller03]. As shown in the figure, the velocity can be approximated reasonably well with the following expression [Toh88, Takeuchi94]:

$$v = \begin{cases} \dfrac{\mu_{\text{eff}} E}{1 + \dfrac{E}{E_c}} & E < E_c \\ v_{\text{sat}} & E \geq E_c \end{cases} \tag{2.24}$$

where, by continuity, the *critical electric field* is

$$E_c = \frac{2v_{\text{sat}}}{\mu_{\text{eff}}} \tag{2.25}$$

The *critical voltage* V_c is the drain-source voltage at which the critical effective field is reached: $V_c = E_c L$.

Example 2.4

Find the critical voltage for fully ON nMOS and pMOS transistors using the effective mobilities from Example 2.3.

SOLUTION: Using EQ (2.25)

$$V_{c-n} = \frac{2\left(10^7 \frac{\text{cm}}{\text{s}}\right)}{96 \frac{\text{cm}^2}{\text{V} \cdot \text{s}}}\left(5 \times 10^{-6} \text{cm}\right) = 1.04 \text{ V}$$

$$V_{c-p} = \frac{2\left(8 \times 10^6 \frac{\text{cm}}{\text{s}}\right)}{36 \frac{\text{cm}^2}{\text{V} \cdot \text{s}}}\left(5 \times 10^{-6} \text{cm}\right) = 2.22 \text{ V}$$

The nMOS transistor is velocity saturated in normal operation because V_{c-n} is comparable to V_{DD}. The pMOS transistor has lower mobility and thus is not as badly velocity saturated.

Using a derivation similar to that of Section 2.2 with the new carrier velocity expression in EQ (2.24) gives modified equations for linear and saturation currents [Sodini84].

$$I_{ds} = \begin{cases} \dfrac{\mu_{\text{eff}}}{1+\dfrac{V_{ds}}{V_c}} C_{\text{ox}} \dfrac{W}{L}\left(V_{GT} - V_{ds}/2\right)V_{ds} & V_{ds} < V_{\text{dsat}} \quad \text{Linear} \\ C_{\text{ox}} W \left(V_{GT} - V_{\text{dsat}}\right) v_{\text{sat}} & V_{ds} > V_{\text{dsat}} \quad \text{Saturation} \end{cases} \tag{2.26}$$

Note that μ_{eff} is a decreasing function of V_{gs} because of mobility degradation. Observe that the current in the linear regime is the same as in EQ (2.5) except that the mobility term is reduced by a factor related to V_{ds}. At sufficiently high lateral fields, the current saturates at some value dependent on the maximum carrier velocity. Equating the two parts of EQ (2.26) at $V_{ds} = V_{\text{dsat}}$ lets us solve for the saturation voltage

$$V_{\text{dsat}} = \frac{V_{GT} V_c}{V_{GT} + V_c} \tag{2.27}$$

Noting that EQ (2.27) is in the same form as a parallel resistor equation, we see that V_{dsat} is less than the smaller of V_{GT} and V_c. Finally, substituting EQ (2.27) into EQ (2.26) gives a simplified expression for saturation current accounting for velocity saturations:

$$I_{\text{dsat}} = W C_{\text{ox}} v_{\text{sat}} \frac{V_{GT}^2}{V_{GT} + V_c} \qquad V_{ds} > V_{\text{dsat}} \tag{2.28}$$

If $V_{GT} << V_c$, velocity saturation effects are negligible and EQ (2.28) reduces to the square-law model. This is also called the *long-channel regime*. But if $V_{GT} >> V_c$, EQ (2.28) approaches the velocity-saturated limit

$$I_{\text{dsat}} \approx W C_{\text{ox}} v_{\text{sat}} V_{GT} \qquad V_{ds} > V_c \tag{2.29}$$

Observe that the drain current is quadratically dependent on voltage in the long-channel regime and linearly dependent when fully velocity saturated. For moderate supply voltages, transistors operate in a region where the velocity neither increases linearly with field, nor is completely saturated. The *α-power law model* given in EQ (2.30) provides a simple approximation to capture this behavior [Sakurai90]. α is called the *velocity saturation index* and is determined by curve fitting measured I-V data. Transistors with long channels or low V_{DD} display quadratic I-V characteristics in saturation and are modeled with $\alpha = 2$. As transistors become more velocity saturated, increasing V_{gs} has less effect on current and α decreases, reaching 1 for transistors that are completely velocity saturated. For simplicity, the model uses a straight line in the linear region. Overall, the model is based on three parameters that can be determined empirically from a curve fit of I-V characteristics: α, βP_c, and P_v.

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t \quad \text{Cutoff} \\ I_{\text{dsat}} \dfrac{V_{ds}}{V_{\text{dsat}}} & V_{ds} < V_{\text{dsat}} \quad \text{Linear} \\ I_{\text{dsat}} & V_{ds} > V_{\text{dsat}} \quad \text{Saturation} \end{cases} \tag{2.30}$$

where

$$\begin{aligned} I_{\text{dsat}} &= P_c \frac{\beta}{2} V_{GT}^{\alpha} \\ V_{\text{dsat}} &= P_v V_{GT}^{\alpha/2} \end{aligned} \tag{2.31}$$

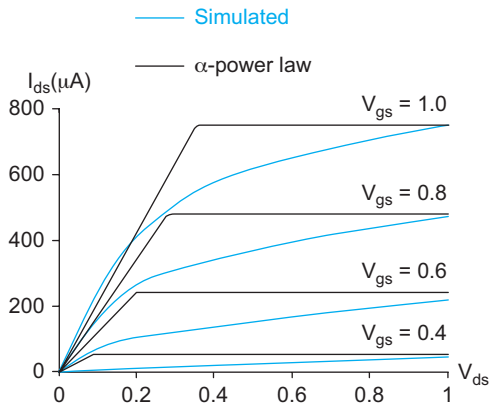


FIGURE 2.16 Comparison of α-power law model with simulated transistor behavior

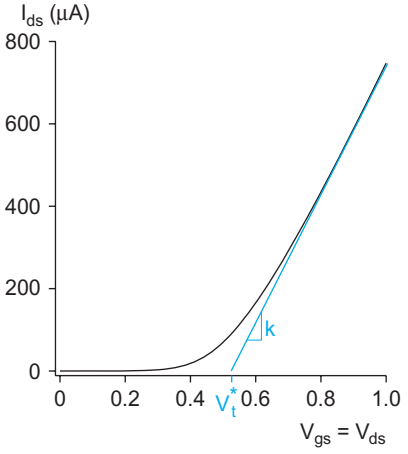


FIGURE 2.17 I_{ds} vs. V_{gs} in saturation, showing good linear fit at high V_{gs}

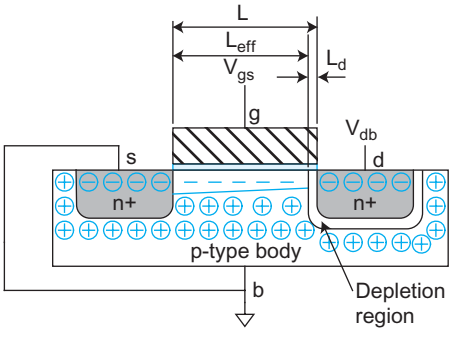


FIGURE 2.18 Depletion region shortens effective channel length

Figure 2.16 compares the α-power law model against simulated results, using $\alpha = 1.3$. The fit is poor at low V_{ds}, but the current at $V_{ds} = V_{DD}$ matches simulation fairly well across the full range of V_{gs}.

The low-field mobility of holes is much lower than that of electrons, so pMOS transistors experience less velocity saturation than nMOS for a given V_{DD}. This shows up as a larger value of α for pMOS than for nMOS transistors.

These models become too complicated to give much insight for hand calculations. A simpler approach is to observe, in velocity-saturated transistors, I_{ds} grows linearly rather than quadratically with V_{gs} when the transistor is strongly ON. Figure 2.17 plots I_{ds} vs. V_{gs} (holding $V_{ds} = V_{gs}$). This is equivalent to plotting I_{on} vs. V_{DD}. For V_{gs} significantly above V_t, I_{ds} fits a straight line quite well. Thus, we can approximate the ON current as

$$I_{ds} = k\left(V_{gs} - V_t^*\right) \tag{2.32}$$

where V_t^* is the x-intercept.

2.4.2 Channel Length Modulation

Ideally, I_{ds} is independent of V_{ds} for a transistor in saturation, making the transistor a perfect current source. As discussed in Section 2.3.3, the p–n junction between the drain and body forms a depletion region with a width L_d that increases with V_{db}, as shown in Figure 2.18. The depletion region effectively shortens the channel length to

$$L_{\text{eff}} = L - L_d \tag{2.33}$$

To avoid introducing the body voltage into our calculations, assume the source voltage is close to the body voltage so $V_{db} \approx V_{ds}$. Hence, increasing V_{ds} decreases the effective channel length. Shorter channel length results in higher current; thus, I_{ds} increases with V_{ds} in saturation, as shown in Figure 2.18. This can be crudely modeled by multiplying EQ (2.10) by a factor of $(1 + V_{ds} / V_A)$, where V_A is called the *Early voltage* [Gray01]. In the saturation region, we find

$$I_{ds} = \frac{\beta}{2} V_{\text{GT}}^2 \left(1 + \frac{V_{ds}}{V_A}\right) \tag{2.34}$$

As channel length gets shorter, the effect of the channel length modulation becomes relatively more important. Hence, V_A is proportional to channel length. This channel length modulation model is a gross oversimplification of nonlinear behavior and is more useful for conceptual understanding than for accurate device modeling.

Channel length modulation is very important to analog designers because it reduces the gain of amplifiers. It is generally unimportant for qualitatively understanding the behavior of digital circuits.

2.4.3 Threshold Voltage Effects

So far, we have treated the threshold voltage as a constant. However, V_t increases with the source voltage, decreases with the body voltage, decreases with the drain voltage, and increases with channel length [Roy03]. This section models each of these effects.

2.4.3.1 Body Effect Until now, we have considered a transistor to be a three-terminal device with gate, source, and drain. However, the body is an implicit fourth terminal. When a voltage V_{sb} is applied between the source and body, it increases the amount of charge required to invert the channel, hence, it increases the threshold voltage. The threshold voltage can be modeled as

$$V_t = V_{t0} + \gamma\left(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s}\right) \tag{2.35}$$

where V_{t0} is the threshold voltage when the source is at the body potential, ϕ_s is the *surface potential* at threshold (see a device physics text such as [Tsividis99] for further discussion of surface potential), and γ is the *body effect coefficient*, typically in the range 0.4 to 1 $V^{1/2}$. In turn, these depend on the doping level in the channel, N_A. The body effect further degrades the performance of pass transistors trying to pass the weak value (e.g., nMOS transistors passing a '1'), as we will examine in Section 2.5.4. Section 5.3.4 will describe how a body bias can intentionally be applied to alter the threshold voltage, permitting trade-offs between performance and subthreshold leakage current.

$$\phi_s = 2v_T \ln\frac{N_A}{n_i} \tag{2.36}$$

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A} = \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}} \tag{2.37}$$

For small voltages applied to the source or body, EQ (2.35) can be linearized to

$$V_t = V_{t0} + k_\gamma V_{sb} \tag{2.38}$$

where

$$k_\gamma = \frac{\gamma}{2\sqrt{\phi_s}} = \frac{\sqrt{\dfrac{q\varepsilon_{si}N_A}{v_T \ln\frac{N_A}{n_i}}}}{2C_{ox}} \tag{2.39}$$

Example 2.5

Consider the nMOS transistor in a 65 nm process with a nominal threshold voltage of 0.3 V and a doping level of 8×10^{17} cm^{-3}. The body is tied to ground with a substrate contact. How much does the threshold change at room temperature if the source is at 0.6 V instead of 0?

SOLUTION: At room temperature, the thermal voltage $v_T = kT/q = 26$ mV and $n_i = 1.45 \times 10^{10}$ cm^{-3}. The threshold increases by 0.04 V.

$$\phi_s = 2(0.026\text{ V})\ln\frac{8\times10^{17}\text{ cm}^{-3}}{1.45\times10^{10}\text{ cm}^{-3}} = 0.93\text{ V}$$

$$\gamma = \frac{10.5\times10^{-8}\text{ cm}}{3.9\times8.85\times10^{-14}\ \frac{\text{F}}{\text{cm}}}\sqrt{2\left(1.6\times10^{-19}\text{C}\right)\left(11.7\times8.85\times10^{-14}\ \frac{\text{F}}{\text{cm}}\right)\left(8\times10^{17}\text{ cm}^{-3}\right)} = 0.16 \quad \textbf{(2.40)}$$

$$V_t = 0.3 + \gamma\left(\sqrt{\phi_s + 0.6\text{ V}} - \sqrt{\phi_s}\right) = 0.34\text{ V}$$

2.4.3.2 Drain-Induced Barrier Lowering The drain voltage V_{ds} creates an electric field that affects the threshold voltage. This *drain-induced barrier lowering* (DIBL) effect is especially pronounced in short-channel transistors. It can be modeled as

$$V_t = V_{t0} - \eta V_{ds} \quad \textbf{(2.41)}$$

where η is the DIBL coefficient, typically on the order of 0.1 (often expressed as 100 mV/V).

Drain-induced barrier lowering causes I_{ds} to increase with V_{ds} in saturation, in much the same way as channel length modulation does. This effect can be lumped into a smaller Early voltage V_A used in EQ (2.34). Again, this is a bane for analog design but insignificant for most digital circuits. More significantly, DIBL increases subthreshold leakage at high V_{ds}, as we will discuss in Section 2.4.4.

2.4.3.3 Short Channel Effect The threshold voltage typically increases with channel length. This phenomenon is especially pronounced for small L where the source and drain depletion regions extend into a significant portion of the channel, and hence is called the *short channel effect*[8] or V_t *rolloff* [Tsividis99, Cheng99]. In some processes, a *reverse short channel effect* causes V_t to decrease with length.

There is also a *narrow channel effect* in which V_t varies with channel width; this effect tends to be less significant because the minimum width is greater than the minimum length.

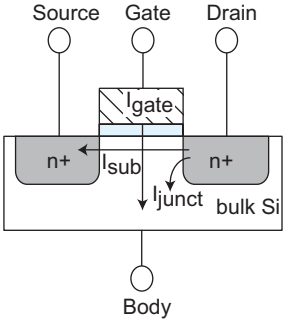


FIGURE 2.19
Leakage current paths

2.4.4 Leakage

Even when transistors are nominally OFF, they leak small amounts of current. Leakage mechanisms include subthreshold conduction between source and drain, gate leakage from the gate to body, and junction leakage from source to body and drain to body, as illustrated in Figure 2.19 [Roy03, Narendra06]. Subthreshold conduction is caused by thermal emission of carriers over the potential barrier set by the threshold. Gate leakage is a quantum-mechanical effect caused by tunneling through the extremely thin gate dielectric. Junction leakage is caused by current through the p-n junction between the source/drain diffusions and the body.

[8]The term *short-channel effect* is overused in the CMOS literature. Sometimes, it refers to any behavior outside the long-channel models. Other times, it refers to a range of behaviors including DIBL that are most significant for very short channel lengths [Muller03]. In this text, we restrict the term to describe the sensitivity of threshold voltage to channel length.

In processes with feature sizes above 180 nm, leakage was typically insignificant except in very low power applications. In 90 and 65 nm processes, threshold voltage has reduced to the point that subthreshold leakage reaches levels of 1s to 10s of nA per transistor, which is significant when multiplied by millions or billions of transistors on a chip. In 45 nm processes, oxide thickness reduces to the point that gate leakage becomes comparable to subthreshold leakage unless high-k gate dielectrics are employed. Overall, leakage has become an important design consideration in nanometer processes.

2.4.4.1 Subthreshold Leakage The long-channel transistor I-V model assumes current only flows from source to drain when $V_{gs} > V_t$. In real transistors, current does not abruptly cut off below threshold, but rather drops off exponentially, as seen in Figure 2.20. When the gate voltage is high, the transistor is strongly ON. When the gate falls below V_t, the exponential decline in current appears as a straight line on the logarithmic scale. This regime of $V_{gs} < V_t$ is called *weak inversion*. The *subthreshold leakage current* increases significantly with V_{ds} because of drain-induced barrier lowering (see Section 2.4.3.2). There is a lower limit on I_{ds} set by drain junction leakage that is exacerbated by the negative gate voltage (see Section 2.4.4.3).

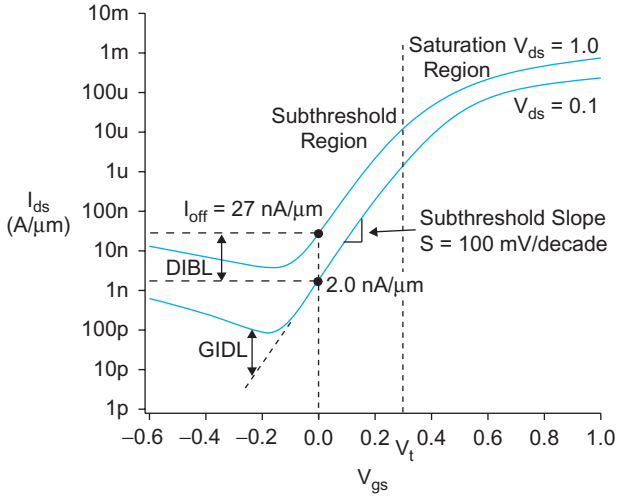


FIGURE 2.20 I-V characteristics of a 65 nm nMOS transistor at 70 °C on a log scale

Subthreshold leakage current is described by EQ (2.42). I_{ds0} is the current at threshold and is dependent on process and device geometry. It is typically extracted from simulation but can also be calculated from EQ (2.43); the $e^{1.8}$ term was found empirically [Sheu87]. n is a process-dependent term affected by the depletion region characteristics and is typically in the range of 1.3–1.7 for CMOS processes. The final term indicates that leakage is 0 if $V_{ds} = 0$, but increases to its full value when V_{ds} is a few multiples of the thermal voltage v_T (e.g., when $V_{ds} > 50$ mV). More significantly, drain-induced barrier lowering effectively reduces the threshold voltage, as indicated by the ηV_{ds} term. This can increase leakage by an order of magnitude for $V_{ds} = V_{DD}$ as compared to small V_{ds}. The body effect also modulates V_t when $V_{sb} \approx 0$.

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_{t0} + \eta V_{ds} - k\gamma V_{sb}}{n v_T}} \left(1 - e^{\frac{-V_{ds}}{v_T}}\right) \quad \textbf{(2.42)}$$

$$I_{ds0} = \beta v_T^2 e^{1.8} \tag{2.43}$$

Subthreshold conduction is used to advantage in very low-power circuits, as will be explored in Section 9.6. It afflicts dynamic circuits and DRAMs, which depend on the storage of charge on a capacitor. Conduction through an OFF transistor discharges the capacitor unless it is periodically refreshed or a trickle of current is available to counter the leakage. Leakage also contributes to power dissipation in idle circuits. Subthreshold leakage increases exponentially as V_t decreases or as temperature rises, so it is a major problem for chips using low supply and threshold voltages and for chips operating at high temperature.

As shown in Figure 2.20, subthreshold current fits a straight line on a semilog plot. The inverse of the slope of this line is called the *subthreshold slope*, S

$$S = \left[\frac{d\left(\log_{10} I_{ds}\right)}{dV_{gs}}\right]^{-1} = n v_T \ln 10 \tag{2.44}$$

The subthreshold slope indicates how much the gate voltage must drop to decrease the leakage current by an order of magnitude. A typical value is 100 mV/decade at room temperature. EQ (2.42) can be rewritten using the subthreshold slope as

$$I_{ds} = I_{\text{off}} 10^{\frac{V_{gs} + \eta\left(V_{ds} - V_{dd}\right) - k_\gamma V_{sb}}{S}} \left(1 - e^{\frac{-V_{ds}}{v_T}}\right) \tag{2.45}$$

where I_{off} is the subthreshold current at $V_{gs} = 0$ and $V_{ds} = V_{DD}$.

Example 2.6

What is the minimum threshold voltage for which the leakage current through an OFF transistor ($V_{gs} = 0$) is 10^3 times less than that of a transistor that is barely ON ($V_{gs} = V_t$) at room temperature if $n = 1.5$? One of the advantages of silicon-on-insulator (SOI) processes is that they have smaller n (see Section 9.5). What threshold is required for SOI if $n = 1.3$?

SOLUTION: $v_T = 26$ mV at room temperature. Assume $V_{ds} >> v_T$ so leakage is significant. We solve

$$\begin{aligned} I_{ds}\left(V_{gs} = 0\right) &= 10^{-3} I_{ds0} = I_{ds0} e^{\frac{-V_t}{n v_T}} \\ V_t &= -n v_T \ln 10^{-3} = 270\text{mV} \end{aligned} \tag{2.46}$$

In the CMOS process, leakage rolls off by a factor of 10 for every 90 mV V_{gs} falls below threshold. This is often quoted as a subthreshold slope of $S = 90$ mV/decade. In the SOI process, the subthreshold slope S is 78 mV/decade, so a threshold of only 234 mV is required.

2.4.4.2 Gate Leakage According to quantum mechanics, the electron cloud surrounding an atom has a probabilistic spatial distribution. For gate oxides thinner than 15–20 Å,

there is a nonzero probability that an electron in the gate will find itself on the wrong side of the oxide, where it will get whisked away through the channel. This effect of carriers crossing a thin barrier is called tunneling, and results in leakage current through the gate.

Two physical mechanisms for gate tunneling are called *Fowler–Nordheim* (FN) *tunneling* and *direct tunneling*. FN tunneling is most important at high voltage and moderate oxide thickness and is used to program EEPROM memories (see Section 12.4). Direct tunneling is most important at lower voltage with thin oxides and is the dominant leakage component.

The direct gate tunneling current can be estimated as [Chandrakasan01]

$$I_{\text{gate}} = WA\left(\frac{V_{DD}}{t_{\text{ox}}}\right)^2 e^{-B\frac{t_{\text{ox}}}{V_{DD}}} \tag{2.47}$$

where A and B are technology constants.

Transistors need high C_{ox} to deliver good ON current, driving the decrease in oxide thickness. Tunneling current drops exponentially with the oxide thickness and has only recently become significant. Figure 2.21 plots gate leakage current density (current/area) J_G against voltage for various oxide thicknesses. Gate leakage increases by a factor of 2.7 or more per angstrom reduction in thickness [Rohrer05]. Large tunneling currents impact not only dynamic nodes but also quiescent power consumption and thus limits equivalent oxide thicknesses t_{ox} to at least 10.5 Å to keep gate leakage below 100 A/cm^2. To keep these dimensions in perspective, recall that each atomic layer of SiO_2 is about 3 Å, so such gate oxides are a handful of atomic layers thick. Section 3.4.1.3 describes innovations in gate insulators with higher dielectric constants that offer good C_{ox} while reducing tunneling.

Tunneling current can be an order of magnitude higher for nMOS than pMOS transistors with SiO_2 gate dielectrics because the electrons tunnel from the conduction band while the holes tunnel from the valence band and see a higher barrier [Hamzaoglu02]. Different dielectrics may have different tunneling properties.

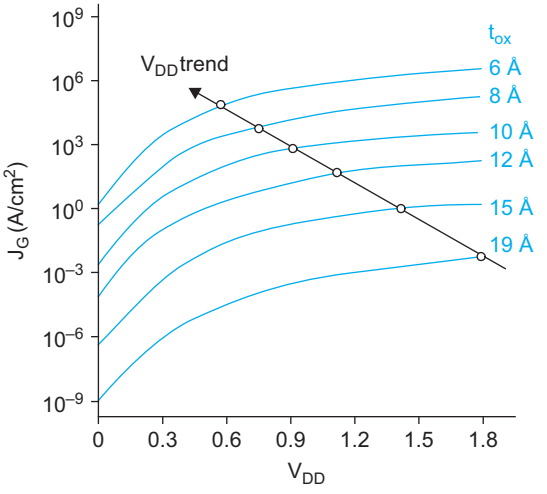


FIGURE 2.21 Gate leakage current from [Song01]

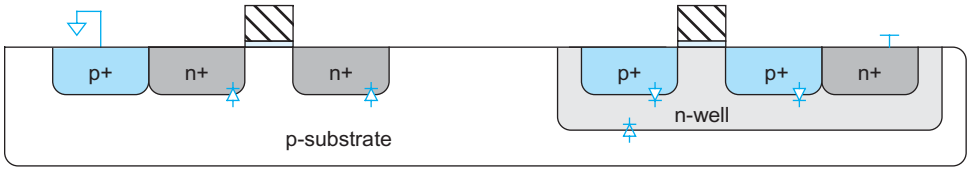


FIGURE 2.22 Substrate to diffusion diodes in CMOS circuits

2.4.4.3 Junction Leakage The p–n junctions between diffusion and the substrate or well form diodes, as shown in Figure 2.22. The well-to-substrate junction is another diode. The substrate and well are tied to GND or V_{DD} to ensure these diodes do not become forward biased in normal operation. However, reverse-biased diodes still conduct a small amount of current I_D.[9]

$$I_D = I_S\left(e^{\frac{V_D}{v_T}} - 1\right) \tag{2.48}$$

where I_S depends on doping levels and on the area and perimeter of the diffusion region and V_D is the diode voltage (e.g., $-V_{sb}$ or $-V_{db}$). When a junction is reverse biased by significantly more than the thermal voltage, the leakage is just $-I_S$, generally in the 0.1–0.01 fA/μm^2 range, which is negligible compared to other leakage mechanisms.

More significantly, heavily doped drains are subject to *band-to-band tunneling* (BTBT) and *gate-induced drain leakage* (GIDL).

BTBT occurs across the junction between the source or drain and the body when the junction is reverse-biased. It is a function of the reverse bias and the doping levels. High halo doping used to increase V_t to alleviate subthreshold leakage instead causes BTBT to grow. The leakage is exacerbated by *trap-assisted tunneling* (TAT) when defects in the silicon lattice called traps reduce the distance that a carrier must tunnel. Most of the leakage occurs along the sidewall closest to the channel where the doping is highest. It can be modeled as

$$I_{BTBT} = WX_j A\frac{E_j}{E_g^{0.5}}V_{dd}\, e^{-B\frac{E_g^{1.5}}{E_j}} \tag{2.49}$$

where X_j is the junction depth of the diffusion, E_g is the bandgap voltage, and A and B are technology constants [Mukhopadhyay05]. The electric field along the junction at a reverse bias of V_{DD} is

$$E_j = \sqrt{\frac{2qN_{halo}N_{sd}}{\varepsilon\left(N_{halo}+N_{sd}\right)}}\left(V_{DD} + v_T \ln\frac{N_{halo}N_{sd}}{n_i^2}\right) \tag{2.50}$$

GIDL occurs where the gate partially overlaps the drain. This effect is most pronounced when the drain is at a high voltage and the gate is at a low voltage. GIDL current is proportional to gate-drain overlap area and hence to transistor width. It is a strong function of the electric field and hence increases rapidly with the drain-to-gate voltage. How-

[9]Beware that I_D and I_S stand for the diode current and diode reverse-biased saturation currents, respectively. The D and S are not related to drain or source.

ever, it is normally insignificant at $|V_{gd}| \le V_{DD}$ [Mukhopadhyay05], only coming into play when the gate is driven outside the rails in an attempt to cut off subthreshold leakage.

2.4.5 Temperature Dependence

Transistor characteristics are influenced by temperature [Cobbold66, Vadasz66, Tsividis99, Gutierrez01]. Carrier mobility decreases with temperature. An approximate relation is

$$\mu(T) = \mu(T_r)\left(\frac{T}{T_r}\right)^{-k_\mu} \tag{2.51}$$

where T is the absolute temperature, T_r is room temperature, and k_μ is a fitting parameter with a typical value of about 1.5. v_{sat} also decreases with temperature, dropping by about 20% from 300 to 400 K.

The magnitude of the threshold voltage decreases nearly linearly with temperature and may be approximated by

$$V_t(T) = V_t(T_r) - k_{vt}(T - T_r) \tag{2.52}$$

where k_{vt} is typically about 1–2 mV/K.

I_{on} at high V_{DD} decreases with temperature. Subthreshold leakage increases exponentially with temperature. BTBT increases slowly with temperature, and gate leakage is almost independent of temperature.

The combined temperature effects are shown in Figure 2.23. At high V_{gs}, the current has a *negative temperature coefficient*; i.e., it decreases with temperature. At low V_{gs}, the current has a positive temperature coefficient. Thus, OFF current increases with temperature. ON current I_{dsat} normally decreases with temperature, as shown in Figure 2.24, so circuit performance is worst at high temperature. However, for systems operating at low V_{DD} (typically < 0.7 – 1.1 V), I_{dsat} increases with temperature [Kumar06].

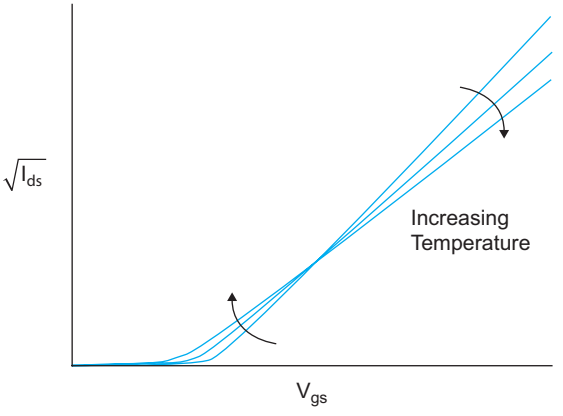


FIGURE 2.23 I–V characteristics of nMOS transistor in saturation at various temperatures

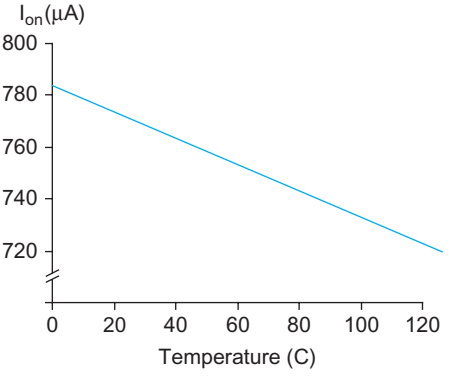


FIGURE 2.24 I_{dsat} vs. temperature

Conversely, circuit performance can be improved by cooling. Most systems use natural convection or fans in conjunction with heat sinks, but water cooling, thin-film refrigerators, or even liquid nitrogen can increase performance if the expense is justified. There are many advantages of operating at low temperature [Keyes70, Sun87]. Subthreshold leakage is exponentially dependent on temperature, so lower threshold voltages can be used. Velocity saturation occurs at higher fields, providing more current. As mobility is also higher, these fields are reached at a lower power supply, saving power. Depletion regions become wider, resulting in less junction capacitance.

Two popular lab tools for determining temperature dependence in circuits are a can of freeze spray and a heat gun. The former can be used to momentarily "freeze" a chip to see whether performance alters and the other, of course, can be used to heat up a chip. Often, these tests are done to quickly determine whether a chip is prone to temperature effects. Be careful—sometimes the sudden temperature change can fracture chips or their packages.

2.4.6 Geometry Dependence

The layout designer draws transistors with width and length W_{drawn} and L_{drawn}. The actual gate dimensions may differ by some factors X_W and X_L. For example, the manufacturer may create masks with narrower polysilicon or may overetch the polysilicon to provide shorter channels (negative X_L) without changing the overall design rules or metal pitch. Moreover, the source and drain tend to diffuse laterally under the gate by L_D, producing a shorter effective channel length that the carriers must traverse between source and drain. Similarly, W_D accounts for other effects that shrink the transistor width. Putting these factors together, we can compute effective transistor lengths and widths that should be used in place of L and W in the current and capacitance equations given elsewhere in the book. The factors of two come from lateral diffusion on both sides of the channel.

$$\begin{aligned} L_{\text{eff}} &= L_{\text{drawn}} + X_L - 2L_D \\ W_{\text{eff}} &= W_{\text{drawn}} + X_W - 2W_D \end{aligned} \tag{2.53}$$

Therefore, a transistor drawn twice as long may have an effective length that is more than twice as great. Similarly, two transistors differing in drawn widths by a factor of two may differ in saturation current by more than a factor of two. Threshold voltages also vary with transistor dimensions because of the short and narrow channel effects.

Combining threshold changes, effective channel lengths, channel length modulation, and velocity saturation effects, I_{dsat} does not scale exactly as $1/L$. In general, when currents must be precisely matched (e.g., in sense amplifiers or A/D converters), it is best to use the same width and length for each device. Current ratios can be produced by tying several identical transistors in parallel.

In processes below 0.25 μm, the effective length of the transistor also depends significantly on the orientation of the transistor. Moreover, the amount of nearby polysilicon also affects etch rates during manufacturing and thus channel length. Transistors that must match well should have the same orientation. Dummy polysilicon wires can be placed nearby to improve etch uniformity.

2.4.7 Summary

Although the physics of nanometer-scale devices is complicated, the impact of nonideal I-V behavior is fairly easy to understand from the designer's viewpoint.

Threshold drops Pass transistors suffer a threshold drop when passing the wrong value: nMOS transistors only pull up to $V_{DD} - V_{tn}$, while pMOS transistors only pull down to $|V_{tp}|$. The magnitude of the threshold drop is increased by the body effect. Therefore, pass transistors do not operate very well in nanometer processes where the threshold voltage is a significant fraction of the supply voltage. Fully complementary transmission gates should be used where both 0s and 1s must be passed well.

Leakage current Ideally, static CMOS gates draw zero current and dissipate zero power when idle. Real gates draw some leakage current. The most important source at this time is subthreshold leakage between source and drain of a transistor that should be cut off. The subthreshold current of an OFF transistor decreases by an order of magnitude for every 60–100 mV that V_{gs} is below V_t. Threshold voltages have been decreasing, so subthreshold leakage has been increasing dramatically. Some processes offer multiple choices of V_t: low-V_t devices are used for high performance in critical circuits, while high-V_t devices are used for low leakage elsewhere.

The transistor gate is a good insulator. However, significant tunneling current flows through very thin gates. This has limited the scaling of gate oxide and led to new high-k gate dielectrics.

Leakage current causes CMOS gates to consume power when idle. It also limits the amount of time that data is retained in dynamic logic, latches, and memory cells. In nanometer processes, dynamic logic and latches require some sort of feedback to prevent data loss from leakage. Leakage increases at high temperature.

V_{DD} Velocity saturation and mobility degradation result in less current than expected at high voltage. This means that there is no point in trying to use a high V_{DD} to achieve fast transistors, so V_{DD} has been decreasing with process generation to reduce power consumption. Moreover, the very short channels and thin gate oxides would be damaged by high V_{DD}.

Delay Transistors in series drop part of the voltage across each transistor and thus experience smaller fields and less velocity saturation than single transistors. Therefore, series transistors tend to be a bit faster than a simple model would predict. For example, two nMOS transistors in series deliver more than half the current of a single nMOS transistor of the same width. This effect is more pronounced for nMOS transistors than pMOS transistors because nMOS transistors have higher mobility to begin with and thus are more velocity saturated.

Matching If two transistors should behave identically, both should have the same dimensions and orientation and be interdigitated if possible.

2.5 DC Transfer Characteristics

Digital circuits are merely analog circuits used over a special portion of their range. The DC transfer characteristics of a circuit relate the output voltage to the input voltage, assuming the input changes slowly enough that capacitances have plenty of time to charge or discharge. Specific ranges of input and output voltages are defined as valid 0 and 1 logic levels. This section explores the DC transfer characteristics of CMOS gates and pass transistors.

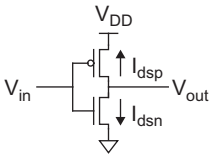


FIGURE 2.25
A CMOS inverter

2.5.1 Static CMOS Inverter DC Characteristics

Let us derive the DC transfer function (V_{out} vs. V_{in}) for the static CMOS inverter shown in Figure 2.25. We begin with Table 2.2, which outlines various regions of operation for the n- and p-transistors. In this table, V_{tn} is the threshold voltage of the n-channel device, and V_{tp} is the threshold voltage of the p-channel device. Note that V_{tp} is negative. The equations are given both in terms of V_{gs}/V_{ds} and V_{in}/V_{out}. As the source of the nMOS transistor is grounded, $V_{gsn} = V_{in}$ and $V_{dsn} = V_{out}$. As the source of the pMOS transistor is tied to V_{DD}, $V_{gsp} = V_{in} - V_{DD}$ and $V_{dsp} = V_{out} - V_{DD}$.

TABLE 2.2 Relationships between voltages for the three regions of operation of a CMOS inverter

	Cutoff	Linear	Saturated
nMOS	$V_{gsn} < V_{tn}$	$V_{gsn} > V_{tn}$	$V_{gsn} > V_{tn}$
	$V_{in} < V_{tn}$	$V_{in} > V_{tn}$	$V_{in} > V_{tn}$
		$V_{dsn} < V_{gsn} - V_{tn}$	$V_{dsn} > V_{gsn} - V_{tn}$
		$V_{out} < V_{in} - V_{tn}$	$V_{out} > V_{in} - V_{tn}$
pMOS	$V_{gsp} > V_{tp}$	$V_{gsp} < V_{tp}$	$V_{gsp} < V_{tp}$
	$V_{in} > V_{tp} + V_{DD}$	$V_{in} < V_{tp} + V_{DD}$	$V_{in} < V_{tp} + V_{DD}$
		$V_{dsp} > V_{gsp} - V_{tp}$	$V_{dsp} < V_{gsp} - V_{tp}$
		$V_{out} > V_{in} - V_{tp}$	$V_{out} < V_{in} - V_{tp}$

The objective is to find the variation in output voltage (V_{out}) as a function of the input voltage (V_{in}). This may be done graphically, analytically (see Exercise 2.16), or through simulation [Carr72]. Given V_{in}, we must find V_{out} subject to the constraint that $I_{dsn} = |I_{dsp}|$. For simplicity, we assume $V_{tp} = -V_{tn}$ and that the pMOS transistor is 2–3 times as wide as the nMOS transistor so $\beta_n = \beta_p$. We relax this assumption in Section 2.5.2.

We commence with the graphical representation of the simple algebraic equations described by EQ (2.10) for the two transistors shown in Figure 2.26(a). The plot shows I_{dsn} and I_{dsp} in terms of V_{dsn} and V_{dsp} for various values of V_{gsn} and V_{gsp}. Figure 2.26(b) shows the same plot of I_{dsn} and $|I_{dsp}|$ now in terms of V_{out} for various values of V_{in}. The possible operating points of the inverter, marked with dots, are the values of V_{out} where $I_{dsn} = |I_{dsp}|$ for a given value of V_{in}. These operating points are plotted on V_{out} vs. V_{in} axes in Figure 2.26(c) to show the inverter DC transfer characteristics. The supply current $I_{DD} = I_{dsn} = |I_{dsp}|$ is also plotted against V_{in} in Figure 2.26(d) showing that both transistors are momentarily ON as V_{in} passes through voltages between GND and V_{DD}, resulting in a pulse of current drawn from the power supply.

The operation of the CMOS inverter can be divided into five regions indicated on Figure 2.26(c). The state of each transistor in each region is shown in Table 2.3. In region *A*, the nMOS transistor is OFF so the pMOS transistor pulls the output to V_{DD}. In region *B*, the nMOS transistor starts to turn ON, pulling the output down. In region *C*, both transistors are in saturation. Notice that ideal transistors are only in region *C* for $V_{in} = V_{DD}/2$ and that the slope of the transfer curve in this example is $-\infty$ in this region, corresponding to infinite gain. Real transistors have finite output resistances on account of channel length modulation, described in Section 2.4.2, and thus have finite slopes over a broader region *C*. In region *D*, the pMOS transistor is partially ON and in region *E*, it is completely

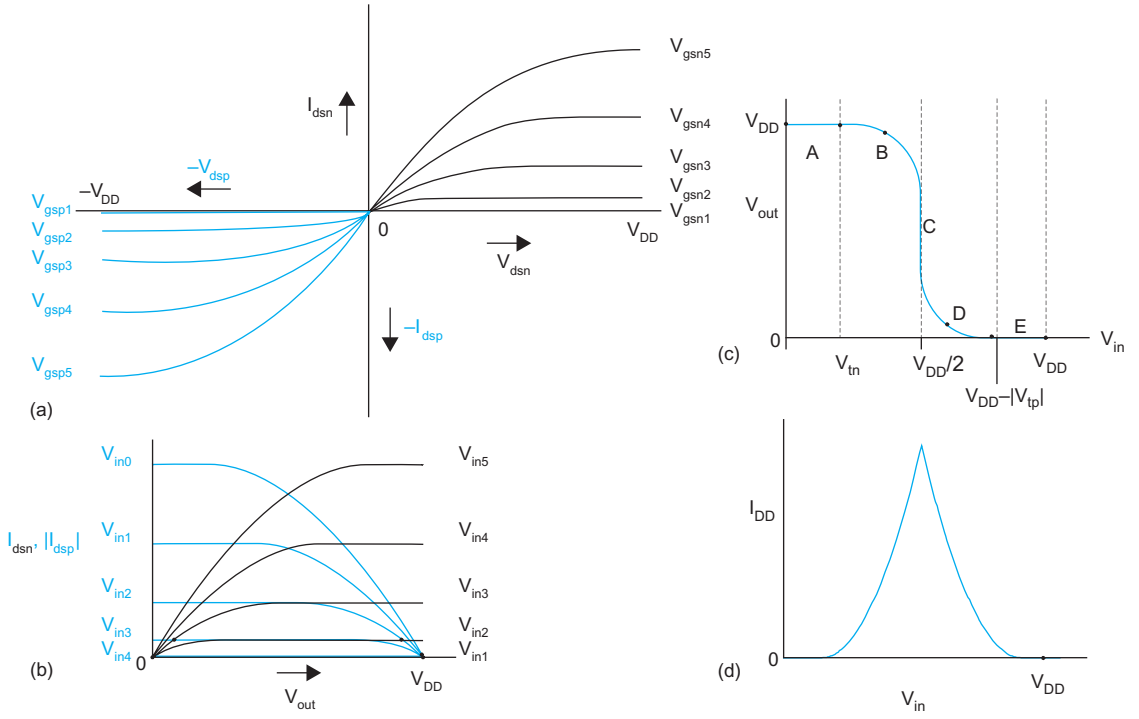


FIGURE 2.26 Graphical derivation of CMOS inverter DC characteristic

OFF, leaving the nMOS transistor to pull the output down to GND. Also notice that the inverter's current consumption is ideally zero, neglecting leakage, when the input is within a threshold voltage of the V_{DD} or GND rails. This feature is important for low-power operation.

TABLE 2.3 Summary of CMOS inverter operation

Region	Condition	p-device	n-device	Output
A	$0 \le V_{in} < V_{tn}$	linear	cutoff	$V_{out} = V_{DD}$
B	$V_{tn} \le V_{in} < V_{DD}/2$	linear	saturated	$V_{out} > V_{DD}/2$
C	$V_{in} = V_{DD}/2$	saturated	saturated	V_{out} drops sharply
D	$V_{DD}/2 < V_{in} \le V_{DD} - \|V_{tp}\|$	saturated	linear	$V_{out} < V_{DD}/2$
E	$V_{in} > V_{DD} - \|V_{tp}\|$	cutoff	linear	$V_{out} = 0$

Figure 2.27 shows simulation results of an inverter from a 65 nm process. The pMOS transistor is twice as wide as the nMOS transistor to achieve approximately equal betas. Simulation matches the simple models reasonably well, although the transition is not quite as steep because transistors are not ideal current sources in saturation.

The crossover point where $V_{inv} = V_{in} = V_{out}$ is called the *input threshold*. Because both mobility and the magnitude of the threshold voltage decrease with temperature for nMOS and pMOS transistors, the input threshold of the gate is only weakly sensitive to temperature.

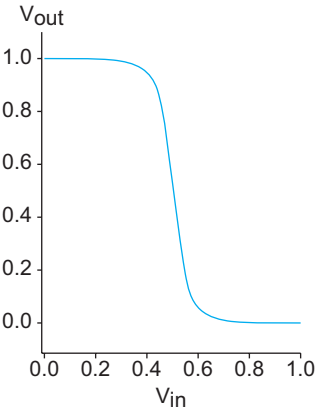


FIGURE 2.27 Simulated CMOS inverter DC characteristic

2.5.2 Beta Ratio Effects

We have seen that for $\beta_p = \beta_n$, the inverter threshold voltage V_{inv} is $V_{DD}/2$. This may be desirable because it maximizes noise margins (see Section 2.5.3) and allows a capacitive load to charge and discharge in equal times by providing equal current source and sink capabilities (see Section 4.2). Inverters with different beta ratios $r = \beta_p/\beta_n$ are called *skewed* inverters [Sutherland99]. If $r > 1$, the inverter is *HI-skewed*. If $r < 1$, the inverter is *LO-skewed*. If $r = 1$, the inverter has normal skew or is *unskewed*.

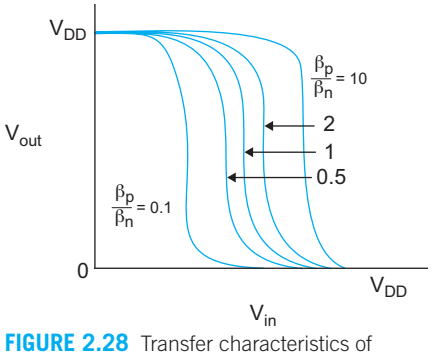


FIGURE 2.28 Transfer characteristics of skewed inverters

A HI-skew inverter has a stronger pMOS transistor. Therefore, if the input is $V_{DD}/2$, we would expect the output will be greater than $V_{DD}/2$. In other words, the input threshold must be higher than for an unskewed inverter. Similarly, a LO-skew inverter has a weaker pMOS transistor and thus a lower switching threshold.

Figure 2.28 explores the impact of skewing the beta ratio on the DC transfer characteristics. As the beta ratio is changed, the switching threshold moves. However, the output voltage transition remains sharp. Gates are usually skewed by adjusting the widths of transistors while maintaining minimum length for speed.

The inverter threshold can also be computed analytically. If the long-channel models of EQ (2.10) for saturated transistors are valid:

$$I_{dn} = \frac{\beta_n}{2}\left(V_{\text{inv}} - V_{tn}\right)^2$$
$$I_{dp} = \frac{\beta_p}{2}\left(V_{\text{inv}} - V_{DD} - V_{tp}\right)^2 \tag{2.54}$$

By setting the currents to be equal and opposite, we can solve for V_{inv} as a function of r:

$$V_{\text{inv}} = \frac{V_{DD} + V_{tp} + V_{tn}\sqrt{\frac{1}{r}}}{1 + \sqrt{\frac{1}{r}}} \tag{2.55}$$

In the limit that the transistors are fully velocity saturated, EQ (2.29) shows

$$I_{dn} = W_n C_{\text{ox}} v_{\text{sat}-n}(V_{\text{inv}} - V_{tn})$$
$$I_{dp} = W_p C_{\text{ox}} v_{\text{sat}-p}(V_{\text{inv}} - V_{DD} - V_{tp}) \tag{2.56}$$

Redefining $r = W_p v_{\text{sat-}p} / W_n v_{\text{sat-}n}$, we can again find the inverter threshold

$$V_{\text{inv}} = \frac{V_{DD} + V_{tp} + V_{tn}\frac{1}{r}}{1 + \frac{1}{r}} \tag{2.57}$$

In either case, if $V_{tn} = -V_{tp}$ and $r = 1$, $V_{\text{inv}} = V_{DD}/2$ as expected. However, velocity saturated inverters are more sensitive to skewing because their DC transfer characteristics are not as sharp.

DC transfer characteristics of other static CMOS gates can be understood by collapsing the gates into an equivalent inverter. Series transistors can be viewed as a single transistor of greater length. If only one of several parallel transistors is ON, the other

transistors can be ignored. If several parallel transistors are ON, the collection can be viewed as a single transistor of greater width.

2.5.3 Noise Margin

Noise margin is closely related to the DC voltage characteristics [Wakerly00]. This parameter allows you to determine the allowable noise voltage on the input of a gate so that the output will not be corrupted. The specification most commonly used to describe noise margin (or *noise immunity*) uses two parameters: the *LOW* noise margin, NM_L, and the *HIGH* noise margin, NM_H. With reference to Figure 2.29, NM_L is defined as the difference in maximum LOW input voltage recognized by the receiving gate and the maximum LOW output voltage produced by the driving gate.

$$NM_L = V_{IL} - V_{OL} \tag{2.58}$$

The value of NM_H is the difference between the minimum HIGH output voltage of the driving gate and the minimum HIGH input voltage recognized by the receiving gate. Thus,

$$NM_H = V_{OH} - V_{IH} \tag{2.59}$$

where

V_{IH} = minimum HIGH input voltage
V_{IL} = maximum LOW input voltage
V_{OH} = minimum HIGH output voltage
V_{OL} = maximum LOW output voltage

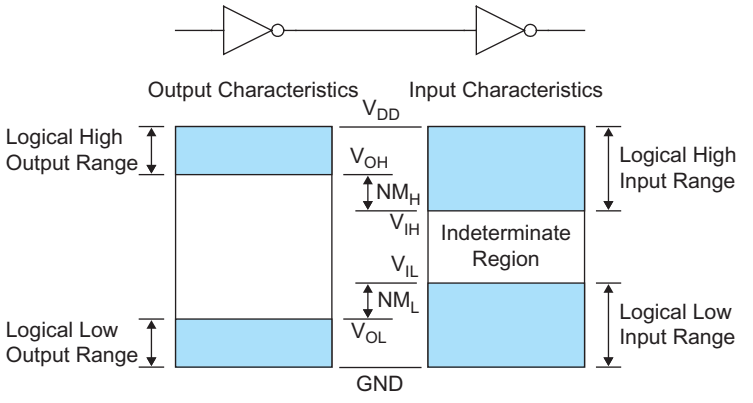


FIGURE 2.29 Noise margin definitions

Inputs between V_{IL} and V_{IH} are said to be in the *indeterminate region* or *forbidden zone* and do not represent legal digital logic levels. Therefore, it is generally desirable to have V_{IH} as close as possible to V_{IL} and for this value to be midway in the "logic swing," V_{OL} to V_{OH}. This implies that the transfer characteristic should switch abruptly; that is, there should be high gain in the transition region. For the purpose of calculating noise margins, the transfer characteristic of the inverter and the definition of voltage levels V_{IL}, V_{OL}, V_{IH}, and V_{OH} are shown in Figure 2.30. Logic levels are defined at the unity gain point where

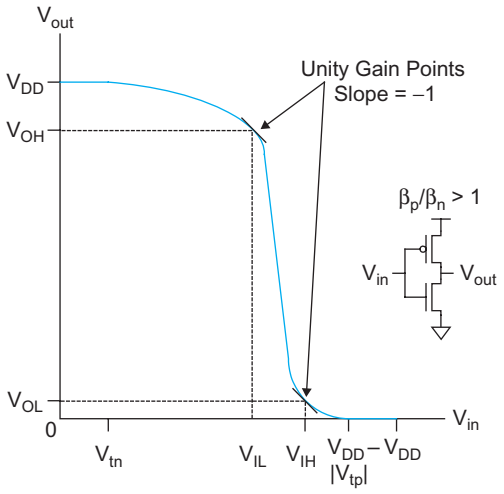


FIGURE 2.30 CMOS inverter noise margins

the slope is −1. This gives a conservative bound on the worst case static noise margin [Hill68, Lohstroh83, Shepard99]. For the inverter shown, the NM_L is 0.46 V_{DD} while the NM_H is 0.13 V_{DD}. Note that the output is slightly degraded when the input is at its worst legal value; this is called *noise feedthrough* or *propagated noise*. The exercises at the end of the chapter examine graphical and analytical approaches of finding the logic levels and noise margins.

If either NM_L or NM_H for a gate are too small, the gate may be disturbed by noise that occurs on the inputs. An unskewed gate has equal noise margins, which maximizes immunity to arbitrary noise sources. If a gate sees more noise in the high or low input state, the gate can be skewed to improve that noise margin at the expense of the other. Note that if $|V_{tp}| = V_{tn}$, then NM_H and NM_L increase as threshold voltages are increased.

Quite often, noise margins are compromised to improve speed. Circuit examples in Chapter 9 will illustrate this trade-off. Noise sources tend to scale with the supply voltage, so noise margins are best given as a fraction of the supply voltage. A noise margin of 0.4 V is quite comfortable in a 1.8 V process, but marginal in a 5 V process.

DC analysis gives us the *static noise margins* specifying the level of noise that a gate may see for an indefinite duration. Larger noise pulses may be acceptable if they are brief; these are described by *dynamic noise margins* specified by a maximum amplitude as a function of the duration [Lohstroh79, Somasekhar00]. Unfortunately, there is no simple amplitude-duration product that conveniently specifies dynamic noise margins.

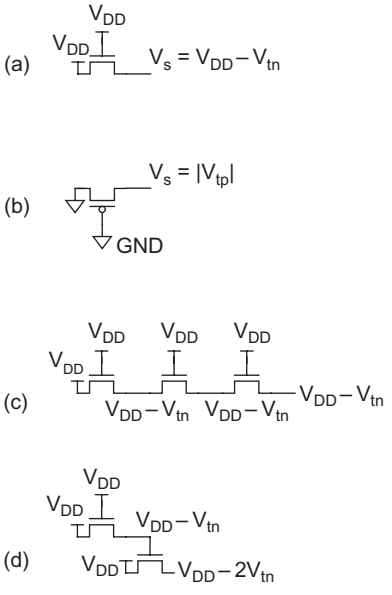


FIGURE 2.31 Pass transistor threshold drops

2.5.4 Pass Transistor DC Characteristics

Recall from Section 1.4.6 that nMOS transistors pass '0's well but 1s poorly. We are now ready to better define "poorly." Figure 2.31(a) shows an nMOS transistor with the gate and drain tied to V_{DD}. Imagine that the source is initially at $V_s = 0$. $V_{gs} > V_{tn}$, so the transistor is ON and current flows. If the voltage on the source rises to $V_s = V_{DD} - V_{tn}$, V_{gs} falls to V_{tn} and the transistor cuts itself OFF. Therefore, nMOS transistors attempting to pass a 1 never pull the source above $V_{DD} - V_{tn}$.[10] This loss is sometimes called a *threshold drop*.

Moreover, when the source of the nMOS transistor rises, V_{sb} becomes nonzero. As described in Section 2.4.3.1, this nonzero source to body potential introduces the body effect that increases the threshold voltage. Using the data from the example in that section, a pass transistor driven with $V_{DD} = 1$ V would produce an output of only 0.65 V, potentially violating the noise margins of the next stage.

Similarly, pMOS transistors pass 1s well but 0s poorly. If the pMOS source drops below $|V_{tp}|$, the transistor cuts off. Hence, pMOS transistors only pull down to within a threshold above GND, as shown in Figure 2.31(b).

[10] Technically, the output can rise higher very slowly by means of subthreshold leakage.

As the source can rise to within a threshold voltage of the gate, the output of several transistors in series is no more degraded than that of a single transistor (Figure 2.31(c)). However, if a degraded output drives the gate of another transistor, the second transistor can produce an even further degraded output (Figure 2.31(d)).

If we attempt to use a transistor as a switch, the threshold drop degrades the output voltage. In old processes where the power supply voltage was high and V_t was a small fraction of V_{DD}, the drop was tolerable. In modern processes where V_t is closer to 1/3 of V_{DD}, the threshold drop can produce an invalid or marginal logic level at the output. To solve this problem, CMOS switches are generally built using transmission gates.

Recall from Section 1.4.6 that a transmission gate consists of an nMOS transistor and a pMOS transistor in parallel with gates controlled by complementary signals. When the transmission gate is ON, at least one of the two transistors is ON for any output voltage and hence, the transmission gate passes both 0s and 1s well. The transmission gate is a fundamental and ubiquitous component in MOS logic. It finds use as a multiplexing element, a logic structure, a latch element, and an analog switch. The transmission gate acts as a voltage-controlled switch connecting the input and the output.

2.6 Pitfalls and Fallacies

This section lists a number of pitfalls and fallacies that can deceive the novice (or experienced) designer.

Blindly trusting one's models

Models should be viewed as only approximations to reality, not reality itself, and used within their limitations. In particular, simple models like the Shockley or RC models aren't even close to accurate fits for the I-V characteristics of a modern transistor. They are valuable for the insight they give on trends (i.e., making a transistor wider increases its gate capacitance and decreases its ON resistance), not for the absolute values they predict. Cutting-edge projects often target processes that are still under development, so these models should only be viewed as speculative. Finally, processes may not be fully characterized over all operating regimes; for example, don't assume that your models are accurate in the subthreshold region unless your vendor tells you so. Having said this, modern SPICE models do an extremely good job of predicting performance well into the GHz range for well-characterized processes and models when using proper design practices (such as accounting for temperature, voltage, and process variation).

Using excessively complicated models for manual calculations

Because models cannot be perfectly accurate, there is little value in using excessively complicated models, particularly for hand calculations. Simpler models give more insight on key trade-offs and more rapid feedback during design. Moreover, RC models calibrated against simulated data for a fabrication process can estimate delay just as accurately as elaborate models based on a large number of physical parameters but not calibrated to the process.

Assuming a transistor with twice the drawn length has exactly half the current

To first order, current is proportional to W/L. In modern transistors, the effective transistor length is usually shorter than the drawn length, so doubling the drawn length reduces current by more than a factor of two. Moreover, the threshold voltage tends to increase for longer transistors, resulting in less current. Therefore, it is a poor strategy to try to ratio currents by ratioing transistor lengths.

Assuming two transistors in series deliver exactly half the current of a single transistor

To first order, this would be true. However, each series transistor sees a smaller electric field across the channel and hence are each less velocity saturated. Therefore, two series transistors in a nanometer process will deliver more than half the current of a single transistor. This is more pronounced for nMOS than pMOS transistors because of the higher mobility and the higher degree of velocity saturation of electrons than holes at a given field. Hence, NAND gates perform better than first order estimates might predict.

Ignoring leakage

In contemporary processes, subthreshold and gate leakage can be quite significant. Leakage is exacerbated by high temperature and by random process variations. Undriven nodes will not retain their state for long; they will leak to some new voltage. Leakage power can account for a large fraction of total power, especially in battery-operated devices that are idle most of the time.

Using nMOS pass transistors

nMOS pass transistors only pull up to $V_{DD} - V_t$. This voltage may fall below V_{IH} of a receiver, especially as V_{DD} decreases. For example, one author worked with a scan latch containing an nMOS pass transistor that operated correctly in a 250 nm process at 2.5 V. When the latch was ported to a 180 nm process at 1.8 V, the scan chain stopped working. The problem was traced to the pass transistor and the scan chain was made operational in the lab by raising V_{DD} to 2 V. A better solution is to use transmission gates in place of pass transistors.

Summary

In summary, we have seen that MOS transistors are four-terminal devices with a gate, source, drain, and body. In normal operation, the body is tied to GND or V_{DD} so the transistor can be modeled as a three-terminal device. The transistor behaves as a voltage-controlled switch. An nMOS switch is OFF (no path from source to drain) when the gate voltage is below some threshold V_t. The switch turns ON, forming a channel connecting source to drain, when the gate voltage rises above V_t. This chapter has developed more elaborate models to predict the amount of current that flows when the transistor is ON. The transistor operates in three modes depending on the terminal voltages:

- $V_{gs} < V_t$ Cutoff $I_{ds} \approx 0$
- $V_{gs} > V_t$, $V_{ds} < V_{dsat}$ Linear I_{ds} increases with V_{ds} (like a resistor)
- $V_{gs} > V_t$, $V_{ds} > V_{dsat}$ Saturation I_{ds} constant (like a current source)

In a long-channel transistor, the saturation current depends on V_{GT}^2. pMOS transistors are similar to nMOS transistors, but have the signs reversed and deliver about half the current because of lower mobility.

In a real transistor, the I-V characteristics are more complicated. Modern transistors are extraordinarily small and thus experience enormous electric fields even at low voltage. The high fields cause velocity saturation and mobility degradation that lead to less current than you might otherwise expect. This can be modeled as a saturation current dependent on V_{GT}^{α}, where the velocity saturation index α is less than 2. Moreover, the saturation current does increase slightly with V_{ds} because of channel length modulation. Although simple hand calculations are no longer accurate, the general shape does not change very much and the transfer characteristics can still be derived using graphical or simulation methods.

Even when the gate voltage is low, the transistor is not completely OFF. Subthreshold current through the channel drops off exponentially for $V_{gs} < V_t$, but is nonnegligible for transistors with low thresholds. Junction leakage currents flow through the reverse-biased p–n junctions. Tunneling current flows through the insulating gate when the oxide becomes thin enough.

We can derive the DC transfer characteristics and noise margins of logic gates using either analytical expressions or a graphical load line analysis or simulation. Static CMOS gates have excellent noise margins.

Unlike ideal switches, MOS transistors pass some voltage levels better than others. An nMOS transistor passes 0s well, but only pulls up to $V_{DD} - V_{tn}$ when passing 1s. The pMOS passes 1s well, but only pulls down to $|V_{tp}|$ when passing 0s. This threshold drop is exacerbated by the body effect, which increases the threshold voltage when the source is at a different potential than the body.

There are too many parameters in a modern BSIM model for a designer to deal with intuitively. Instead, CMOS transistors are usually characterized by the following basic figures of merit:

- V_{DD} Target supply voltage
- L_{gate} / L_{poly} Effective channel length (< feature size)
- t_{ox} Effective oxide thickness (a.k.a. EOT)
- I_{dsat} I_{ds} @ $V_{gs} = V_{ds} = V_{DD}$
- I_{off} I_{ds} @ $V_{gs} = 0$, $V_{ds} = V_{DD}$
- I_g Gate leakage @ $V_{gs} = V_{DD}$

[Muller03] and [Tsividis99] offer comprehensive treatments of device physics at a more advanced level. [Gray01] describes MOSFET models in more detail from the analog designer's point of view.

Exercises

2.1 Consider an nMOS transistor in a 0.6 μm process with $W/L = 4/2\ \lambda$ (i.e., 1.2/0.6 μm). In this process, the gate oxide thickness is 100 Å and the mobility of electrons is 350 cm^2/V · s. The threshold voltage is 0.7 V. Plot I_{ds} vs. V_{ds} for $V_{gs} = 0, 1, 2, 3, 4,$ and 5 V.

2.2 Show that the current through two transistors in series is equal to the current through a single transistor of twice the length if the transistors are well described by the Shockley model. Specifically, show that $I_{DS1} = I_{DS2}$ in Figure 2.32 when the transistors are in their linear region: $V_{DS} < V_{DD} - V_t$, $V_{DD} > V_t$ (this is also true in saturation). *Hint*: Express the currents of the series transistors in terms of V_1 and solve for V_1.

2.3 In Exercise 2.2, the body effect was ignored. If the body effect is considered, will I_{DS2} be equal to, greater than, or less than I_{DS1}? Explain.

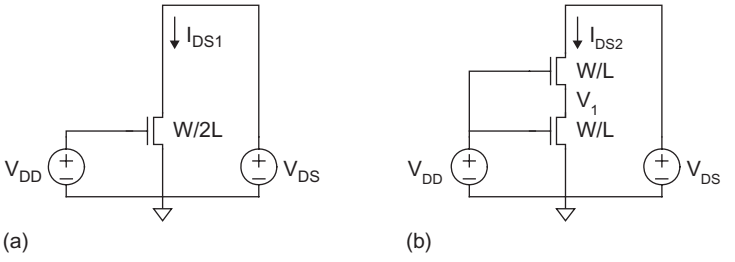


FIGURE 2.32 Current in series transistors

2.4 A 90 nm long transistor has a gate oxide thickness of 16 Å. What is its gate capacitance per micron of width?

2.5 Calculate the diffusion parasitic C_{db} of the drain of a unit-sized contacted nMOS transistor in a 0.6 μm process when the drain is at 0 and at $V_{DD} = 5$ V. Assume the substrate is grounded. The transistor characteristics are $CJ = 0.42$ fF/μm^2, $MJ = 0.44$, $CJSW = 0.33$ fF/μm, $MJSW = 0.12$, and $\psi_0 = 0.98$ V at room temperature.

2.6 Prove EQ (2.27).

2.7 Consider the nMOS transistor in a 0.6 μm process with gate oxide thickness of 100 Å. The doping level is $N_A = 2 \times 10^{17}$ cm^{-3} and the nominal threshold voltage is 0.7 V. The body is tied to ground with a substrate contact. How much does the threshold change at room temperature if the source is at 4 V instead of 0?

2.8 Does the body effect of a process limit the number of transistors that can be placed in series in a CMOS gate at low frequencies?

2.9 Sometimes the substrate is connected to a voltage called the substrate bias to alter the threshold of the nMOS transistors. If the threshold of an nMOS transistor is to be raised, should a positive or negative substrate bias be used?

2.10 An nMOS transistor has a threshold voltage of 0.4 V and a supply voltage of $V_{DD} = 1.2$ V. A circuit designer is evaluating a proposal to reduce V_t by 100 mV to obtain faster transistors.

a) By what factor would the saturation current increase (at $V_{gs} = V_{ds} = V_{DD}$) if the transistor were ideal?

b) By what factor would the subthreshold leakage current increase at room temperature at $V_{gs} = 0$? Assume $n = 1.4$.

c) By what factor would the subthreshold leakage current increase at 120 °C? Assume the threshold voltage is independent of temperature.

2.11 Find the subthreshold leakage current of an inverter at room temperature if the input $A = 0$. Let $\beta_n = 2\beta_p = 1$ mA/V^2, $n = 1.0$, and $|V_t| = 0.4$ V. Assume the body effect and DIBL coefficients are $\gamma = \eta = 0$.

2.12 Repeat Exercise 2.11 for a NAND gate built from unit transistors with inputs $A = B = 0$. Show that the subthreshold leakage current through the series transistors is half that of the inverter if $n = 1$.

2.13 Repeat Exercises 2.11 and 2.12 when $\eta = 0.04$ and $V_{DD} = 1.8$ V, as in the case of a more realistic transistor. γ has a secondary effect, so assume that it is 0. Did the leakage currents go up or down in each case? Is the leakage through the series transistors more than half, exactly half, or less than half of that through the inverter?

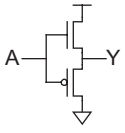


FIGURE 2.33
Noninverting buffer

2.14 Peter Pitfall is offering to license to you his patented noninverting buffer circuit shown in Figure 2.33. Graphically derive the transfer characteristics for this buffer. Assume $\beta_n = \beta_p = \beta$ and $V_{tn} = |V_{tp}| = V_t$. Why is it a bad circuit idea?

2.15 A novel inverter has the transfer characteristics shown in Figure 2.34. What are the values of V_{IL}, V_{IH}, V_{OL}, and V_{OH} that give best noise margins? What are these high and low noise margins?

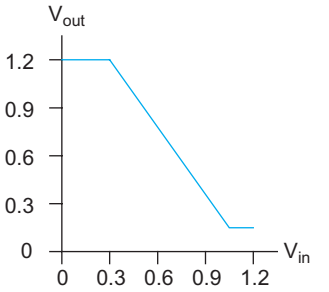


FIGURE 2.34
Transfer characteristics

2.16 Section 2.5.1 graphically determined the transfer characteristics of a static CMOS inverter. Derive analytic expressions for V_{out} as a function of V_{in} for regions *B* and *D* of the transfer function. Let $|V_{tp}| = V_{tn}$ and $\beta_p = \beta_n$.

2.17 Using the results from Exercise 2.16, calculate the noise margin for a CMOS inverter operating at 1.0 V with $V_{tn} = |V_{tp}| = 0.35$ V, $\beta_p = \beta_n$.

2.18 Repeat Exercise 2.16 if the thresholds and betas of the two transistors are not necessarily equal. Also solve for the value of V_{in} for region *C* where both transistors are saturated.

2.19 Using the results from Exercise 2.18, calculate the noise margin for a CMOS inverter operating at 1.0 V with $V_{tn} = |V_{tp}| = 0.35$ V, $\beta_p = 0.5\beta_n$.

2.20 Give an expression for the output voltage for the pass transistor networks shown in Figure 2.35. Neglect the body effect.

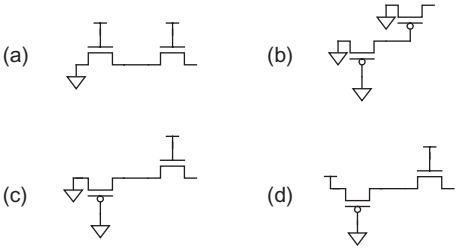


FIGURE 2.35 Pass transistor networks

2.21 Suppose $V_{DD} = 1.2$ V and $V_t = 0.4$ V. Determine V_{out} in Figure 2.36 for the following. Neglect the body effect.

a) $V_{\text{in}} = 0$ V

b) $V_{\text{in}} = 0.6$ V

c) $V_{\text{in}} = 0.9$ V

d) $V_{\text{in}} = 1.2$ V.

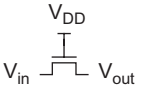


FIGURE 2.36
Single pass transistor